11.4

A LEAD-TIME METRIC FOR ASSESSING SKILL IN FORECASTING THE ONSET OF IFR EVENTS[†]

Andrew F. Loughe,[1,3*] Sean Madine,[2,3] Jennifer Mahoney,[3] Mike Graf [4]

[1]Cooperative Institute for Research in Environmental Sciences, Boulder, CO
[2]Cooperative Institute for Research in the Atmosphere, Fort Collins, CO
[3]NOAA/Earth System Research Laboratory, Boulder, CO
[4]NOAA/NWS Headquarters/Aviation Services Branch, Silver Spring, MD

1. INTRODUCTION

The Real-time Verification System (RTVS), developed by NOAA/ESRL/GSD's Aviation Branch, is principally focused on supporting and advancing verification activities within the FAA's Aviation Weather Research Program (AWRP). RTVS provides performance metrics for new forecast algorithms that are being transitioned into National Weather Service (NWS) operations, and gives direct feedback to forecasters and decision-makers alike. Increasingly, the system is being utilized to supply information to individual users of aviation-related weather forecasts. One example of such user-specific verification is a new lead-time metric that has been developed for assessing skill in forecasting the onset of Instrument Flight Rules events (IFR events). The method utilized by RTVS compares the onset from observed IFR events (via METARs) with the onset from NWS-issued Terminal Aerodrome Forecasts (TAFs). A relational database of these events has been created, consisting of 3.2 million individual TAFs, covering a 3-year time period, at over 600 U.S. airports.

An important by-product of this database development is a web-based verification tool that provides users a simple interrogation mechanism for comparing TAF weather elements and observed weather conditions obtained from METAR reports. The main purpose in gathering statistics on such a large number of events— over 1 million TAF IFR events are analyzed— is to highlight the capability of forecasts that have a profound impact on strategic planning of air travel throughout the United Sates. The approach used to associate forecast and observed IFR events is event driven, and yields results that are stratified by time period, station, region, weather category, forecast issuance hour, and other TAF attributes.

The purpose of this paper is to outline an approach for comparing such a large number of forecasts and observations for IFR conditions, and to highlight some of the capabilities of the web tool that is used to interrogate this extensive database (Fig. 1). All of the graphical results presented in this paper, and that appear in an accompanying poster at this conference, are generated using this web-based resource.

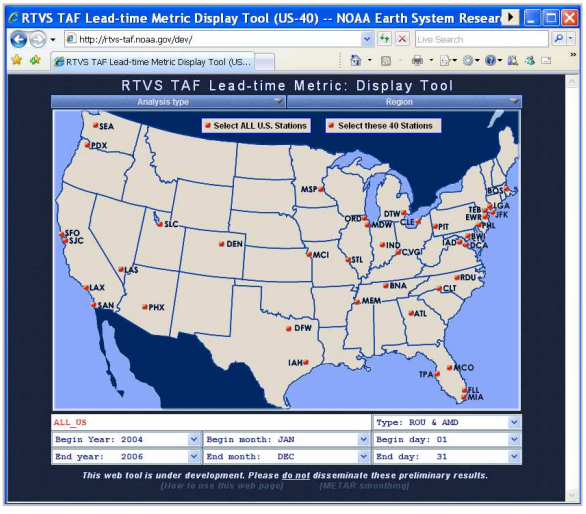


Fig. 1. *Web resource integrated into the RTVS for interrogating a database of IFR forecast lead times at over 600 U.S. airports from 2004-2006. Results can be stratified by time period, station, region, weather category, and issuance hour.*

[†]This work is sponsored by the National Weather Service's Aviation Services Branch, and the Federal Aviation Administration's Aviation Weather Research Program.

*Corresponding author address: Andrew Loughe
325 Broadway, Boulder, CO 80305-3328
e-mail: Andrew.Loughe@noaa.gov

2. UTILIZING TAFs TO IDENTIFY PREDICTED VFR/IFR TRANSITIONS

In this section we outline a method for identifying forecast transitions to and from VFR/IFR conditions, with the aim of matching forecast transitions with corresponding events from the METAR record. The intent of this work is to compute the lead time in forecasting the onset of IFR conditions, as these forecasts affect aviation interests throughout the entire country. It is important to realize that for this particular study, we consider only IFR and non-IFR conditions, which we loosely refer to as VFR. This is strictly a binary approach for identifying IFR events.

We begin by providing background information on the forecast product itself. TAFs are human-generated, with guidance provided from many sources, including numerical weather models. TAFs provide a concise statement on expected weather conditions at commercial and military airports throughout the world. Routine, NWS TAFs are issued 4 times daily (00Z, 06Z, 12Z, 18Z) for forecast periods usually extending out to 24 hours. Amended and corrected TAFs are issued throughout the day as needed. These forecasts are utilized by all airports in the National Plan of Integrated Airport Systems (NPIAS), and are an important strategic aid to major users of the National Aviation System (NAS), including large air carriers, air taxi/commuters, general aviation pilots, and military aircraft (FAA 2007).

