U.S. DEPARTMENT OF COMMERCE
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
NATIONAL WEATHER SERVICE
OFFICE OF SCIENCE AND TECHNOLOGY
METEOROLOGICAL DEVELOPMENT LABORATORY

MDL OFFICE NOTE 02-04

DETERMINING THE EFFECT OF IFPS IMPLEMENTATION ON VERIFICATION METRICS

Jerry L. Gorline

August 2002

DETERMINING THE EFFECT OF IFPS IMPLEMENTATION ON VERIFICATION METRICS

1. INTRODUCTION

In this report, we investigated the impact of implementing the Interactive Forecast Preparation System (IFPS) on accuracy and skill of maximum and minimum (max/min) temperature and Probability of Precipitation (PoP) forecasts. For this study, we found 29 stations that had sufficient data to compare scores for a pre-IFPS and post-IFPS period. We will refer to these sites as "IFPS stations." Because each IFPS station did not implement IFPS at the same time, obtaining a large matched sample for pre- and post-IFPS periods was not possible. Instead, for each individual site, we chose the 12 months prior to IFPS implementation and verified the forecasts made during this time (henceforth, termed the "pre-IFPS period"). Subsequent to the IFPS implementation, we eliminated the first 3 months as a transition period; we then chose the next 12 months and verified the forecasts made during this time (henceforth, termed the "post-IFPS period"). Comparing scores computed over 12-mo periods eliminates the effect of seasonal variations, although the inherent difference in the difficulty of forecasting during the pre- and post-IFPS periods is still present. In the final analysis, we combined verification scores for all 29 stations even though the pre- and post-IFPS periods did not coincide.

Table 1 lists the 29 stations we used and includes the IFPS start date for each station and the pre-IFPS/post-IFPS periods used to calculate the performance metrics. The start date is the month a WFO began regularly producing their full suite of products (e.g., zones, CCFs, RDFs) via IFPS, day in and day out. These dates were verified by cross-checking MDL IFPS status records against those of the regional IFPS focal points.

2. COMPARISON OF SCORES

For max/min temperature forecasts, we calculated the Mean Absolute Error (MAE) and for PoPs we calculated the Brier Score which represents the Mean Squared Error (MSE) of the probability forecasts. For both performance measures, a smaller value corresponds to a decrease in error or improvement in forecast performance. We computed the pre-IFPS and post-IFPS scores and used the paired t-test to test the hypothesis that differences were not significant. For each pair of values, we calculated the difference $D_i = preIFPS_i - postIFPS_i$, where the subscript i represents the score for the i^{th} station. The paired t-test is:

$$t = \left(\frac{\overline{D} - D_o}{\hat{\sigma}/\sqrt{n}} \right),$$

where $\overline{D}$ is the mean of the differences, $\hat{\sigma}$ is an estimate of the standard deviation of D_i, n is the sample size (29 for the study), and $D_o = 0$, namely, the null hypothesis that the two sample means are the same.

A positive t-value corresponds to an improvement in post-IFPS forecast performance compared with pre-IFPS performance and a negative t-value corresponds to a degradation. For a sample size of 29, if the absolute value of the t-value exceeds 1.701, we can say with 95% confidence that the difference between post-IFPS and pre-IFPS is significant and not the result of random

fluctuations. At the 90% confidence level, the t-value must exceed 1.313 for the difference to be considered significant.

Table 1. Twenty-nine IFPS stations including IFPS start dates, pre-IFPS, and post-IFPS periods.

Call Letters	Site Name	IFPS Start Date	Pre-IFPS Period	Post-IFPS Period
KPIT	Pittsburgh, PA	04/2000	04/1999 - 03/2000	07/2000 - 06/2001
KALB	Albany, NY	"	"	"
KCAE	Columbia, SC	05/2000	05/1999 - 04/2000	08/2000 - 07/2001
KDCA	Washington, DC	06/2000	06/1999 - 05/2000	09/2000 - 08/2001
KROA	Roanoke, VA	07/2000	07/1999 - 06/2000	10/2000 - 09/2001
KEWN	New Bern, NC	08/2000	08/1999 - 07/2000	11/2000 - 10/2001
KCLE	Cleveland, OH	09/2000	09/1999 - 08/2000	12/2000 - 11/2001
KILM	Wilmington, NC	"	"	"
KPHL	Philadelphia, PA	"	"	"
KRDU	Raleigh, NC	"	"	"
KERI	Erie, PA	"	"	"
KORF	Norfolk, VA	10/2000	10/1999 - 09/2000	01/2001 - 12/2001
KBOS	Boston, MA	"	"	"
KPWM	Portland, ME	"	"	"
KATL	Atlanta, GA	"	"	"
KPVD	Providence, RI	"	"	"
KCON	Concord, NH	"	"	"
KEWR	Newark, NJ	11/2000	11/1999 - 10/2000	02/2001 - 01/2002
KBGM	Binghamton, NY	"	"	"
KBUF	Buffalo, NY	"	"	"
KGSP	Greenville, SC	"	"	"
KSYR	Syracuse, NY	"	"	"
KAVP	Scranton, PA	"	"	"
KCLT	Charlotte, NC	"	"	"
KLGA	New York, NY	"	"	"
KBTV	Burlington, VT	12/2000	12/1999 - 11/2000	03/2001 - 02/2002
KCAR	Caribou, ME	"	"	"
KCHS	Charleston, SC	"	"	"
KSAV	Savannah, GA	"	"	"

The results for max/min temperature forecasts are shown in Tables 2a and 2b for the 0000 and 1200 UTC forecast cycles, respectively. In these tables and subsequent tables, an asterisk next to a t-test value denotes a difference that exceeds the 90% significance threshold and a double asterisk denotes a difference that is significant at the 95% confidence level. For the 0000 UTC cycle, all four forecast projections show improvement in post-IFPS performance. For two of the four projections, namely, the 48-h max and 60-h min, the improvement is significant at the 95% level. For the 1200 UTC cycle, all forecast projections show improvement with the 36 and 60-h max showing significant improvement at the 95% level.

