An Investigation of Reforecasting Applications for NGGPS Aviation Weather Prediction: An Initial Study of Cloud and Visibility Prediction

Kathryn L. Verlinden
College of Earth, Oceanic, and Atmospheric Sciences
Oregon State University
Corvallis, Oregon

David R. Bright
NOAA/National Weather Service
Portland, Oregon

Abstract

This study is the first known aviation-based application of NOAA's second-generation Global Ensemble Forecast System (GEFS) reforecast (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. This type of post-processing could one day form the backbone of the National Weather Service's common operating procedure (COP) for aviation.

Results from this study show that even at the very coarse resolution of the GEFS reforecast dataset, the analog approach yielded skillful probabilistic forecasts of IFR and VFR flight conditions at most of the FAA's Core 30 airports. This was particularly true over the central and eastern U.S., 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 the modeling system and reforecast archive would be very useful for aviation forecasting applications.

1.0 Introduction

The Next Generation Global Prediction System (NGGPS) is a 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 project utilizes NOAA's 2nd-Generation Global Ensemble Forecasting System (GEFS) to explore cloud ceiling and visibility prediction at major airports across the United States. While numerous studies have demonstrated the value of reforecasting for ensemble post-processing and decision support (e.g., Hamill et al. 2006, 2013, 2015; Wilks and Hamill 2007; Hagedorn et al. 2008), none have been specific to aviation.

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 feet above ground level and/or a visibility less than 3 miles, 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 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 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.

This study takes a preliminary look at downscaling NOAA's 2nd-generation global ensemble reforecast dataset to the FAA's Core 30 airports (Table 1). The Core 30 airports 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 (HNL).. A model climatology and probabilistic ceiling and visibility forecasts through 30 hours are created using an analog reforecast approach. Thirty hours was chosen because it encompasses the 24-hour period of most Terminal Aerodrome Forecasts (TAFs), and in the aviation community is a reasonable traffic flow management outlook period. Similar to the work of Hamill et al. (2004), the ensemble mean reforecasts are used for determining analogs at all airports; additionally, an approach using all 11 individual ensemble members for Hartsfield-Jackson Atlanta International Airport (ATL) is tested. Historical METAR observations of ceiling and visibility at each airport serve as ground truth.

This report provides a description of the datasets, an overview of the data post-processing including the analog forecast approach, and the statistical methods employed. This is followed by a summary of the analog forecast system results for predicting IFR and VFR, broken down regionally and seasonally. A description of the results for the individual members of the ensemble forecast for ATL follows in the same manner. This is followed by a general discussion of possible reasons for observed differences between the regional subsets, and suggestions for future refinements and potential applications for the aviation weather community.

2.0 Data

2.1 Model Data

This effort utilizes the entire 30 years of NOAA's 2nd-generation global ensemble reforecast data set, which uses the identical modeling system as NOAA/NCEP 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 present, these reforecasts have been made available at 3-hourly intervals for lead times of 0-72 hours, and then 6-hourly intervals out to 16 days. Global fields at 1ºx1º latitude-longitude resolution for 98 different fields are forecast; many of these fields are surface fields, but temperature, specific humidity, wind components, and geopotential height are also available at isobaric and/or hybrid vertical levels. This dataset can be accessed at: http://www.esrl.noaa.gov/psd/forecasts/reforecast2/ . For further information regarding the reforecast dataset see Hamill et al. (2013).

For this study, daily forecasts from1 December 1984 through 31 May 2015 made every 3 hours with lead times out to 30 hours are used. Utilized fields include: surface pressure, temperature at 2m and available isobaric pressure levels (1000mb, 850mb, 700mb, 500mb), and specific humidity at 2m and available isobaric pressure levels (1000mb, 850mb, 700mb, 500mb). The ensemble mean is used at all Core 30 airports, and for comparison the ensemble members are also used at ATL.

2.2 Observational Data

For the same time period, METARs at the Core 30 airports are used as ground truth for ceiling height and surface visibility. Data were accessed through NCDC's Climate Data Online portal at www.ncdc.noaa.gov.

3.0 Methods

3.1 Data Preparation

For each airport, the four surrounding datapoints are stripped. Bilinear interpolation is applied to the four points to estimate the forecast value at the airport's location. Relative humidity (RH) is calculated at every pressure level from the forecast temperature (T), saturation mixing ratio (w_s) and specific humidity (q) fields via

$$RH \approx 100\frac{q}{w_s} \approx (100q)\frac{100p}{.622 * 611e^{\frac{17.67(T-273K)}{T-29.65}}}.$$

Vertical profiles of dew point temperature (T_d) are derived from the moisture and temperature fields via

$$T_d = \frac{243.04°C\left[ln\left(\frac{RH}{100}\right) + \frac{17.625T}{243.04°C+T}\right]}{17.625 - ln\left(\frac{RH}{100} - \frac{17.625T}{243.04°C+T}\right)}.$$

METAR observations at the forecast valid times are used for analog downscaling and verification (i.e., every third hour from 00 UTC through forecast hour 30). In very rare instances, when reported observations were not available on the hour, values were linearly interpolated in time to create an on-the-hour observation. The primary METAR observations of interest are cloud ceiling height and visibility. The reported ceiling height and visibility observations are then classified into flight regulation categories (Table 2).

3.2 Analog Forecasts

An analog approach is used to identify similar historical reforecasts to downscale the global reforecast to a point (in the manner of, e.g., Toth 1989; Van den Dool 1989). Vertical profiles ("soundings") of temperature and dew point temperature are created by concatenating model output grids at 2-meters above the surface, 1000mb, 850mb, 700mb, and 500mb. If the modeled surface pressure is less than any of the isobaric grid levels, then those levels are removed from the sounding. Every fifth day starting with 1 December 1984, the forecast sounding at a given lead time is compared to all historical reforecast soundings at the same lead time via a normalized root-mean-squared difference (RMSD). Variables are normalized at each pressure level by typical errors assigned in the Eta Data Assimilation/Forecast System (EDAS) to rawinsonde observations (see Zapotocny et al. 2000 for values). The equation for determining the normalized RMSD is

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

where T is temperature, Td is dew point temperature, p is the vertical level, m is the model (reference sounding) value, r is a reforecast value, and e is the representative measurement error. The 50 soundings corresponding to the smallest normalized RMSD are considered analog forecast matches. The quantity of 50 analog soundings is chosen so as to provide reasonable sample size without causing over-filtering, and has previously been identified as adequate for the short forecast lead times considered here (Hamill et al. 2015). This process is repeated every fifth day through the entire reforecast period. Every fifth day is used to avoid oversampling any single weather regime. Data denial is employed for verification and validation of the technique, and 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 category at the airport. Probability of observed flight categories are then determined using these observations. For example, if 20 of the top 50 matched soundings' METARs report IFR conditions, 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 were not tested here. Because 50 matches is a somewhat arbitrary choice, perfectly reliable probabilities are unlikely without further calibration.

3.3 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 probability forecast for a dichotomous event normalized by the same for a reference

forecast, in this case the observed sample climatology constructed for December 1984 through May 2015 (Wilks 2011). Assuming that each forecast is equally likely, the Brier Score of the forecast, BS_f, is calculated as

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

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

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

where BS_r is the Brier Score of the reference probability forecast, in this case the constructed (sample) climatology. A BSS of 0.0 indicates that the forecast has the same skill as climatology, a BSS of 1.0 indicates a perfect forecast, and negative BSS indicates less skill than climatology. A BSS can be interpreted as a percent improvement over the reference dataset.

3.4 Ensemble members

At ATL a deeper analysis considering all 11 ensemble members is tested. The above steps described in 3.2 and 3.3 are completed as before, but rather than using the ensemble mean, analogs for each ensemble member are found. For each date-lead time forecast, the 50 analogs for each ensemble member are combined for a total of 550 days (allowing repeats) when calculating probabilities. From this set of 550 days, the subset consisting of only the unique analog dates for each forecast lead time is also considered.

4.0 Results

4.1 Ensemble Mean

BSS versus forecast lead time for each airport is shown in Figures 1-6 and 7-12 for IFR conditions and VFR conditions, respectively. Note: MVFR conditions were examined but are not included in these results due to the coarse vertical resolution of the model and archive, and, thus, inability to resolve thin slices of lower tropospheric fields. Recall that all forecasts are initialized at 0000 UTC. To more easily examine the results, the Core 30 airports are categorized into regions: West, Midwest, New England, and South (Table 1). Forecast skill relative to the sample climate of the study period (December 1984 through May 2015 for all seasons, or for only the applicable months within this time period for each particular season; hereafter simply referred to as "climatology") is examined for each forecast lead time for the entire record as well as for each season.

4.1.1 IFR

4.1.1.1 All seasons

The GEFS Reforecast 2 considered through this analog downscaling method shows skillful improvement over climatology (IOC) for forecasting IFR conditions at the majority of the Core 30 airports for all forecast lead times (Figures 1 and 2). This is particularly the case for airports in the Midwest, New England, and 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 South, except Florida and Memphis, show a diurnal cycle in skill with maximum occurring during late afternoon and early evening local time. Forecast skill for the Florida airport cluster remain 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.

4.1.1.2 Winter (DJF)

BSS versus forecast lead time for just the winter months (DJF) for each airport separated by geographic region are shown in Figure 3. Throughout all lead times forecasts for the majority of airports show IOC with the exception of three in the West (SEA, DEN, PHX) and three in Florida (FLL, MIA, TPA). In particular opposition to all seasons, DEN at a forecast lead time of 9 hours has a BSS of -0.15. In contrast to all seasons, only considering the winter months, BSSs generally show a very low amplitude diurnal cycle, with the exception of forecasts for airports in the Midwest. Additionally, BSSs generally do not show a marked decrease in skill between 0-hour and 30-hour lead times as they did in all seasons.

4.1.1.3 Spring (MAM)

BSSs for IFR conditions versus lead time are shown in Figure 4 for spring months (MAM). For the majority of airports, BSSs are highest during spring at early lead times compared to other or all seasons, and particularly New England and the Midwest show the greatest decrease (~20%) in BSS by 30-hour lead times. In the South, BSSs for the Texas airports more closely match the continued low and nearly flat BSSs of the Florida cluster (lower right panel, dashed lines). BSSs for CLT show the greatest IOC for all airports during spring, and compared to all times. Forecasts for airports in the West are nearly constant throughout lead times during spring, except for SLC which has large jumps in BSS during early lead times and some of the largest negative skill scores reported in this study. The saw tooth pattern at SLC is characteristic of a small sample size (i.e., very little IFR) consistent with the climatological frequency of 0.5-2%, depending on time of day. PHX and LAS both report intermittent BSSs. This is a result of IFR not being observed during these lead times, causing a zero denominator in the BSS equation and an undefined value being reported, and thus no value plotted.

4.1.1.4 Summer (JJA)

Among the seasons, summer contains the most variability and least distinct patterns of forecast skill scores throughout the lead times (Figure 5). In New England, slight negative BSSs are observed in IAD, DCA, and PHL during local

evening hours (forecast hours 15-24), while the remaining airports remain positive throughout, but lower than 'all season' values. In the West, PHX has no observed IFR during the majority of daylight hours (12-24 UTC) and both SLC and LAS have no observed IFR at any time. In the South, similar to spring, the Texas airports, MEM, and the Florida airports dip in and out of negative BSSs.

4.1.1.5 Autumn (SON)

Figure 6 is as above, but for autumn months (SON). On the whole, autumn is a clear transition in values and pattern of BSS between the low, nearly constant values of summer and higher, slightly cyclic BSSs of winter. For all airports in New England, forecasts show skill over climatology for all lead times. In addition, BSSs for New England show a strong diurnal cycle with the greatest skill during late afternoon. Forecast skill for airports in the South show separation between the cluster of airports in Florida and the other southern airports starting around 12-hour lead times, although all airports show a decrease (~7%) in BSS with increasing lead time. BSSs for airports in the West appear unorganized, with IFR only observed in LAS during autumn during 21- and 24-hour lead times (local evening hours).