TAFs are text products that are similar in style to METAR reports. They provide specific information on the timing of expected weather conditions that are likely to occur during a given time period, including: ceiling, visibility, weather, and low-level wind shear. NWS TAFs are comprised of change groups, including: **FM** (from), **TEMPO** (temporary), and **PROB** (probability of an event being 30% or greater), which indicate what weather changes can be expected during specified time periods. The first line of a TAF, as well as any subsequent FM group, represents the (expected) **PREVAILING** weather condition.

For this study, which focuses on forecast and observed transitions to and from VFR/IFR, TAFs were analyzed at all major U.S. airports, including those located in Alaska and Hawaii. Each TAF change group was compared with the previous group's weather condition, and an IFR event was recorded whenever the ceiling was predicted to drop below 1,000 feet, or the visibility to fall below 3 statute miles. Since TEMPO and PROB groups overlap with PREVAILING and FM groups within a full TAF, each time segment of the TAF is compared to the one before it, to reduce the predicted weather condition to the worst possible condition that is forecast at any particular point in time. This orientation toward worst weather was done in the context of Flight Rules for aviation, with a particular emphasis, for this study, on VFR and IFR conditions.

The process of rendering worst weather criteria from individual TAF groups (so-called *flattening* of a TAF) was accomplished following this simple rule set applied to NWS-issued (non-international, and non-military) TAFs:

(1) The ending time of one FM group is determined by the beginning time of the next FM group, or by the ending time of the entire forecast. As a result, the following rules, which apply to FM groups, are absolute:

> a) FM groups never overlap in time, they are always contiguous in time. Taking the FM groups alone, as you string them out together over the full forecast period, there is never a gap in time between FM groups. For the full forecast period, at any one point in time, there is exactly one FM group that is valid for any point in time. If an observation is taken precisely at the time one TAF group ends and another TAF group begins, the observation will be verified against the newer TAF group, which would have a starting time equal to the time of the observation.
>
> b) Forecast weather conditions specified within an FM group are only superseded by worsening weather conditions specified in a TEMPO or a PROB group.

(2) TEMPO and PROB groups never follow each other in direct order; there is always an FM group between

TEMPO and PROB groups. For example, in each of the successive change groups shown here, an FM group must be present at the location of the vertical line:

TEMPO | PROB; TEMPO | TEMPO;
PROB | PROB; PROB | TEMPO.

Once TAFs are *flattened* toward worst weather conditions, and the onset and cessation of IFR conditions is determined, timing and duration information for these forecast conditions is placed into a relational database of forecast IFR events.

3. UTILIZING METARs TO IDENTIFY OBSERVED VFR/IFR TRANSITIONS

In our approach, hourly and special METAR reports are compared with TAF reports at over 600 individual stations. Transitions are noted whenever the ceiling drops below 1,000 feet or the visibility falls below 3 statute miles (IFR conditions) at a time when the previously observed condition was above these threshold criteria (VFR). Again, for this particular study, we consider only IFR and non-IFR conditions, which we loosely refer to as VFR. This is strictly a binary approach for identifying IFR events.

When identifying these VFR/IFR transitions, a requirement was imposed that there be at least one METAR report two hours prior to the observed IFR transition, one report two hours following the transition, and an average of one report every 90 minutes throughout the time period of the event. The same requirement was imposed upon IFR event trailing edges, marking the transition from IFR back to VFR conditions. Once METARs were analyzed in the context of Flight Rules for aviation, their timing information was stored in a METAR IFR events database.

An interesting nuance with METAR data is the high frequency with which these VFR/IFR fluctuations sometimes occur in a very short time period (Fig. 2). One method for dealing with this characteristic of automated METAR (ASOS) reports, is to smooth over those transitions that do not persist longer than a given threshold, like 30 minutes. Ignoring these short-lived events places the observations on a more level playing field with the forecasts against which they are compared.

Smoothing high-frequency, automated observations is necessary, mostly because TAFs are

