Using the Second-Generation GEFS Reforecasts to Predict Ceiling, Visibility, and Aviation Flight Category

KATHRYN L. VERLINDEN[a]

Cooperative Institute for Research in the Atmosphere, Fort Collins, Colorado

DAVID R. BRIGHT

NOAA/National Weather Service, Portland, Oregon

(Manuscript received 30 November 2016, in final form 16 April 2017)

ABSTRACT

This study is an aviation-based application of NOAA's second-generation medium-range Global Ensemble Forecast System Reforecast (GEFS/R; i.e., hindcast or retrospective forecast) dataset. The study produced a downscaled probabilistic prediction of instrument flight conditions at major U.S. airports using an analog approach. This represents an initial step toward applications of reforecast data to probabilistic aviation decision support services. Results from this study show that even at the very coarse resolution of the GEFS/R dataset, the analog approach yielded skillful probabilistic forecasts of flight conditions (i.e., instrument flight rules vs visual flight rules) at most of the Federal Aviation Administration (FAA)'s Core 30 airports. This was particularly true over the central and eastern United States, including the important Golden Triangle, where aircraft flow affects traffic flow management across the entire national airspace system. Additionally, the results suggest that reforecast systems utilizing better horizontal and vertical resolution, in both the modeling system and the reforecast archive, would be very useful for aviation forecasting applications.

1. Introduction

The Next Generation Global Prediction System (NGGPS) is a National Oceanic and Atmospheric Administration (NOAA)/National Weather Service (NWS) initiative to expand and accelerate development and implementation of global weather prediction and data assimilation, as well as increase the accuracy of weather forecasts and build foundational forecast guidance for the next several decades. As part of this initiative, this study utilizes NOAA's second-generation medium-range Global Ensemble Forecast System Reforecast (GEFS/R) dataset (Hamill et al. 2013) to explore cloud ceiling and visibility prediction at major airports across the United States. Reforecasts (known synonymously as hindcasts or retrospective forecasts) provide a large sample of historical model/ensemble forecasts available to statistically postprocess real-time forecasts using an identical (i.e., statistically consistent) model. The goal of the postprocessing is to produce calibrated and reliable forecasts, often of rare or infrequent events, by accounting for both chaotic (i.e., initial condition uncertainty) and systematic model errors. The second-generation reforecast dataset is derived from the 2012 version of the 0000 UTC cycle of the NWS Global Ensemble Forecast System (GEFS). While numerous studies have demonstrated the value of reforecasting for ensemble postprocessing and decision support (e.g., Hamill et al. 2006, 2013, 2015; Wilks and Hamill 2007; Hagedorn et al. 2008), only one prior study has been specific to aviation (Herman and Schumacher 2016).

Poor weather conditions have been shown to dramatically increase the rate of aviation fatalities. For example, under instrument flight rules (IFR), defined as a cloud ceiling below 1000 ft above ground level and/or visibility of less than 3 mi, about two-thirds of all general aviation accidents are fatal—a rate much higher than the overall fatality rate for all general aviation incidents (NTSB 2014). Similarly, between 1983 and 2009

[a] Current affiliation: College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Corvallis, Oregon.

Corresponding author: Kathryn L. Verlinden, kverlinden@coas.oregonstate.edu

DOI: 10.1175/WAF-D-16-0211.1

© 2017 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 07/19/23 06:51 PM UTC

TABLE 1. The Core 30 airports across CONUS by call sign and airport name. Golden Triangle airports considered here have been italicized. Intl = international.

Call sign	Airport name and state	Call sign	Airport name and state
ATL	*Hartsfield–Jackson Atlanta Intl*, GA	SFO	San Francisco Intl, CA
CLT	Charlotte Douglas Intl, NC	SLC	Salt Lake City Intl, UT
DFW	Dallas/Fort Worth Intl, TX	DTW	Detroit Metropolitan Wayne County, MI
FLL	Fort Lauderdale/Hollywood Intl, FL	MDW	Chicago Midway, IL
IAH	George Bush Houston Intercontinental, TX	MSP	Minneapolis/St. Paul Intl, MN
MCO	Orlando Intl, FL	*ORD*	*Chicago O'Hare Intl*, IL
MEM	Memphis Intl, TN	BOS	Boston Logan Intl, MA
MIA	Miami Intl, FL	BWI	Baltimore/Washington Intl, MD
TPA	Tampa Intl, FL	DCA	Ronald Reagan Washington National, VA
DEN	Denver Intl, CO	EWR	Newark Liberty Intl, NJ
LAS	Las Vegas McCarran Intl, NV	*JFK*	*John F. Kennedy Intl*, NY
LAX	Los Angeles Intl, CA	IAD	Washington Dulles Intl, VA
PHX	Phoenix Sky Harbor Intl, AZ	LGA	New York LaGuardia, NY
SAN	San Diego Intl, CA	PHL	Philadelphia Intl, PA
SEA	Seattle/Tacoma Intl, WA		

over the Gulf of Mexico, 16% of helicopter accidents and 40% of the resulting fatalities were attributable to poor weather conditions (Baker et al. 2011).

In addition to safety, accurate predictions of ceiling and visibility may have far-reaching economic and traffic flow management implications. Probabilistic forecasts for both visibility conditions and ceilings surrounding airports allow for cost-based critical decision thresholds to be created for fuel loading in accordance with airlines' planning timelines (Keith and Leyton 2007). Increasing skill at longer lead times allows for more efficient and effective planning, with potential savings of tens of millions of dollars annually via fuel cost reductions for many major airlines (Keith and Leyton 2007). Additionally, the ability to adjust flight plans based on predicted ceilings and visibilities that reduce arrival and departure rates could streamline air traffic movement across the United States. While modern instrument landing systems (ILS) allow aircraft to land in weather conditions well below IFR, closely spaced runways and competing ILS capacity (e.g., New York metropolitan area airports) can reduce operations substantially below peak arrival rates. For example, San Francisco International Airport (SFO) requires visual flight rules (VFR) to support its dual-runway-approach configuration. The presence of low clouds in the SFO approach zone reduces arrival capacity by 33%–50% (Reynolds et al. 2012).

This study takes a preliminary look at downscaling NOAA's GEFS/R dataset to the Federal Aviation Administration's (FAA) Core 30 airports (Table 1). In the context of this work, downscaling refers to the postprocessing of the relatively low-resolution reforecast grids to produce a flight condition prediction (i.e., IFR or VFR conditions based on the ceiling and visibility) at a point. The points evaluated are the Core 30 airports, which are 30 of the nation's busiest airports used by the FAA to monitor aviation system safety and efficiency. All Core 30 airports are used in this study, with the exception of Honolulu, Hawaii (HNL). A model climatology and probabilistic ceiling and visibility forecasts through 30 h are created using an analog reforecast approach (Hamill et al. 2006; Hamill and Whitaker 2006). A period of 30 h was chosen because it encompasses the 24-h period of most terminal aerodrome forecasts (TAFs), and it is a reasonable traffic flow management outlook period in the aviation community. Similar to the work of Hamill et al. (2004), the ensemble mean reforecasts are used for determining analogs at all airports. Historical METAR observations of ceiling and visibility at each airport serve as ground truth. Additionally, the analog reforecast approach examined here is compared to TAFs at three airports within the aviation Golden Triangle: John F. Kennedy International (JFK), Chicago O'Hare International (ORD), and Hartsfield–Jackson Atlanta International (ATL).

This next section provides a description of the datasets, while section 3 provides an overview of the data postprocessing, including the analog forecast approach and the statistical methods employed. This is followed by a summary of the analog forecast system results for predicting both IFR and VFR for the continental United States (CONUS); these results are further examined seasonally for the Golden Triangle in section 4. Section 5 contains a general discussion of possible reasons for observed differences across the CONUS and suggestions for future refinements and potential applications for the aviation weather community. Section 6 contains a short summary and a few concluding remarks highlighting potential future paths for this research.

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 07/19/23 06:51 PM UTC

2. Data

a. Model data

This effort utilizes the entire 30 years of NOAA's second-generation GEFS/R dataset, which uses the identical modeling system as GEFS, version 9.0.1. The ensemble forecasts are initialized once daily at 0000 UTC to create 10 perturbed forecast members and one control forecast. Running from December 1984 to the present, these reforecasts have been made available at 3-hourly intervals for lead times of 0–72 h, and then 6-hourly intervals out to 16 days. To build the historical reforecast dataset, the GEFS was run at T254L42 for the first 8 days (42 layers and an equivalent grid spacing of approximately 40 km at 40° latitude), and then T190L42 for days 8–16 (about 54 km at 40° latitude). This study utilized only the first 2 days of the reforecasts. The global fields are at 1° × 1° grid spacing for 98 different variables with an additional 28 variables available at the native resolution. All of the data used in this study are extracted from the 1° × 1° grids. This dataset can be accessed online (http://www.esrl.noaa.gov/psd/forecasts/reforecast2/). See Hamill et al. (2013) for a complete description of the construction and data availability of the GEFS/R dataset.

