

Scale Normalization for IFR-Frequency Effects in Aviation Forecast Performance Statistics

MATTHEW LORENTSON
National Weather Service, Silver Spring, Maryland

(Manuscript received 21 August 2013; review completed 15 November 2013)

ABSTRACT

The National Weather Service uses *probability of detection* and *false alarm ratio* to assess forecast performance. Statistical evidence indicates that a quantitative relationship exists between these forecast performance metrics and the frequency with which a forecasted condition occurs. Current national aviation performance goals do not account for this relationship, which reduces their utility. There is meaningful evidence that indicates the influence of low ceiling/visibility frequency on national aviation forecast performance metric averages can be neutralized through scale normalization.

1. Introduction

The United States' Government Performance and Results Act (GPRA) of 1993 requires government agencies to measure outcome-related goals for major functions and operations (Office of Management and Budget 2013). To meet that requirement, the National Weather Service (NWS) sets forecast performance goals. One such goal within the NWS aviation program targets instrument flight rules (IFR) forecast criteria. This is important because IFR thresholds—ceilings <304.8 m (1000 ft) above ground level and/or surface-based visibility <4.8km (3 mi)—are critical to many aviation operations. National IFR performance scores exhibited steady improvement from 2007 to 2010, and in 2010 the NWS set GPRA goals for the 2011–2016 period based on the improvement trend, as shown in Table 1.

Curiously, steady improvement in national IFR forecast scores was not sustained. Terminal aerodrome forecast (TAF) performance metrics decreased in 2011 and 2012, although there was no obvious reason for the decline. It was noted, however, by both NWS headquarters personnel and forecasters in the field that observed IFR conditions occurred with less frequency during the 2011–2012 period. It was suspected that the diminished percentage of observed IFR frequency

(henceforward termed “IFR Frequency”) was related to the reduction in IFR forecast performance. The NWS Aviation Services Branch responded to this concern and investigated the relationship between IFR forecast metrics and IFR Frequency. It was found that when forecast datasets containing a large number of TAF locations were normalized for IFR Frequency, the 2009–2012 period actually exhibited an improvement trend. This study will describe the procedure used to scale-normalize the forecast performance figures for IFR Frequency and describe new national IFR “total performance index” goals.

2. Data and methods

The NWS uses probability of detection (POD) and false alarm ratio (FAR) to assess TAF performance. POD and FAR are defined using dichotomous, yes–no elements described in “hit” and “miss” terms. Definitions for forecast hits, misses, and false alarms are as follows:

- hit – event forecasted to occur, and did occur
- miss – event forecasted not to occur, but did occur
- false alarm – event forecasted to occur, but did not occur

Table 1. NWS annual TAF POD averages, FAR averages, and national 2010 GPRA goals; data available from “NWS GPRA Measures – Performance Charts” [www.nws.noaa.gov/cfo/program_planning/program_planning.htm, Text Only Version, Aviation Forecasts (Ceiling/Visibility Forecasts)]. Values are percentages expressed as decimals; years refer to fiscal calendar. Actual POD and FAR averages for the nation in 2011–2013 are underlined and were added by the author. The author also modified some terms used in this table to be more representative (i.e., “POD” was substituted for “Accuracy” and “FAR” was substituted for “False Alarm Rate”).

	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
POD: <4.8km (3 mi) / <304.8 m (1000 ft)	0.61	0.62	0.63	0.65	<u>0.63</u>	<u>0.61</u>	<u>0.62</u>			
POD Goals: <4.8km (3 mi) / <304.8 m (1000 ft)				0.65	0.66	0.67	0.68	0.69	0.69	0.70
FAR: <4.8km (3 mi) / <304.8 m (1000 ft)	0.40	0.39	0.38	0.36	<u>0.39</u>	<u>0.39</u>	<u>0.37</u>			
FAR Goals: <4.8km (3 mi) / <304.8 m (1000 ft)				0.42	0.41	0.40	0.39	0.38	0.38	0.38

POD is a fraction that describes the number of hits versus the number of hits and misses; FAR is a fraction that describes the number of false alarms versus the total number of hits and false alarms. These metrics complement each other and should be used together (Centre for Australian Weather and Climate Research 2013). The definitions for each metric can be visualized in Table 2 where POD is equal to $a/(a+c)$ and FAR is equal to $b/(a+b)$. “Inverse FAR,” or $1-b/(a+b)$, is described by Schaefer (1990) as the success ratio (SR) of hits to the total number of event forecasts. SR describes the positive fraction of predicted “yes” events that were *not* false alarms.