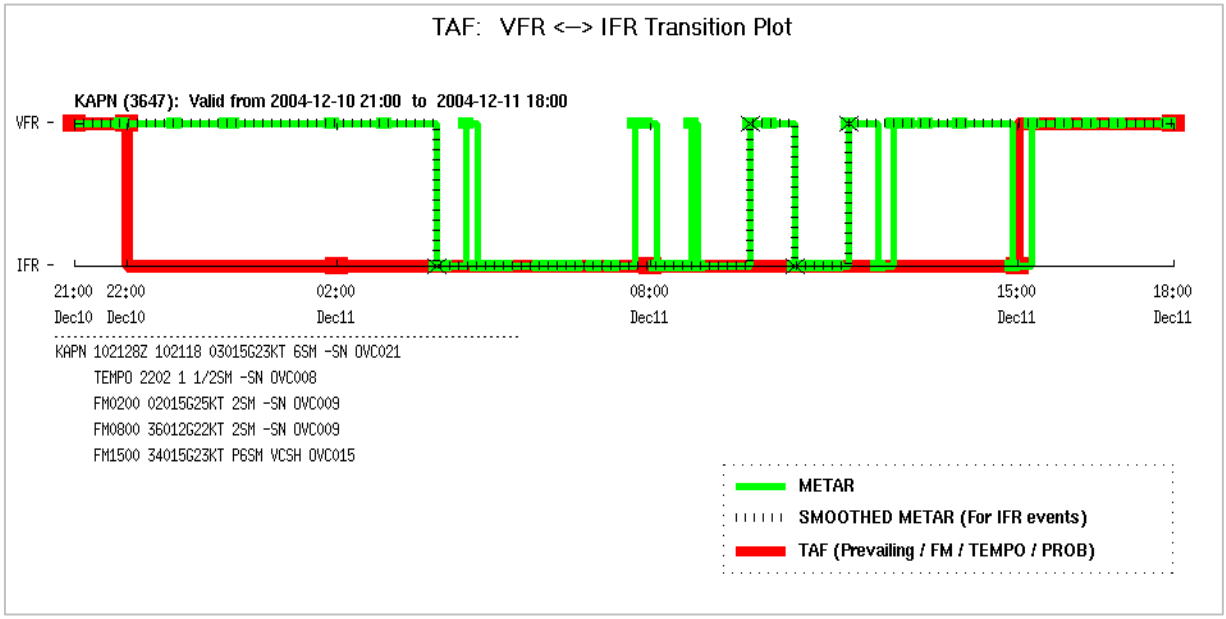


Fig. 2. *Example of a "noisy" METAR IFR event signal (green line) that is smoothed (green line with dashes) by including only those VFR/IFR transitions that persist for 30 minutes or longer. Note on the y-axis of this binary transition plot, that fair weather conditions (VFR) are assigned a value of 1, and poor weather conditions (IFR) are assigned a value of 0.*

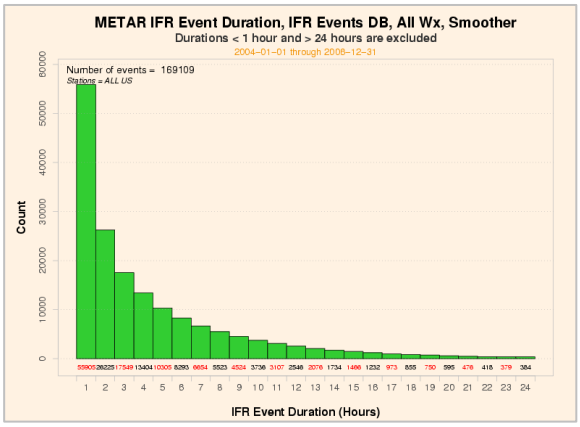


Fig. 3. *METAR-derived IFR event duration for the time period 2004-2006. This analysis contains data from over 600 individual sites throughout all 50 states.*

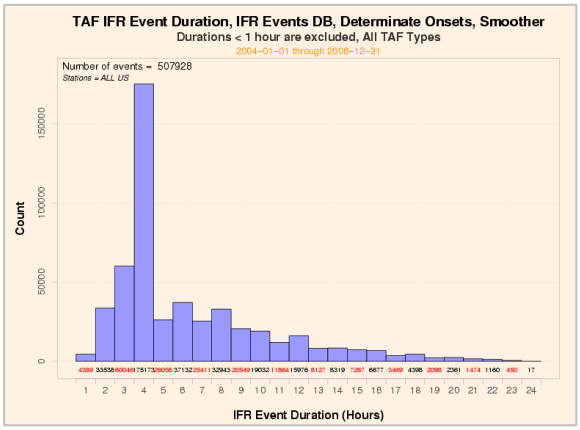


Fig. 4. *TAF-derived IFR event duration for the time period 2004-2006. This analysis uses forecasts obtained from over 600 individual sites throughout all 50 states.*

not also designed to produce such numerous fluctuations over short time periods, and this causes a matching problem between forecasts and observations. Statistically, without smoothing, the number of hits for a single forecast event could be arbitrarily large, and for an incorrect forecast, the number of misses could be arbitrarily large over a relatively short time period. Thus, performance of the TAF would appear to improve greatly in the first case, and appear to worsen significantly in the second case.

Automated METAR sites often report fluctuating VFR/IFR conditions two or more times during a single hour, whereas NWS guidelines encourage forecasters to create TAF change groups that last anywhere from 1 to 6 hours. Figs. 3 and 4 respectively show the climatological event duration of observed and forecast IFR events. These diagrams reveal that the vast majority of METAR IFR events are of duration 0-1 hour, while the vast majority of TAF IFR events are of duration 3-4 hours. This is partly due to TAF issuance guidelines (NWS 2005). Forecasters are instructed that TAF change groups generally last 1 hour or longer, although FM groups may be encoded to the minute if the expected change can be forecast to that degree of accuracy. TEMPO change groups can last no longer than 4 hours, and PROB change groups can last no longer than 6 hours. Moreover, if a TEMPO change group in effect longer than 2 hours does not verify, the forecast is to be amended appropriately. You can see that there are considerable constraints imposed upon TAF forecasts that are not also imposed upon automated METAR reports.