4.1.2 VFR

4.1.2.1 All seasons

BSSs for VFR conditions for all seasons versus lead time are displayed in Figures 7 and 8. 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 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 well separated from 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 with maxima occurring during local late afternoon.

4.1.2.2 Winter (DJF)

During winter the diurnal cycle observed when considering all seasons remains particularly prevalent in New England and the Midwest, with BSSs in New England being nearly twice as large (Figure 9). In the South, forecasts for Florida show increasing skill compared to 'all season' BSSs, thereby decreasing the separation in skill between them and the other southern airports. For all airports a decrease in skill is observed with increasing forecast lead time.

4.1.2.3 Transitional Seasons (MAM, SON)

During spring all airports, except those in Florida and most in the West, show a decrease in BSS with increasing lead time (Figure 10). In New England, there is a dramatic drop in BSS (~25% absolute drop in IOC) between 21- and 27-hour lead times for nearly all airports. A slight diurnal cycle in BSS is apparent for airports in New England, the Midwest, and the South, except FLL and MIA. Negative BSSs are observed in PHX for long lead times, and early leads times for FLL. BSSs are fairly consistent for airports in the West, with a slight increase in skill for California airports (SFO, LAX, SAN) during local early morning hours.

Patterns of BSS during autumn months (Figure 12), similar to spring, show a more organized regime, particularly in New England, with a clustering of BSS values and clear diurnal cycle. In contrast, the largest separation between skill scores for the Florida airports and those in the other southern airports occurs during autumn, similar to the pattern observed when considering all seasons. DEN and LAS, in the West, show similar low BSSs as those calculated for the Florida airports. During autumn, SLC and LAS are the only airports to have negative BSSs. As with the other seasons, BSSs in the West are generally lower than the other regions, and BSSs are generally highest across New England.

4.1.2.4 Summer (JJA)

Forecasts resulting from this analog method show the lowest IOC for VFR during summer in comparison to the other seasons (Figure 11). BSSs for airports in the West and the South show little organization, with negative skill observed at various lead times for SLC, PHX, LAS, MIA, TPA, and IAH. In the West, BSSs for LAS hover around zero (equally as good as using climatology) at all lead times, while PHX, which has nearly 100% VFR climatologically throughout all hours but reference Brier Score of 0, reports an undefined BSS through the late afternoon and evening hours. BSSs for New England and the Midwest airports generally show a decrease in skill with increasing lead time. Additionally, BSSs for airports in the Midwest show a slight diurnal cycle with maximum improvement over climatology seen during afternoon hours, and minima during late night hours.

4.1.3 Attributes Diagrams

4.1.3.1 Golden Triangle

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-hour lead times by compositing forecasts for the five airports in the Golden Triangle for all seasons (ATL, MDW, ORD, JFK, LGA; Figure 13). 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 highest reliability (best calibration) for the Golden Triangle exists for 12-hour lead times. Results at both 12-hour and 24-hour lead times are quite reliable through about 50% and show good resolution through about 70%. As the sample size decreases above about 50%, the results become a little choppy due to decreasing sample size. Unsurprisingly, the number of observations in each forecast probability bin is more evenly distributed across the bins for the 12-hour lead time forecasts than the other two lead times. This would be the early morning hours (local time) when climatologically IFR conditions are more prevalent.

Similarly, forecasts for VFR are more reliable at the higher probability bin (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-hour and 24-hour lead times. Forecasts at 0-hour lead times also show very good reliability for the highest forecast probabilities, but overforecast for 50% probability and below.

4.1.3.2 CONUS

All 29 airports are composited at 0-, 12-, and 24-hour lead times to examine IFR and VFR forecast reliability across the CONUS for all seasons (Figure 14). For both IFR and VFR, forecasts for all three lead times show very good reliability and resolution. Forecasts at 12-hour lead times show the best calibration, followed closely by 24-hour forecasts. 0-hour lead time forecasts for IFR have good reliability for the most populated bins, and start to underforecast at and above 50% forecast probability. Conversely, at 0-hour lead times VFR is overforecasted for probabilities of 20-60%. This composite of all 29 airports shows increased reliability in forecasting IFR versus the Golden Triangle, while reliability is very similar for VFR. Interestingly, both composites show similar over- and underforecasting tendencies for 0-hour lead times.

4.2 Ensemble Members: ATL

For Figures 13 and 14, the forecast BSSs versus forecast lead time for all 550 analogs (50 for each ensemble member) are in the left panel, BSSs resulting from using just the unique analog dates from the larger pool of 550 analogs are in the middle panel, the BSSs from the ensemble mean are on the right, and the observed relative frequencies of the VFR or IFR prediction for the lead time are displayed in the bottom panels. In all panels, BSSs for winter (DJF) are solid black lines, spring (MAM) are dashed, summer (JJA) are dash-dot, autumn (SON) are dotted, and all seasons (entire record) are solid gray.

4.2.1 IFR

Considering forecasts across all seasons (Figure 15, solid gray lines), skill scores from ensemble members forecasting IFR (right and middle panel) are positive at all times with a minimum in skill occurring during local night and maximum skill over climatology at 0- and 18-hour lead times with a generally downward trend in skill with increasing forecast lead time. This pattern is consistent with that of BSSs utilizing the ensemble mean (right panel). The transitional seasons (MAM, dashed line; SON, dotted line) also show this generally downward trend in forecast skill with increasing lead time. BSSs for IFR are greatest during winter in the late afternoon (~35% IOC). There are minor differences in skill between using 550 analogs and utilizing only unique analogs, with the unique analog method showing slight increases in IOC compared to all analogs during the early forecast lead times. Both analog methods utilizing the ensemble members show minor increases in skill over the ensemble mean at longer lead times. IOC does not appear to be entirely related to IFR observed relative frequency (bottom panels) during any or all of the seasons, except perhaps summer, but this relationship, or lack thereof, requires further study.

To further investigate the probabilistic skill of each of these approaches, forecasts for all seasons at ATL are investigated through the use of attributes diagrams at 0-, 12-, and 24-hour lead times (Figure 17). These attributes diagrams show that all three methods show skill across all forecast probabilities with few exceptions (i.e., ensemble members using 550 analogs at 24-hour lead times for 20% and 70% probabilities and using unique analogs at 12-hour lead times with 30% forecast probability, and ensemble mean for 12-hour lead time and 80% forecast probability). For all three investigated lead times each method stays well away from the sample climatology (horizontal dashed lines). As such, all methods have a good degree of resolution and are able to discern events with different frequencies of occurrence through 60% at which point resolution for the ensemble mean and ensemble member unique analog methods start to break down. Generally, all three methods are best calibrated at 12-hour lead times when the

number of observed IFR are more evenly spread across forecast probabilities. At 0- and 12-hour forecast lead times the 550 analogs method shows the best calibration, while the methods using ensemble member unique analogs and the ensemble mean both consistently underforecasting for 0-hour lead times. For 24-hour forecast lead times using the ensemble mean is the most reliable at low forecast probabilities where the highest number of forecast IFR conditions occur (inset bar graphs), although all three methods jump between under and over forecasting as forecast probability increases.

4.2.2 VFR

For VFR conditions (Figure 16), utilizing this analog post-processing method with ensemble members over all seasons (solid gray line) there is skill in forecasts over climatology for all lead times (left and middle panels), with a slight decrease (~5%) in IOC with increasing forecast lead time. BSSs computed for autumn (dotted lines) and winter (solid black lines) show similar patterns but with greater magnitude to that of all seasons, with autumn having slightly higher BSSs, particularly for the shortest lead times. The spring months (MAM, dashed lines) have BSSs that drop off steeply after a maximum skill at a lead time of 18 hours (valid time of 18 UTC) for all three post-processing methods. Summer months show the least skill in forecasting VFR. Although BSSs remain positive, skill is often at half of that observed for other seasons at similar times, particularly during short forecast lead times. For all seasons, and through each season, the ensemble member approaches slightly outperform the ensemble mean approach (right panel) by approximately 2-3%. There do not appear to be any obvious correlations between BSSs and observed relative frequencies of VFR.

Attributes diagrams (Figure 18) support all three analog forecast approaches showing skill in forecasting VFR at 0-, 12-, and 24-hour lead times. The exception to this general observation is 60% or 70% forecast probability for 12-hour lead times for all three methods. It is unclear why this consistent deviation in skill occurs. The method utilizing 550 analogs from the ensemble members has nearly perfect

calibration at 24-hour forecast lead time and good resolution at the other forecast lead times, with less resolution when the least VFR forecasts are present. Interestingly, the ensemble member unique analog method and the ensemble mean tend to slightly overforecast VFR at 0-hour lead times, and in general, compared to the 550 analogs from ensemble members method at all three forecast lead times.

5.0 Discussion

Overall, the analog approach demonstrated here provided skillful IOC. 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 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.

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. Due to the coarse resolution of the available reforecast dataset (1^{o} by 1^{o} horizontally, and surface/mandatory levels vertically), we postulate that most of these issues arise due to 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 Cascade Mountains 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 mountainous regions such as those surrounding LAS and DEN. Also, some locations in the Southwest and Intermountain West 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 post-processing 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.

As mentioned, the low vertical resolution also impacts the skill of the model and post-processing to predict low clouds and visibility. Reforecast grids are archived at most mandatory isobaric levels, but this leaves the surface fields and large gaps in the atmospheric column. Due to 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 be reliably identified due to the narrow band of the atmosphere that defines the flight rule category and, as such, was necessarily left out of the reported results. Increasing vertical resolution (of the native model and archive) will allow for identifying analogs through the inclusion of more levels providing better identification of moist atmospheric. 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 post-processing resource for aviation forecasting.

While this study only takes a cursory look at forecasts considering all ensemble members at a single airport (ATL), the individual member approach may provide a slight improvement over utilizing just the ensemble mean. This is based on the improvement in VFR predictions annually and seasonally, and IFR predictions when considered annually. Given this is only one airport, more extensive research would need to be undertaken to determine whether individual members or the ensemble mean is superior.

6.0 Conclusions

This research makes an initial foray into analog-type post-processing of NOAA's 2nd-Generation Global Ensemble Forecast System Reforecast for aviation applications. Results show this post-processing method yields skillful predictions discerning IFR and VFR flight conditions out to 30-hours for the majority of Core 30 airports. This is particularly true for those airports in the central and eastern U.S., 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 post-processing. Based on this study, the reforecast dataset is suitable for aviation decision support services, and underscores the importance of ensemble and reforecast post-processing 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 post-processing. 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 members vs. 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, but more competitive skill metrics such as those based on actual TAFS and 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.

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.

Hagedorn, R. 2008: Using the ECMWF reforecast data set to calibrate EPS reforecasts. *ECMWF Newsletter*, **117**, ECMWF, Reading, United Kingdom, 8-13.

Hamill, T., J. S. Whitaker, and X. Wei, 2004: Ensemble re-forecasting: Improving medium-range forecast skill using retrospective forecasts. *Monthly Weather Review*, 132, 1434-1447.

Hamill, T. M., J. S. Whitaker, and S. L. Mullen, 2006: Reforecasts: An important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.*, **87**, 33-46.

Hamill, T. M., 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 meadium-range ensemble reforecast dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553-1565.

Hamill, T. M., M. Scheuerer, and G. T. Bates, 2015: Analog probabilistic precipitation forecasts using GEFS reforecast and climatology-calibrated precipitation analyses. *Monthly Weather Review*, **143**, 3300-3309.

NTSB, 2014: General aviation: Identify and communicate hazardous weather. NTSB most wanted list. available at: https://app.ntsb.gov/safety/mwl2014/07_MWL_GAweather.pdf

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

Toth, Z., 1989: Long-range weather forecasting using an analog approach. *J. Climate*, **2**, 594-607.

Van den Dool, H. M., 1989: A new look at weather forecasting through analogues. *Monthly Weather Review*, **117**, 2230-2247.