For this study, output at 3-h intervals through forecast hour 30 from the daily 0000 UTC 1 December 1984–31 May 2015 reforecasts are used. Utilized fields include surface pressure and temperature at 2 m and available isobaric pressure levels (1000, 850, 700, 500 hPa), and specific humidity at 2 m and available isobaric pressure levels (1000, 850, 700, 500 hPa).

b. Observational data

For the same period, aviation routine weather reports (METARs) at the Core 30 airports are used as ground truth for ceiling height and surface visibility. METARs are generated every hour by observations made at each airport by an Automated Surface Observing System (ASOS). For more information on ASOS instrumentation and METAR generation, see NOAA (1998). Data were accessed through NOAA's National Centers for Environmental Information (NCEI) website (www.ncei.noaa.gov/).

c. Forecast data

NWS TAFs (NWS 2016) for three Golden Triangle airports (JFK, ATL, and ORD) from January 2010 through May 2015 are used as a more competitive test of the analog forecast method than climatology. NWS TAFs are used in a variety of applications throughout the aviation industry, including general and commercial aviation, and civilian and military operations. TAFs are also a key component of flight planning and aircraft movement within the National Airspace System (NWS 2016). The NWS TAF represents forecast conditions at an airport (within 5 statute miles of the center of an airport's runway complex) and is issued four times per day (amended as necessary), typically covering the next 24 h (some international airports require 30 h). The TAF is a formatted forecast consisting of weather, surface wind (speed and direction), visibility, cloud layers and ceiling, and low-level wind shear. By design, TAFs may include more detail in the first 12 h and less thereafter, as the latter portion is primarily for strategic planning. Historical TAFs were accessed through Ogimet (www.ogimet.com), which collects, stores, and makes available public data from sources including NOAA. The truncated period of study is due to data availability on the site. The NWS uses its Stats-On-Demand software (NWS 2016) for quality assurance. Lorentson (2013) shows statistical results for several years of NWS TAF forecasts of IFR, along with the Government Performance and Results Act (GPRA) goals for 2007–16.

3. Methods

a. Data preparation

For each airport, the four surrounding model forecast grid points are identified and bilinear interpolation is applied to estimate the forecast value at an airport's location. Relative humidity is calculated at every pressure level from the available grids of forecast temperature, saturation mixing ratio, and specific humidity. Vertical profiles of dewpoint temperatures are derived from the calculated relative humidity and temperature fields.

METAR observations at the forecast valid times are used for analog downscaling and verification (i.e., every third hour from 0000 UTC through forecast hour 30). In very rare instances when reported observations are not available on the hour but an observation was recorded within 10 min of the hour, values are linearly interpolated in time to create an on-the-hour observation. If an observation was both not on the hour and not reported within 10 min of the hour, then the observation is marked as missing and not included in the study. The primary METAR observations of interest are cloud ceiling height and visibility. The reported ceiling height and visibility observations are then classified into flight conditions (Table 2).

b. Analog forecasts

An analog approach is used to identify similar historical reforecasts to downscale the global reforecast

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 07/19/23 06:51 PM UTC

TABLE 2. Flight condition definitions. Conditions are defined on an and/or basis with the lowest visibility or ceiling defining the current flight conditions. Ceiling is evaluated relative to airport elevation.

Flight condition	Ceiling (ft)	Visibility (statute miles)
IFR	<1000	<3
MVFR	≥1000 and ≤3000	≥3 and ≤5
VFR	>3000	>5

to a point [in the manner of, e.g., Toth (1989) and Van den Dool (1989)]. Vertical profiles ("soundings") of temperature and dewpoint temperature are created by concatenating model output grids at 2 m above the surface, 1000, 850, 700, and 500 hPa (Fig. 1b). If the modeled surface pressure is less than any of the isobaric grid levels, then those levels are removed from the sounding. Starting with 1 December 1984, every fifth day the forecast sounding at a given lead time is compared to all historical reforecast soundings at the same lead time (Fig. 1c) via a normalized root-mean-square difference (RMSD). Every fifth day is used to avoid oversampling any single weather regime. Variables are normalized at each pressure level by representative measurement errors assigned in the Eta Data Assimilation System (EDAS) to rawinsonde observations [see Table 2 in Zapotocny et al. (2000) for values]. The equation for determining the normalized RMSD is

$$\mathrm{RMSD} = \sqrt{\frac{1}{2N_p}\sum_{p=1}^{N}\left[\frac{(T_{m_p}-T_{r_p})^2}{T_{e_p}}+\frac{(\mathrm{Td}_{m_p}-\mathrm{Td}_{r_p})^2}{T_{e_p}}\right]}, \quad (1)$$

where T is temperature, Td is dewpoint temperature, subscript p is the vertical level, subscript m denotes a model (reference sounding) value, subscript r denotes a reforecast value, and e is the representative measurement error from Zapotocny et al. (2000). The 50 soundings corresponding to the smallest normalized RMSD are considered analog forecast matches (Fig. 1d). The quantity of 50 analog soundings is chosen so as to provide reasonable sample size without causing overfitting, and it has previously been identified as adequate for the short forecast lead times considered here (Hamill et al. 2015). Data denial is employed for verification and validation of the technique; as such, the original forecast sounding is removed from the comparison such that the date of interest is never included as an analog. It should be noted that we also considered analogs including wind profiles, but this did not significantly change the results. METAR observations at the verifying time for each of the 50 analog reforecast matches provide ceiling and visibility observations to determine the downscaled flight condition at the airport (Fig. 1e). Probability of observed flight conditions are then determined using these observations. For example, if 20 of the top 50 matched soundings' METARs report IFR conditions, then the probability of IFR is 40%. This process is repeated for every airport for each of the 11 forecast lead times. More sophisticated methods of ranking or weighting the analog matches may improve results, but they were not tested here. Because 50 matches is a somewhat arbitrary choice, perfectly reliable probabilities are unlikely without further calibration.

c. *Terminal aerodrome forecasts*

TAFs issued at 0600 UTC are used, since they are informed by the 0000 UTC model runs, the same information included in the GEFS/R dataset. Forecasts for ceiling and visibility are stripped from the text TAFs for the appropriate forecast valid time (from 0600 through 2100 UTC). TAFs may contain a temporary (TEMPO) change indicator group and a probability group, which must be addressed differently from the change indicator group. For the period considered in this study, only the 30% probability (Prob30) group is standard practice. When a probability group or temporary change indicator group is encountered, its probability is recorded. However, if the prevailing flight condition is the same as the Prob30 or TEMPO group (e.g., if visibility is temporarily forecasted to decrease but remains within the bounds of the prevailing flight condition), then the probability is marked as 100%, as with standard forecasts. Only dates that intersect with the analog forecasts (i.e., every fifth day during the period subset) are considered—again, to avoid oversampling any specific weather event.

According to NWS (2016, p. 23), the TEMPO group represents "a high percentage (greater than 50%) probability of occurrence." As such, we calculated Brier skill scores designating TAF TEMPO forecasts as 50%, 75%, and 100% probabilities. Designating the TEMPO group as 100% probability resulted in the most skillful forecasts and thus proved to be the strongest competition for the analog forecast method. Likewise, Keith and Leyton (2007) incorporated TAFs as either 0% or 100% to study the economic value of weather forecasts. Values reported in this paper reflect designating the TEMPO group forecasts as 100% probability.

d. *Model verification*

Brier skill scores (BSS) are employed here as a metric of skill in forecasting flight condition categories. The BSS is a measure of the mean-square error of a