4. MATCHING TAF VFR/IFR TRANSITIONS TO OBSERVED TRANSITIONS

Once timing characteristics are derived for the onset and cessation of forecast and observed IFR events, one must determine which forecast events match *close-enough-in-time* to register a hit on accompanying observed IFR events. The criteria used for matching TAFs with METARs takes into account that the acceptable proximity in time increases as the lead time of the forecast increases. The function utilized to include forecasts as hits on a given METAR event has the shape indicated in Fig. 5. Regardless of lead time, our approach registers a hit for all forecasts that are within one hour of the observed event. It also registers a hit for those forecasts that are sufficiently aligned in time, according to the prescribed function, out to a maximum four-hour timing error for those

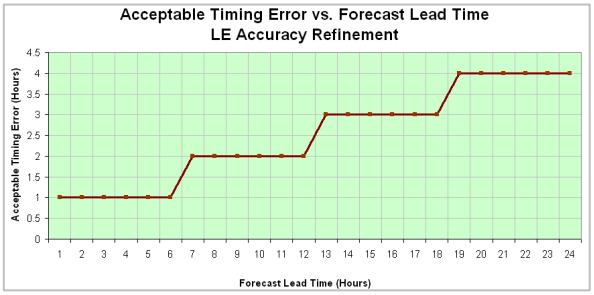


Fig. 5: *Accuracy function utilized to determine acceptable timing error with lead time, for registering a "hit" from a TAF on an observed IFR event.*

forecasts issued from 18-24 hours prior to the onset of the observed event. For the timing attributes indicated, we have recorded over 160,000 observed IFR events (Fig. 3), and over 500,000 forecast IFR events (Fig. 4). Since forecasts are usually issued 4 times per day, it is not surprising that there is a greater number of forecast events than observed events.

It is important for us to determine a single lead time in forecasting the onset of each observed IFR event; therefore, we must choose which of the multiple forecast hits on an observed event is to be counted in the tally of "best hits" (or earliest hits) on these events. To begin, we define the lead time as the difference in time between the onset of an observed event, and the issuance of a forecast that is associated with the observed event (Fig. 6). The timing error is the difference between the forecast and actual onset time of the IFR event. In this system, a negative timing error is associated with an event that is forecast to occur too early, so that in the aggregate, a large timing error indicates that you typically forecast events to occur after they have already been realized.

We record a hit on an observed event when the forecast is within an acceptable timing range of the observation (see Fig. 5). Because multiple TAFs may register a hit on the same observation, it is prudent to record only the *earliest forecast hit* (EFH) on a given event— giving credit to the forecaster who early-on realizes the likelihood of IFR conditions, and correctly forecasts the event to occur. Another acceptable method for registering hits on observed events, is to give credit only for those early forecasts that do not subsequently retreat from that position— those that do not subsequently forecast non-IFR conditions during a period of observed IFR. This type of hit is referred to as the *earliest uninterrupted forecast hit* (EUFH): it is the earliest forecast that associates itself with the onset of an observed event, while all subsequent forecasts continue to associate themselves with the same observed event. This method rewards consistency in forecasting the occurrence of IFR events.

In analyzing lead-time statistics from the large number of events that were identified, both the EFH and EUFH approaches can elucidate much about the forecasting process. One would expect the EUFH statistics to have a lower lead time, since longer-range forecasts are more likely to be "interrupted" by a subsequent forecast that backs off from the original forecast stance before the event actually occurs. Such theories can be tested with the web-based interrogation tool developed by RTVS staff members at NOAA/ESRL.

