LEVEL II

AWS/TN-79/005

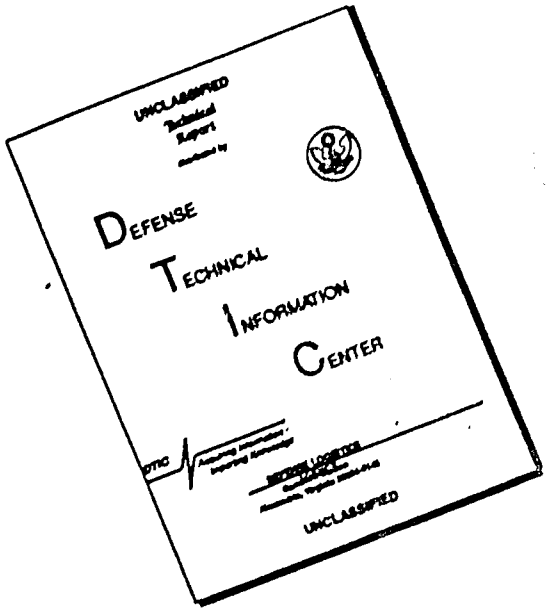
FORECAST SKILL SCORE TEST - FINAL REPORT DECEMBER 1978

William F. Johnson, Lt Col, USAF
Arthur C. Kyle, Maj, USAF
Paul B. Knutson, Capt, USAF

November 1979

AIR WEATHER SERVICE (MAC)
Scott AFB, Illinois 62225

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

REVIEW AND APPROVAL STATEMENT

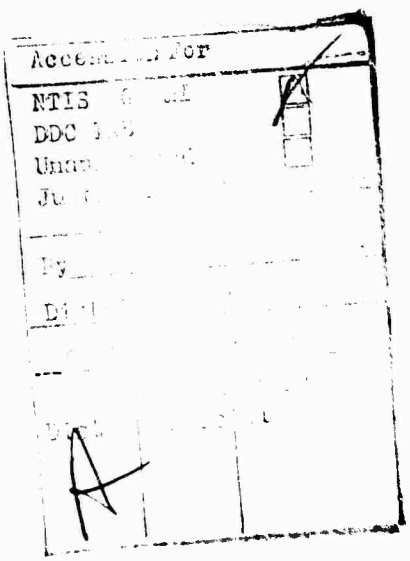
This publication approved for public release. There is no objection to unlimited distribution of this document to the public at large, or by the Defense Documentation Center (DDC) to the National Technical Information Service (NTIS).

This technical publication has been reviewed and is approved for publication.

KENNETH E. GERMAN, Colonel, USAF
Asst DCS/Aerospace Sciences
Reviewing Officer

FOR THE COMMANDER

THOMAS A. STUDER, Colonel, USAF
DCS/Aerospace Sciences
Air Weather Service



UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: AWS/TN-79/005 (14)	2. GOVT ACCESSION NO. / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): Forecast Skill Score Test (6)	5. TYPE OF REPORT & PERIOD COVERED: Final, 1 Oct 77 - 31 Mar 78
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): ~~Lt Col~~ William F. Johnson, Arthur C Kyle, ~~Maj~~ ~~Capt~~ Paul B. Knutson (10)	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: AWS/DOA, Scott AFB, IL 62225	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: (12) 35
11. CONTROLLING OFFICE NAME AND ADDRESS: AWS/DNDA, Scott AFB, IL 62225 (11)	12. REPORT DATE: December 1978 / 13. NUMBER OF PAGES: 32
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): AWS/DNDA, Scott AFB, IL 62225	15. SECURITY CLASS. (of this report): Unclassified / 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE: None

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

(9) Final rept. 1 Oct 77 - 31 Mar 78

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

AWS Skill Score
Brier Score
Gringorten Skill Score
Log Skill Score
Probability Forecasts

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Publication contains results of test of different skill scores to be used by ~~AWS~~ Air Weather Service (AWS). It also presents the first organized attempt by AWS units to issue probability forecasts. This test was conducted from 1 Oct 77 - 31 Mar 78.

014 670

DD FORM 1 JAN 73 1473

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

FORECAST SKILL SCORE TEST

1. Background:

a. For the last ten years, Air Weather Service has used the following skill score for the Terminal Forecast Verification (TAFVER) Program: SS = (S-P)/(100%-P). In this formula, S = Station (Det/Sq/Wing) percent correct, and P = Persistence percent correct. This skill score is simple and straightforward, but also has its limitations.

(1) Since the AWS Skill Score (SS) is sensitive to the quantity P as well as the difference (S-P), climatology has a big effect on a station's score. For example, if S beats P by 5% with P = 85%, then a skill score of .33 results. However, if P = 60%, the skill score would be .125.

(2) The AWS SS rewards forecast hits equally regardless of category/difficulty. Present TAFVER contingency tables show that AWS forecasters predict the low categories less frequently than they occur.

b. One of Dr Robert Miller's first projects, when he became AWS Chief Scientist in Sep 76, was to review TAFVER procedures. He noted that performance in the low categories (below 200/1/2, and 200/1/2 to 1000/2) needed improvement. He attributed this deficiency to the fact that since Persistence is the AWS SS baseline, forecasters tend to wait until they have a good chance of beating Persistence before they go against it (a "tie" with Persistence is better than a "loss"). Consequently, this verification system results in a reluctance to forecast the low categories. To correct this problem, a system is needed that encourages forecasters to forecast low categories as often as they occur. One approach would be a system that gives more credit for hitting the climatologically rare categories. To this end, Dr Miller led discussions which resulted in the proposal of a test of the Gringorten Score. The test was designed so that, in addition to the Gringorten Score, the Log Score, developed by McDonald of NWS, could also be examined. After considering test costs, the AWS commander approved the test plan.

c. As the test plan was being developed, it was realized that with no extra cost or effort the test results could be used to determine our capability to produce skillful and reliable forecasts in probabilistic terms (probability forecasts). Consequently, this objective was also added to the test.

2. Test Objectives:

a. To determine which skill score, if any, should replace the present AWS SS.

b. To assess the participating units' capability to prepare reliable and skillful probability forecasts.

c. To compare subjective probability forecasts prepared by AWS forecasters with objective probability forecasts prepared by the National Weather Service Techniques Development Laboratory (NWS/TDL).

3. Test Schedule:

Sep 77	AWS/DN presented probability forecasting seminar to participating units.
1 Oct 77 - 31 Mar 78	Test conducted.
Jun 78	Analyze and present results.

4. Participating Units: Forecasting units in CONUS Regions 43 and 47 participated in the test and are listed in Attachment 1. On 1 Dec 77, seven more units were added. These units did not receive the AWS/DN probability seminar; the objective was to see if these units were able to prepare probability forecasts as well as the 23 units that received the seminar.

5. Forecast Unit Tasks:

a. Field Units:

(1) Prepared ceiling and visibility probability forecasts at each regularly scheduled forecast time (02Z, 08Z, 14Z, and 20Z), for each ceiling category ($<$200 ft, $\geq$ 200 ft to $<$1000 ft, $\geq$1000 ft to$<$3000 ft, $\geq$ 3,000 ft) and each

visibility category (< 1/2 mi, ≥ 1/2 mi to < 2 mi, ≥ 2 mi to < 3 mi, ≥ 3 mi) for the 3- and 6-hour verifying times. The probability for a particular category could range from 0.00 to 1.00 and the sum of the four categories for both ceiling or visibility had to equal one. Increments of 0.01 were used.

(2) Sent completed test forms (see Attachment 2) to AFGWC twice a month.

b. AFGWC:

(1) At each regularly scheduled forecast time (02Z, 08Z, 14Z, and 20Z), subjectively assigned probabilities to each ceiling category and each visibility category for the 12- and 24-hour verifying times. Categories and probability value instructions were the same as for field units.

(2) Computed Brier Score and reliability and sharpness diagrams for each unit and forecast length for: ceiling, visibility, and ceiling/visibility combined forecasts (the combined probabilities were calculated as in Atch 3); conditional climatology forecast; sample climatology forecast; and persistence forecast. Sent monthly verification feedback to each unit.

(3) Using the probability forecasts and weighting matrices, derived categorical forecasts that would maximize each of the test skill scores. For example, to maximize the AWS SS, the category with the highest probability was selected as the forecast. These were then known as categorical forecasts by "PROB." Categorical forecasts by "GRING" maximized the Gringorten skill score and were determined by multiplying the categorical probabilities by the inverse of the long-term climatology probability for the same category. The highest product was the categorical forecast. Forecasts by "LOG" were determined by multiplying the same probabilities by a matrix that tailored the Log Score to the AWS categories. Then, these three sets of categorical forecasts, each chosen to maximize a skill score, were verified using each of the three skill scores and percent correct. Results were computed monthly and sent to AWS/DOA for analysis.

(4) Verified NWS/TDL model output statistics (MOS) 12- and 24-hour forecasts for the test units. AFGWC's liaison staff at TDL provided tapes of MOS ceiling and visibility forecasts precise to two digits.

6. Test Results:

a. Attachment 3 defines each of the skill scores used and Attachment 4 summarizes verification of the three categorical forecasts by each skill score. Findings are:

(1) Forecasts by PROB were best for percent correct and AWS Skill Score (which is based on percent correct). Forecasts by LOG were best for Log Score. Forecasts by GRING did not always score best for Gringorten Score.

(2) For all skill measures, PROB and LOG were nearly the same.

(3) All skill measures showed forecast skill deteriorates with increasing forecast length.