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 07/19/23 06:51 PM UTC

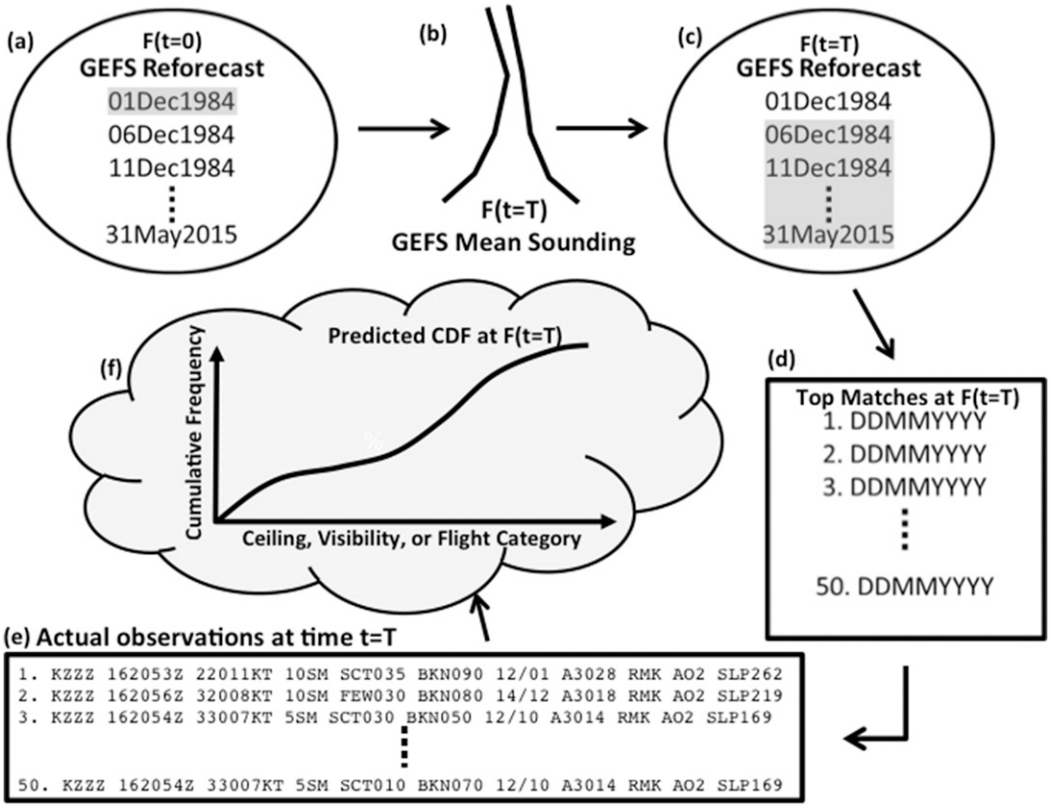

FIG. 1. Flowchart depicting the analog forecast methodology. Using cross validation, every fifth day of the reforecast (to avoid oversampling dependent weather events) is tested. For example, (a) the forecast day is 1 Dec 1984 (shaded). (b) Based on the ensemble mean at forecast hour T, vertical profiles of temperature and dewpoint are constructed for each of the Core 30 airports. (c) For each airport, the ensemble mean vertical profile is then compared to the reforecast database at forecast hour T (excluding the day being forecast), searching for similar ensemble mean vertical profiles at that airport (the shading depicts eligible days in the database). (d) The nearest 50 matches based on Eq. (1) (see text for details) are identified at forecast hour T. For the sake of illustration, assume the airport is ZZZ at forecast hour $T = 21$ (which would be valid at 2100 UTC, since the reforecast always begins at 0000 UTC). (e) The actual METARs valid at the matching day and time (again, time in this example would be 2100 UTC) are collected to form the possible outcomes at the actual station based on similar ensemble mean vertical profiles. (f) From this collection of 50 METARs, the frequency of actual surface conditions that occurred based on similar vertical profiles is used to build probabilistic forecasts of ceiling, visibility, and flight condition. Given a robust sample of quality matches, the collection of information represented by the cumulative distribution function in (f) should account for systematic model bias, chaotic error growth, and downscaling of the gridded values to a point.

probability forecast for a dichotomous event normalized by the same for a reference forecast (Wilks 2011). Tested separately were the observed sample climatology constructed from December 1984 through May 2015, and TAFs from January 2010 through May 2015. Assuming that each forecast is equally likely, the Brier score of the forecast BS_f is calculated as

$$BS_f = \frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2, \tag{2}$$

where P_i is the forecasted event probability, and O_i is either 1 or 0 if the event was observed or not. A BSS is then calculated as

$$BSS = 1 - \frac{BS_f}{BS_r}, \tag{3}$$

where BS_r is the Brier score of the reference probability forecast. A BSS of 0.0 indicates that the forecast has the same skill as the reference forecast, a BSS of 1.0 indicates a perfect forecast, and a negative BSS indicates less skill than the reference. A BSS can be interpreted as a percent improvement over the reference dataset.

Attributes diagrams address how the predicted probabilities of an event correspond to their observed frequencies via a plot of observed frequency versus the forecast probability (Wilks 2011). The range of forecast probabilities has been divided into 10 bins (0%–10%, 10%–20%,

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 07/19/23 06:51 PM UTC

etc.). Deviation of the curve from the major diagonal 1:1 line ("perfect reliability" line) gives the conditional bias. If the curve lies above the perfect reliability line, then underforecasting (probabilities too low) is indicated, and conversely points on the curve that fall below this line indicate overforcasting (probabilities too high). The horizontal and vertical dashed lines indicate climatological frequency, where the horizontal line can be considered the "no resolution" line. The flatter the curve, the less resolution it has, since it approaches climatology, which inherently does not discriminate between events and nonevents. The diagonal line halfway between the no resolution line and the perfect reliability line can be considered the no skill line, since points that fall along it do not add to the BSS, while points that fall below contribute negatively to the BSS, and points that lie above it contribute positively to the BSS. Included on each attributes diagram is an inset histogram displaying the frequency of forecasts in each probability bin. Diagrams were created for a composite of the 29 CONUS airports and a composite of the three Golden Triangle airports at 0-, 12-, and 24-h lead times for the period from December 1984 through May 2015.

Relative operating characteristic (ROC) curves convey a measure of discrimination, or classifying event occurrence versus nonoccurrence. (Mason 1982). ROC curves were created for the three Golden Triangle airports for the analog forecast method and for TAFs by plotting the probability of detection (hit rate) versus the probability of false detection (false alarm rate). A perfect curve will travel from the bottom-left corner to the top-left corner and continue across the top to the top-right corner. Conversely, a curve depicting no skill (i.e., a random forecast) will follow the major diagonal from (0, 0) to (1, 1). The area under the ROC curve is typically between 0.5 and 1.0. An area greater than or equal to 0.70 is considered to have a reasonable amount of skill in discriminating between events and nonevents (Stanski et al. 1989), with 1.0 indicative of a perfect classifier. Areas under the ROC curves were calculated using a trapezoidal method, although this method may underestimate the area under curves defined by few points.

4. Results

a. Ensemble mean versus climatology

BSS versus forecast lead time for each airport is shown in Fig. 2 for IFR conditions (see Fig. 4 below for VFR conditions). Marginal visual flight rule (MVFR; defined in Table 2) conditions were examined but are not included in these results because of the coarse vertical resolution of the reanalysis grids (which are available on mandatory isobaric levels only). With so few vertical levels available in the reanalysis to define thin layers of the lower troposphere, there is insufficient vertical resolution to differentiate cloud ceilings that in reality differ by only hundreds of meters. Hence, because of the coarse vertical resolution of the reforecast dataset, only the results for VFR and IFR predictions are shown. Recall all forecasts are initialized at 0000 UTC. To more concisely examine the results, we briefly discuss patterns across the Core 30 airports but focus on three airports in the aviation Golden Triangle: ATL, ORD, and JFK. The forecast skill relative to the sample climate of the study period (from December 1984 through May 2015 for all seasons, or for only the applicable months within this period for each particular season; referred to as "climatology") is examined for each forecast lead time for the entire record as well as for each season.

1) IFR

(i) All seasons

The GEFS/R data considered through this analog downscaling method show skillful improvement over climatology (IOC) for forecasting IFR conditions at the majority of the Core 30 airports for all forecast lead times (Fig. 2). As a convention, IOC implies positive (skillful) improvement throughout the results presented here unless otherwise noted. This is particularly the case for airports in the Midwest, New England, and the South (sans several in Florida) with a 15%–25% IOC. For nearly all airports, skill decreases with increasing forecast lead time. Airports in New England, the Midwest, and the South, except Florida and Memphis, Tennessee, show a diurnal cycle in skill with the maximum occurring during late afternoon and early evening local time. Forecast skill for the Florida airport cluster remains well separated from the rest of the airports in the South, particularly during these late afternoon/early evening times with IOC of 1%–10%. This analog downscaling method for airports in the West and Florida shows the least improvement over climatology, with PHX, MIA, and FLL showing negative skill during short lead times, with the addition of DEN and TPA showing negative skill at long lead times (see Table 1 for airport identifiers).

The three Golden Triangle airports show IOC at all lead times with localized peaks at 0000 and 2100 UTC, and otherwise a slight decrease in skill with increasing lead time (Fig. 3, top-left panel). A bootstrapping procedure with 10 000 samples was employed to obtain confidence intervals (not shown) and confirmed that the IOC is significant at all lead times at the 95% confidence level.