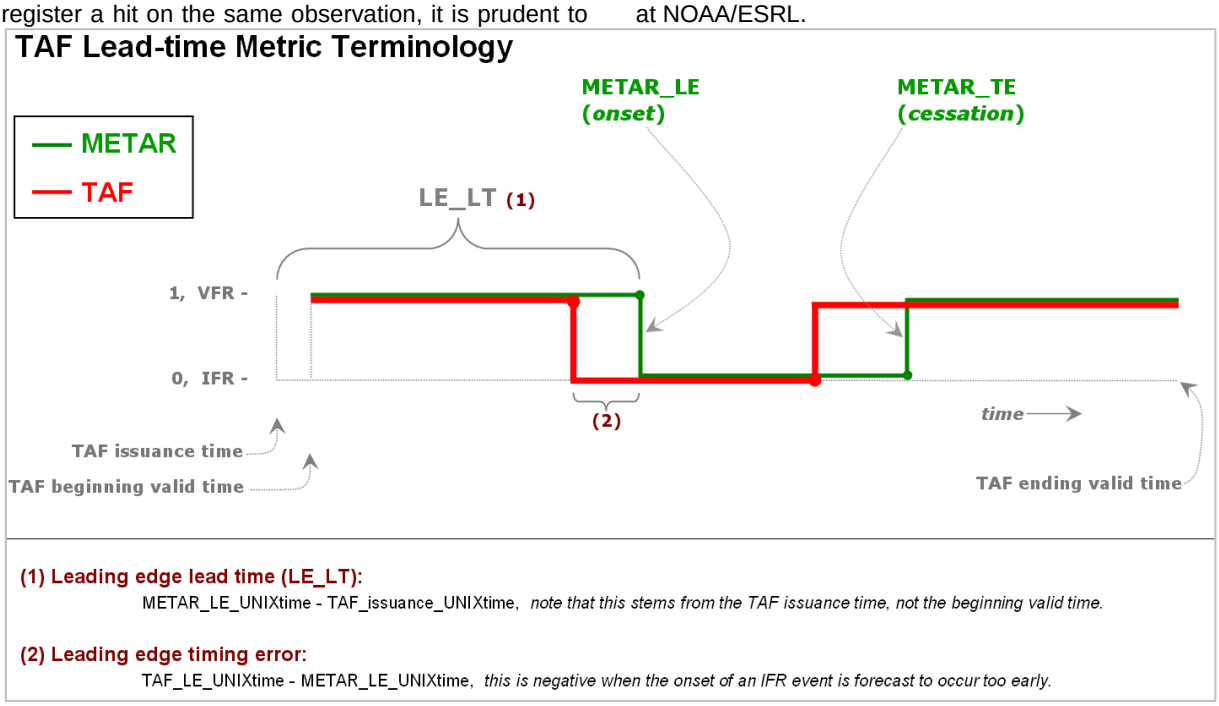


Fig. 6. *Terminology used to describe the lead time of a forecast IFR event on an observed event. Note that the TAF* ***issuance time*** *is used when computing the forecast lead time. The timing error is the difference between the forecast and actual onset time of the IFR event. Negative timing errors are associated with events that are forecast to occur too early.*

5. STRATIFICATION OF RESULTS BY WEATHER CATEGORY

One important aspect of our research is to gain an understanding of those weather conditions that typically cause IFR events to be forecast, and to determine how such conditions affect the overall forecast lead time. Clearly, adverse weather conditions such as fog, rain, snow, and the formation of stratus can lead to a reduction in ceiling and/or visibility to the point where conditions deteriorate from VFR to IFR. The approach used to stratify results by weather category follows closely the recommendation of Clark (1995). In his study, the author developed a hierarchical method for typifying weather, based upon METAR present weather elements, time of day for the beginning of an event, time of day for the cessation, and the minimum wind speed (see appendix).

Using this approach, one is able to characterize the weather that is occurring at any given time, based upon available METAR reports. For example, when the visibility drops below 3 statute miles, during a time period when fog is reported in a METAR, but there is no precipitation occurring, and the weather event begins in the evening and ends during the day under low wind conditions, the event is classified as radiation fog. If the fog event had ended at night with a substantial wind, it would have been classified as advection fog. Following this decision tree approach, we classified observed IFR events into 8 broad weather categories (Fig. 7). Note that event types 1, 4, and 8 are associated with precipitation, and represent nearly 60% of all IFR events. The largest single category is for visibility reduced by rain or snow with no fog present. There are also a large percentage of observed IFR events associated with stratus ending during the day without any precipitation occurring.

We have computed climatologies of IFR events from 2004-2006, which show a strong diurnal signal for events associated with stratus or fog, but no precipitation (Fig. 8c). These climatological results have a great impact on the lead time computed for TAFs. Note that the distribution of IFR events for all weather categories (including unidentifiable types, Fig. 8a) resembles a broad, flat signal with a peak imposed around mid-day UTC. For precipitation cases (Fig. 8b), the distribution of IFR events is quite uniform throughout the day. For fog and stratus cases (Fig. 8c), we see a strong diurnal signal from about 06Z-12Z. This peaked distribution for non-precipitation cases is, of course, the same peak manifest in the top figure for all weather cases. Such a distribution of events can greatly impact a forecaster's ability to register long lead times on IFR events.

To illustrate this point, assume that you typically issue TAFs at 06Z, while a colleague issues TAFs at