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

Wilks, D. S. and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Monthly Weather Review*, **135**, 2379-2390.

Zapotocny, T. H. and Coauthors, 2000: A case study of the sensitivity of the Eta Data Assimilation System. *Weather and Forecasting*. **15**, 603-621.

Region	Callsign	Airport Name
South	ATL	Hartsfield-Jackson Atlanta Intl
	CLT	Charlotte Douglas Intl
	DFW	Dallas/Fort Worth Intl
	FLL	Fort Lauderdale/Hollywood Intl
	IAH	George Bush Houston Intercontinental
	MCO	Orlando Intl
	MEM	Memphis Intl
	MIA	Miami Intl
	TPA	Tampa Intl
West	DEN	Denver Intl
	LAS	Las Vegas McCarran Intl
	LAX	Los Angeles Intl
	PHX	Phoenix Sky Harbor Intl
	SAN	San Diego Intl
	SEA	Seattle/Tacoma Intl
	SFO	San Francisco Intl
	SLC	Salt Lake City Intl
Midwest	DTW	Detroit Metropolitan Wayne County
	MDW	Chicago Midway
	MSP	Minneapolis/St. Paul Intl
	ORD	Chicago O'Hare Intl
New England	BOS	Boston Logan Intl
	BWI	Baltimore/Washington Intl
	DCA	Ronald Reagan Washington National
	EWR	Newark Liberty Intl
	JFK	New York John F. Kennedy Intl
	IAD	Washington Dulles Intl
	LGA	New York LaGuardia
	PHL	Philadelphia Intl

Table 1: The Core 30 airports across CONUS by region. Golden Triangle airports are highlighted in gray.

Flight Condition	Ceiling (feet)	Visibility (SM)
IFR	< 1000	< 3
MVFR	≥ 1000 & ≤ 3000	≥ 3 & ≤ 5
VFR	> 3000	> 5

Table 2: Flight rule condition definitions. Conditions are defined on an and/or basis with the lowest visibility or ceiling defining the current flight rule conditions.

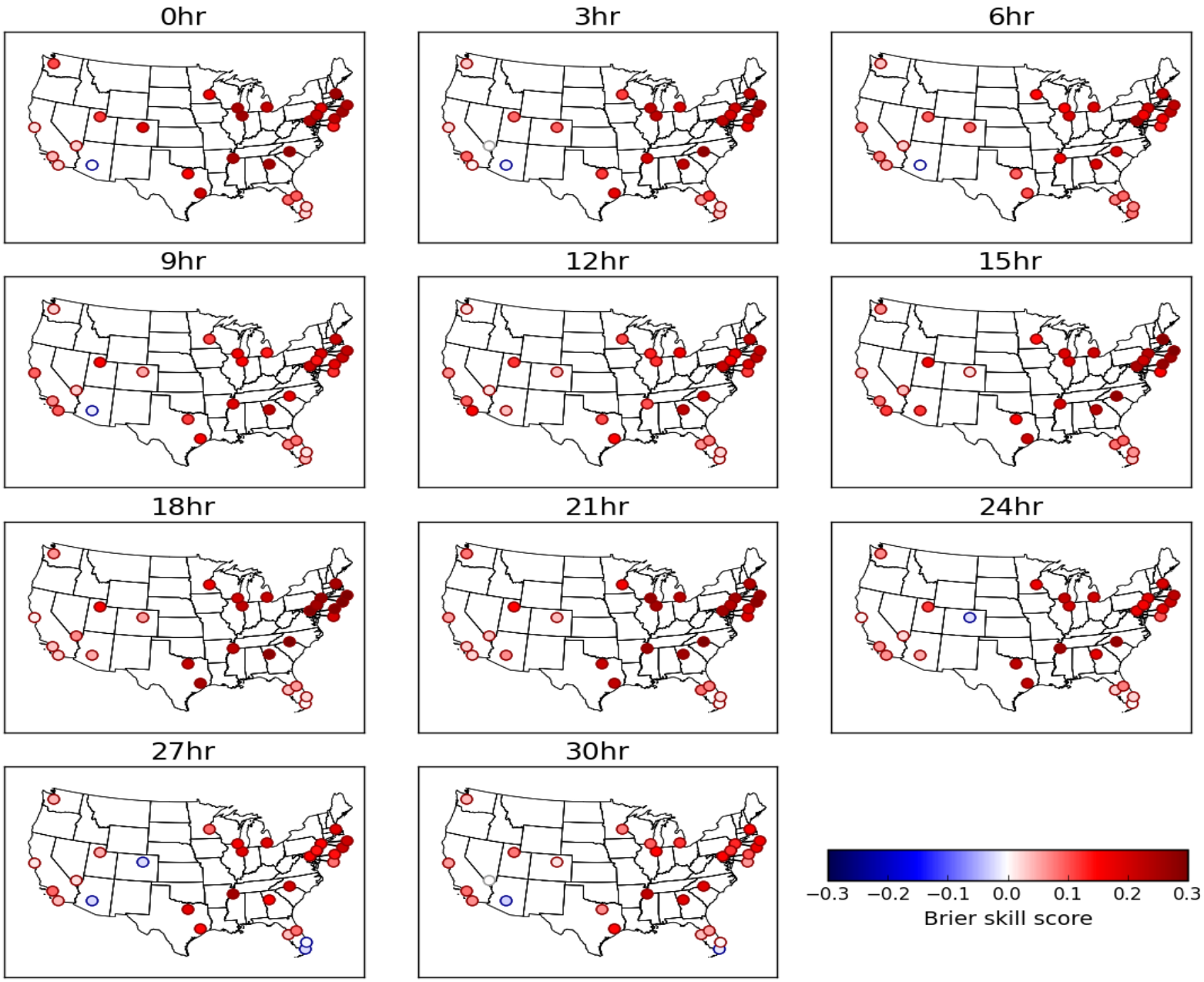

Figure 1: Brier skill scores computed for all seasons for IFR at the Core 30 airports across CONUS at each forecast lead time. Perimeters of the circles denote positive (red) or negative (blue) skill. Shading within each circle denotes the skill score magnitude.

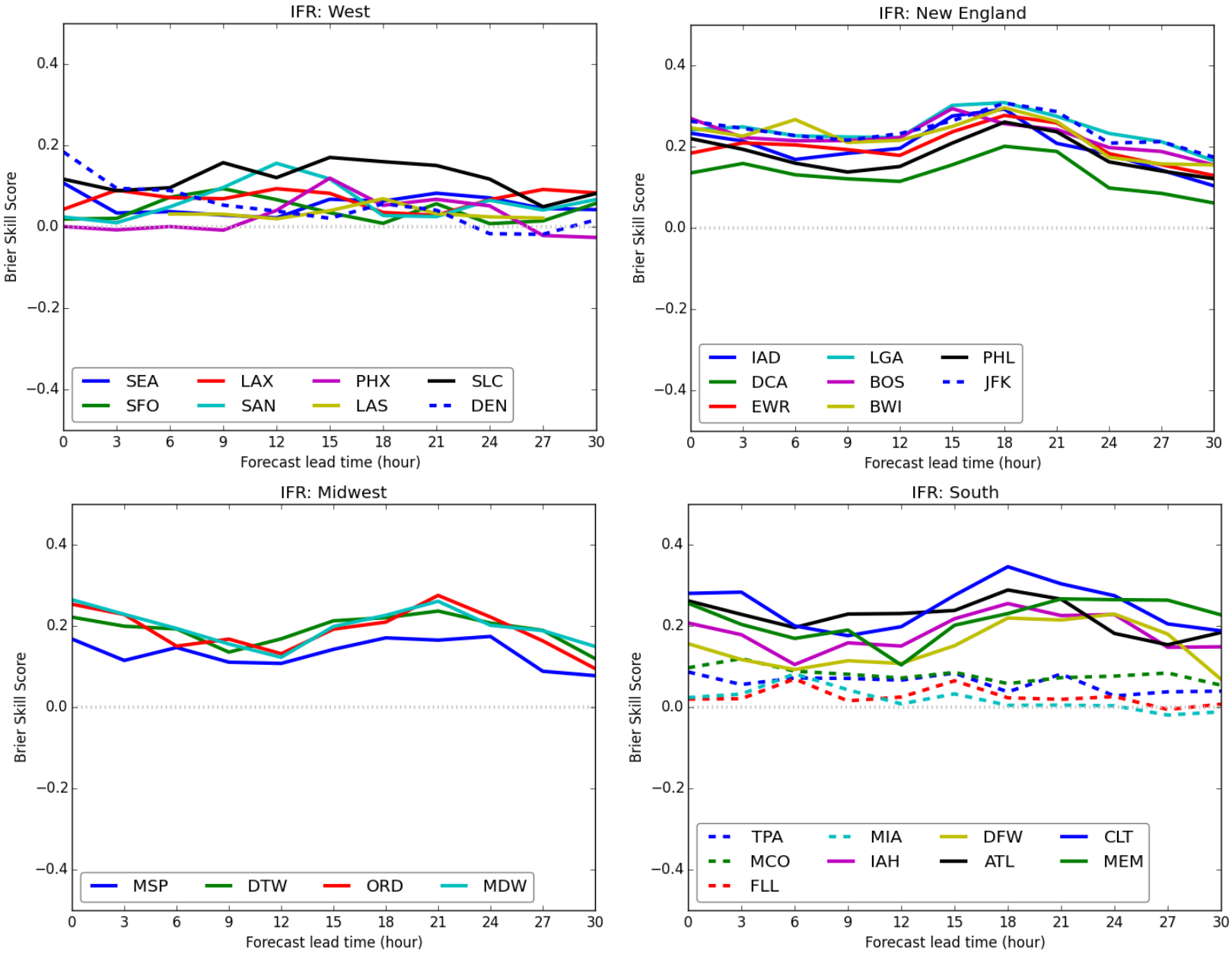

Figure 2: Brier skill scores for IFR for all seasons versus forecast lead time by region: West (upper left), New England (upper right), Midwest (lower left), South (lower right).