(ii) Seasons

For the discussion of seasonal differences, only the skill scores of the three Golden Triangle airports and their composite are examined. The composite (solid line) of all three Golden Triangle airports show IOC for all lead times

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 07/19/23 06:51 PM UTC

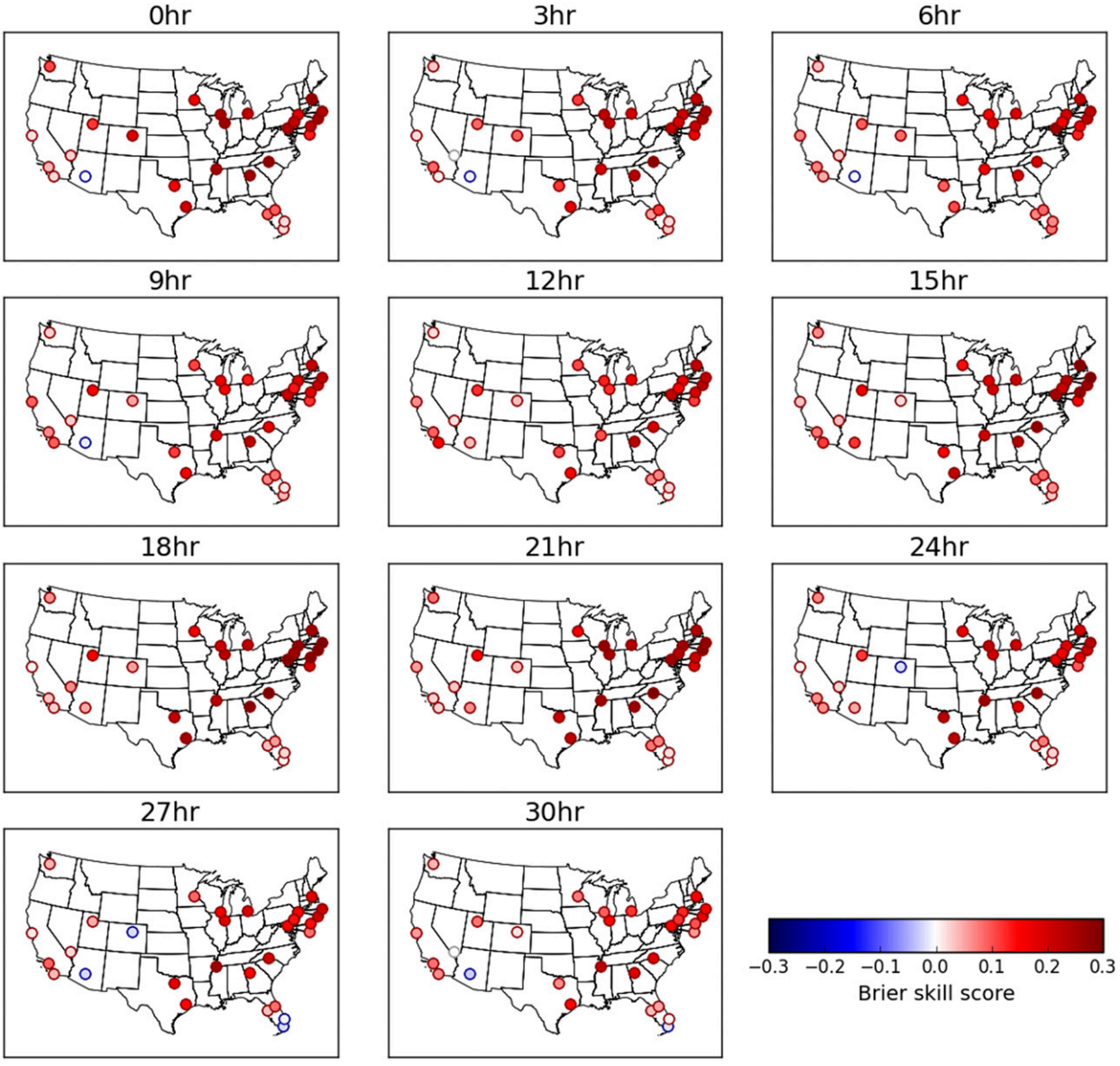

FIG. 2. BSSs computed for all seasons for IFR at the Core 30 airports across the CONUS at each forecast lead time. Perimeters of the circles denote positive (red) or negative (blue) skill. Shading within each circle denotes the skill score magnitude.

during each season (Fig. 3). The transitional seasons (MAM and SON) show the greatest decrease in IOC with increasing lead time, with Chicago (dashed–dotted line), even having a negative IOC at 30-h lead time during spring months. IOC is consistently highest across the three airports during winter (DJF), when IFR conditions are most prevalent. In contrast, IOC is consistently lowest during summer (JJA).

2) VFR

(i) All seasons

BSSs for VFR conditions across CONUS for all seasons versus lead time are displayed in Fig. 4. Considering all seasons, forecasts for VFR utilizing this analog approach show IOC for all lead times at all airports. Generally, forecasts of VFR show greater improvement over climatology than are seen for IFR. Regionally, the greatest skill is seen in New England, with high skill also seen in the Midwest and much of the South. These areas have many Core 30 airports with values around 30%–40% IOC. As with IFR, BSSs for the Florida airports are much lower than the rest of the airports in the South. The low and fairly constant BSSs seen in the Florida airports are similar to those observed in LAS and DEN. Excluding the West and Florida airports, a diurnal cycle in BSS is observed

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 07/19/23 06:51 PM UTC

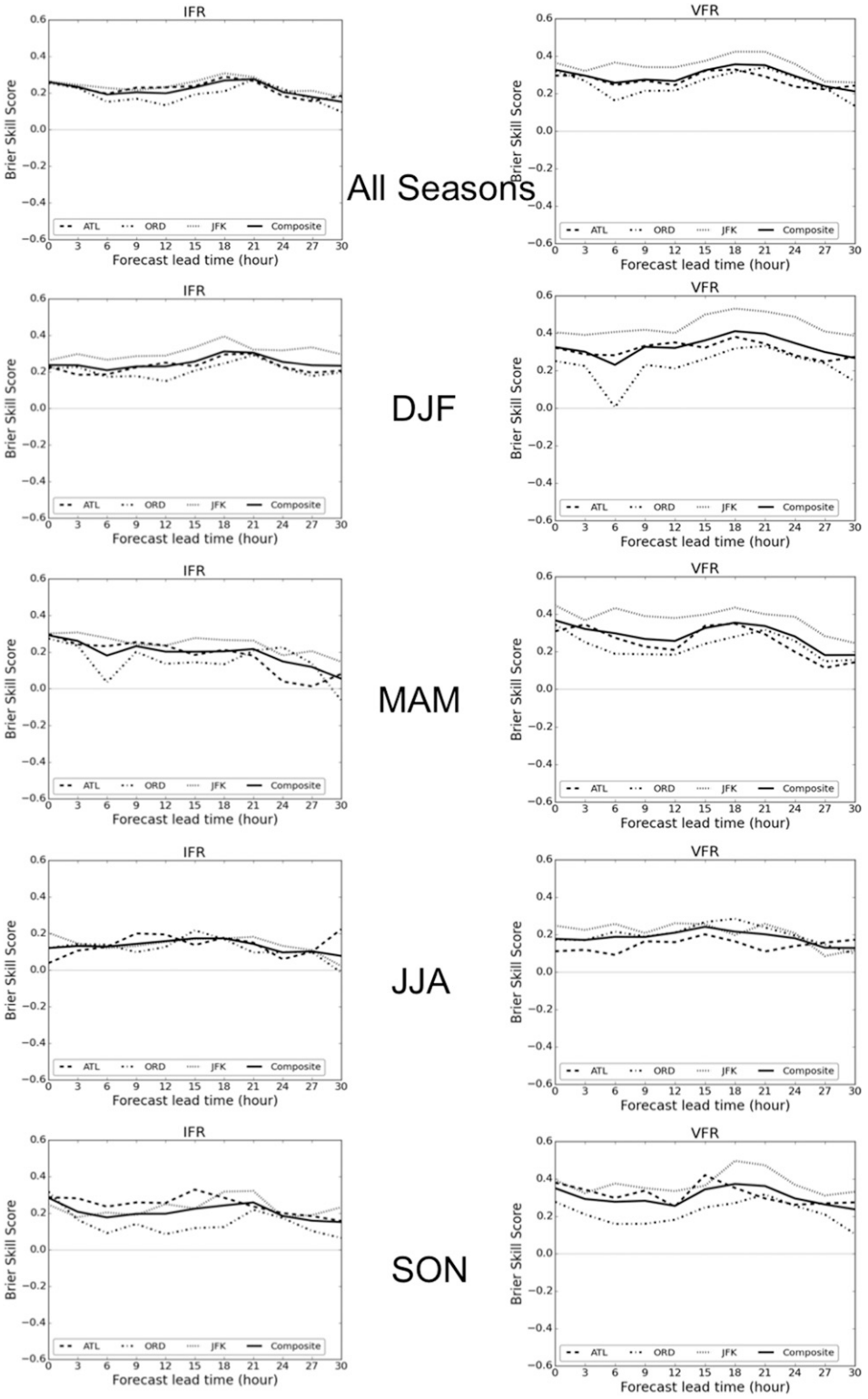

FIG. 3. BSSs (IOC) for the Golden Triangle airports vs forecast lead time for (left) IFR and (right) VFR for (top) all seasons, as well as for each season. All panels show ATL (dashed line), ORD (dashed–dotted line), JFK (dotted line), and a composite of all three airports (solid line).