(4) MOS represents verification of the category with the highest probability from MOS 12- and 24-hour probability forecasts. Thus, MOS is analogous to PROB (12- and 24-hour forecasts from AFGWC), and scored better than PROB for all measures. Later results will show that MOS scored better than AFGWC in the Brier Score also.

b. Attachment 5 is the six month summary of Brier Scores for 3- and 6-hour forecasts by 23 field units who participated in the entire test, for 12- and 24-hour forecasts of 22 stations by AFGWC and TDL MOS, and for two "controls" - - conditional climatology and sample climatology. Approximately 12,000 forecasts were made for each verifying hour. This summary shows that the field units beat both conditional and sample climatology for all categories and AFGWC beat them for the 12- and 24-hour combined CIG/VSBY forecasts as well as the 12- and 24-hour CIG forecasts. Also, MOS beat both AFGWC and climatology for all 12- and 24-hour forecasts.

c. Attachment 6 shows the percent improvement of the average of the station forecast Brier Scores over the average of the station conditional climatology Brier Scores for each month of the test. Also shown is the percentage of stations whose Brier Score exceeded conditional climatology Brier Score in each month. Field units started high at 3-hours and maintained this level; performance at 6-hours was more erratic. AFGWC showed little skill with respect to conditional climatology in October but improved rapidly after receiving verification feedback.

d. Attachment 7 shows the month-to-month percent improvement over conditional climatology for 3- and 6-hour ceiling and visibility forecasts. Attachment 8 shows the same data for 12- and 24-hour forecasts.

e. Attachment 9 shows a comparison of subjective (AFGWC) and objective (MOS) probability forecasts. MOS performance was relatively consistent throughout the period; as noted earlier, AFGWC showed improvement in the first four months. However, for Jan-Mar, MOS still maintained an edge over AFGWC forecast skill.

f. Attachments 10-18 are probability forecast, reliability, and distribution plots of forecasts by the 23 field units, AFGWC, and MOS for category D ($\geq$ 3000 fcst) ceilings, and category D ($\geq$3 miles) visibilities and the 23 field unit forecasts for category A, B, and C ceilings. The MOS results cover the entire six month test period and are thus not directly comparable to the field units and AFGWC results shown in the attachments, which only cover the final two months of the test. The forecasts were grouped into 11 probability intervals (0-5%, 5-15%, 15-25%, ...85-95%, 95-100%) and the results plotted at the midpoints of the intervals. The dashed line on the reliability plots shows the locus of perfectly reliable forecasts and the points connected by the solid line show the actual reliability results. The fraction of the forecasts falling within each probability interval is indicated by the length of the horizontal lines in the distribution plots. Note that different horizontal scales are used in the distribution plots. The short vertical lines on the forecast distribution plots indicate a modeled distribution which assumes the forecasts are perfectly reliable and the correlation between forecast probabilities and observations is given by $R = 0.98^t$ where t is in hours (for a 12-hour forecast R = 0.785). The total number of forecasts in each sample, the fractior of the time the category occurred, and the overall forecast bias are also shown. The bias was calculated by:

$$\text{Bias} = \frac{\left(\sum_{i=1}^{11} P_i N_i\right) - O}{O}$$

Where O is the total number of times the category was observed, N_i is the number of forecasts in the ith probability interval, and P_i is the mean probability for the ith probability interval (0.025, 0.1, 0.2, ...0.9, or 0.975).

(1) The 3-and 6-hour category D ceiling results in Atch 10 show generally good reliability and excellent sharpness. As indicated by the reliability plots and the bias there was a tendency to underforecast (assign too low a probability) the occurrence of D ceilings for both forecast periods. The probabilities forecast most frequently (0-5%, 85-95%, and 95-100%) were very reliable. The nearly 20% reliability error at 60% probability in the 3-hour forecasts was based on less than 2% of the total forecasts. The AFGWC category D ceiling results shown in Atch 11 are outstanding. Some breakdown from perfect reliability occurs for the less frequently used low probabilities. MOS tended to underforecast category D ceilings (Atch 12). For probabilities above 50%, the AFGWC forecasts were more reliable than those for MOS while the opposite was true below 50%. (Remember that different periods of record are plotted for the AFGWC and MOS results). Somewhat of a surprise was the size of the negative bias for these objective MOS forecasts.

(2) The category D visibility results (Atch 13-15) are poorer than those for ceilings. The field units generally underforecasted this event, MOS overforecasted it, and AFGWC forecasts definitely exhibited the characteristic of overconfidence (overforecasting at high probabilities and underforecasting at low probabilities, an attempt to forecast with greater sharpness than warranted by forecast skill). The broken line for the 95-100% interval in the distribution plots in Atch 13 and 14 indicates the extension of the line beyond the end of the horizontal scale to the value shown at the tip of the arrow. The value in parenthesis is that for the model distribution. The erratic reliability results for AFGWC at probabilities below 55% were based on less than 5% of the total forecasts. The AFGWC forecasts were too sharp, especially at 24-hours. The MOS category D visibility forecasts were very reliable, had little overall bias, and showed a good match to the model distributions. The large reliability error in the 5-15% interval for 12-hour D visibility forecasts was the result of three occurrences of category D out of four forecasts. This error is mostly likely due to sampling effects and some basic instability in the MOS equations at the less frequently used low probabilities; i.e., insufficient low visibility cases available for equation development. This reliability error occurred for just four forecasts out of 10,838. The MOS probability distributions for D visibility (Atch 15) fit a model distribution generated using $R = 0.97^t$ (not shown) better than that using 0.96^t. This is indicative of the basically lower skill in predicting visibility which is also seen in the Brier Scores and other verification results. This effect of lower skill is not easily detectable in the field unit and AFGWC distributions because of overriding reliability problems.

(3) The field unit results for category A, B, and C ceiling forecasts (Atch 16, 17, and 18) all show a basic tendency to overforecast. This is seen most clearly in the large, positive overall biases and the departures from the model distributions. The reliability results also show this. The erratic reliability plot for category A is the result of event rarity (less than 1% frequency) and sample size problems in the higher probability intervals. In particular, the forecasters at the individual units did not have enough cases to adequately identify their overforecasting problems with category A. Using the model distributions as guidance only 14 forecasts out of 1000 (1 - .986) should be for probabilities greater than 5% for a 3-hour forecast and 30 out of 1000 for a 6-hour forecast. By contrast the units placed 53 and 61 forecasts per 1000 for 3-and 6-hours respectively at probabilities above 5%, approximately 4 and 2 times the model amounts. The category B results (Atch 17) are quite good. The erratic reliability at 6-hours again reflects sample size effects rather true reliability problems. The distribution plots for both A and B indicate a forecaster perference for 60, 80, and greater than 95% probabilities vice 50, 70, and 90% values. The reliability pattern at high probability values for category C ceiling (Atch 18) is rather puzzling. It appears to be the result of forecaster overuse of 5 to 85% probability values; i.e., forecasting with less sharpness than skill would dictate, as well as an overforecasting problem. It may also be that with four ceiling categories insufficient attention is given to the assessment of the probabilities of each of the three, rarer low ceiling categories after the assignment of a probability for category D. The overforecasting and strong positive bias for categories A, B, and C are direct results of underforecasting and negative bias for category D.

g. Attachment 19 summarizes a comparision of the original 23 field units and the 7 field units added on 1 Dec 77. The 23 units which received the seminar scored better than did the 7 units which did not, regardless of the period of comparison.

h. Attachments 20-23 show the 3-, 6-, 12- and 24-hour contingency tables for persistence and the categorical forecasts which maximize AWS, Log, and Gringorten skill scoring methods. Attachments 24-26 summarize these tables. These data indicate that forecasts maximized for AWS and Log Skill Scores were best and nearly equal for all hours and categories. Forecasts maximized for AWS Skill Score had more correct hits but forecasts maximized for Log Skill Score were less biased between optimistic and pessimistic forecasts. Additionally, forecasts maximized for AWS Skill Score were better for Category A; forecasts maximized for Log Skill Score were better for Category B. The Log Skill Score (when compared to AWS Skill Score) does encourage the forecaster to make more Category B forecasts. Forecasts to maximize the Gringorten Score show more forecasts for the lower categories but these forecasts were usually pessimistic.

i. During the test, problems collecting and processing the forecast data resulted in about a 25% data loss. Some of these problems were not all information recorded on the form, forms were misplaced, and data were incorrectly entered on punch cards. However, we believe the overall impact on the test was negligible and should not bias the results.

7. Summary of Test Results:

a. Categorical forecasts made to maximize the Gringorten Score or LOG Score are not significantly better than categorical forecasts made to maximize the AWS Skill Score.

b. Forecasts made to maximize the Gringorten Score were more pessimistic.

c. AWS forecasters can, with training and verification feedback, issue skillful probability forecasts.

d. AFGWC 12- and 24-hour probability forecasts almost equal TDL MOS probability forecasts.

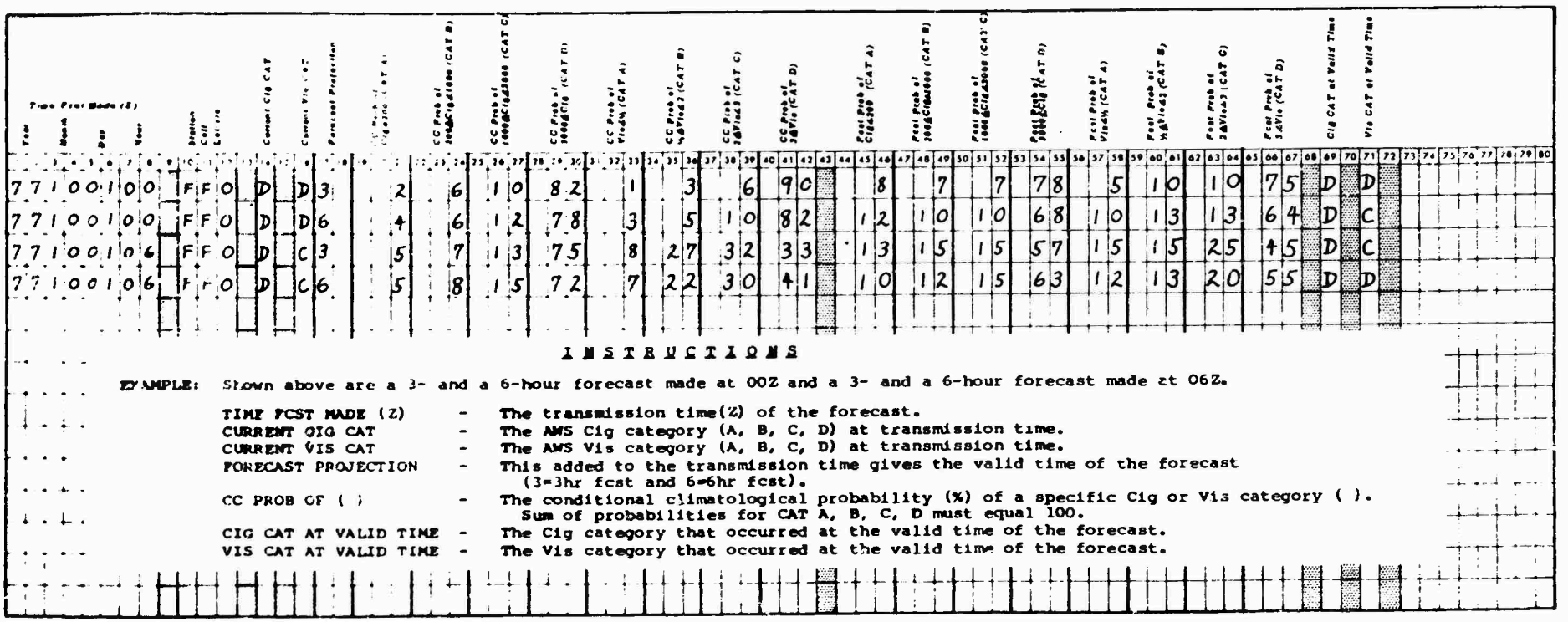
e. The AWS/DN probability forecasting seminar is of value to novice probability forecasters.