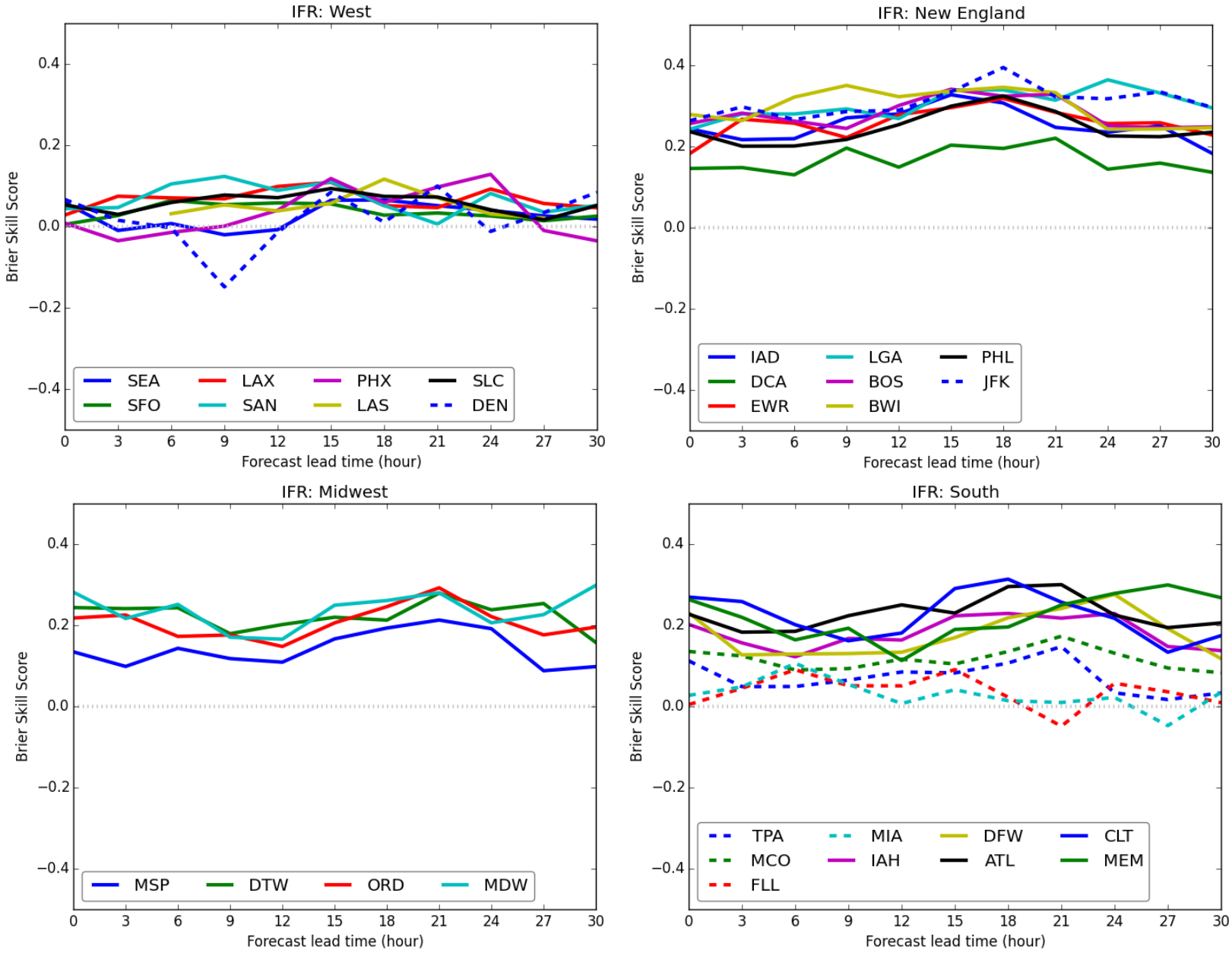

Figure 3: As in Figure 2, but only considering winter months (DJF).

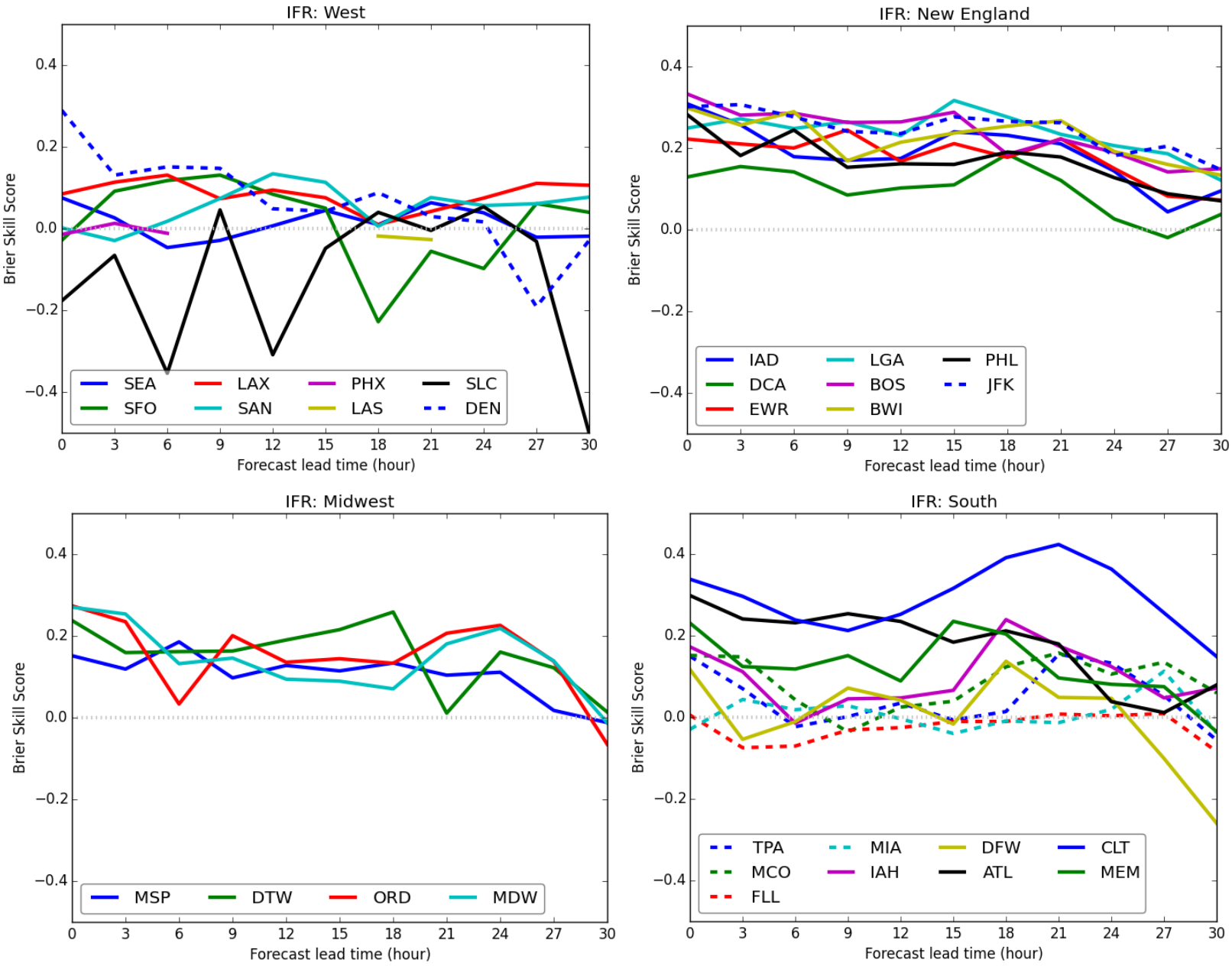

Figure 4: As in Figure 2, but only considering spring months (MAM).

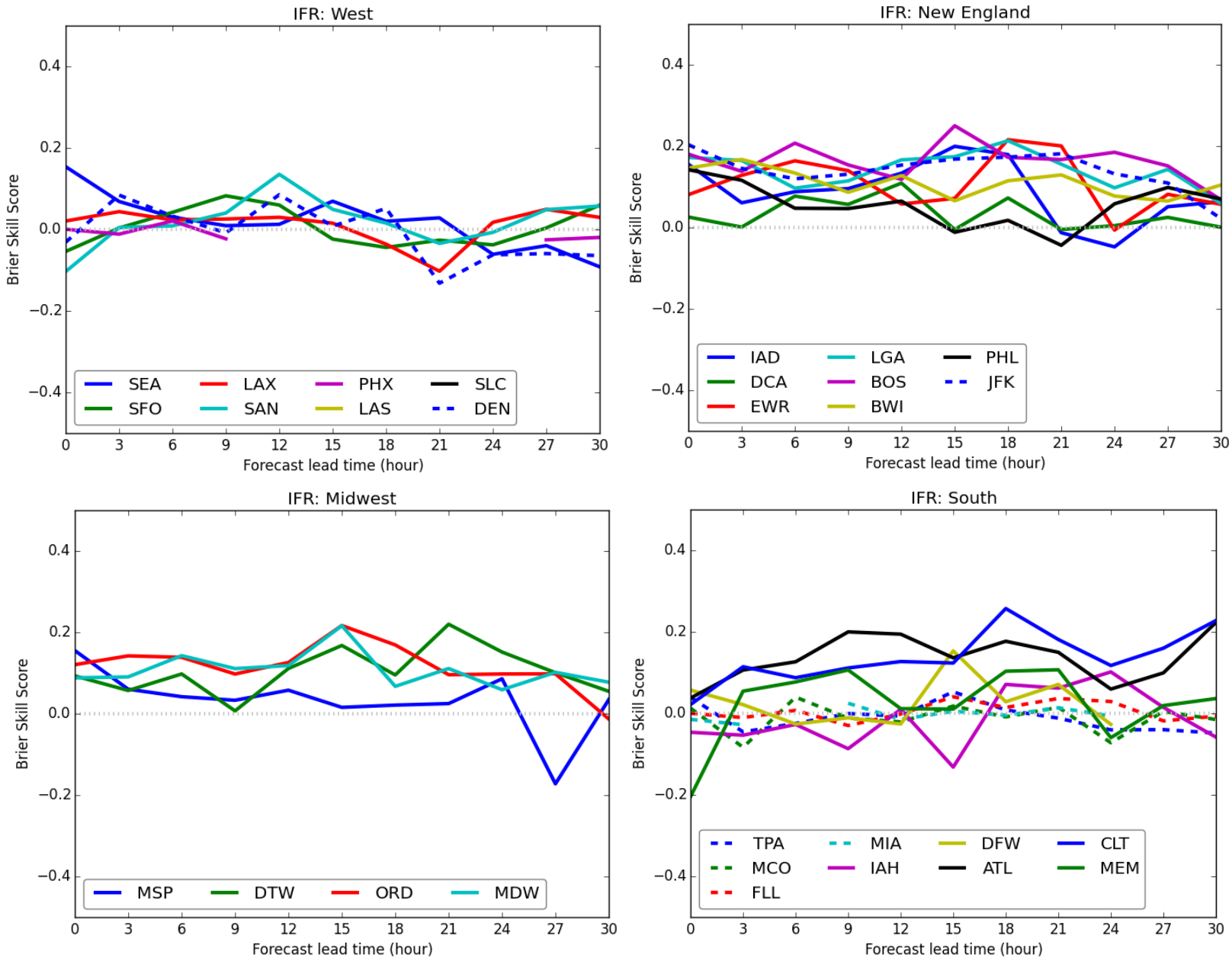

Figure 5: As in Figure 2, but only considering summer months (JJA).

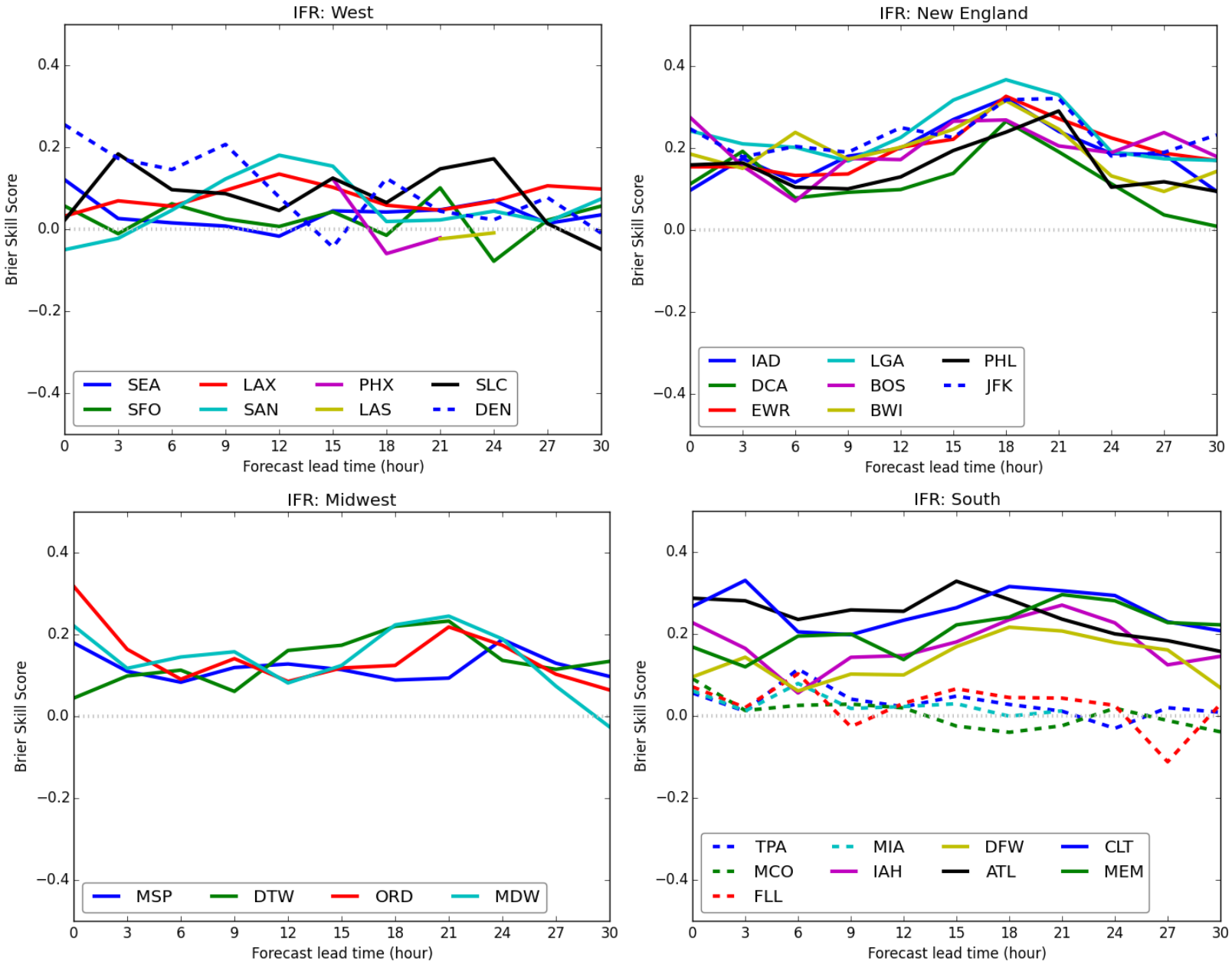

Figure 6: As in Figure 2, but only considering autumn months (SON).