Table 2. Contingency table for forecast verification. $POD = a/(a+c)$ and $FAR = b/(a+b)$.

Contingency Table			
	Observed		Total
Forecast	<i>a. hits</i>	<i>b. false alarms</i>	<i>Forecast yes</i>
	<i>c. misses</i>	<i>b. correct negatives</i>	<i>Forecast no</i>
Total	<i>Observed yes</i>	<i>Observed no</i>	Total

Observations used in the calculation of POD and FAR are recorded by equipment such as the Automated Surface Observing System (ASOS). ASOS equipment at each TAF airport records and disseminates cloud ceiling (base) and surface visibility—descriptors of IFR conditions—as either a regularly scheduled meteorological aerodrome report (METAR) or special report (SPECI) for weather changes that meet certain criteria (NWS 2013a). NWS Statistics-on-Demand (SOD), the official NWS TAF verification tool, evaluates TAFs every 5 min (12 times per hour), which is 288 times per day. The 5-min interval times end in either a “0” or “5.” Forecast conditions at the end of each 5-min interval are matched with the most recently reported METAR/SPECI, and each element (e.g., ceiling) is verified separately. Routine hourly

METARs not received just before the hour are assumed to be missing, and all 5-min verification intervals following that scheduled METAR are discarded as missing until a new METAR or SPECI is reported (NWS 2013b). Frequency ratios of IFR observations and forecast POD/FAR may be calculated in SOD for any TAF location throughout the United States, its territories, and Micronesia.

Table 3. National NWS IFR GPRA criteria used in this study to generate observed IFR Frequency and TAF POD/FAR information on the NWS SOD site. Criteria selections in the table are expressed in query terminology and have the following meaning (in sequential order from the top): time range start and end dates for the specific period of analysis—example given is the first month of the data sample used in this paper, October 2005; the pertinent month is identified for the report; area selected is the entire “National” TAF set; element type “Flight Category” separates results into categories such as IFR and VFR; “Operational Impact” forecast type verifies the forecast in effect that is most likely to have the largest impact on operations (generally the least favorable condition); Global Forecast System Model Output Statistics (GFS MOS) guidance type is included for comparison in this report only to reproduce the official NWS GPRA statistics, not to use the model performance figures; ceiling and visibility criteria set to IFR flight category—ceiling <304.8 m (1000 ft) and/or visibility <4.8 km (3 mi); only scheduled TAF types analyzed (amendments excluded); no TAF begin times excluded; and only the first two forecast projection periods of evaluated TAFs analyzed (0–3 and 3–6 h).

Criteria	Selection
From date:	10/01/2005
To end date:	10/31/2005
Months to report:	OCT
Selection area:	National
Element type:	Flight Category
Forecast type:	Operational Impact
Guidance type:	GFS MOS
Ceilings below:	1000 ft
Visibilities below:	3 Statute Miles
TAF type:	Scheduled
TAF begin times:	Select All
Forecast projections:	>0–3, >3–6

IFR forecast information from TAFs may be compared to IFR observations in SOD for multiple locations to derive a POD and FAR report for regional TAF-location groups or even the entire NWS TAF dataset. IFR Frequency is generated in these same reports. The SOD database contains records from 1 September 2005 forward and is available to registered users (NWS 2013c). The SOD criteria used to generate IFR Frequency, IFR POD, and IFR FAR data in this

study are outlined in Table 3. These parameters are identical to those used by the NWS for official national aviation GPRA analyses and goals. The selection area criterion “National” described on line #4 specifies the use of data from all 600+ TAF/observation sites in the SOD database—a large body of data. For example, a total of over 64 million 5-min observation and forecast comparisons were generated in total for the 2012 national sample alone.