Participating Units

3WW:		
Det 9, 12WS	Tyndall AFB, FL	PAM
*Det 4, 26WS	Loring AFB, ME	LIZ
*Det 6, 26WS	Pease AFB, NH	PSM
*Det 8, 26WS	Griffiss AFB, NY	RME
*Det 12, 26WS	Plattsburg AFB, NY	PBG
Det 14, 26WS	Blytheville AFB, AR	BYH
Det 18, 26WS	Rickenbacker AFB, OH	LCK
*Det 19, 26WS	Whiteman AFB, MO	SZL
Det 20, 26WS	Barksdale AFB, LA	BAD
*Det 22, 26WS	Carswell AFB, TX	FWH
*Det 23, 26WS	McConnell AFB, KS	IAB
Det 24, 26WS	K. I. Sawyer AFB, MI	SAW
Det 26, 26WS	Grissom AFB, IN	GUS
Det 28, 26WS	Wurtsmith AFB, MI	OSC
5WW:		
Det 5, 3WS	England AFB, LA	AEX
Det 12, 3WS	Selfridge ANGB, MI	MTC
Det 31, 3WS	Dobbins AFB, GA	MGE
Det 75, 3WS	Hurlburt AFB, FL	HRT
Det 1, 5WS	Ft Campbell, KY	HOP
Det 5, 5WS	Ft Knox, KY	FTK
Det 10, 5WS	Ft Benning, GA	LSF
**Det 31, 5WS	Ft Polk, LA	POE
Det 2, 24WS	Columbus AFB, MS	CBM
Det 9, 24WS	Maxwell AFB, AL	MXF
Det 22, 24WS	Keesler AFB, MS	BIX
7WW:		
Det 9, 7WW	Scott AFB, IL	BLV
Det 20, 7WW	Little Rock AFB, AR	LRF
Det 13, 15WS	Robins AFB, GA	WRB
Det 15, 15WS	Wright-Patterson AFB, OH	FFO
AFGWC:		
Det 10, 2WS	Eglin AFB, FL	VPS

*Added on 1 Dec 77.

**No 12- or 24-hour forecasts were made for Ft Polk.



Time Fcst Made (Z)				Station Call Letters	Current Cig CAT	Current Vis CAT	Forecast Projection	CC Prob of 1000Cig (CAT A)	CC Prob of 1000Cig2000 (CAT B)	CC Prob of 1000Cig2000 (CAT C)	CC Prob of 1000Cig (CAT D)	CC Prob of 3AVis (CAT A)	CC Prob of 3AVis3 (CAT B)	CC Prob of 3AVis3 (CAT C)	CC Prob of 3AVis (CAT D)	Fcst Prob of Cig2000 (CAT A)	Fcst Prob of 2000Cig1000 (CAT B)	Fcst Prob of 1000Cig2000 (CAT C)	Fcst Prob of 1000Cig (CAT D)	Fcst Prob of 3AVis (CAT A)	Fcst Prob of 3AVis3 (CAT B)	Fcst Prob of 3AVis3 (CAT C)	Fcst Prob of 3AVis (CAT D)	Cig CAT at Valid Time	Vis CAT at Valid Time
7	7	1	0	0	1	0	0	F	F	O	D	D	3	2	6	1	0	8	2	1	3	6	9	0	8	7	7	7	8	5	1	0	1	0	7	5	D	D
7	7	1	0	0	1	0	0	F	F	O	D	D	6	4	6	1	2	7	8	3	5	1	0	8	2	1	2	1	0	1	0	6	8	1	0	1	3	1	3	6	4	D	C
7	7	1	0	0	1	0	6	F	F	O	D	C	3	5	7	1	3	7	5	8	2	7	3	2	3	3	1	3	1	5	1	5	5	7	1	5	1	5	2	5	4	5	D	C
7	7	1	0	0	1	0	6	F	F	O	D	C	6	5	8	1	5	7	2	7	2	2	3	0	4	1	1	0	1	2	1	5	6	3	1	2	1	3	2	0	5	5	D	D

INSTRUCTIONS

EXAMPLE: Shown above are a 3- and a 6-hour forecast made at 00Z and a 3- and a 6-hour forecast made at 06Z.

TIME FCST MADE (Z)	-	The transmission time(Z) of the forecast.
CURRENT CIG CAT	-	The AWS Cig category (A, B, C, D) at transmission time.
CURRENT VIS CAT	-	The AWS Vis category (A, B, C, D) at transmission time.
FORECAST PROJECTION	-	This added to the transmission time gives the valid time of the forecast (3=3hr fcst and 6=6hr fcst).
CC PROB OF ( )	-	The conditional climatological probability (%) of a specific Cig or Vis category ( ). Sum of probabilities for CAT A, B, C, D must equal 100.
CIG CAT AT VALID TIME	-	The Cig category that occurred at the valid time of the forecast.
VIS CAT AT VALID TIME	-	The Vis category that occurred at the valid time of the forecast.

ATCM 2

SKILL SCORES

The AWS Skill Score = (Unit percent correct - persistence percent correct)/(100% - persistence percent correct). This score weights all correct forecasts equally, a hit from predicting the difficult to forecast bad weather categories (A and B) is worth the same as a correct prediction of easier to forecast good weather. This score can range from $-\infty$ to a maximum possible of +1. A negative score indicates the absence of skill. The greater the number above zero, the greater the skill.

The Gringorten Skill Score (GSS)[1] gives greater weight to correct forecasts of the harder to predict bad weather categories. The weight for each category is inversely proportional to the climatological frequency of occurrence of the category. A correct forecast of a weather category which occurs 2% of the time would be given a weight of 50, (1/0.02); whereas, a correct forecast of a category which occurs 80% of the time would be given a weight of 1.25, (1/0.8). The GSS is calculated as follows:

$$GSS = \frac{\left(\sum_{i=1}^{G} H_i W_i\right) - N}{\left(\sum_{i=1}^{G} O_i W_i\right) - N}$$

Where N is the number of forecasts, H_i is the number of forecast hits in category i, O_i is the number of observations in category i, W_i is the weighting factor for category i (1/climatological frequency of category i), and G is the greater of: (a) the number of categories in which at least one observation occurred, or (b) the number of categories for which at least one forecast was issued. This score can also range from $-\infty$ to + 1 where + 1 is perfect forecasting. For the test, the weighting factors were calculated using the observed frequencies of the occurrence, N/O_i, rather than the climatological frequencies. With this change, the Gringorten Skill Score becomes

$$GSS = \frac{\left(\sum_{i=1}^{G} \frac{H_i}{O_i}\right) - 1}{G - 1}$$

The Log Skill Score[2] is a penalty score; i.e., correct forecasts are given a weight of zero and missed forecasts are given "penalty points." The Log Score takes the "closeness" of incorrect forecasts into account by giving relatively few penalty points to one category busts compared to the maximum penalties assessed for three category busts. The penalty matrix for ceiling forecasts is:

		FORECAST			
		A	B	C	D
OBSERVED	A	0	23	58	81
	B	35	0	15	39
	C	63	16	0	10
	D	89	38	16	0

A similar penalty matrix is used for visibility forecasts. The Log Score is computed by multiplying the elements of the verification matrix by the corresponding elements of the penalty matrix and summing of the products, i.e.,

$$LS = \frac{1}{N} \sum_{i=1}^{4} \sum_{j=1}^{4} N_{ij} M_{ij}$$

Where N_{ij} are the elements of the verification matrix and M_{ij} are the elements of the penalty matrix. The lower the score, the greater the forecast skill. A perfect score is zero and the maximum score (for ceiling forecasts) is 89 (forecast category A every time and observe only category D).

1. Gringorten, I. I., 1967 Journal of Applied Meteorology, 6, pp 742-747.

2. MacDonald, A.E., 1977, Western Region Technical Attachment No. 77-18.

Combining ceiling-visibility probabilities: The probabilities for combined ceiling-visibility categories were calculated by AFGWC using the relation proposed by Capt Al Boehm:

$P_{cv} = (1 - \rho) P_c P_v + \rho \, \text{MIN} (P_c, P_v)$ where P_{cv} is the combined probability for ceiling-visibility category, P_v is the assigned probability for visibility in the category, P_c is the assigned probability for ceiling in the category, ρ is the correlation between ceiling and visibility. 0.3 was used for the value of the correlation.

ATCH 3A

SKILL SCORE COMPARISON

	PERCENT CORRECT*					AWS SCORE				GRINGORTEN				LOG SCORE			
	PROB	LOG	GRING	MOS	PERS	PROB	LOG	GRING	MOS	PROB	LOG	GRING	MOS	PROB	LOG	GRING	MOS
3 HR	88.5#	88.4	77.9		80.8	.406#	.395	-.198		.62#	.59	.57		1.90	1.87#	4.77	
6 HR	84.0#	83.7	71.8		75.0	.360#	.348	-.128		.47#	.46	.47#		2.81	2.72#	5.86	
12 HR	75.2	74.3	52.4	77.1#	69.5	.186	.157	-.561	.249#	.24	.23	.25#	.24	4.68	4.65	14.28	4.03#
24 HR	72.7	71.3	49.4	74.3#	65.3	.213	.187	-.458	.277#	.15	.16	.17	.20#	5.39	5.34	14.89	4.94#

*Percent correct categorical forecasts maximized for the AWS skill score (PROB), Log skill score (LOG), Gringorten skill score (GRING) and model output statistics MOS; percent correct for persistence (PERS) is also shown.