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 07/19/23 06:51 PM UTC

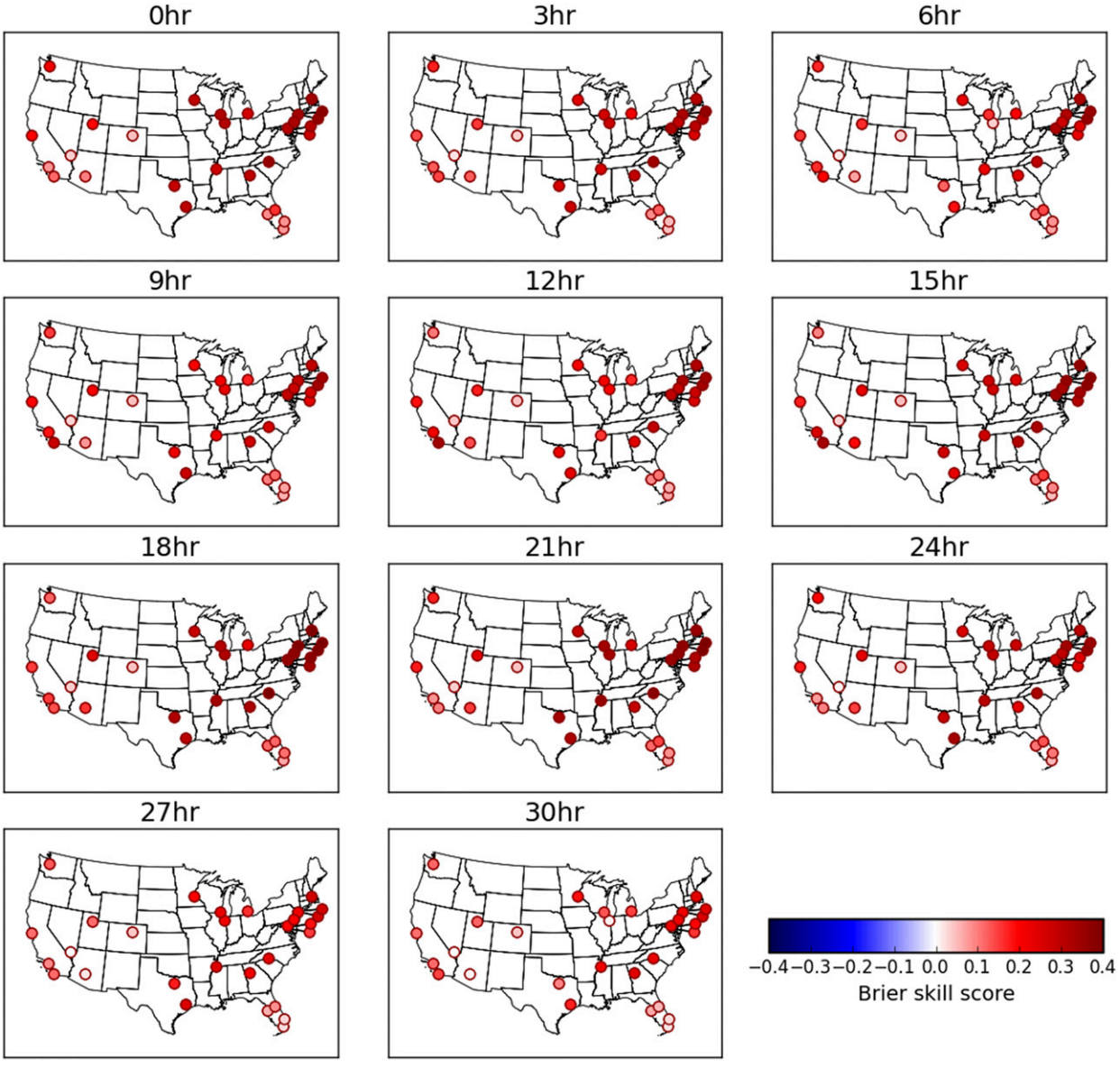

FIG. 4. As in Fig. 2, but for VFR.

with maxima occurring during local late afternoon. Employing a bootstrapping procedure, BSSs for the Golden Triangle airports are determined to have significant improvement over climatology at all lead times at the 95% confidence level.

(ii) Seasons

For the Golden Triangle airports (Fig. 3) every season shows a slight increase in IOC for VFR between 12- and 18-h lead times before decreasing through a 30-h lead time. Skill is most spread among the three airports during winter months (DJF) with JFK performing consistently better than the other two airports. During winter in Chicago at a 6-h lead time IOC dips to nearly 0. As with IFR, the Golden Triangle airports consistently have the lowest IOC for VFR during summer months (JJA).

b. Attributes diagrams

1) CONUS

Attributes diagrams are utilized to further examine the probabilistic skill of this analog forecast approach to forecast IFR and VFR at 0-, 12-, and 24-h lead times by compositing forecasts for the 29 airports across the CONUS for all seasons (Fig. 5). For both IFR and VFR, forecasts for all three lead times show very good