Table 4. An excerpt of monthly National NWS IFR POD, IFR FAR, and IFR SR performance averages for TAFs (i.e., the first and last fiscal years of the sample), expressed as percentages $\times 100$. The full dataset used in this paper includes all 96 months from fiscal year 2006 through 2013. The TPIX in the fifth column from the left is the product of multiplying POD and SR. IFR Frequency is included in the sixth column. All data were derived from the NWS SOD database using GPRA criteria (see Table 3).

Month	POD	FAR	SR (1 – FAR)	TPIX (POD \times SR)	IFR Freq.
October-05	64.30	40.00	60.00	3858.00	7.69%
November-05	62.30	40.70	59.30	3694.39	6.95%
December-05	67.40	34.80	65.20	4394.48	11.51%
January-06	67.40	37.20	62.80	4232.72	10.41%
February-06	61.00	39.30	60.70	3702.70	7.42%
March-06	59.20	43.70	56.30	3332.96	6.15%
April-06	51.90	51.00	49.00	2543.10	4.15%
May-06	61.10	45.10	54.90	3354.39	4.72%
June-06	55.80	51.20	48.80	2723.04	4.50%
July-06	56.80	53.30	46.70	2652.56	4.04%
August-06	58.20	46.70	53.30	3102.06	5.45%
September-06	60.70	43.10	56.90	3453.83	6.65%
...
October-12	60.10	36.90	63.10	3792.31	6.52%
November-12	56.90	37.60	62.40	3550.56	5.46%
December-12	65.20	34.40	65.60	4277.12	12.22%
January-13	69.20	31.80	68.20	4719.44	12.32%
February-13	64.40	34.40	65.60	4224.64	10.58%
March-13	62.40	35.70	64.30	4012.32	7.09%
April-13	63.20	37.80	62.20	3931.04	7.48%
May-13	58.90	39.50	60.50	3563.45	6.06%
June-13	55.50	44.00	56.00	3108.00	5.02%
July-13	58.40	41.60	58.40	3410.56	5.63%
August-13	57.50	41.00	59.00	3392.50	5.45%
September-13	52.70	42.50	57.50	3030.25	4.98%

Table 4 shows an excerpt of the monthly national TAF IFR POD, IFR FAR, and IFR Frequency data generated from IFR GPRA criteria, fiscal years 2006–2013. SR is calculated in column 4 and the total performance index (TPIX) is in column 5. TPIX, the product of multiplying POD and SR, is introduced to (1) force the use of POD and FAR in conjunction with each other, (2) streamline performance analysis, and (3) create efficiency and simplicity in NWS reporting. As mentioned previously, POD and FAR should be used together; each metric can be enhanced at the expense of the other. Emphasis on POD performance improvement should not occur at the expense of FAR

performance; use of a $\text{POD} \times \text{SR}$ index is a means to eliminate singular focus on either metric. Roebber (2009) described two other metrics, bias and critical success index (CSI), that together with POD and SR all move toward unity when forecasts are good. The relationship among these metrics is represented graphically in Fig. 1. Note that TPIX, the product of POD and SR described by the area of a quadrilateral, is maximized in the form of a square when bias = 1.0. As bias is equal to POD/SR , both POD and SR contribute equal influence on TPIX when bias = 1.0. Bias can therefore be used to qualify the goodness of TPIX scores. For the sake of simplicity, bias will not be used