# Indicates best forecast score.

ATCH 4

BRIER SCORES[1]

6 MONTH SUMMARY (Oct 77 - Mar 78)

	3 HOUR			6 HOUR			12 HOUR				24 HOUR			
CIG/VSBY	STN	COND CLIMO[2]	SAMPLE CLIMO[3]	STN	COND CLIMO[2]	SAMPLE CLIMO[3]	STN	MOS	COND CLIMO[2]	SAMPLE CLIMO[3]	STN	MOS	COND CLIMO[2]	SAMPLE CLIMO[3]
COMB	.18	.43	.42	.24	.42	.42	.35	.33	.42	.42	.39	.36	.43	.43
CIG	.16	.25	.40	.22	.29	.40	.31	.30	.33	.39	.35	.32	.37	.40
VSBY	.13	.15	.19	.17	.18	.20	.22	.19	.20	.21	.23	.21	.22	.22

Footnotes:

1. The Brier Score = $\frac{1}{N} \sum_{j=1}^{K} \sum_{i=1}^{N} (R_{ij} - D_{ij})^2$

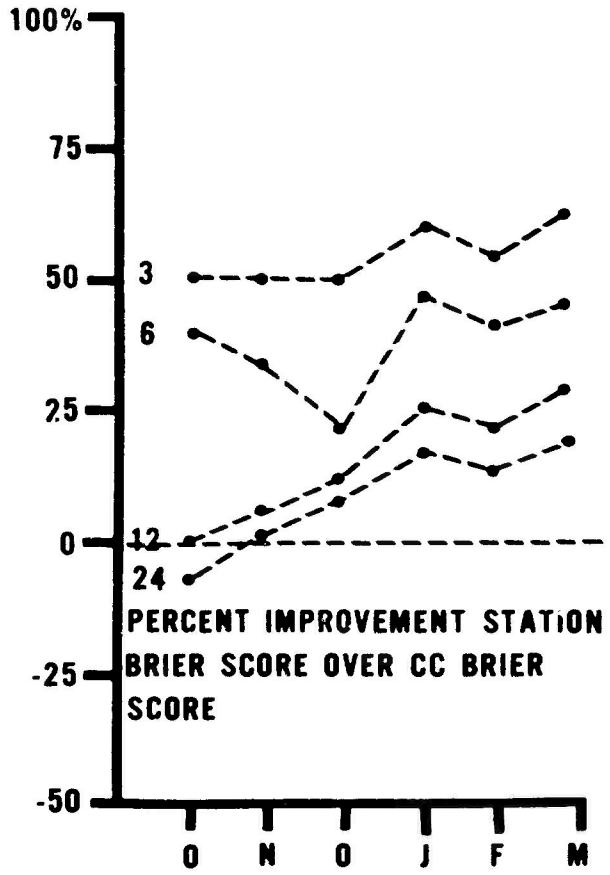
where: N is the number of forecasts verified; K is the number of categories for each forecast; R_{ij} is the forecast probability value assigned to category j of the ith forecast; D_{ij} is the observed probability (0 for miss, 1 for hit) for category j of the ith forecast. The score varies from 0 (perfect forecast) to 2 (worst possible forecast); the lower number indicates greater skill.

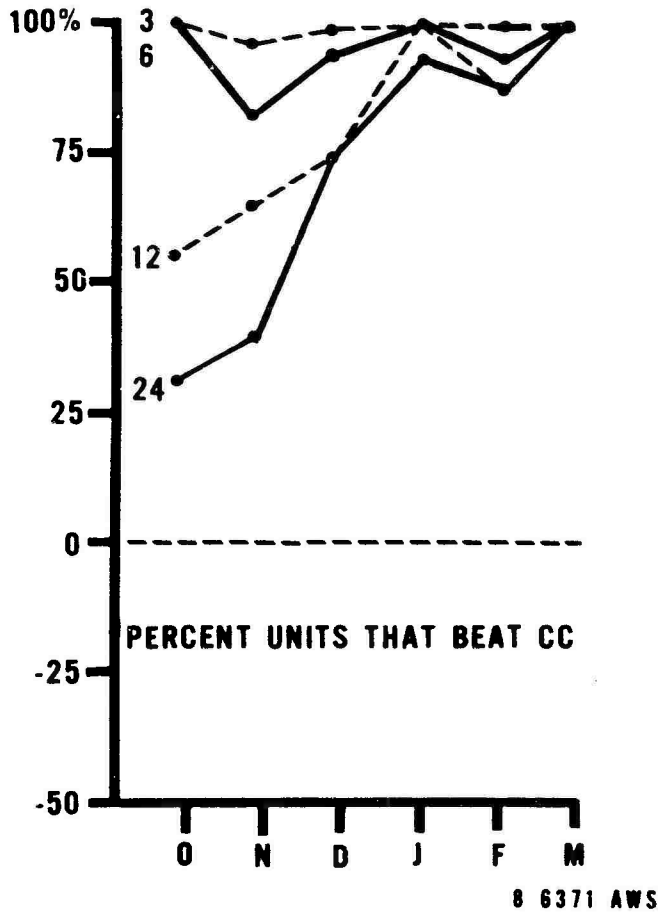
2. The climatological probability that a weather event will occur based on historical observations divided according to time of day, wind direction, and month of year.

3. The climatological probability that a weather event will occur based on historical observations that occurred only during the sample period.

ATCH 5

PERFORMANCE TREND BY FORECAST LENGTH (HOURS)
COMBINED CIG/VSBY

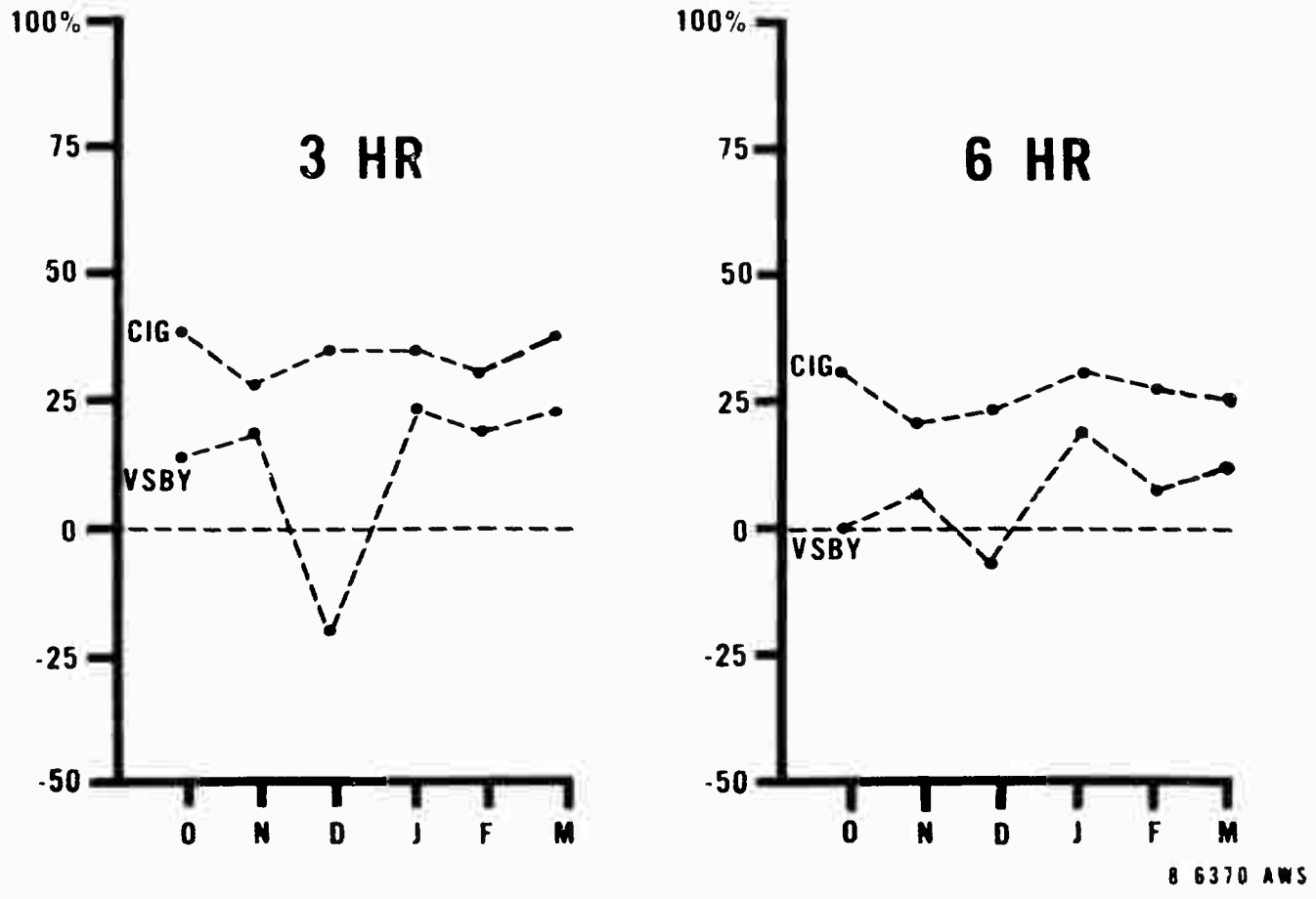




8 6371 AWS

ATCH 6

PERCENT IMPROVEMENT STATION BRIER SCORE OVER CC BRIER SCORE

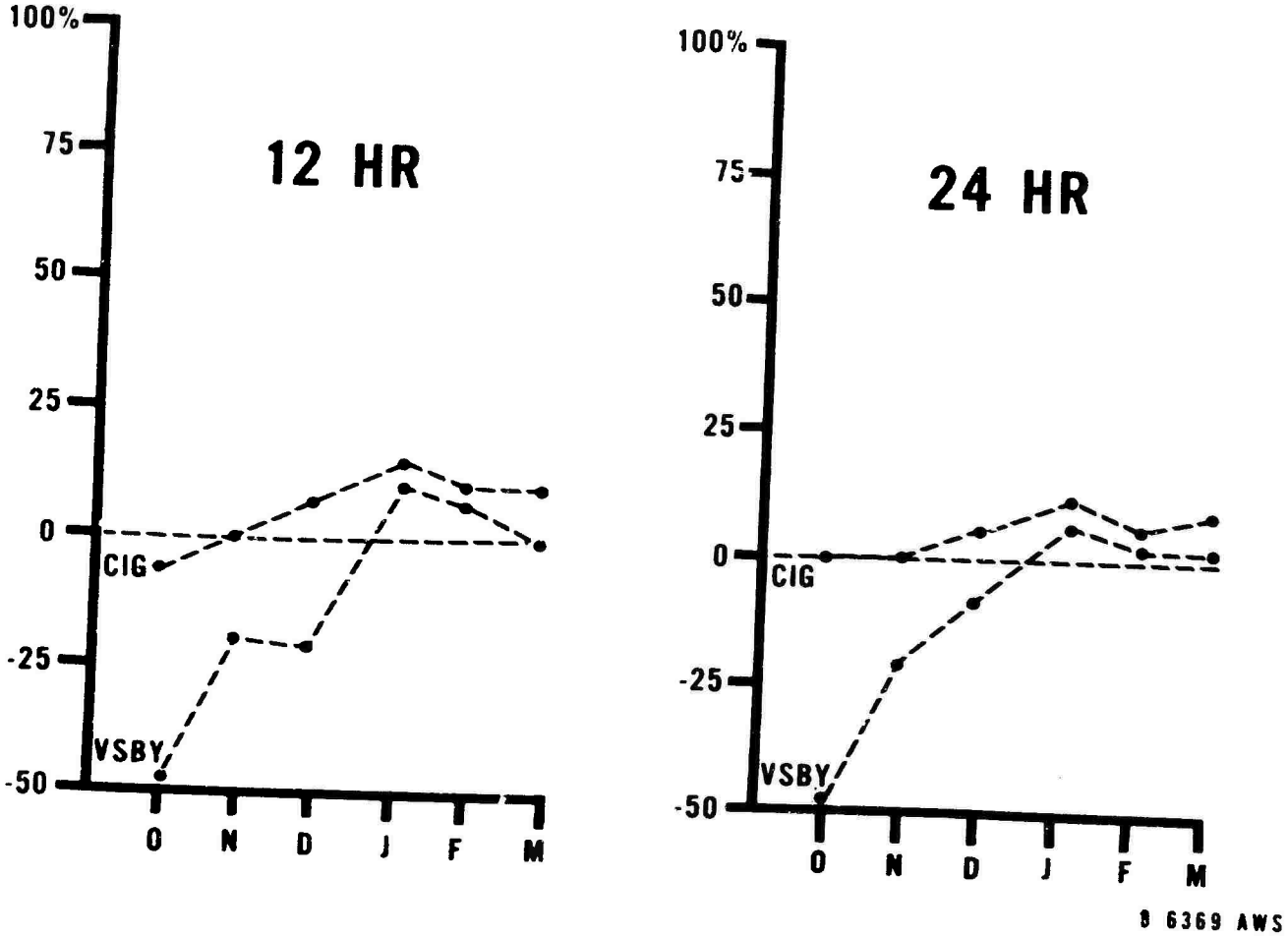


8 6370 AWS

ATCH 7

ATCH 8

PERCENT IMPROVEMENT STATION BRIER SCORE
OVER CC BRIER SCORE



Comparison

of

Subjective (AFGWC) and Objective (MOS) Probability Forecasts

Percent Improvement Forecast Brier Score over Conditional Climatology Brier Score