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 07/19/23 06:51 PM UTC

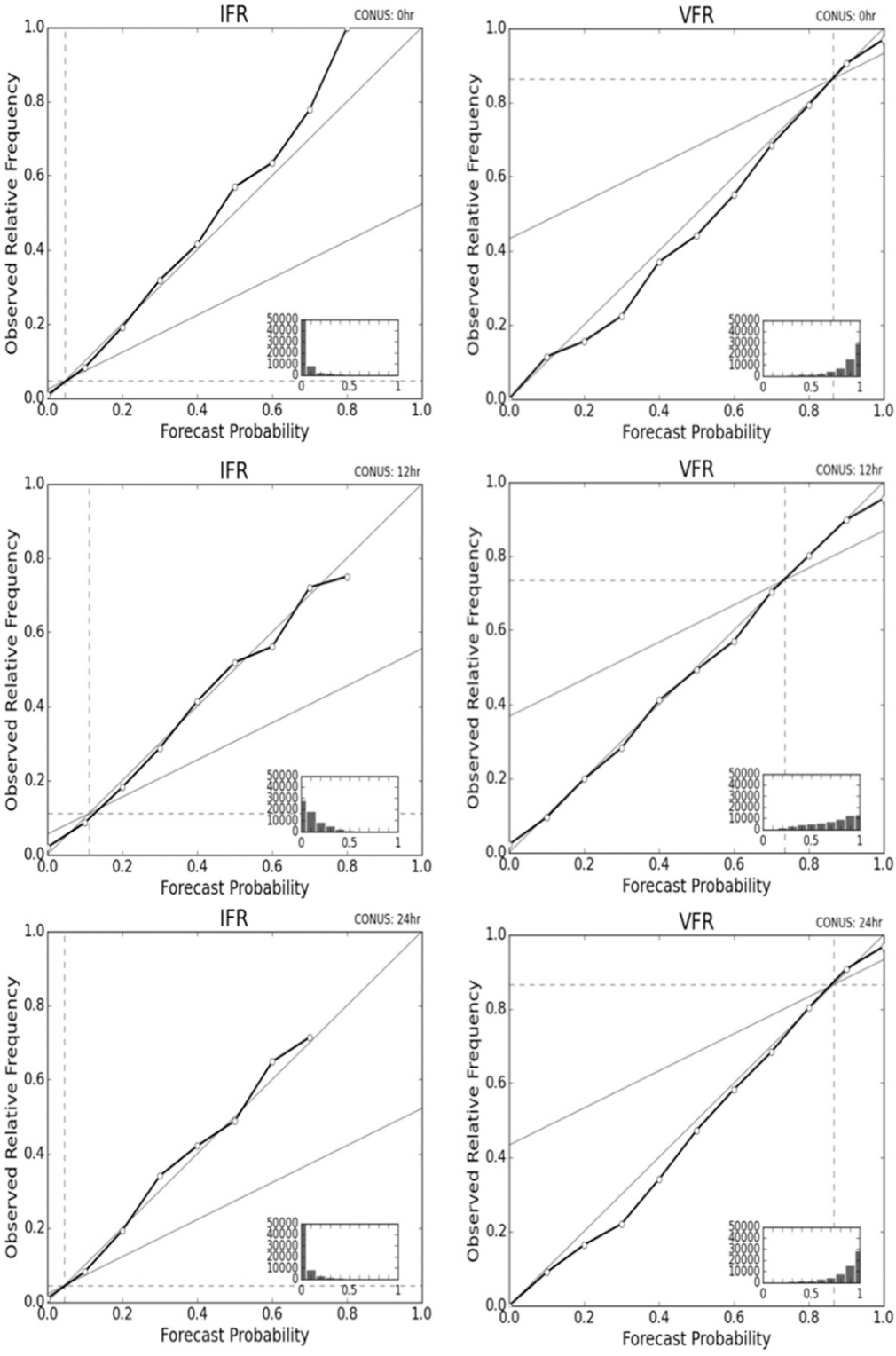

FIG. 5. Attributes diagrams from compositing forecasts for the 29 CONUS Core 30 airports for (top) 0-, (middle) 12-, and (bottom) 24-h lead times for (left) IFR and (right) VFR. Shown is the observed relative frequency (black line) with white filled circles denoting the center of the forecast probability bins. Perfect forecast reliability (1:1) is plotted as a solid gray line for reference. Climatological frequency of conditions is plotted as the horizontal dashed line. The solid gray line falling halfway between the climatological frequency line and perfect reliability line is the ''no skill'' line. Any points that lie on the no skill line do not add to the BSS. The inset bar graph displays the number of observations in each forecast probability bin.

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 07/19/23 06:51 PM UTC

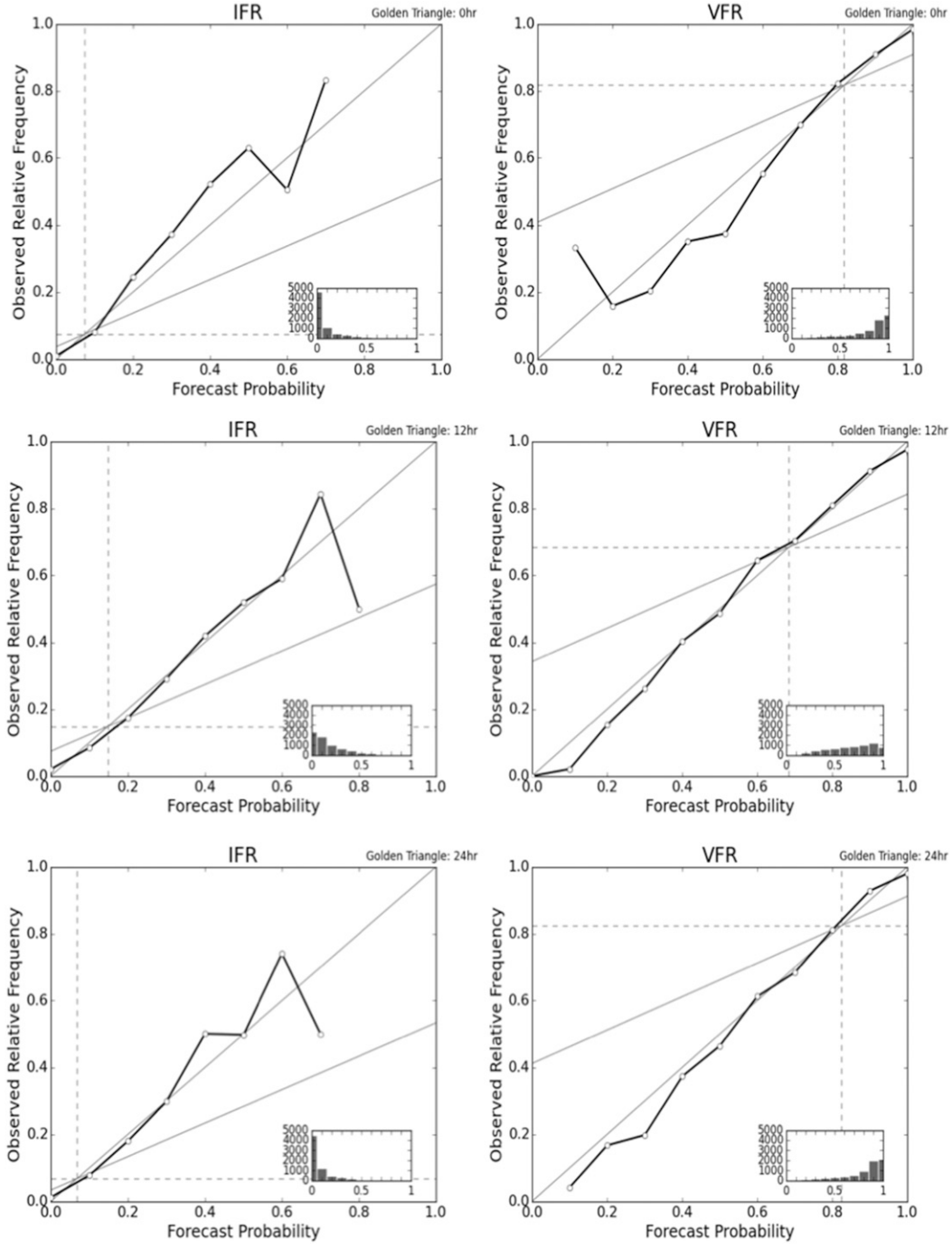

FIG. 6. As in Fig. 5, but for attributes diagrams from compositing the three airports in the Golden Triangle.

reliability and resolution. Forecasts at 12-h lead times show the best calibration, followed closely by 24-h forecasts. Forecasts with a 0-h lead time for IFR have good reliability for the most populated bins and start to underforecast at and above 50% forecast probability. It is unclear why this underforecast is observed at higher probabilities at the initial time, but it could be attributable to overperturbing the initial conditions of the ensemble, the horizontal and vertical resolution of the reforecast data, or some combination thereof. Conversely, at 0-h lead times VFR is slightly overforecasted for probabilities of 20%–60%, which again may have something to do with the perturbations in the GEFS/R.

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 07/19/23 06:51 PM UTC

2) GOLDEN TRIANGLE

Forecasts for three Golden Triangle airports are composited at 0-, 12-, and 24-h lead times to examine IFR and VFR forecast reliability for all seasons (Fig. 6). Generally, forecasts for IFR are fairly reliable at all three lead times. This is a little surprising in that choosing the top 50 matches was somewhat arbitrary, and no further calibration was performed to improve reliability. In forecasting IFR, the reliability (calibration) for the Golden Triangle is best for 12-h lead times. Results at both 12- and 24-h lead times are quite reliable through about 50% and show good resolution through about 70%. As the sample size decreases probabilities above about 50%, the results become a little choppy as a result of decreasing sample size. Unsurprisingly, the number of observations in each forecast probability bin is more evenly distributed across the bins for the 12-h lead time forecasts than the other two lead times. This valid time would be the early morning hours (local time), when climatologically IFR conditions are more prevalent.

Forecasts for VFR are more reliable at the higher probability bins (higher forecast probabilities for VFR) and tend to overforecast the occurrence of VFR at lower probabilities. Overall, this analog method shows very good calibration for forecasting VFR at both 12- and 24-h lead times. Forecasts at 0-h lead times also show very good reliability for the highest forecast probabilities.

c. *Ensemble mean versus TAFs*

Forecast skill relative to TAFs for the period from January 2010 through May 2015 is examined for forecast lead times every 3 h from 6 through 21 h. Recall that all reforecasts are initialized at 0000 UTC; thus, we compare these reforecasts to TAFs issued at 0600 UTC so as to reflect the same model information. BSS versus forecast lead time for our subsample of Golden Triangle airports is shown in Fig. 7 for IFR conditions and VFR conditions.

1) IFR

The composite of the three Golden Triangle airports (Fig. 7, solid line) shows that TAFs outperform the analog forecast method through 1200 UTC, but starting at 1500 UTC the analog forecast method shows an improvement over the TAFs. JFK (dotted line) has a very similar pattern to the composite but with even lower Brier skill scores. Atlanta (dashed line) initially at 0600 UTC has a negative BSS (i.e., TAFs perform better than the analog forecast method); however, just 3 h later through the end of our study period, BSSs become positive with a general increase at the end at 2100 UTC.