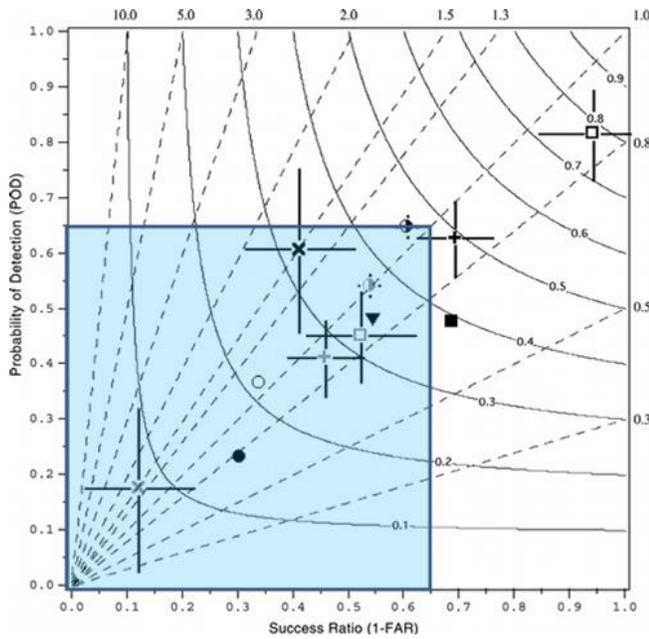


Figure 1. The geometric relationship among POD, SR, bias, and CSI (image from Roebber 2009). Dashed lines represent bias scores with labels on the outward extension of the line, whereas labeled solid contours are CSI. Cross and shape figures represented on the graph are not discussed here. TPIX can be visualized on this graph as a quadrilateral area calculated by multiplying POD by SR. For example, using whole numbers, the blue square area with POD and SR scores of 65 (bias = 1.0) produces a TPIX of 4225. TPIX is maximized in the form of a square when bias = 1.0. *Click image for an external version; this applies to all figures hereafter.*

in the calculation of new IFR GPRA goals in this paper, but it should be considered for more in-depth TPIX analysis.

Annual national IFR TPIX and national observed IFR Frequency averages show similar distributions when graphed (Fig. 2). As IFR Frequency increased from 2007 to 2010, there was a commensurate increase in TPIX. Then, as IFR Frequency decreased in 2011 and 2012, TPIX followed. The proportional rate of change among these coefficients suggests that IFR Frequency should predict TPIX in a linear manner. A scatterplot of monthly IFR Frequency and TPIX data (Fig. 3a) visually verifies a positive relationship between the coefficients that approximates a straight line. Regression can be used to find an equation that best estimates the relationship.

3. Analyses and discussion

This approach is similar to that used by Berk and Carey (2000, pp. 299–367). The least squares method of simple linear regression is an attempt to find the line that best estimates the relationship between two

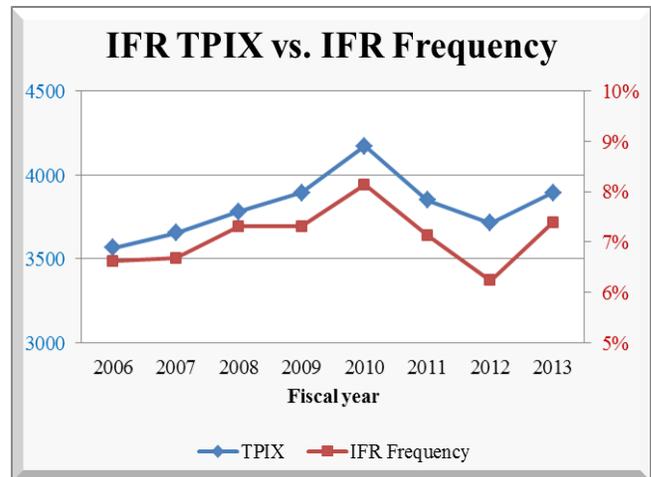


Figure 2. IFR TPIX and IFR Frequency averages for all NWS TAF locations by fiscal year. Data were derived from NWS SOD database using GPRA criteria (see Table 3). IFR Frequency range (right axis) is magnified compared to IFR TPIX range (left axis) to emphasize comparable distribution shapes.