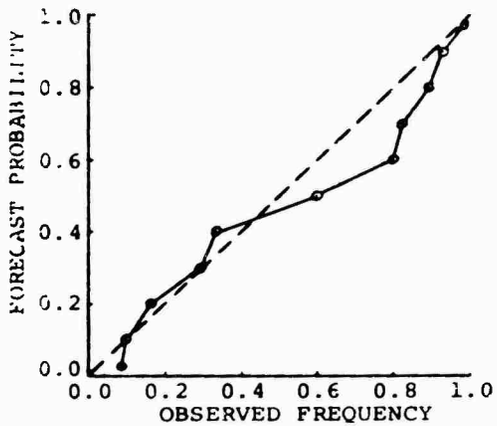
	12 HOUR						24 HOUR					
	CIG/VSBY		CIG		VSBY		CIG/VSBY		CIG		VSBY	
	GWC	MOS	GWC	MOS	GWC	MOS	GWC	MOS	GWC	MOS	GWC	MOS
OCT	0	24	-6	17	-45	9	-8	17	0	20	-50	10
NOV	6	20	0	9	-15	8	0	15	0	11	-23	0
DEC	12	21	6	14	-17	0	9	17	5	15	-8	4
JAN	26	20	12	7	7	-3	20	16	12	10	9	3
FEB	21	23	10	13	6	11	13	13	6	6	5	5
MAR	27	32	9	15	0	5	18	27	11	22	4	18
6 MONTH SUMMARY	19	23	6	9	-10	5	11	18	5	14	-4	4
AVERAGE LAST 3 MONTHS	26	26	9	11	4	4	17	19	10	13	8	8

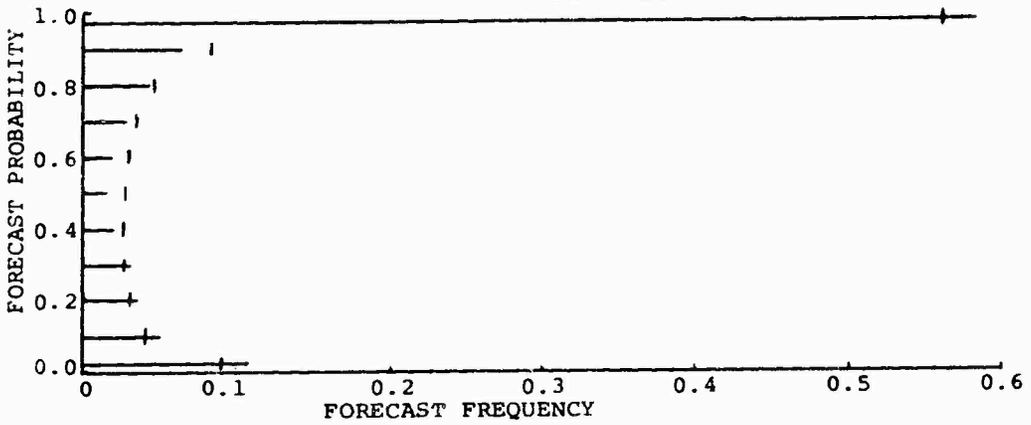
DETACHMENT CATEGORY D CEILING FEB-MAR

3 HR FORECASTS

SAMPLE SIZE: 3814

EVENT FREQUENCY: 0.759
FORECAST BIAS: -0.037

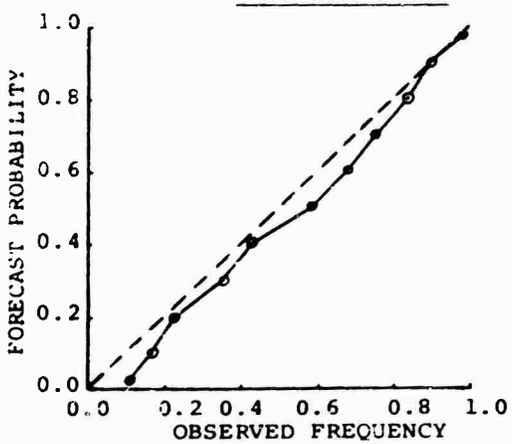


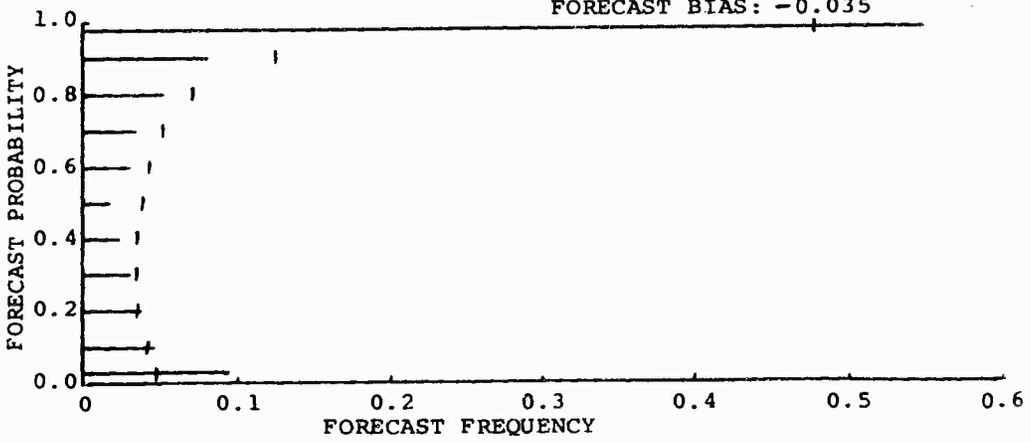


6 HR FORECASTS

SAMPLE SIZE: 3812

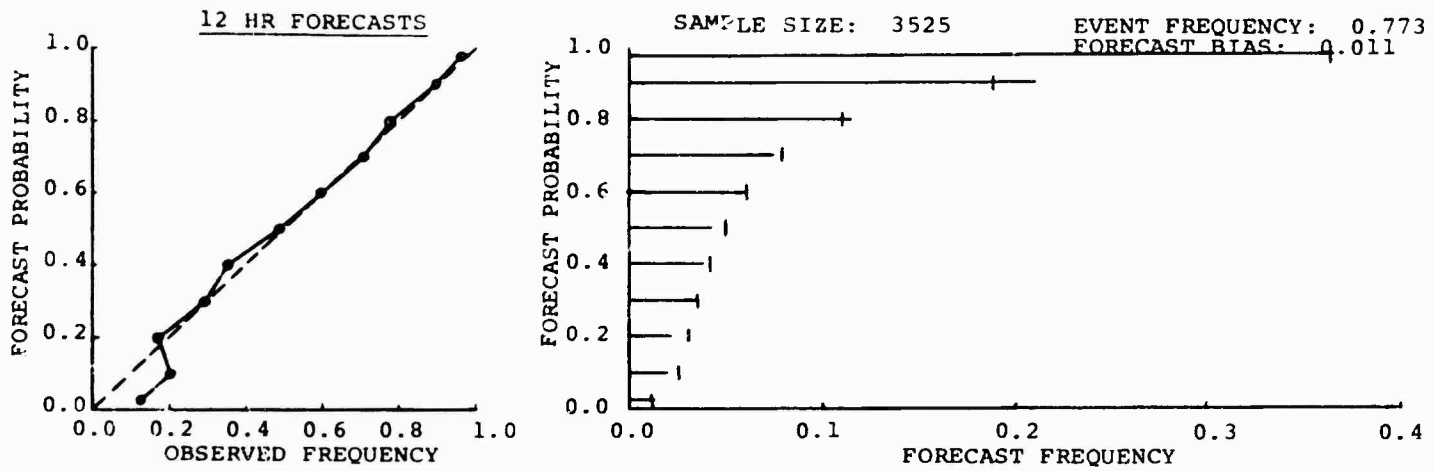
EVENT FREQUENCY: 0.763
FORECAST BIAS: -0.035