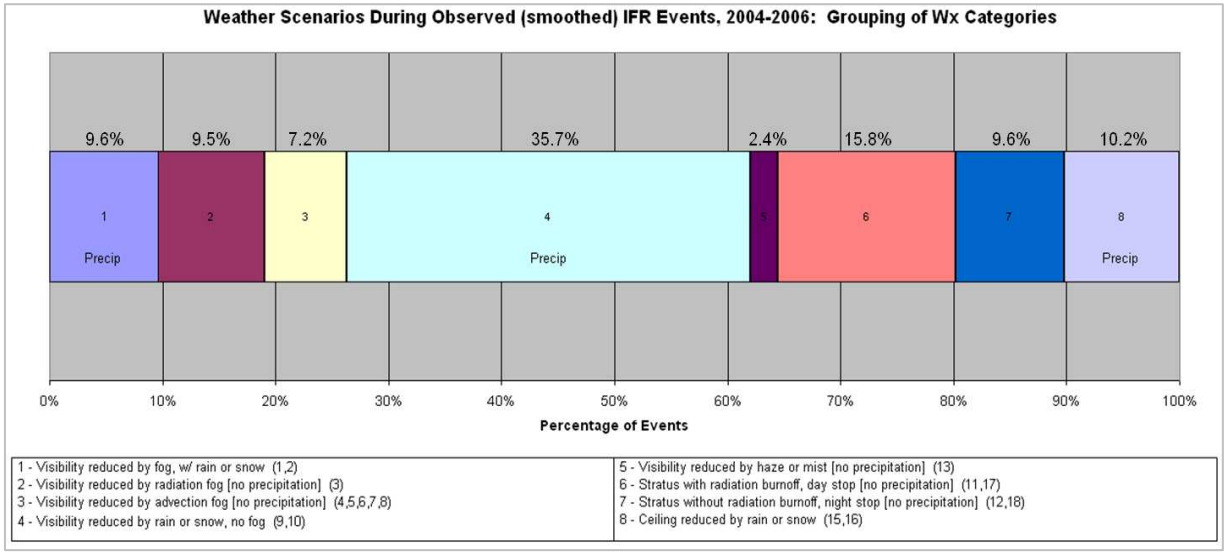


Fig. 7. *Weather categories derived from observed IFR events analyzed from over 600 U.S. airports during the time period 2004-2006 (see appendix).*

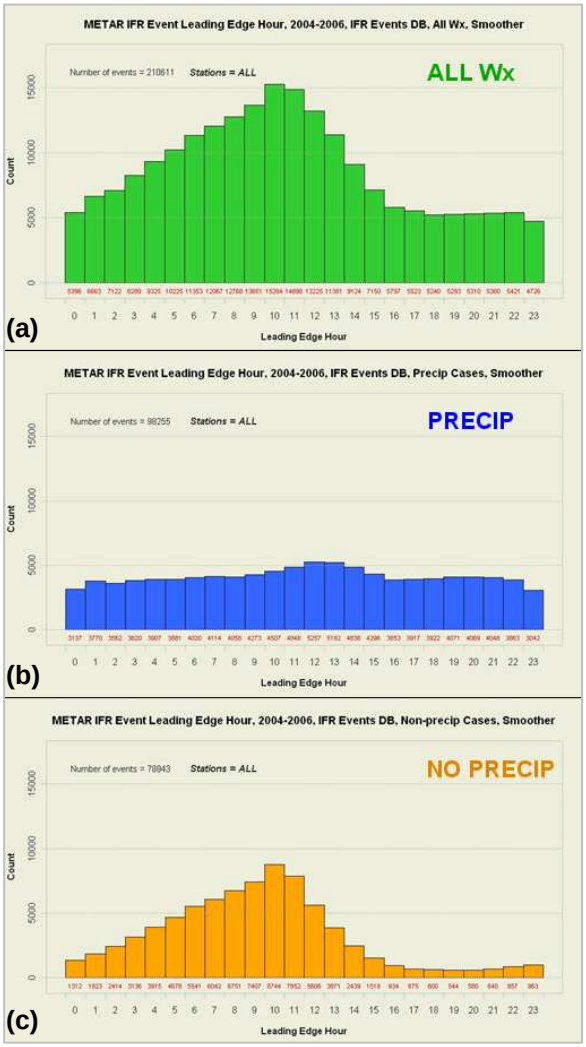


Fig. 8. *Climatology of METAR IFR event leading edge hour of day for all weather cases (including unidentifiable types), precipitation-only cases, and non-precipitation cases. Note the strong diurnal signal for non-precipitation cases.*

12Z. Since TAFs generally extend out to 24 hours, it will be difficult for the 06Z forecaster to register long-lead-time hits on events that typically occur from 06Z-12Z, whereas the 12Z forecaster will routinely register hits 18-24 hours out into the future, because the valid time of his forecasts extends into the next peak period of occurrence for these events. When it comes to recording long lead times in forecasting IFR events, one might think it fortuitous to be working the 12Z shift, rather than the more-difficult 06Z shift.

This type of climatological analysis is helpful when performing lead-time analyses for weather forecasts. It is difficult or impossible to understand why some forecasters outperform their colleagues without first understanding how and why one forecasting shift is more difficult than another. It is important to remember one fundamental truth regarding the computation of skill scores and hit rates: one cannot record a hit where there is no accompanying observation of the event. Hits are certainly more likely when (and where) the event has a higher climatological frequency of occurrence. In statistics, this is referred to as the base rate of an event (Wilks 2006).

6. A LEAD TIME METRIC FOR IFR EVENTS

The distribution of forecast lead time (EFH context) is presented in Fig. 9. Note that this particular lead time distribution does not take into account the large number of events that are not correctly forecast to occur— when no warning is given of an observed IFR event. When incorporating missed events as a zero-hour lead time, the results are more sobering than is presented here. To highlight exactly what is occurring with the forecast hits, we ignore from this graphic the missed events, and simply analyze results in an *earliest forecast hit* (EFH) context with lead-time greater than 0. It can be useful, however, to numerically compute the value of the lead time both with and without the inclusion of missed events, to gain understanding about the overall impact of these missed events.