The PoP verification results are shown in Tables 3a and 3b for the 0000 and 1200 UTC cycles, respectively. For the 0000 UTC cycle, all three of the forecast projections show improvement. For the 36-h projection, the improvement exceeds the 95% significance threshold, while the 48-h projection exceeds the 90% significance threshold. For the 1200 UTC cycle, the improvement for the 36-h forecast projection exceeds the 90% significance threshold, and the improvement for the 48-h projection exceeds the 95% significance threshold.

Table 2a. Paired t-test results for max/min temperature forecasts, 0000 UTC cycle.

Projection	Mean of Differences	Paired T-Test Value
24-h Max	0.0059	0.165
48-h Max	0.1217	2.642**
36-h Min	0.0034	0.107
60-h Min	0.1510	3.850**

Table 2b. Same as Table 2a, except for the 1200 UTC cycle.

Projection	Mean of Differences	Paired T-Test Value
24-h Min	0.0131	0.393
48-h Min	0.0452	1.303
36-h Max	0.0803	2.492**
60-h Max	0.2579	4.725**

Table 3a. Paired t-test results for PoP forecasts, 0000 UTC cycle.

Projection	Mean of Differences	Paired T-Test Value
24-h PoP	0.0008	0.277
36-h PoP	0.0057	1.726**
48-h PoP	0.0044	1.435*

Table 3b. Same as Table 3a, except for the 1200 UTC cycle.

Projection	Mean of Differences	Paired T-Test Value
24-h PoP	0.0033	1.095
36-h PoP	0.0049	1.447*
48-h PoP	0.0083	2.551**

In summary, of the 14 comparisons made, six demonstrated improvement at the 95% level; two more indicated improvement at the 90% level.

3. POSSIBLE IMPROVEMENT WITHOUT IFPS

There is considerable yearly variation in temperature and PoP scores, but the scores have been gradually improving over the years. It cannot be known what the post-IFPS scores, at the IFPS stations, would have been if IFPS had not been used. Because of the gradual improvement over the years, one might expect that the scores at the IFPS stations would have improved whether or not IFPS were used.

To address this question, we computed the trend of the scores over a 5-yr period by using the results from the AFOS-Era forecast Verification (AEV) program. For this purpose, we did not stagger the verification periods as we did in computing the pre-IFPS and post-IFPS scores. Annual scores (April 1 - March 31) from the 1995-2000 period were used to compute these trends. A trend was computed for each weather element and projection. Six of the 29 stations did not have sufficient data to obtain a trend, so the trends are based on combined data from 23 stations. All trends showed that the forecasts were improving over time. We assume these trends to be a reasonable estimate

of what would have happened without IFPS. We then applied these trends to the pre-IFPS scores and estimated what the post-IFPS scores might have been for the 12-mo period that followed the 3-mo transition. These estimates were compared to the actual post-IFPS scores. The results are shown in Tables 4 and 5.

Table 4a. Paired t-test results for max/min temperature forecasts after correcting for the expected trend, 0000 UTC cycle.

Projection	Mean of Differences	Paired T-Test Value
24-h Max	-0.0688	-1.932**
48-h Max	0.0206	0.447
36-h Min	-0.0208	-0.643
60-h Min	0.0997	2.541**

Table 4b. Same as Table 4a, except for the 1200 UTC cycle.

Projection	Mean of Differences	Paired T-Test Value
24-h Min	0.0014	0.041
48-h Min	0.0269	0.777
36-h Max	-0.0160	-0.497
60-h Max	0.1572	2.879**

Table 5a. Paired t-test results for PoP forecasts after correcting for the expected trend, 0000 UTC cycle.

Projection	Mean of Differences	Paired T-Test Value
24-h PoP	-0.0045	-1.523*
36-h PoP	0.0008	0.256
48-h PoP	-0.0002	-0.069

Table 5b. Same as Table 5a, except for the 1200 UTC cycle.

Projection	Mean of Differences	Paired T-Test Value
24-h PoP	-0.0013	-0.451
36-h PoP	-0.0009	-0.227
48-h PoP	0.0034	1.048

The mean differences have, of course, decreased from those shown in Tables 2 and 3 and half are even negative, indicating that if these trends represent what would have happened without IFPS, the scores would have been better. The t-test now indicates some post-IFPS scores are slightly better than the trend would suggest and some indicate the scores are slightly worse. Overall, we conclude there is little difference in the post-IFPS scores compared with the trend estimates.

4. CONCLUSIONS

Our purpose in doing this study was to judge whether IFPS had a negative impact on performance measures. The data available showed improvement in verification metrics during a 1-yr post-IFPS period compared to a 1-yr pre-IFPS period. The natural variability in yearly scores is of the same order of magnitude as the differences we computed for the two periods. This variability is largely due to difficulty in forecasting for one year compared to another. Therefore, it is not possible to say definitely, even when considering average trends, what effect IFPS had on forecast performance. However, we can say there was no evidence to indicate general degradation.

ACKNOWLEDGMENTS

I would like to thank Paul Dallavalle, Bob Glahn, Dave Ruth, and Wilson Shaffer for their advice and guidance throughout this investigation. I would also like to thank Valery Dagostaro for her advice and guidance in the calculation of the forecast performance measures. Thanks also goes out to Michael Schenk and Arthur Taylor for their valuable contributions.