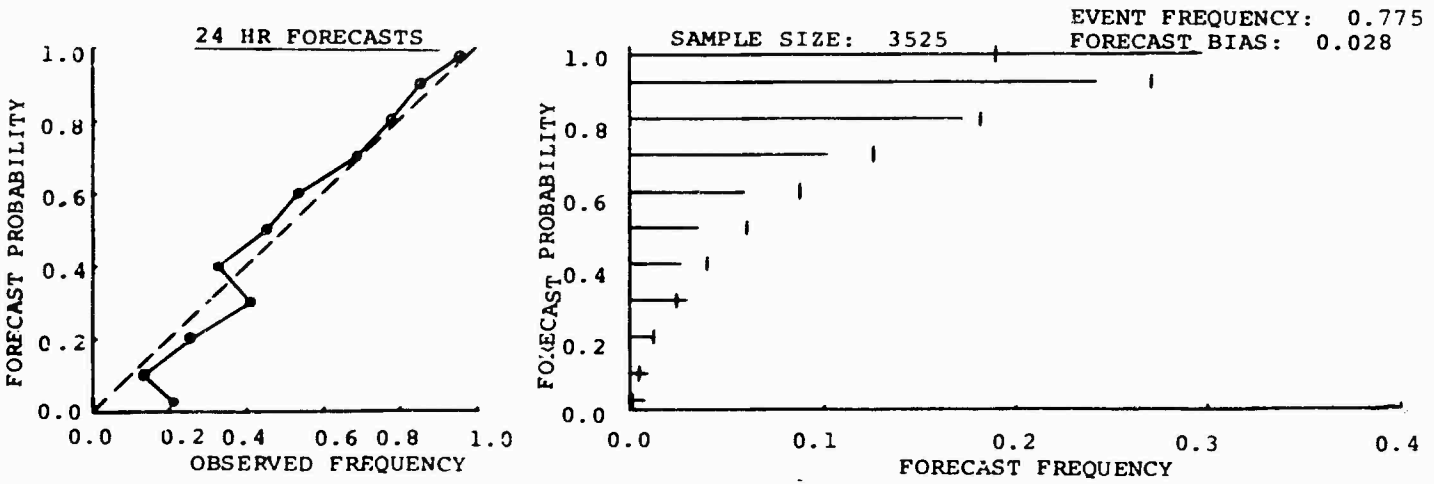




MARCH 10

AFGWC CATEGORY D CEILING FEB-MAR



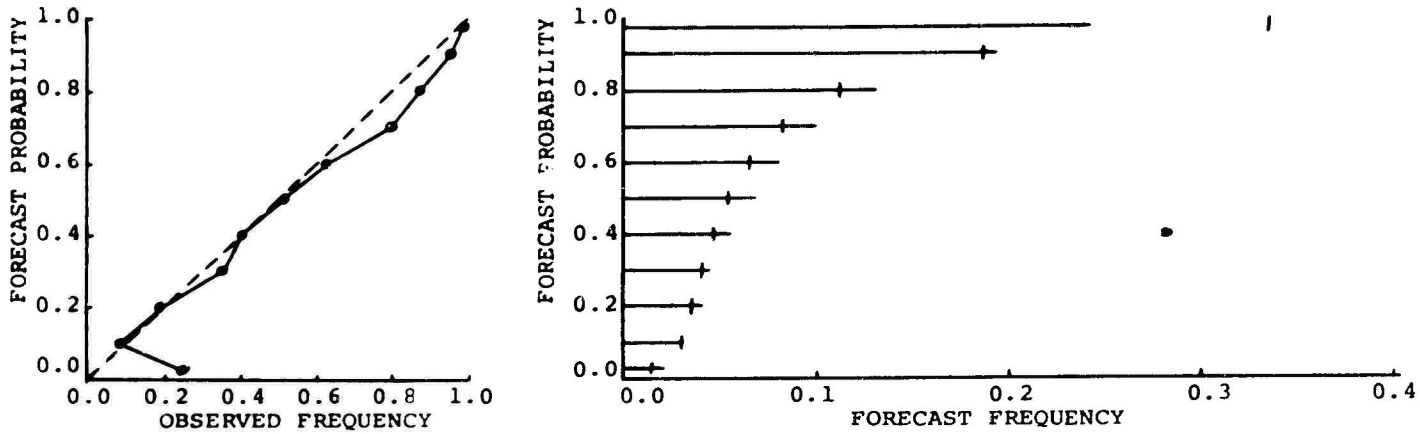


MOS CATEGORY D CEILING OCT-MAR

12 HR FORECASTS

SAMPLE SIZE: 10841

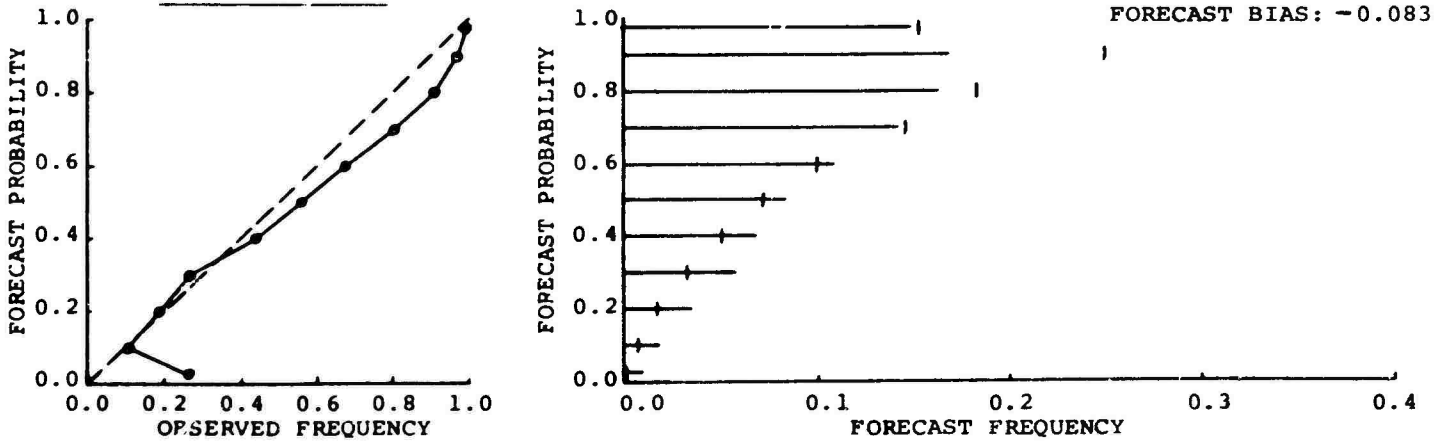
EVENT FREQUENCY: 0.755
FORECAST BIAS: -0.055



24 HR FORECASTS

SAMPLE SIZE: 10824

EVENT FREQUENCY: 0.748
FORECAST BIAS: -0.083



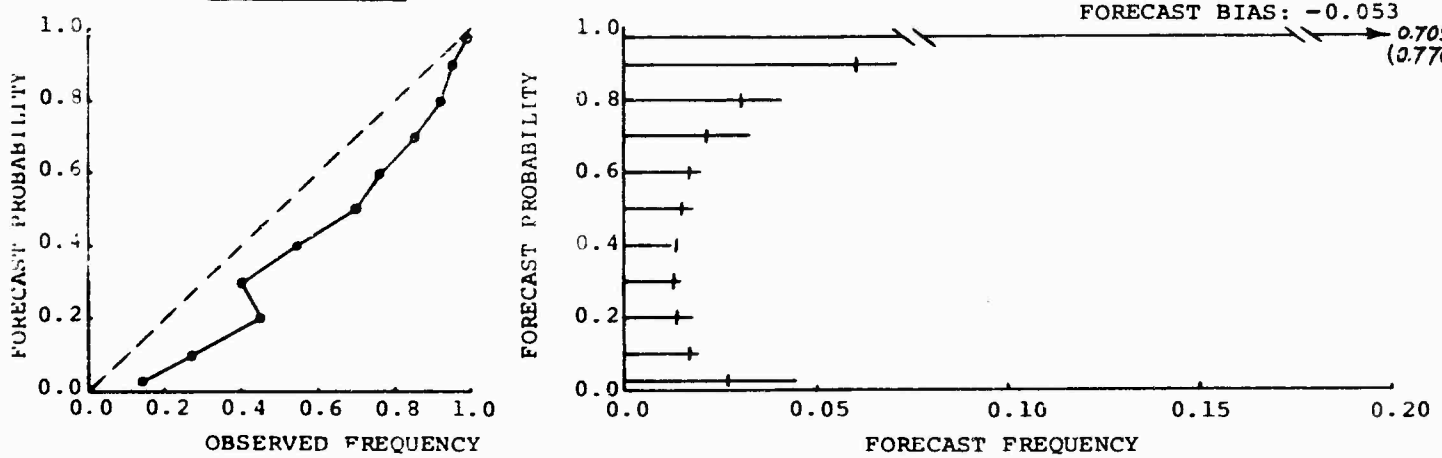
DETACHMENT CATEGORY D VISIBILITY FEB-MAR