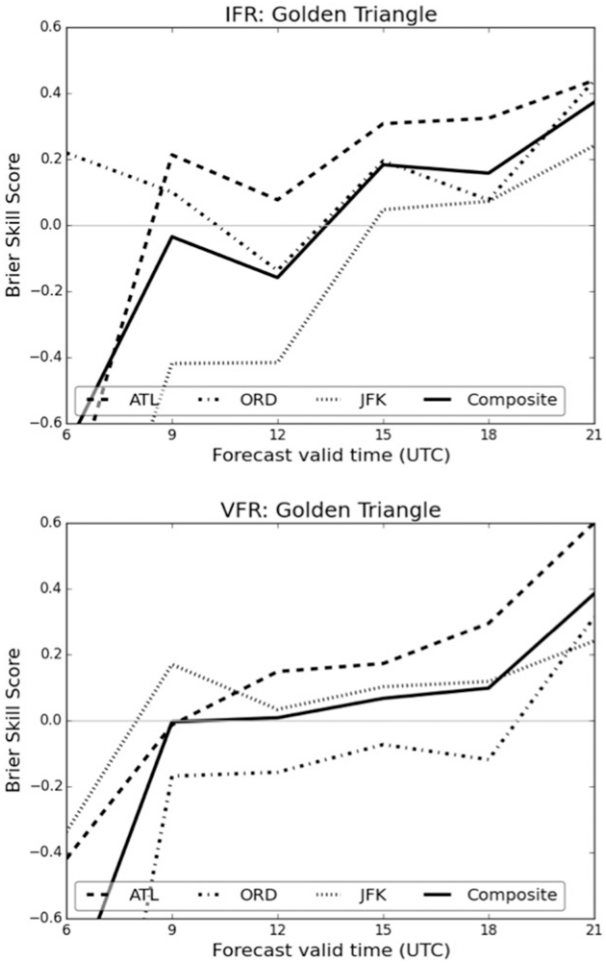

FIG. 7. BSSs (improvement over TAFs) vs forecast valid time calculated over all seasons for (top) IFR and (bottom) VFR. Shown are ATL (dashed line), ORD (dashed–dotted line), JFK (dotted line), and a composite of the three airports (solid line). Sample size is 382 days for each airport.

Unlike the other two airports, Chicago O'Hare (dashed–dotted line) initially has a positive BSS decreasing to a negative value at 1200 UTC then bouncing back to positive again through 2100 UTC. Focusing on the composite BSS, TAFs outperform the analog forecast method through 1200 UTC, but then the analog forecast method begins to outperform the forecasts. BSSs are statistically significant at the 95% confidence level (not shown) for 0600 and 2100 UTC.

2) VFR

For VFR conditions, the composite of our three Golden Triangle airports have a negative BSS at 0600 UTC, then negligible (i.e., TAFs and the analog forecast method are equally skillful at forecasting VFR at 0900 and 1200 UTC) with increasing improvement

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 07/19/23 06:51 PM UTC

over the TAFs through 2100 UTC (Fig. 7, bottom panel). VFR TAFs at ORD (dashed–dotted) outperform the analog forecast method through 1800 UTC until the analog forecast method shows improvement over TAFs at 2100 UTC. In contrast, both ATL and JFK have initially negative BSSs, but starting with forecasts from 1200 through 2100 UTC the BSSs become positive, showing the analog forecast method outperforms the TAFs for VFR forecasts during these forecast times. As with IFR, BSSs for VFR are statistically significant at the 95% confidence level (not shown) for 0600 and 2100 UTC as determined by bootstrapping.

d. *ROC curves*

ROC curves are created for our three Golden Triangle airports for lead times of 6–21 h (Fig. 8). The analog forecasts for VFR (red lines) and IFR (blue lines) and their corresponding area under the curve (corresponding number in color in each panel) show that in all cases the analog forecast method clearly has the ability to discriminate events, whether IFR or VFR, from nonevents. For all periods and each of the three airports, the area under the curve is greater than 0.7 and often greater than 0.8, reflecting a useful predictive discrimination ability. Forecasts for VFR at JFK consistently have higher predictive ability than IFR based on the geometric area under the ROC curve, while the opposite is true at ATL.

5. Discussion

Overall, the analog approach demonstrated here provided skillful improvement over climatology. Results were most positive in areas of flat, homogeneous terrain away from strong coastal, convective, and geographic influences. As might be expected, IOC generally decreased with increasing forecast lead time. A distinct seasonal cycle in IOC was seen at airports across the United States, as highlighted by the results presented here for the Golden Triangle airports, with greatest IOC for both IFR and VFR observed during winter months. Composite attributes diagrams for the Golden Triangle and CONUS demonstrated very good resolution and reliability from these analog forecasts.

Generally, forecasts for airports located in the West and in Florida tended to show the least skill, and at times negative skill relative to climatology, as compared to the other Core 30 airports. Because of the coarse resolution of the available reforecast dataset (1° × 1° horizontally, and surface plus mandatory levels vertically), we postulate that most of these issues arise because of where the bounding latitude–longitude points exist for these airports and the lack of similarity of these points and the location of the airport. For example, values for SEA must be interpolated from data points located nearly in the Strait of Juan de Fuca as well as in the Cascades on the slopes of Mt. Rainier. Likewise, the model tends to struggle for many airports that include bounding boxes with vertices in the ocean, such as SAN, FLL, and LAX, and also mountainous regions such as those surrounding LAS and DEN. It is possible that improvement could be gained by refining the interpolation technique to consider only representative grid point(s); however, as a proof-of-concept study in reforecast applications to aviation forecasting, the most notable and obvious recommendation is to build operational reforecast systems on higher-resolution operational ensemble systems. Low and/or negative skill also occurs at some locations in the Southwest and Intermountain West that receive very little IFR, particularly during the warm season, making it such a rare event that climatology becomes extremely competitive, particularly considering the low resolution of the reforecast. Places that are more geographically homogeneous, such as CLT, EWR, and ATL perform much better with typically higher skill scores throughout. Generally, and perhaps of greatest aviation significance, this postprocessing method performs with rather impressive skill for the Golden Triangle (New York–Atlanta–Chicago) for IFR and VFR throughout the year and for all lead times.

For our subset of Golden Triangle airports (JFK, ATL, and ORD), the analog forecast approach tested here begins to show improvements over TAFs for both IFR and VFR starting at 1500 UTC (9 h after TAF issuance) through the end of our study period at 2100 UTC. The TAF is produced by the NWS forecaster taking into account current observations, radar, satellite, short-term trends, deterministic and ensemble modeling systems of various resolutions, and postprocessed guidance. Its superiority in the short lead time is likely due to the mental and numerical assimilation of these higher-resolution datasets. The analog forecast method's improvement over the TAF after about forecast hour 9 is likely due to the systematic, ensemble-based approach. One would hypothesize that a higher-resolution reforecast system would improve the ensemble prediction at shorter lead times. While we examined TAFs issued only at 0600 UTC, since they are informed by the 0000 UTC models, same as the GEFS/R dataset, it should be noted that TAFs are updated operationally every 6 h. ROC curves computed every 3 h confirm there is some improved predictive skill over TAFs in the analog forecast method.

As mentioned, the low vertical resolution impacts the skill of the model and postprocessing to predict low clouds and visibility. Reforecast grids are archived at most mandatory isobaric levels, resulting in data