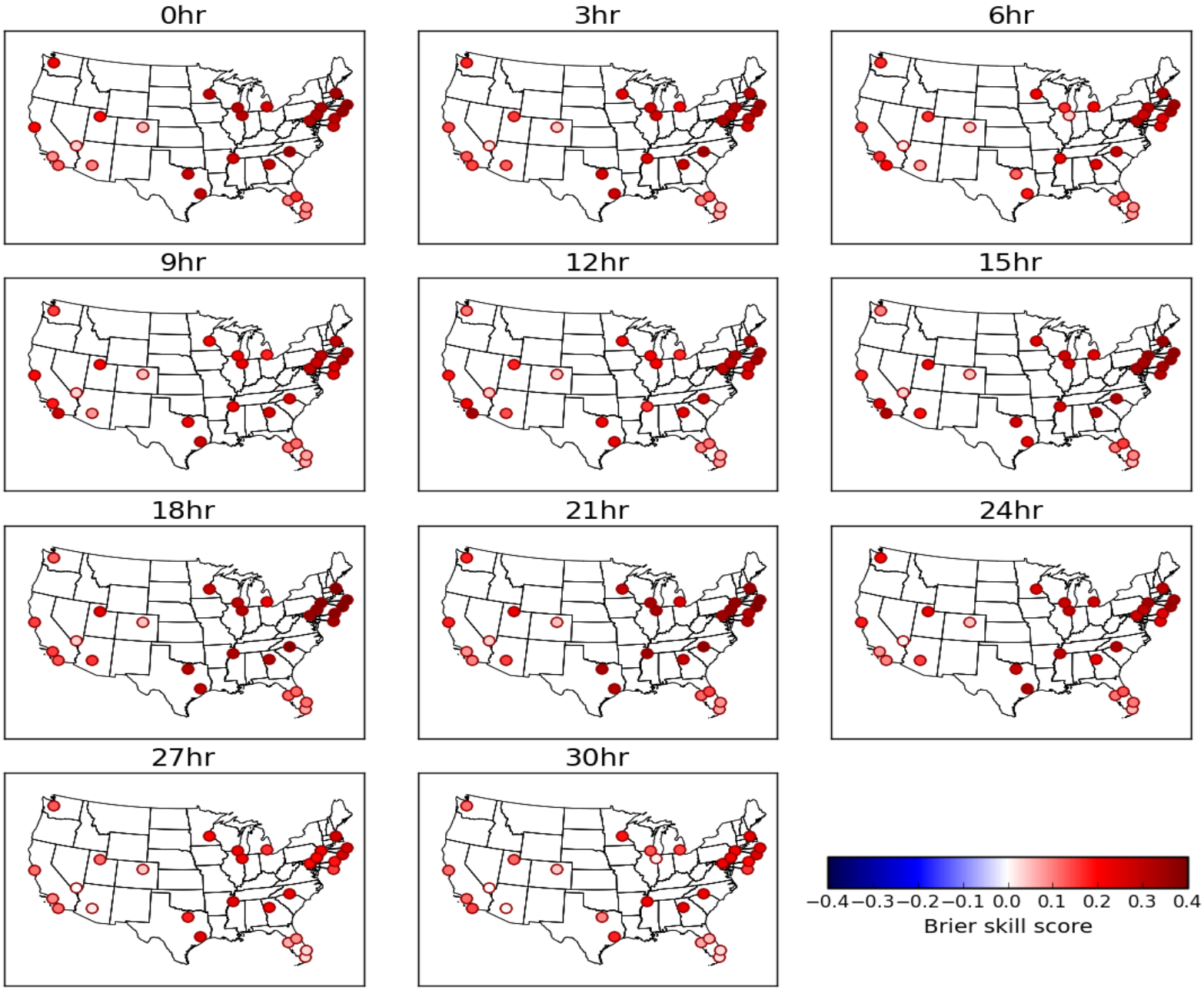

Figure 7: As in Figure 1, but for VFR

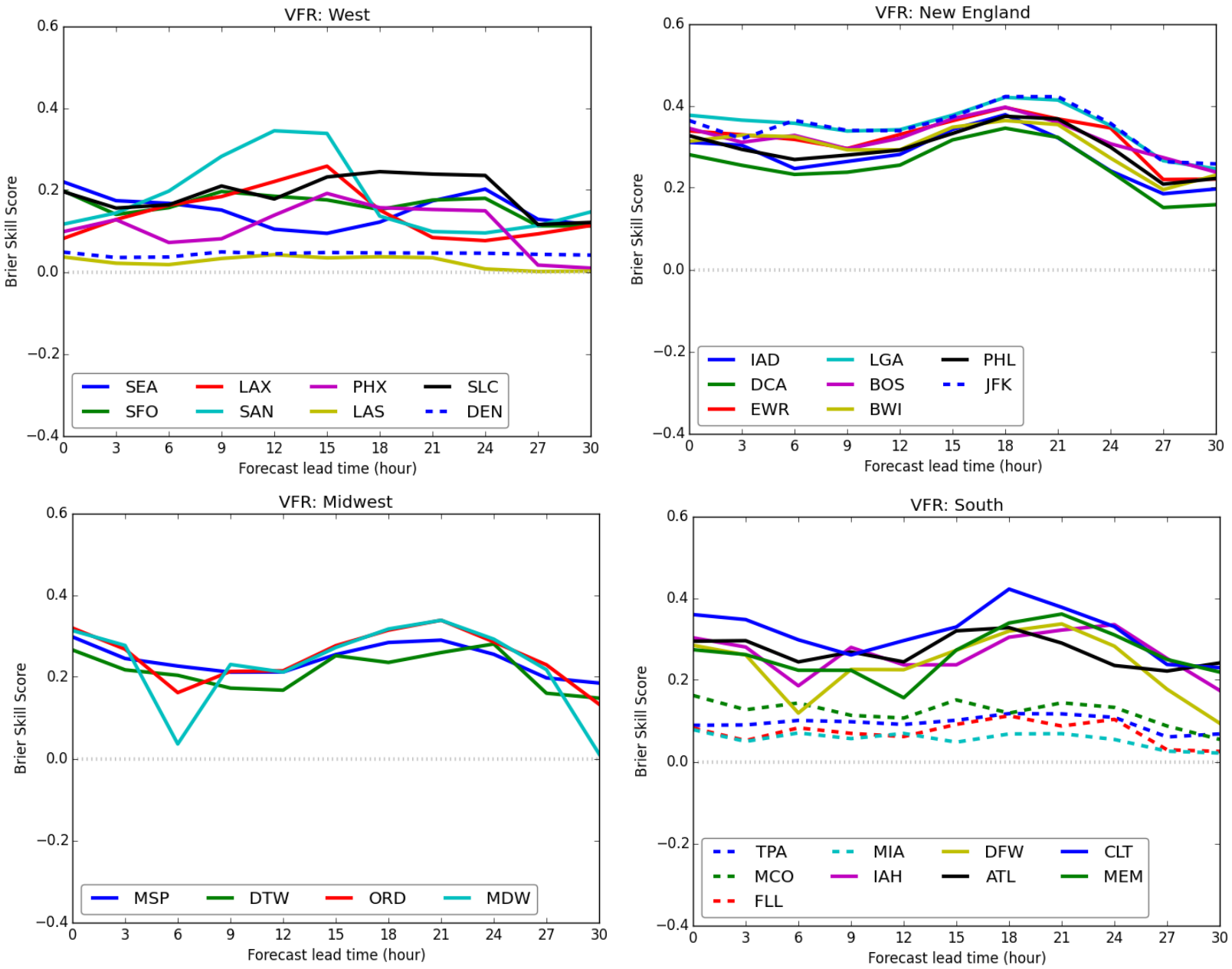

Figure 8: Brier skill scores for VFR for all seasons versus forecast lead time by region: West (upper left), New England (upper right), Midwest (lower left), South (lower right).

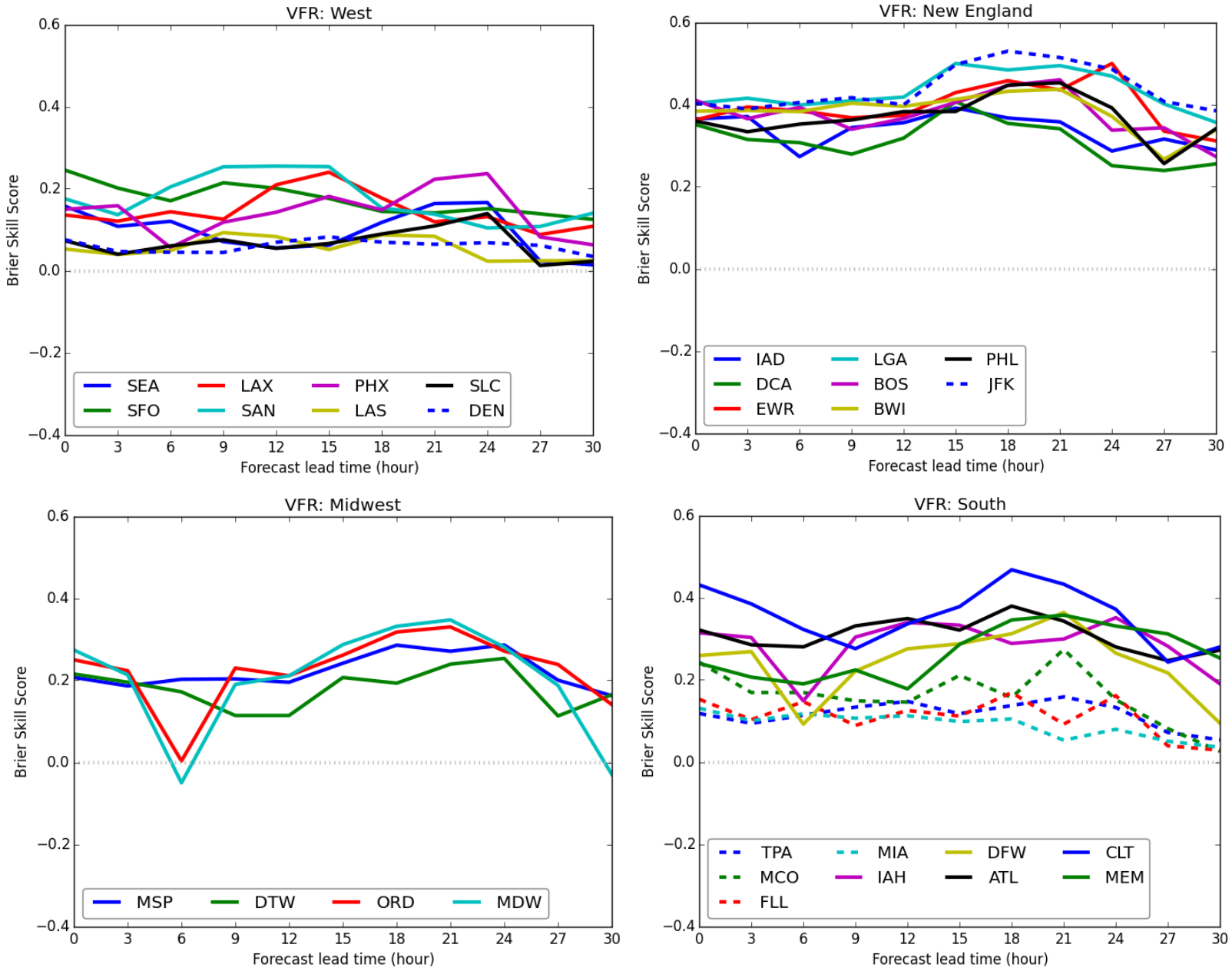

Figure 9: As in Figure 8, but only considering winter months (DJF).