3 HR FORECASTS

SAMPLE SIZE: 3773

EVENT FREQUENCY: 0.895

FORECAST BIAS: -0.053

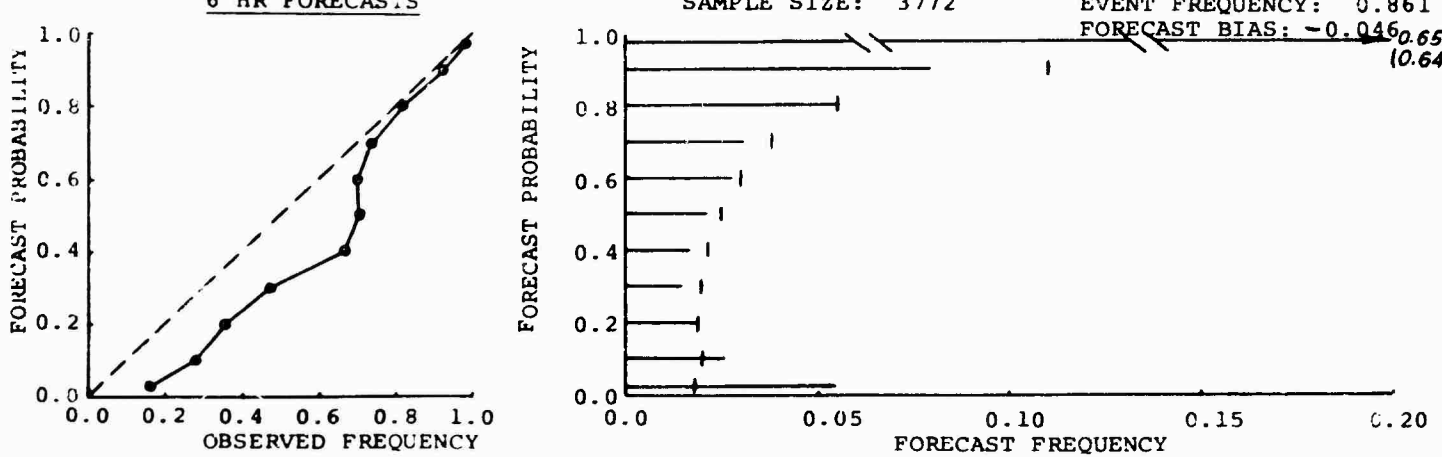


6 HR FORECASTS

SAMPLE SIZE: 3772

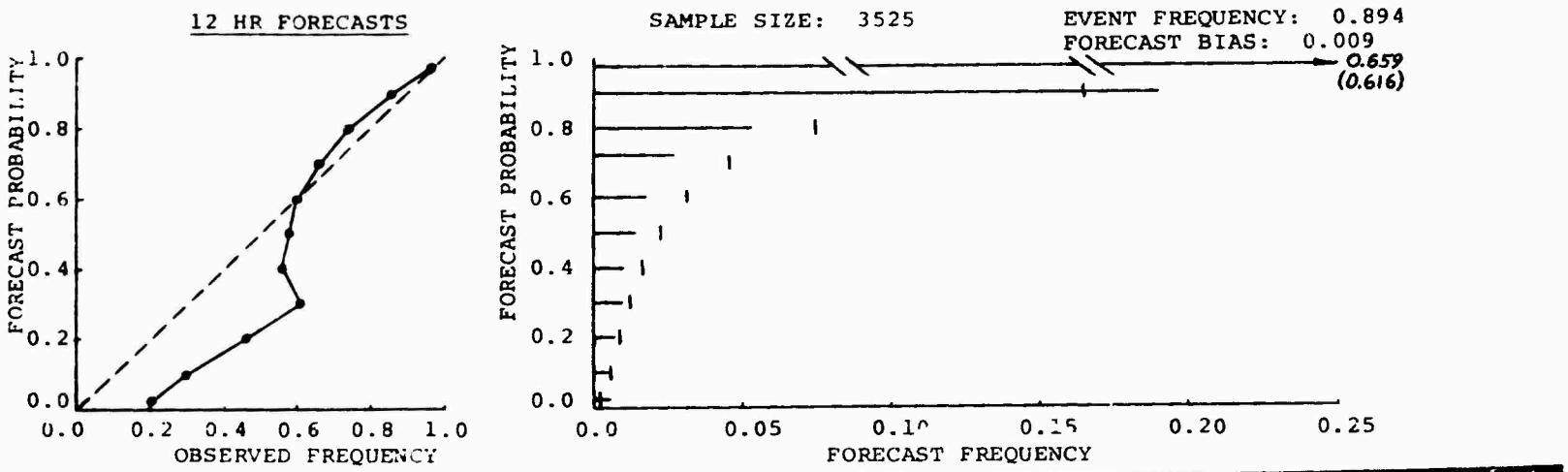
EVENT FREQUENCY: 0.861

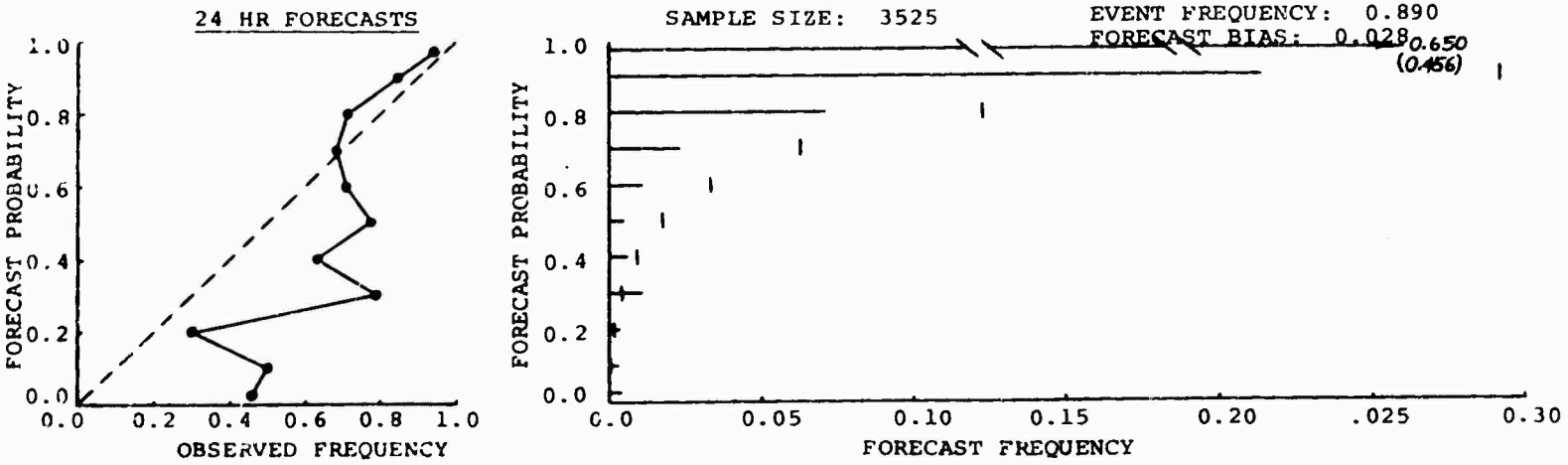
FORECAST BIAS: -0.046



ATCH 13

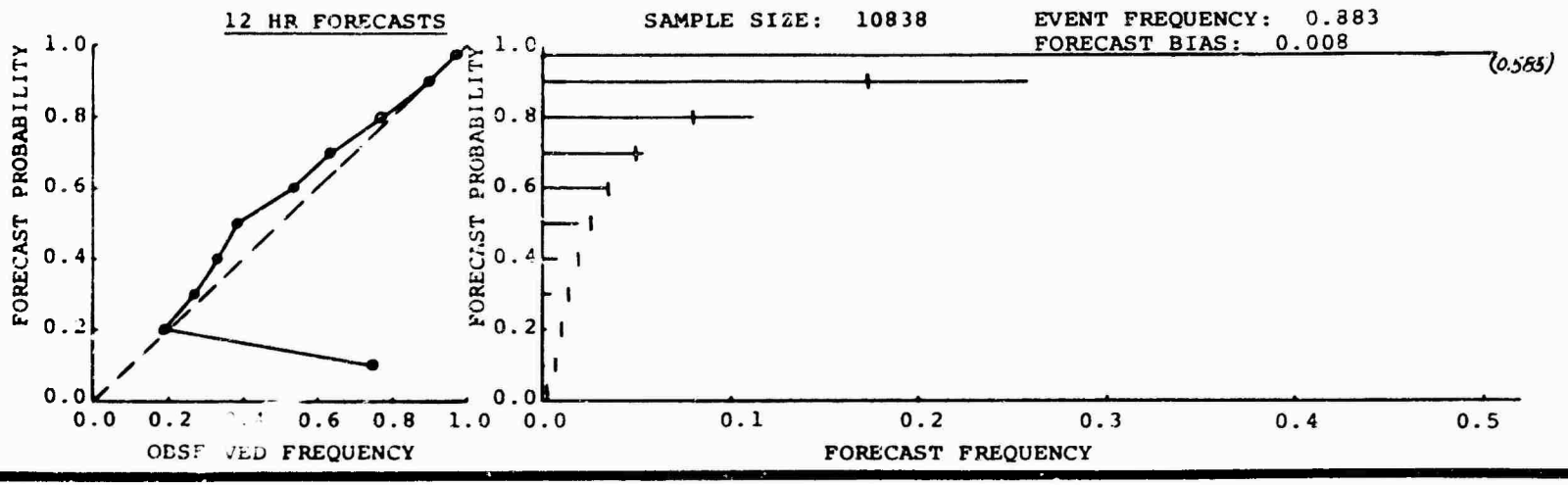
AFGWC CATEGORY D VISIBILITY FEB-MAR