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 07/19/23 06:51 PM UTC

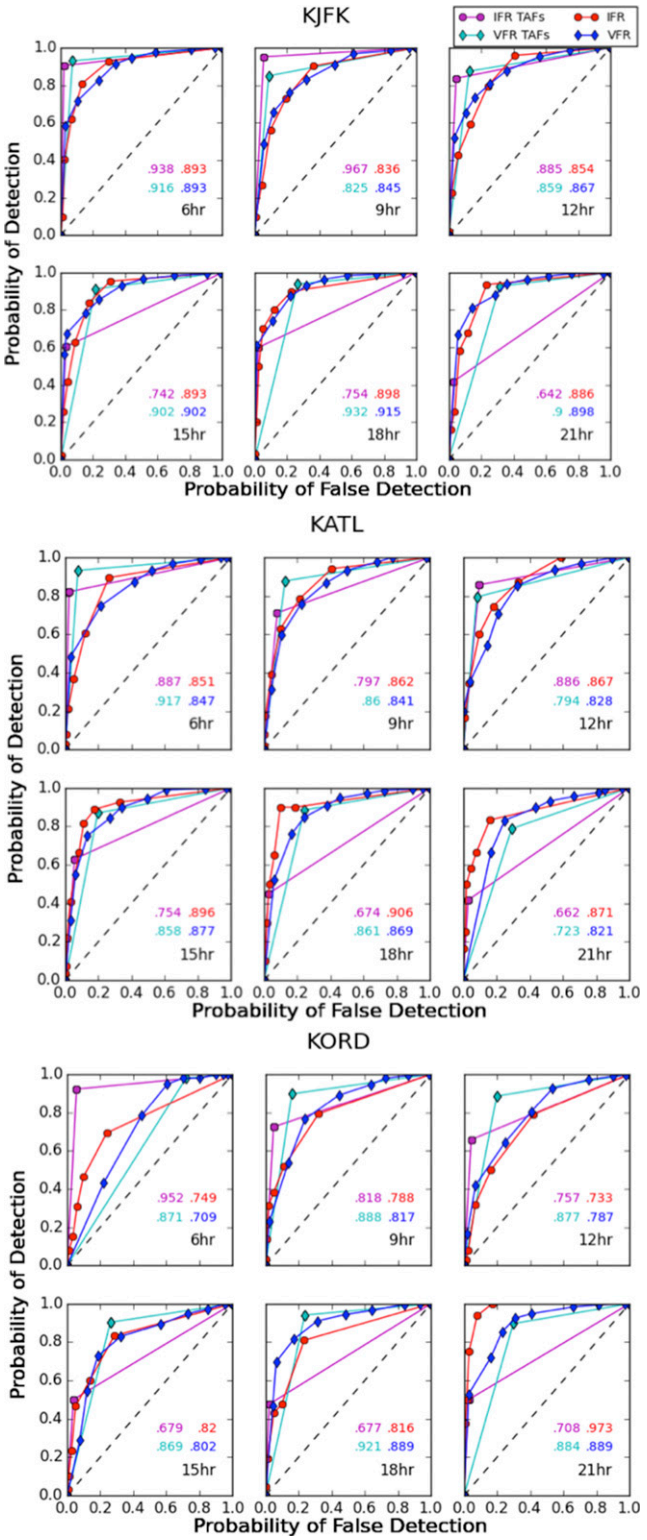

FIG. 8. ROC curves by forecast valid time for (top) JFK, (middle) ATL, and (bottom) ORD showing the curve for the analog forecast method (red) and the area under the curve (blue) for IFR and VFR, and for TAFs the ROC curves (magenta) and the corresponding area under the curve (cyan) for IFR and VFR.

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 07/19/23 06:51 PM UTC

concentrated near the surface but with otherwise large gaps in the atmospheric column. Because of this crude vertical resolution, low cloud layers and inversion height may be poorly simulated and not well represented in the reforecast archive. Likewise, because of the coarse-resolution, MVFR conditions are not reliably identified as a result of the narrow band of the atmosphere that defines the flight condition and, as such, was necessarily left out of the reported results. Increasing vertical resolution (of the native model and archive) would allow for identifying analogs through the inclusion of more levels, providing better identification of moist atmospheric processes. Additionally, increasing vertical resolution would allow for the integration of fog and turbulence models to further aid in the forecasting of surface visibility and low clouds. This would be particularly helpful for air traffic along the West Coast and the Gulf of Mexico, where fog is a major impediment. Considering the low resolution available for constructing historical reforecast analogs and the results presented here, a mesoscale reforecast system with higher-resolution reforecast archives would likely improve results and be a powerful postprocessing resource for aviation forecasting.

6. Conclusions

This research makes an initial foray into analog-type postprocessing of NOAA's second-generation Global Ensemble Forecast System Reforecast for aviation applications. Results show this postprocessing method yields skillful predictions discerning IFR and VFR flight conditions out to 30 h for the majority of Core 30 airports. This is particularly true for those airports in the central and eastern United States, which happen to be most critical to the nation's air traffic flow management.

The overall results are encouraging and suggest reforecasting is a useful approach for aviation forecast postprocessing. Based on this study, the reforecast dataset is suitable for aviation decision support services and underscores the importance of ensemble and reforecast postprocessing as a continuing goal of the NGGPS.

Extrapolating these results beyond this initial study suggests that higher-resolution (i.e., mesoscale or convection allowing) models and accompanying reforecast systems would be of great value to aviation weather postprocessing. Further research should focus on systems with higher vertical and horizontal resolution, optimal methods of analog matching, improved statistical weighting and calibrating of close analogs, ensemble reforecast membership size, and utilizing some or all of the individual members versus using only the ensemble mean. Extensions of the approach could also include additional aviation variables, such as low-level wind shear, mountain waves, icing, and turbulence. In this case, skill was based on sample climatology and a subset of the dataset was compared against TAFs, but expanding the comparison with TAFs and/or calculating skill based on existing statistical guidance would be enlightening. Finally, extracting the most likely deterministic forecast to accompany the probabilistic forecast would be a necessary extension to satisfy the aviation community.

Acknowledgments. The scientific results and conclusions, as well as any views or opinions expressed herein, are those of the authors and do not necessarily reflect the views of the NWS, NOAA, or the Department of Commerce. This research was supported by the NOAA/NWS Next Generation Global Prediction System (NGGPS) project, in cooperation with the NOAA/NWS Aviation Weather Testbed and the NOAA/NWS Western Region. The authors thank Ron Beitel for his invaluable computing insights and enthusiasm and are grateful to three anonymous reviewers who improved the quality of the manuscript. This research would not have been possible without access to the GEFS Reforecast 2 model data provided by the NOAA/Earth System Research Laboratory, METAR data provided by the NOAA/National Centers for Environmental Information, and historical TAF data accessed through Ogimet.

REFERENCES

Baker, S. P., D. F. Shanahan, W. Haaland, J. E. Brady, and G. Li, 2011: Helicopter crashes related to oil and gas operations in the Gulf of Mexico. *Aviat. Space Environ. Med.*, **82**, 885–889, doi:10.3357/ASEM.3050.2011.

Hagedorn, R., T. M. Hamill, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures. *Mon. Wea. Rev.*, **136**, 2608–2619, doi:10.1175/2007MWR2410.1.

Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.*, **134**, 3209–3229, doi:10.1175/MWR3237.1.

——, ——, and X. Wei, 2004: Ensemble re-forecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132**, 1434–1447, doi:10.1175/1520-0493(2004)132<1434:ERIMFS>2.0.CO;2.

——, ——, and S. L. Mullen, 2006: Reforecasts: An important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.*, **87**, 33–46, doi:10.1175/BAMS-87-1-33.

——, G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galameau Jr., Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565, doi:10.1175/BAMS-D-12-00014.1.

——, M. Scheuerer, and G. T. Bates, 2015: Analog probabilistic precipitation forecasts using GEFS reforecasts and climatology-calibrated precipitation analyses. *Mon. Wea. Rev.*, **143**, 3300–3309, doi:10.1175/MWR-D-15-0004.1.

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 07/19/23 06:51 PM UTC

Herman, G. R., and R. S. Schumacher, 2016: Using reforecasts to improve forecasting of fog and visibility for aviation. *Wea. Forecasting*, **31**, 467–482, doi:10.1175/WAF-D-15-0108.1.

Keith, R., and S. M. Leyton, 2007: An experiment to measure the value of statistical probability forecasts for airports. *Wea. Forecasting*, **22**, 928–935, doi:10.1175/WAF988.1.

Lorentson, M., 2013: Scale normalization for IFR-frequency effects in aviation forecast performance statistics. *J. Oper. Meteor.*, **1**, 275–281, doi:10.15191/nwajom.2013.0122.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291–303.

NOAA, 1998: Automated Surface Observing System (ASOS) user's guide. 74 pp., http://www.nws.noaa.gov/asos/pdfs/aum-toc.pdf.

NTSB, 2014: General aviation: Identify and communicate hazardous weather. NTSB most wanted list, https://www.ntsb.gov/safety/mwl/Pages/mwl7_2014.aspx.

NWS, 2016: Terminal aerodome forecasts. National Weather Service Instruction 10-813, NWSPD 10-8, 39 pp., http://www.nws.noaa.gov/directives/.

Reynolds, D. W., D. A. Clark, F. W. Wilson, and L. Cook, 2012: Forecast-based decision support for San Francisco International Airport: A NextGen prototype system that improves operations during summer stratus season. *Bull. Amer. Meteor. Soc.*, **93**, 1503–1518, doi:10.1175/BAMS-D-11-00038.1.

Stanski, H., L. J. Wilson, and W. R. Burrows, 1989: Survey of common verification methods in meteorology. Atmospheric Environmental Service Research Rep. MSRB 89-5, WWW Tech. Rep. 8, WMO/TD-358, 114 pp.

Toth, Z., 1989: Long-range weather forecasting using an analog approach. *J. Climate*, **2**, 594–607, doi:10.1175/1520-0442(1989)002<0594:LRWFUA>2.0.CO;2.

Van den Dool, H. M., 1989: A new look at weather forecasting through analogues. *Mon. Wea. Rev.*, **117**, 2230–2247, doi:10.1175/1520-0493(1989)117<2230:ANLAWF>2.0.CO;2.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences.* 3rd ed. Elsevier, 676 pp.

——, and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.*, **135**, 2379–2390, doi:10.1175/MWR3402.1.

Zapotocny, T. H., and Coauthors, 2000: A case study of the sensitivity of the Eta Data Assimilation System. *Wea. Forecasting*, **15**, 603–621, doi:10.1175/1520-0434(2000)015<0603:ACSOTS>2.0.CO;2.

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 07/19/23 06:51 PM UTC