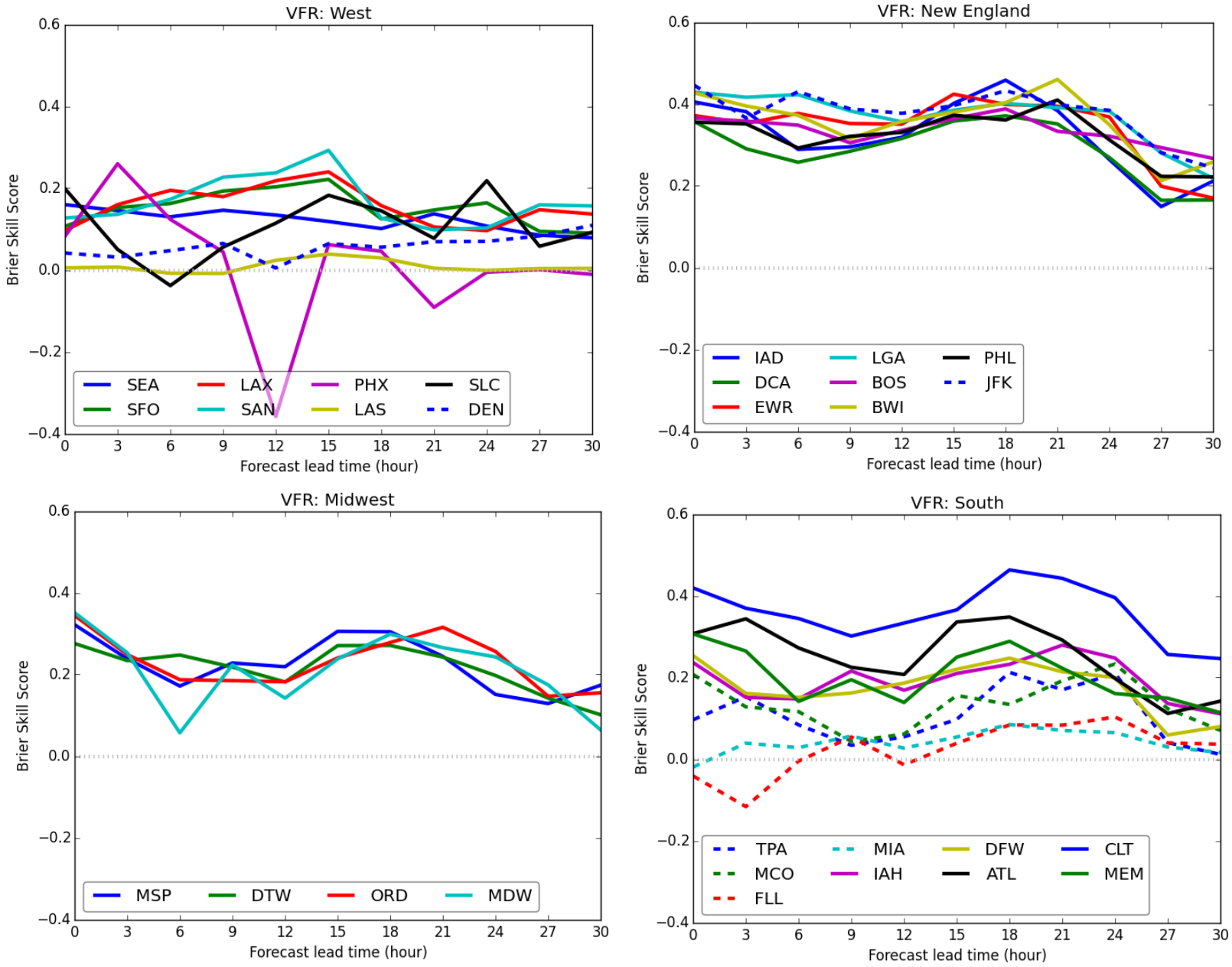

Figure 10: As in Figure 8, but only considering spring months (MAM).

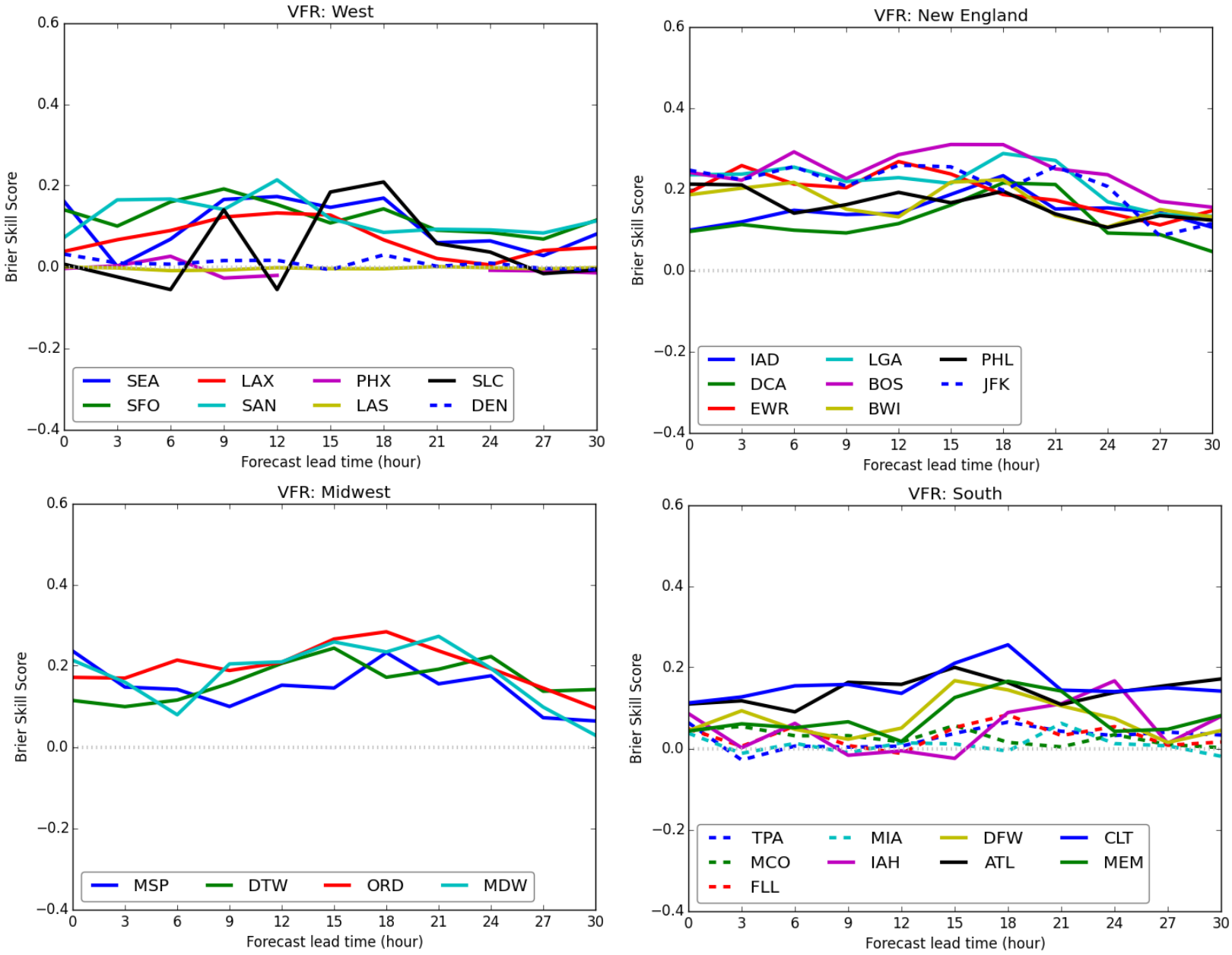

Figure 11: As in Figure 8, but only considering summer months (JJA).

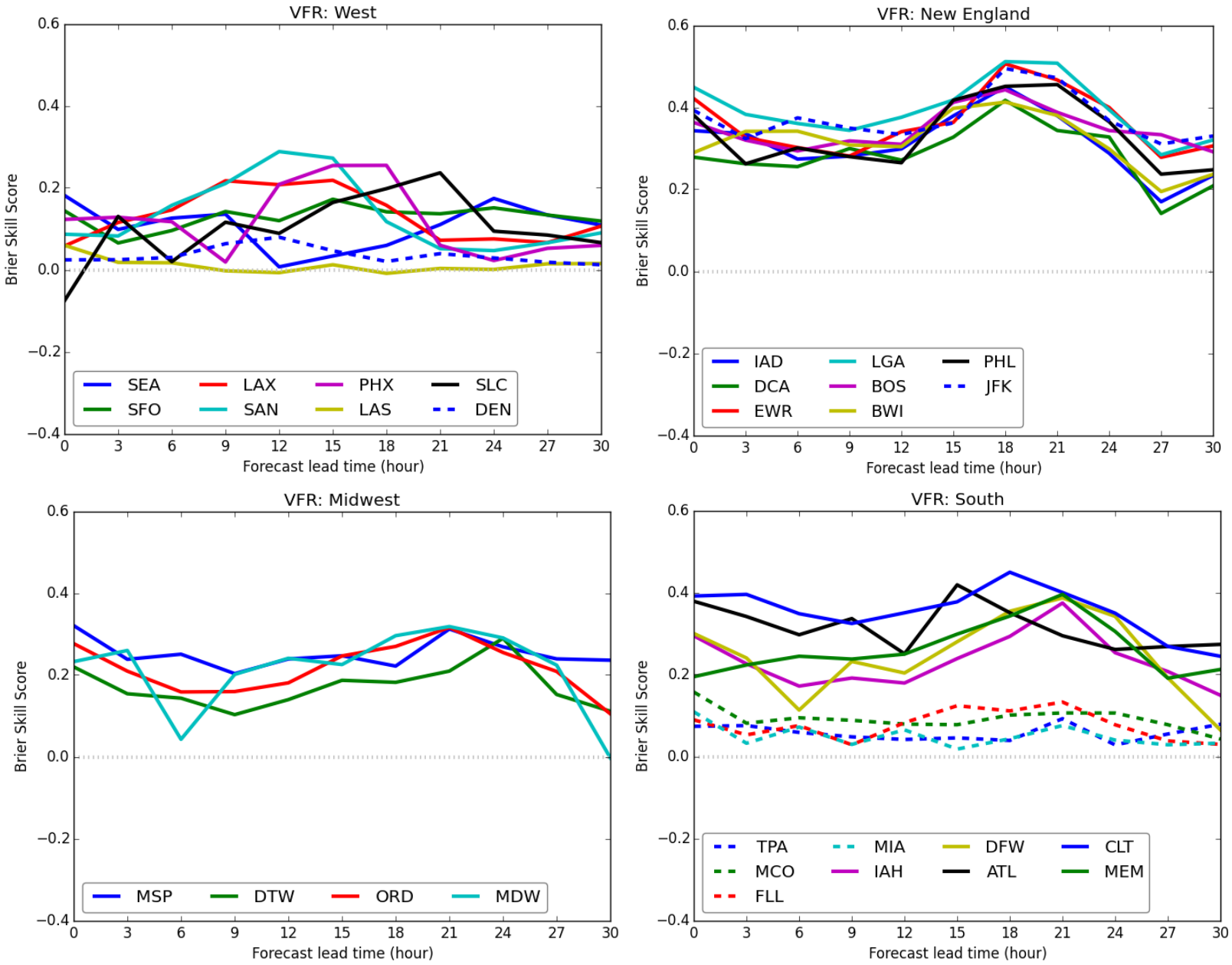

Figure 12: As in Figure 8, but only considering autumn months (SON).