Fig. 9 shows that the mean lead time for all forecasts that register EFH hits on observed IFR events during 2004-2006 is 12 hours. The median lead time is 12.5 hours, and the standard deviation is 5.6 hours. The first and third quartiles are shown on the histogram along the x-axis. As discussed earlier, computing the same measure for the *earliest uninterrupted forecast hit* (EUFH) will cause the lead time to decrease significantly, as those previous long-lead-time forecasts are replaced by later forecasts which successfully record a hit without subsequently backing off from that stance.

Using the RTVS web tool to interrogate the database of TAFs and METARs, we have studied distributions of lead time stratified by warm and cold season, by issuance hour, and by weather category.

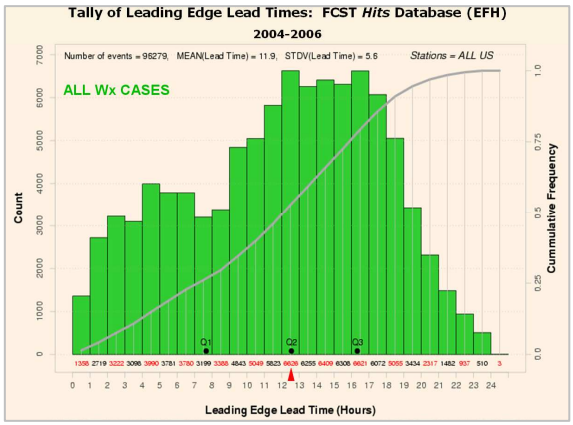


Fig. 9: *Tally of leading edge lead times for forecasting IFR conditions. This result is derived using an earliest forecast hit approach, and by ignoring missed events.*

Many of our expectations about the impact of weather and forecast issuance hour that were discussed regarding Fig. 8, are borne out in the more-complete analysis of lead time presented in Fig. 10. These box plots of lead time distribution are divided into 4 main groups. The first quadrant is for all weather cases (in green), all precipitation cases (in blue), and all non-precipitation cases (in orange), with all issuance hours (00Z, 06Z, 12Z, 18Z) combined into one box plot for each weather classification. In the first quadrant, one sees very little variation in the median lead time for each weather classification over all issuance hours.

If you take each weather classification and study the results by individual issuance hour, more information can be revealed. The second quadrant presents results for all weather cases (green) separated by issuance hour. In the third quadrant, all precipitation cases (blue) are separated by issuance hour. In the fourth quadrant, the same is done for all non-precipitation cases (the fog and stratus events). A careful study of these results shows that most of the variation in lead time by issuance hour that is manifest in the second quadrant (for all weather cases) is actually due to variations apparent in the non-precipitation cases (quadrant 4). There is little variation in lead time for precipitation cases (quadrant 3) for the reasons noted when analyzing results shown in Fig. 8: since there is little variation in the climatological occurrence of these events with time-of-day, one would expect little variation in forecast lead time with issuance hour for the precipitation cases.