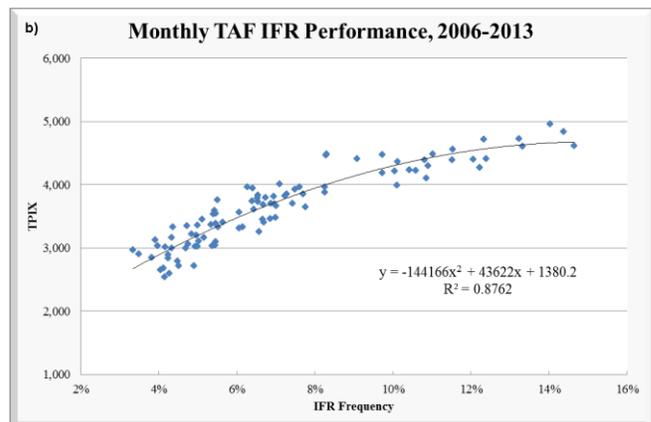
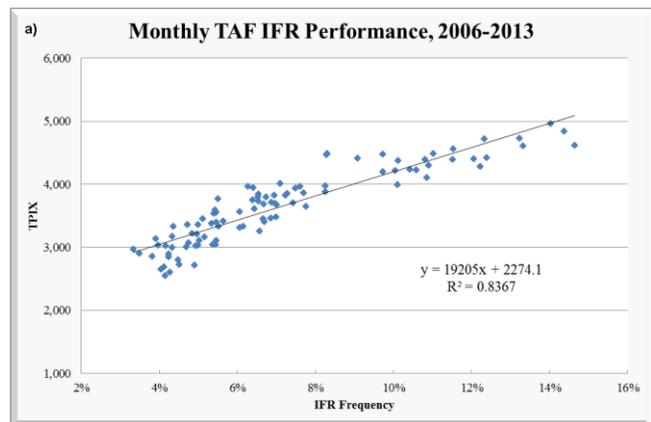


Figure 3. (a) Scatterplot of monthly national IFR TPIX and observed IFR Frequency from fiscal years 2006 to 2013 (columns five and six of Table 4). The trend line and statistics are derived from simple linear regression. (b) Same as (a) except the data were fitted with a 2nd order polynomial trend line. The adjusted R^2 value for this multiple regression is 0.874 (see Table 5).

variables—the x , or independent, variable, and the y , or dependent, variable. In the Fig. 3a scatterplot comparison of national IFR Frequency as the independent variable and national IFR TPIX as dependent variable, simple linear regression generates an R^2 (coefficient of determination) value of 0.84. However, the scatterplot data visually exhibit a slight curve and may be estimated more accurately in a polynomial multiple regression. This technique, which uses both IFR Frequency and squared IFR Frequency as independent variables to predict TPIX, yields the graph in Fig. 3b and regression statistics in Table 5. The multiple regression-derived adjusted R^2 value of 0.87 indicates that almost 87% of the variation in TPIX can be explained by the change in IFR Frequency. In other words, a little less than 13% of the variation in TPIX is presumed to be due to random variability. From the analysis of variance (ANOVA) table, the probability of error p value (*Significance F*) of 6.4×10^{-43} indicates that there is only a 6.4×10^{-43} percent probability the result could occur by random chance. These results are robust and validate the use of scale normalization for IFR Frequency effects in the development of more meaningful IFR forecast performance goals.

Table 5. Multiple regression and analysis of variance (ANOVA) statistics for IFR Frequency-predicted TPIX. The adjusted R^2 value indicates that around 87% of the variation in TPIX is explained by changes in IFR Frequency. Standard Error is expressed in terms of TPIX points.

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.936
R^2	0.876
Adjusted R^2	0.874
Standard Error	211.147
Observations	96
ANOVA	
	<i>Significance F</i>
Regression	6.40×10^{-43}

Using IFR Frequency to predict IFR TPIX in the 2nd order polynomial regression yields the equation

$$y = -144166x^2 + 43622x + 1380.2 \quad (1)$$

where x is IFR Frequency and y is predicted TPIX. The trend line crosses the x axis near 1380. This means that if IFR Frequency decreased to zero, TPIX would be 1380.2—an impossible outcome if IFR conditions did not occur. On the other end of the range, the slope

of the curve turns negative as IFR Frequency increases to around 15.13% and predicted IFR TPIX decreases to zero again at around 31.22%, another impossible outcome. These concerns do not invalidate the regression equation, but they do explicitly remind us to avoid extrapolation of the regression beyond the bounds of the current range of observed data.