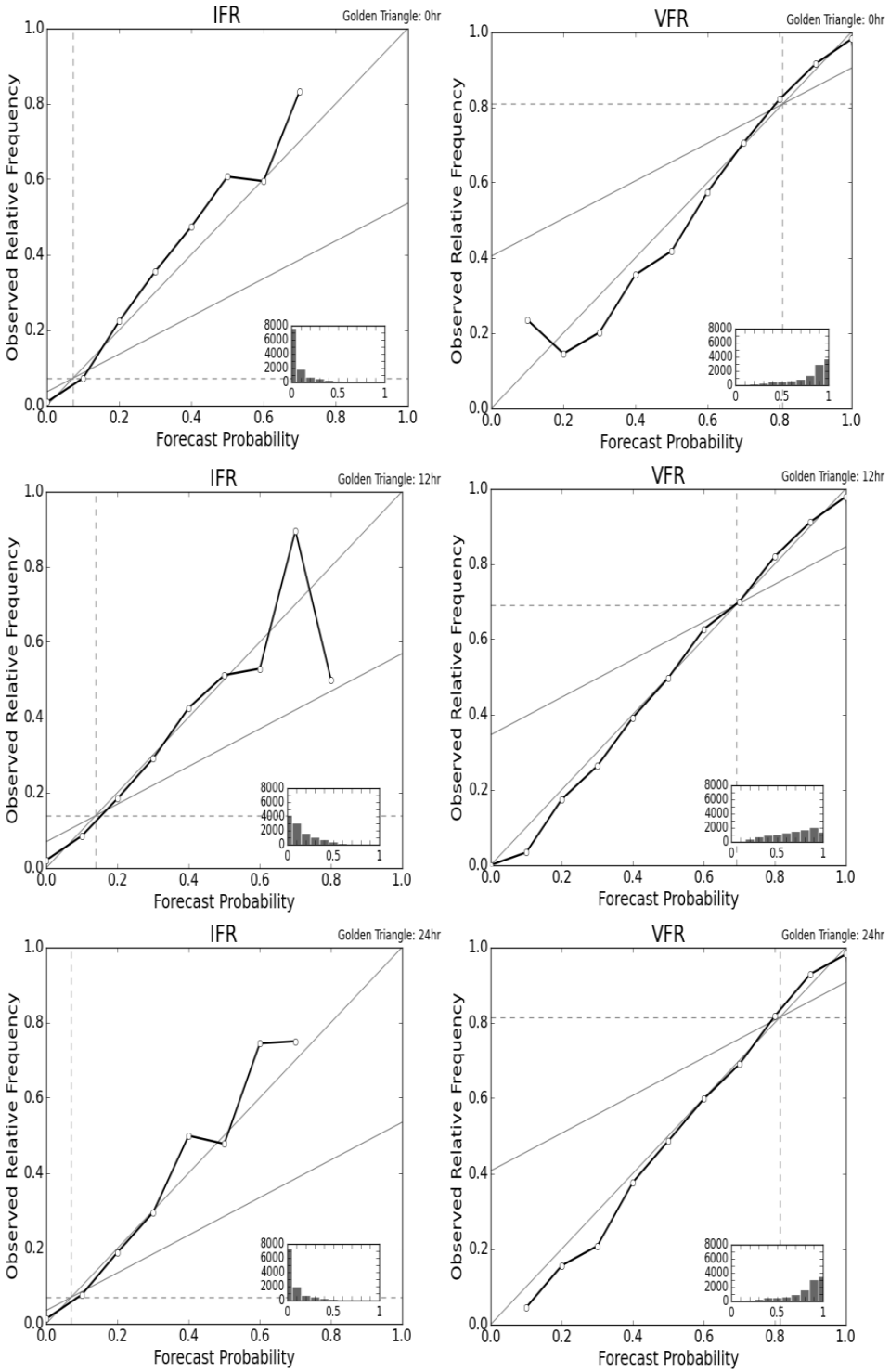

Figure 13: Attributes diagrams from compositing forecasts for the five airports in the Golden Triangle for 0-, 12-, and 24-hour lead times (top, middle, and bottom, respectively) for IFR (left column) and VFR (right column). Observed relative frequency is plotted in the black line with white filled circles 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 are plotted as the horizontal dashed line. The inset bar graph displays the number of observations in each forecast probability bin.

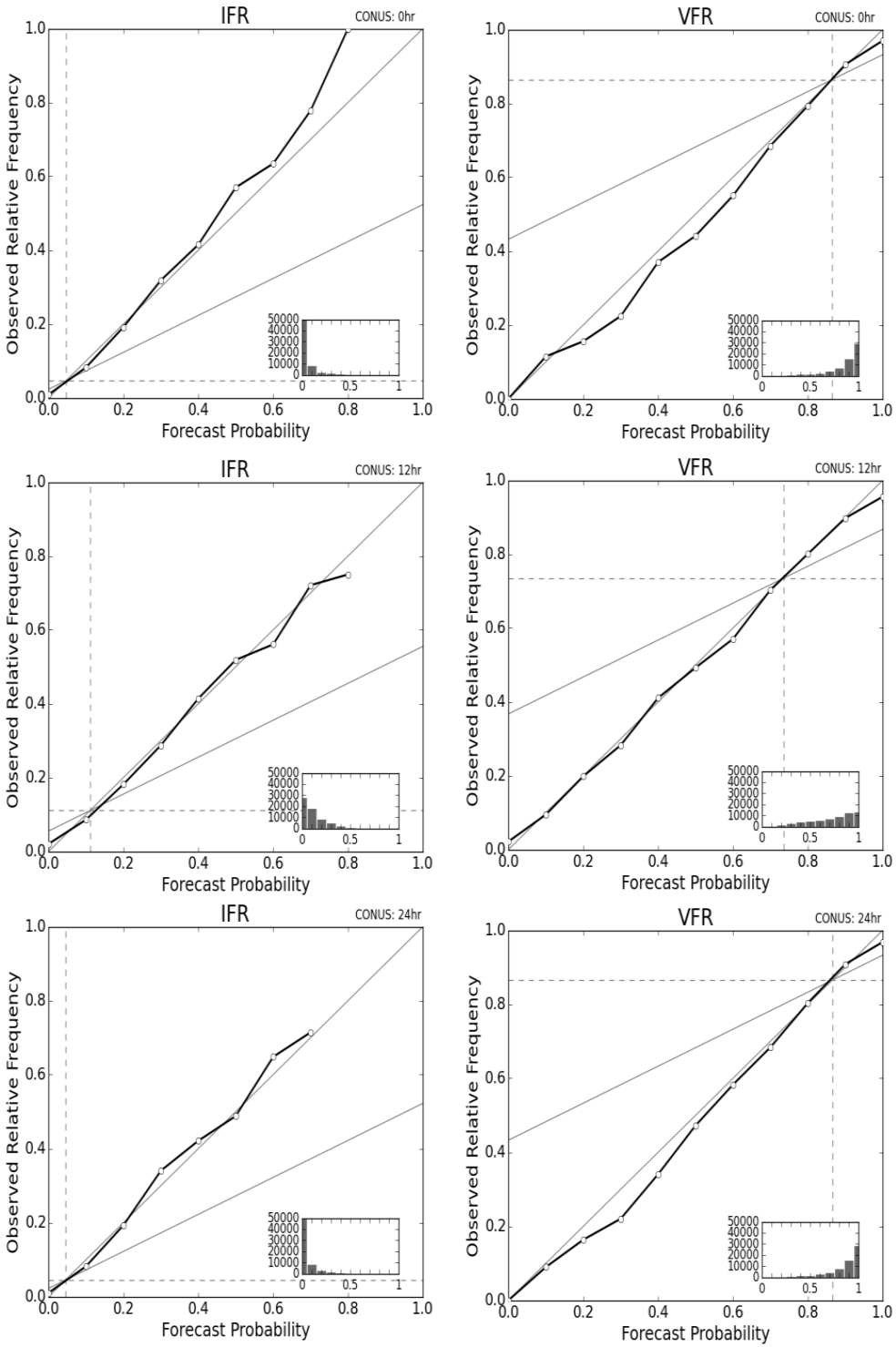

Figure 14: As in Figure 13, but attributes diagrams from compositing the 29 CONUS Core 30 airports for 0- (top), 12- (middle), and 24-hour (bottom) lead times for IFR (left column) and VFR (right column).

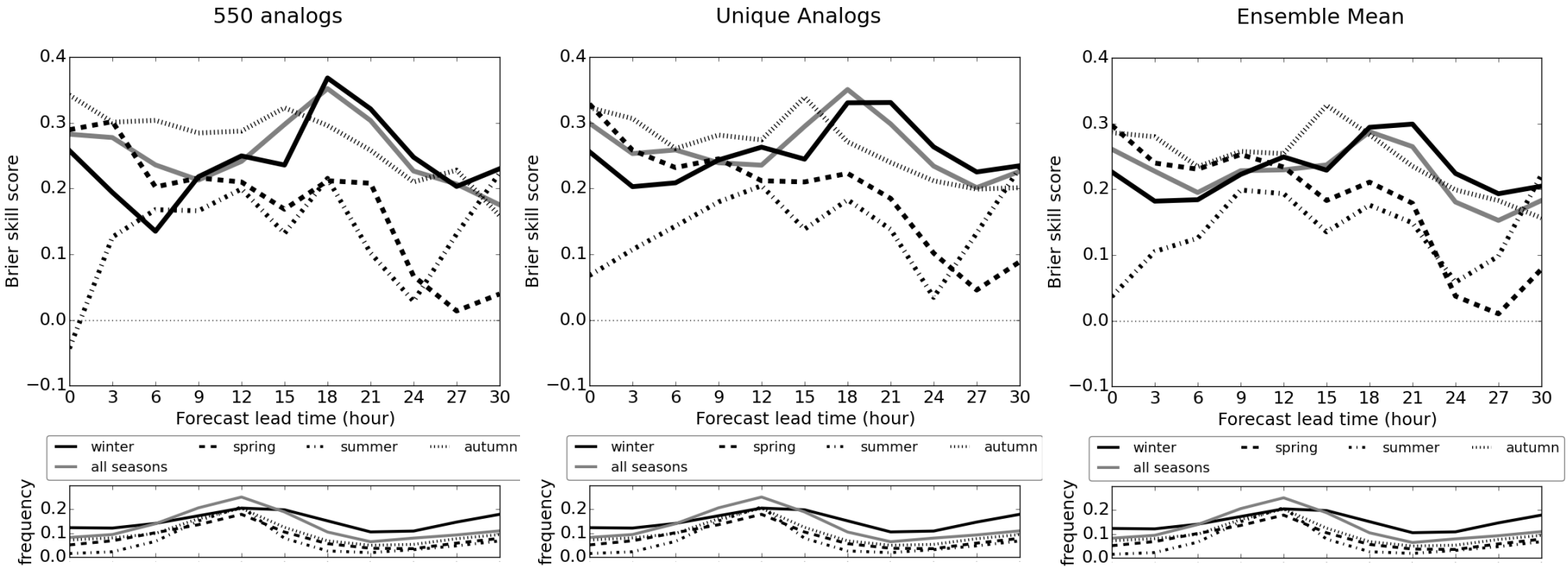

Figure 15: IFR Brier skill scores (BSSs) (upper panels) and observed relative frequency (lower panels) for Hartsfield-Jackson Atlanta International Airport (ATL) versus forecast lead time by season: winter (solid black), spring (dashed), summer (dot-dash), autumn (dotted), and all seasons (solid gray). BSSs are computed by looking at all 550 analog forecasts from ensemble members (upper left panel), and a subset containing the unique analog dates (middle panel). BSSs for the ensemble mean (right panel) are the same as shown in previous figures, but aggregated here for easy reference.