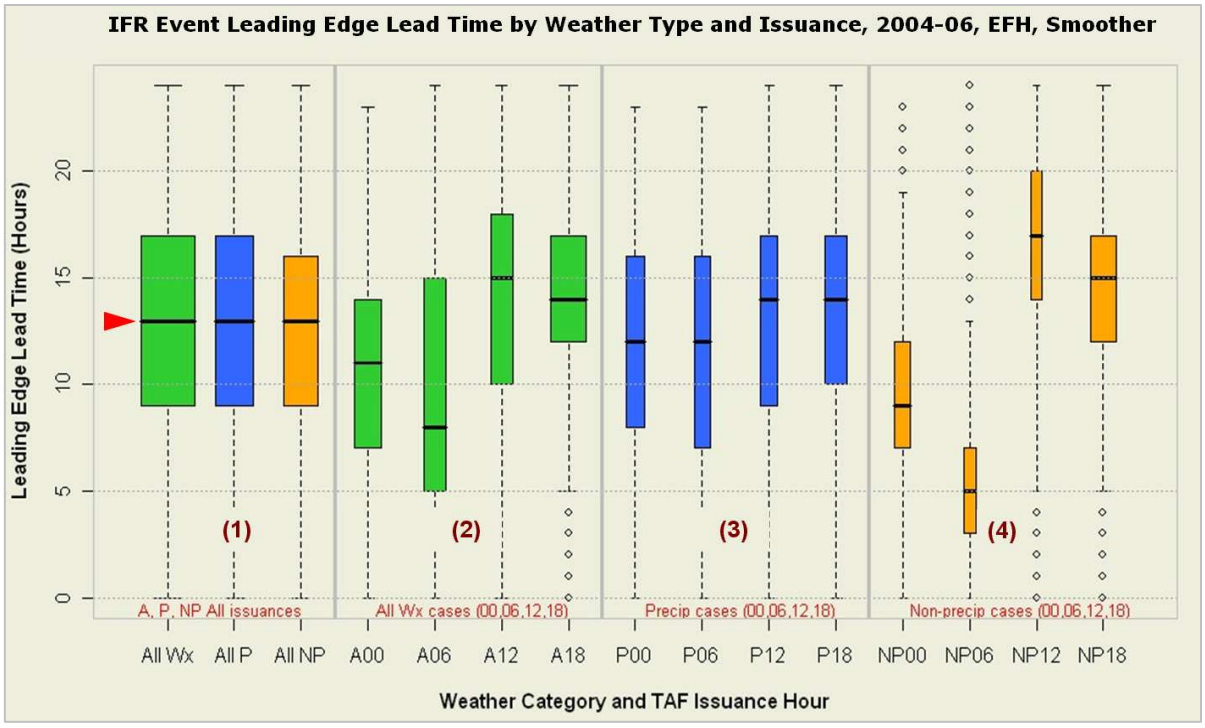


Fig. 10. *IFR event leading edge lead time for (1): all issuance hours combined with all weather cases (green), precipitation cases (blue), and non-precipitation cases (orange); (2) individual issuance hours with all weather cases; (3) individual issuance hours with all precipitation cases; (4) individual issuance hours with all non-precipitation cases. Note that the variation in lead time by issuance hour is largely controlled by the large variation manifest for non-precipitation cases.*

7. TAF LEAD-TIME METRIC WEB INTERROGATION TOOL

The TAF lead-time interrogation tool being developed for the RTVS, provides a convenient mechanism for performing all of the analyses that have been presented in this paper, plus more capabilities that we are unable to show at this time. The NWS is a valuable partner in developing this online tool, because their forecasters provide great insight into operational constraints that can affect the ability to record long lead times when forecasting IFR events. As a development team, we are working to incorporate these insights into our verification tool, with the goal of helping forecasters, decision makers, and mangers understand what steps can be taken to improve the lead time of these important aviation forecasts.

8. ACKNOWLEDGEMENTS

This research is sponsored by the NWS Aviation Services Branch, and in response to requirements and funding from the FAA's Aviation Weather Research Program. We have had numerous productive discussions with NWS forecasters, and with members of the NWS Performance Branch. The views expressed in this paper are those of the authors, and do not necessarily represent the official policy or position of the NWS or the FAA.

9. REFERENCES

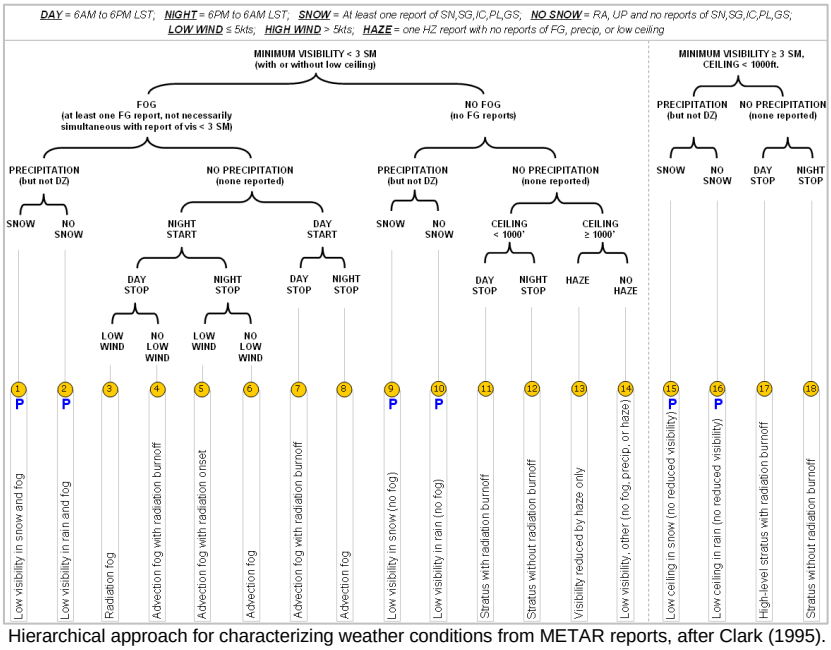
Clark, D.A., 1995: Characterizing the causes of low ceiling and visibility at U.S. airports. *6th Conf. on Aviation Weather Systems*, Dallas, TX, Amer. Meteor. Soc., 325-330.

Federal Aviation Administration, cited 2007: Terminal Area Forecast. [Available online at http://aspm.faa.gov/main/taf.asp.]

National Weather Service Instruction 10-813, 2005: *Aviation Weather Services, NWSPD 10-8, Terminal Aerodrome Forecasts*, 57 pp.

Wilks, D.S., 2006: *Statistical Methods in the Atmospheric Sciences,* Second Edition. International Geophysics Series, Vol. 91, Academic Press, 627 pp.

Appendix:



Hierarchical approach for characterizing weather conditions from METAR reports, after Clark (1995).