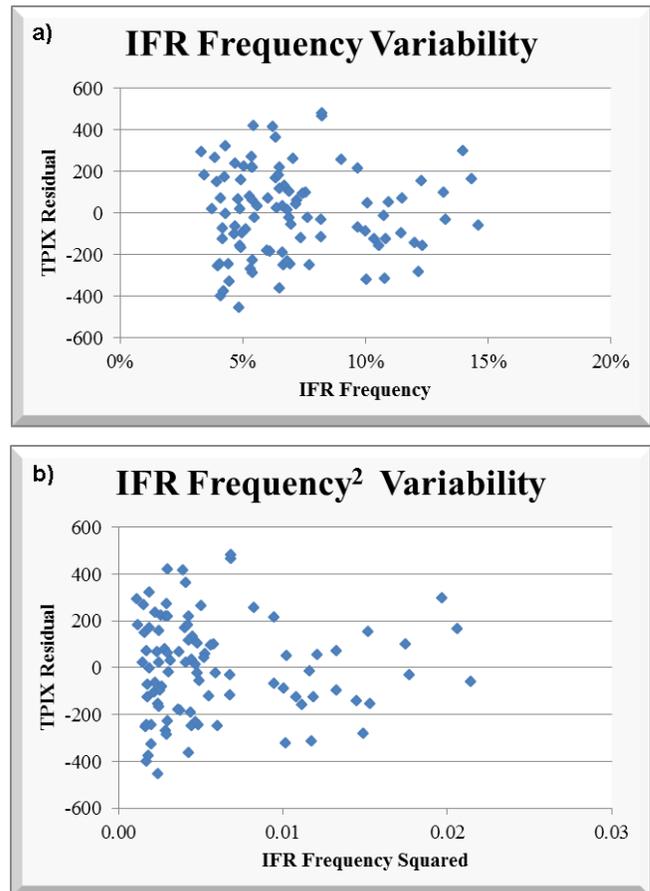


Figure 4. (a) TPIX residuals versus IFR Frequency for the TPIX multiple regression. The residuals appear to exhibit constant variability, without a curve or wedge shape in evidence. (b) TPIX residuals versus squared IFR Frequency for the TPIX multiple regression. Standard deviation does not appear as constant as with the other independent variable in (a), but neither a curve nor wedge shape is clearly evident, either. If future large and positive TPIX performance scores (200–500 overpredicted) are realized in the 9–14% IFR Frequency range, the regression residuals will exhibit more obvious constant variability.

Diagnostics should be used to test the regression model for assumptions in constant variance and normal distribution. Plots of the residuals versus the independent variables (Fig. 4) each demonstrate reasonably constant variability, with no obvious rainbow, u-shape, or wedge pattern in evidence. The squared IFR Frequency coefficient’s standard devia-

tion distribution emphasizes the gap in large, positive TPIX performance scores (200–500 overpredicted) in the 9–14% IFR Frequency range. Presumably, such scores are attainable. If those scores are realized in the future, they will resolve any concerns about constant variability. Figure 5, a normal probability plot, compares regression residuals to scores from a standard normal distribution and shows whether the IFR Frequency/POD regression residuals are normally distributed. The points do not violate linear distribution; the slight departure at the upper limit is not strong enough to invalidate the assumption of normality in this regression. In summary, the diagnostics indicate that this small-sample regression does not violate normal distribution or significant constant-variance assumptions.

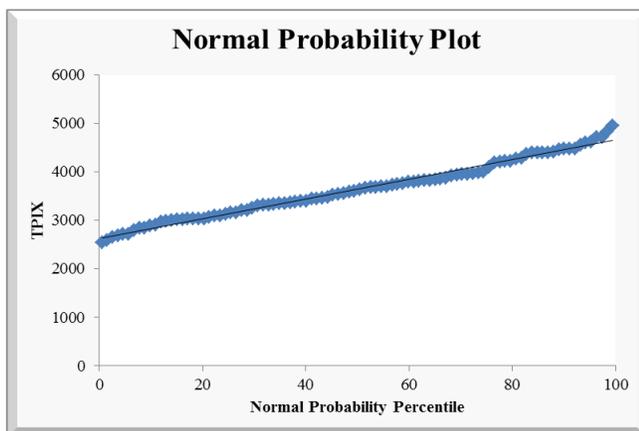


Figure 5. Probability plot of monthly TPIX scores versus standardized distribution position, used to test for normal distribution of residuals. The black line represents the standardized values expected from a standard normal distribution sample. The points fall close to the line—no issues significant enough to invalidate the assumption of normality.