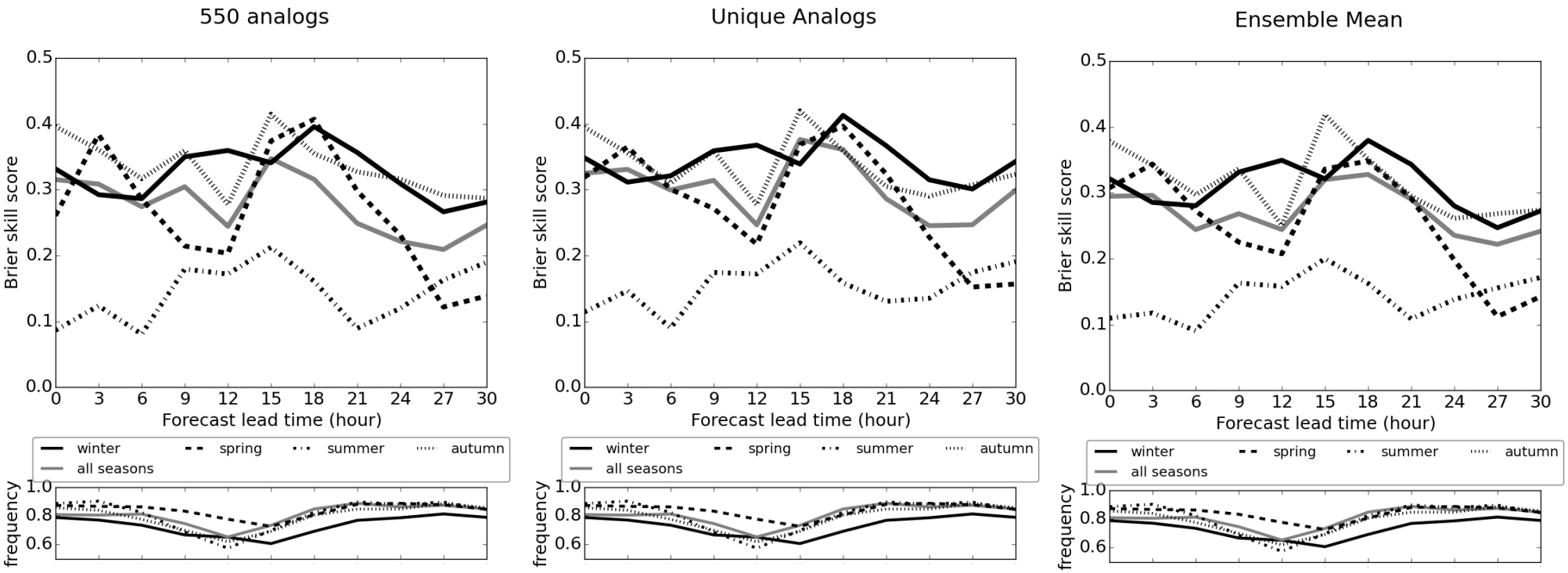

Figure 16: As in Figure 15, but for VFR.

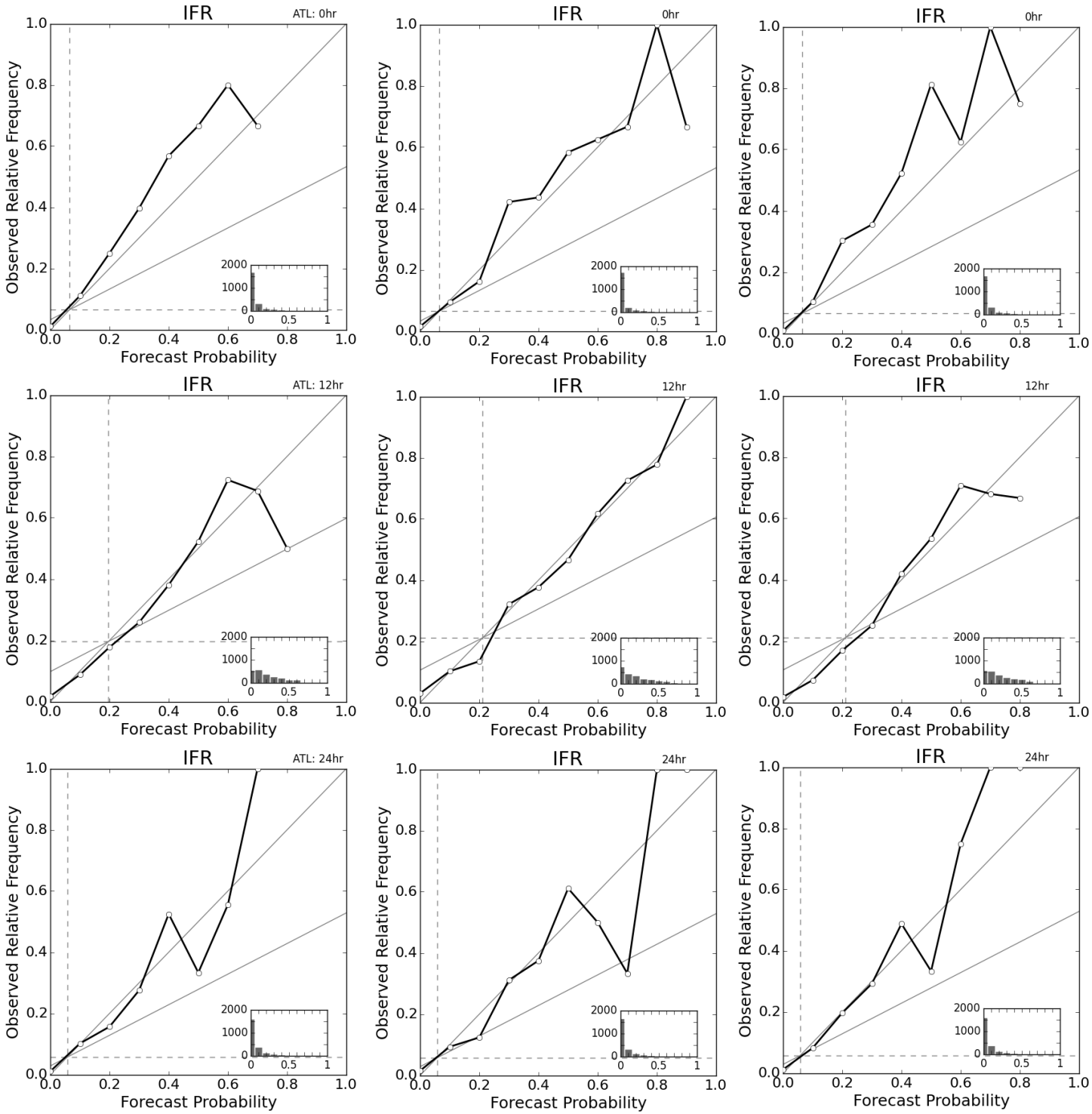

Figure 17: As in Figure 13, but attributes diagrams for IFR at ATL considering forecasts from the ensemble mean (left column), ensemble member 550 analogs (middle column), or ensemble member unique analogs (right column) for 0- (top), 12- (middle), or 24-hour (bottom) lead times.

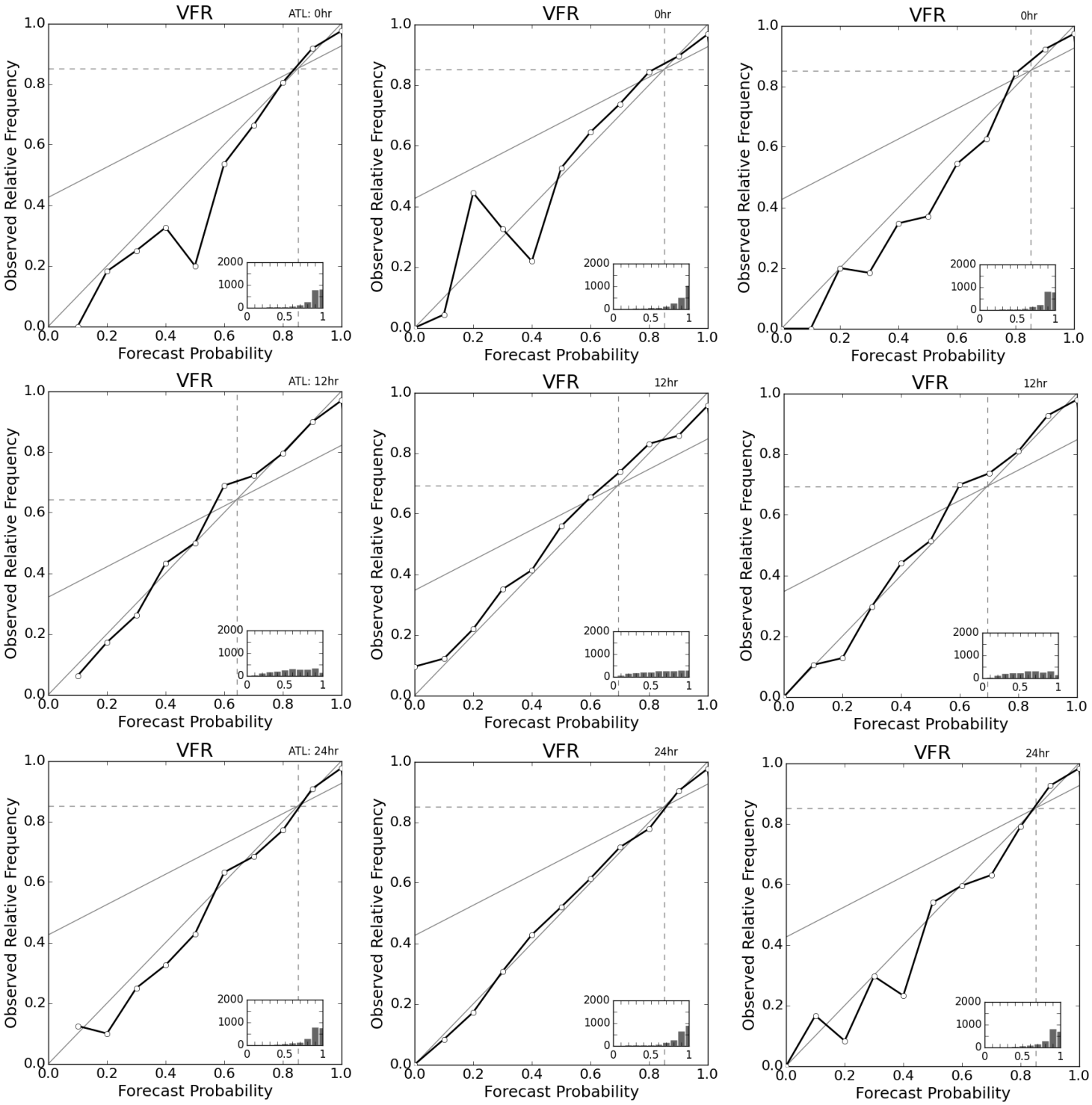

Figure 18: As in Figure 13, but attributes diagrams for VFR at ATL considering forecasts from the ensemble mean (left column), ensemble member 550 analogs (middle column), or ensemble member unique analogs (right column) for 0- (top), 12- (middle), or 24-hour (bottom) lead times.