The regression-predicted TPIX compared to actual TPIX statistics for 2006–2013 yields differences (residuals) that range from -452.75 to 480.39 . The residuals describe scale-normalized national TPIX and account for the gross influence of IFR Frequency. Table 6 shows the variability and shape statistics of the residuals, including standard deviation and standard error. To calculate the standard error and 95% confidence interval, the standard deviation of 208.91 is divided by the square root of the sample size, 96, which yields 21.32. This means that the residual sample average will fall within 21.32 of the residual mean (average) of zero with 95% confidence. The relatively large standard deviation and small standard error suggest that longer-period averages, such as

Table 6. Univariate statistics derived from the TPIX multiple regression residuals sample. Note the large monthly standard deviation, expressed in TPIX points, which indicates significant variability, and that TPIX averages over time should be used to measure performance, not individual months.

Residuals Statistics	
Average	0.00
Median	6.25
Standard Deviation	208.91
Variance	43 644.47
Standard Error	21.32
Skewness	0.11
Kurtosis	-0.52

yearly or running averages, should be verified for performance instead of individual, wide-ranging monthly performances.

A plot of the TPIX residuals (Fig. 6) depicts the scale-normalized monthly performance trend during the past eight fiscal years. The TPIX residuals trend line equation,

$$y = 2.3253t - 112.78, \quad (2)$$

can be used to predict improvement over IFR Frequency-predicted TPIX (y) for a future month (t). For example, inserting month 100 (January 2014—the 100th month since September 2005) into Eq. (2) yields a residual, an improvement-over-predicted TPIX goal, of 119.75. The residual trend provides a reasonable approximation of scale-normalized performance with which to adapt performance goals and control for the gross influence of IFR Frequency. Again, because the standard deviation is large for monthly residuals, a 12-mo average of the monthly goals and actual performance residuals is used for each year. Twelve-month performance averages for fiscal years 2014–2020 are calculated in Table 7 and describe new GPRA performance goals that are scale normalized for IFR Frequency. The annual residual trend also can be used to calculate past performance (Fig. 7). It must be emphasized that the scale-normalization procedure used here relies on national data; the IFR Frequency effect may only be predictable in large verification group samples. The regression should be recalculated when national IFR Frequency values occur outside the 2006–2013 monthly range of 3.34–14.64%.

4. Conclusions

True NWS TAF IFR performance is difficult to assess without accounting for IFR Frequency's gross influence on national averages. The correlation be-

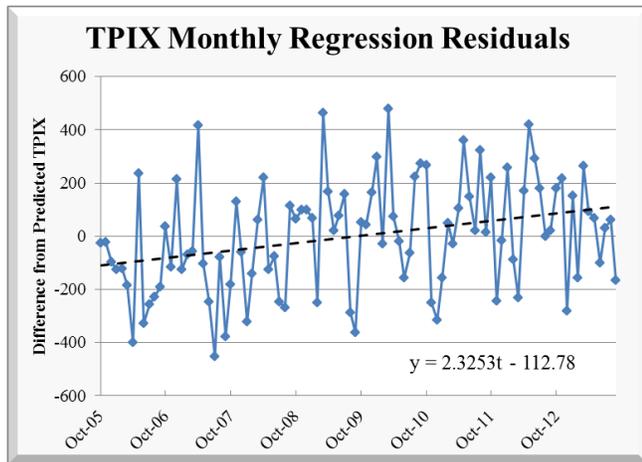


Figure 6. Time series of the differences between observed TPIX and predicted TPIX, or residuals. The residuals represent 2006–2013 monthly performance that is scale-normalized for IFR Frequency effects. The trend line can be used to set new NWS GPRA goals.

Table 7. Scale-normalized IFR TPIX GPRA goals. The values are 12-mo TPIX residual point averages extrapolated from the 2006 to 2013 normalized performance trend (see Fig. 6), and describe performance over IFR Frequency-predicted TPIX. The point goals for 2014–2016 in this table are plotted in Fig. 7 as the extension of the TPIX multiple regression residuals trend.

Fiscal Year	Goal
2014	126
2015	153
2016	181
2017	209
2018	237
2019	265
2020	293

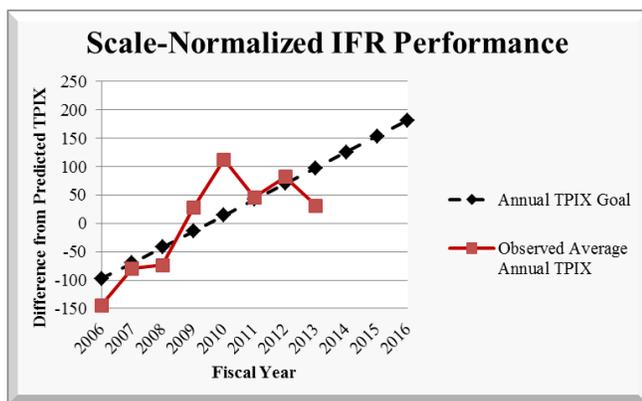


Figure 7. Annual NWS IFR performance and revised GPRA Goals based on scale-normalized TPIX. The annual observed TPIX performance in red is the 12-mo TPIX average for each fiscal year. Annual TPIX goals (in black) are 12-mo averages generated from the trend given in Fig. 6.

tween annual national IFR Frequency and IFR POD/FAR metrics suggests that uncorrected scores are less a measure of forecast performance than they are a measure of IFR Frequency by proxy. The significance of the correlation between IFR Frequency and IFR performance in large data samples indicates that regression-derived scale normalization can be used to neutralize IFR Frequency effects in the national dataset. Scale-normalized IFR TPIX values place greater emphasis on true performance compared to uncorrected metrics, and simplify NWS verification output by combining POD and FAR into a single index.

Acknowledgments. The author gratefully acknowledges Dr. Mike Baldwin (Purdue University), Dr. David Bright (NWS Aviation Weather Center), Dr. Adam Clark (NOAA NSSL CIMMS), Todd Lericos (NWS WFO Las Vegas), Kevin Stone (NWS Aviation Services Branch), and Doug Young (NWS Performance Branch) for their manuscript edits and input. The author also thanks Matthew Bunkers (NWS WFO Rapid City) and Connie Crandall (SD School of Mines and Technology) for technical editing assistance. The views expressed are those of the author and do not necessarily represent those of the National Weather Service.

REFERENCES

Berk, K. N., and P. M. Carey, 2000: *Data Analysis with Microsoft Excel*. Duxbury Press, 587 pp.

Centre for Australian Weather and Climate Research, cited 2013: Forecast verification: Issues, methods and FAQ. [Available online at [www.cawcr.gov.au/projects/verification/#What makes a forecast good.](http://www.cawcr.gov.au/projects/verification/#What%20makes%20a%20forecast%20good)]

NWS, cited 2013a: Automated Surface Observation System (ASOS) User’s Guide. [Available online at [www.nws.noaa.gov/asos/aum-toc.pdf.](http://www.nws.noaa.gov/asos/aum-toc.pdf)]

NWS, cited 2013b: National Weather Service Instruction 10-1601: Verification. [Available online at [www.nws.noaa.gov/directives/sym/pd01016001curr.pdf.](http://www.nws.noaa.gov/directives/sym/pd01016001curr.pdf)]

NWS, cited 2013c: NWS Performance Management. [Available online at [verification.nws.noaa.gov/.](http://verification.nws.noaa.gov/)]

Office of Management and Budget, cited 2013: Government Performance and Results Act of 1993. [Available online at [www.whitehouse.gov/omb/mgmt-gpra/gplaw2m#h2.](http://www.whitehouse.gov/omb/mgmt-gpra/gplaw2m#h2)]

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601–608.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting*, **5**, 570–575.