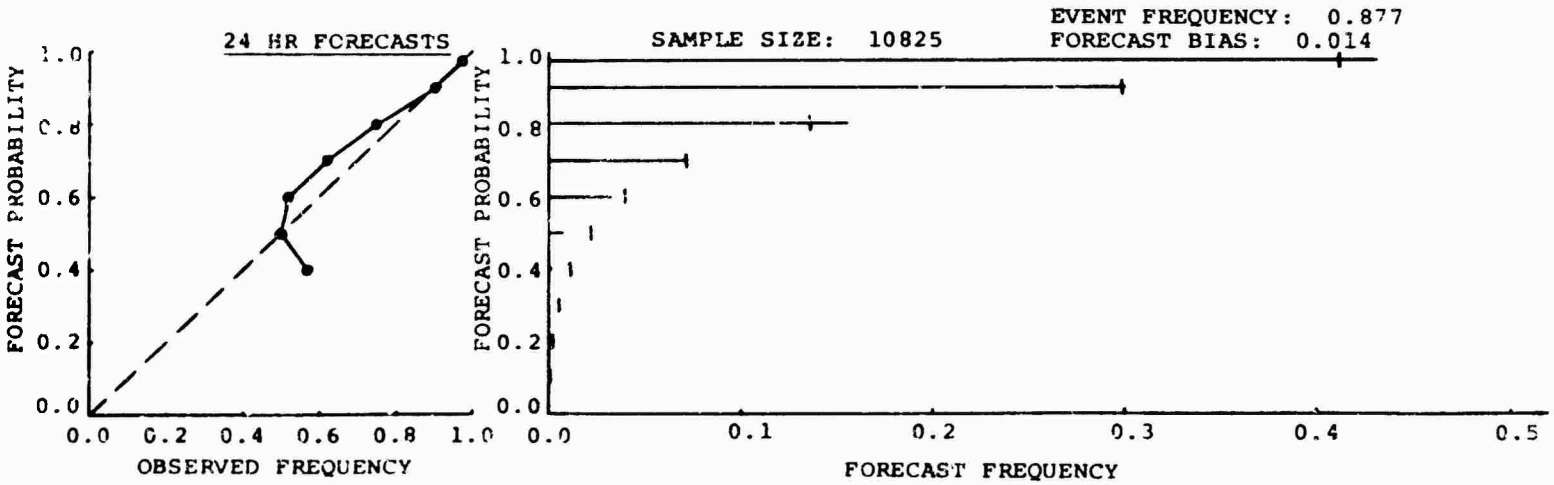




ATCH 14

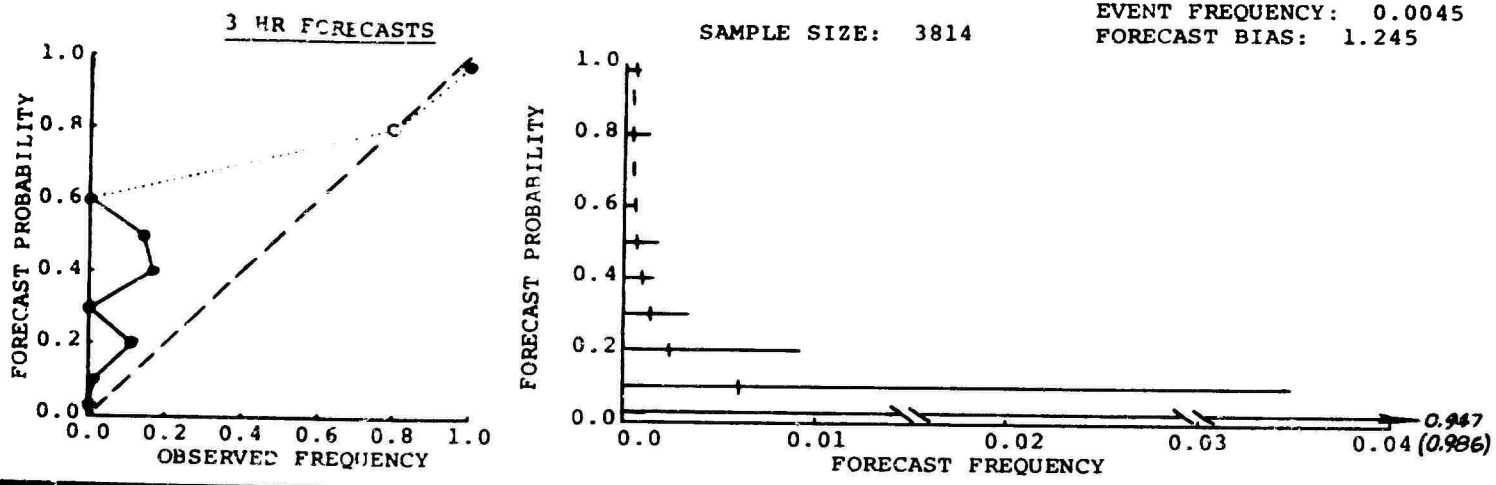
MOS CATEGORY D VISIBILITY OCT-MAR

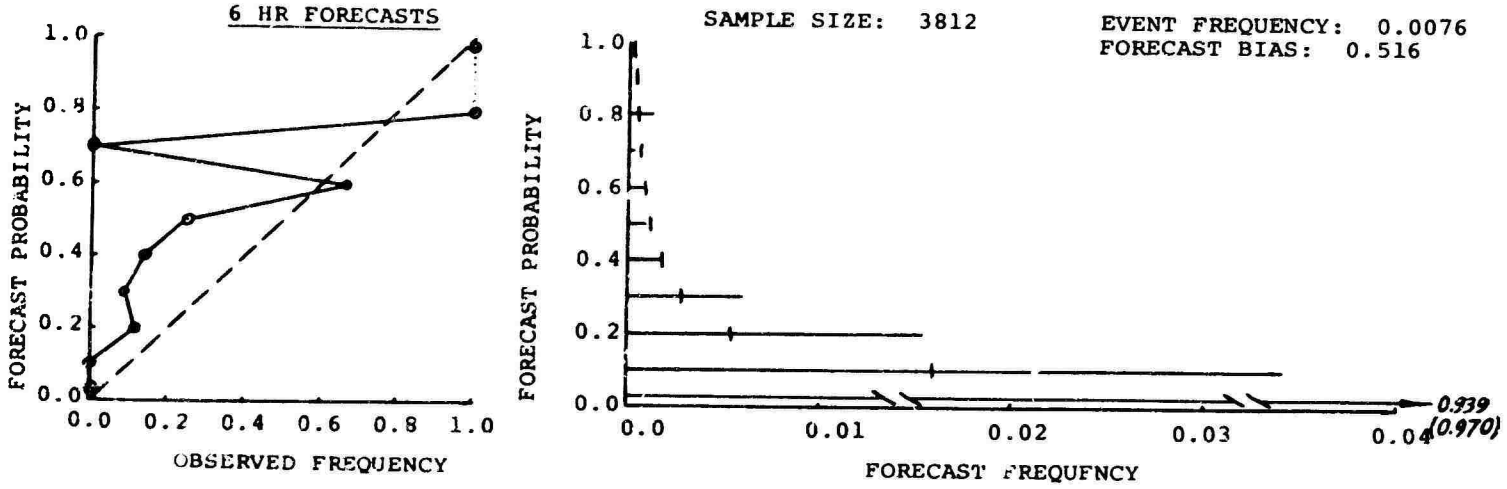




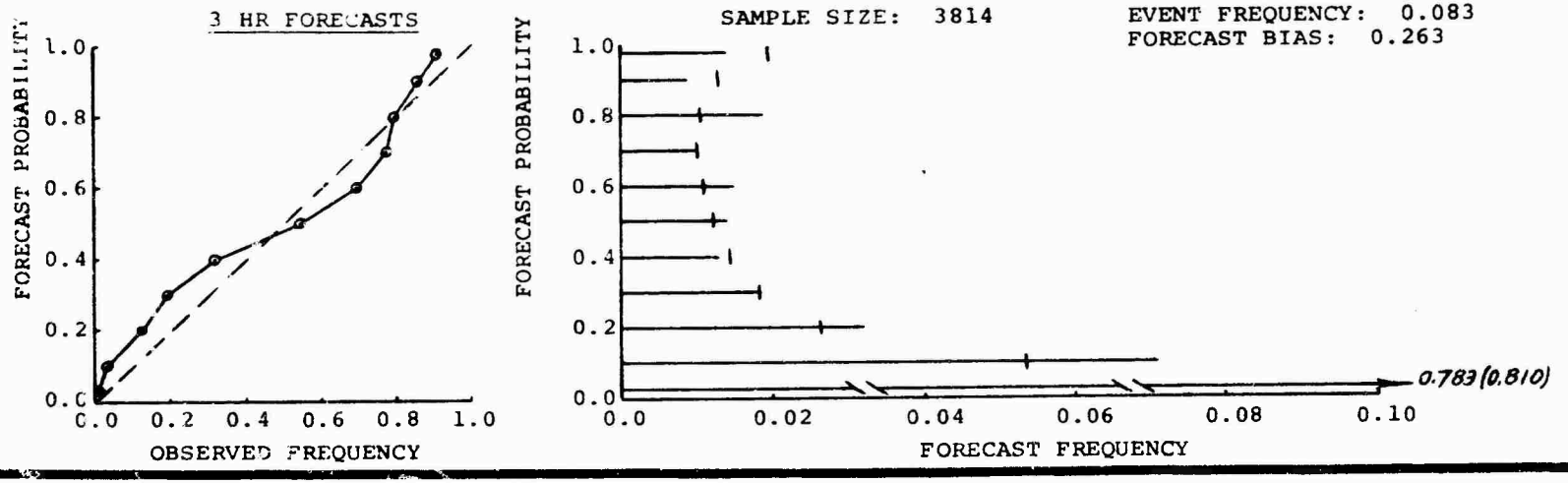
ATCH 15

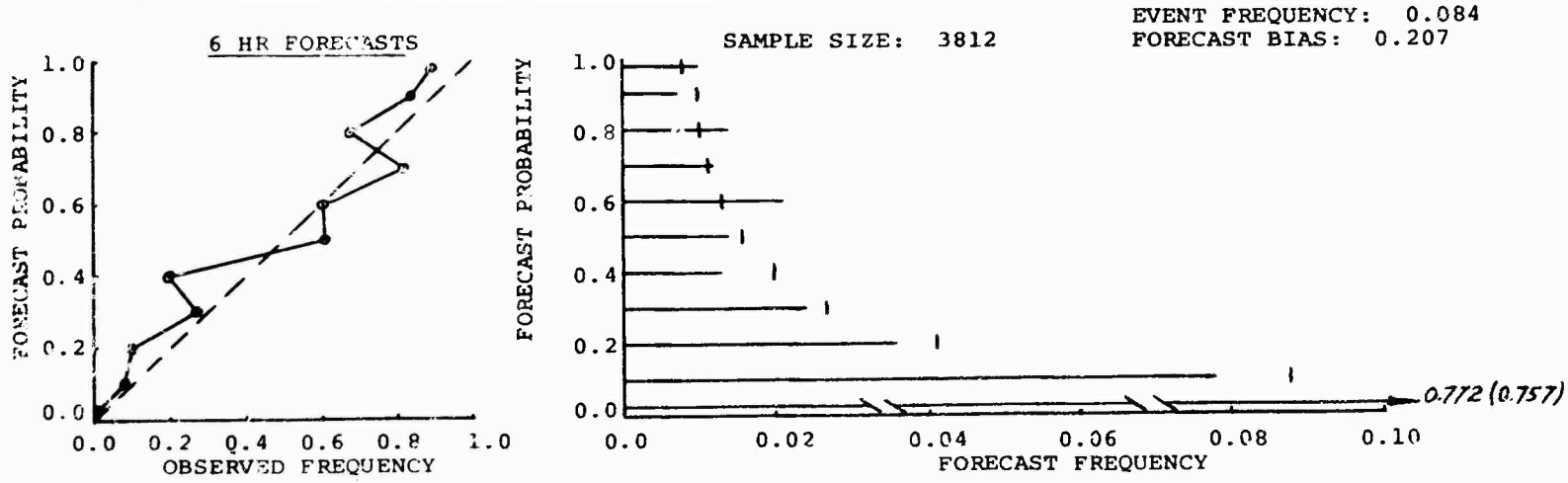
DETACHMENT CATEGORY A CEILING FEB-MAR





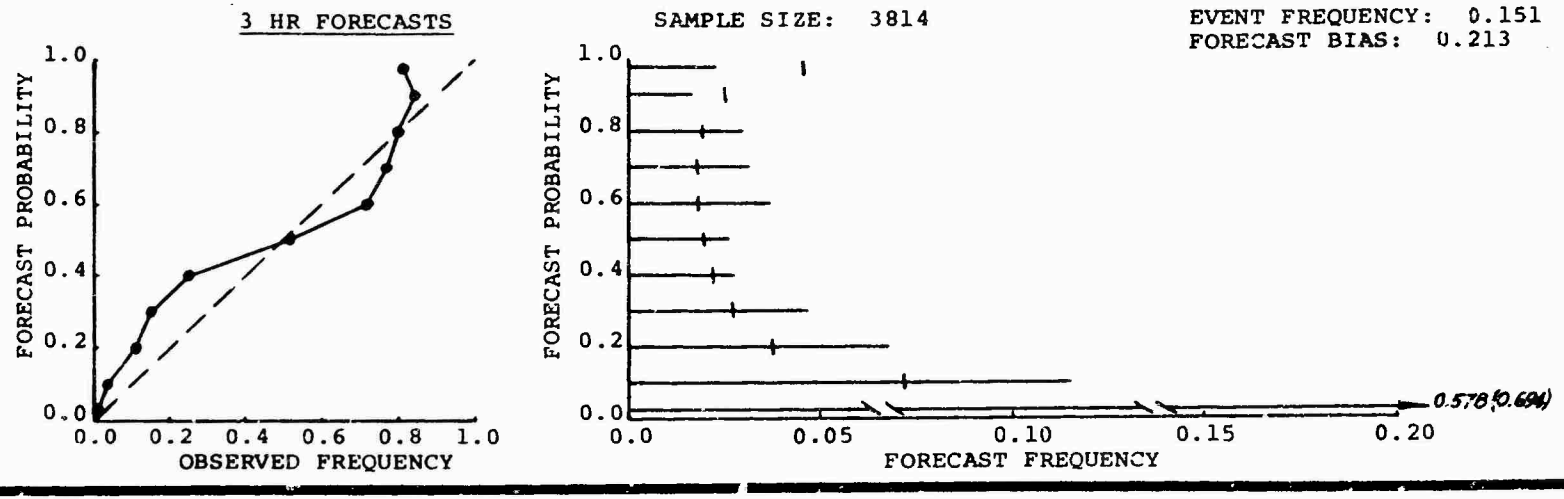
DETACHMENT CATEGORY B CEILING FEB-MAR

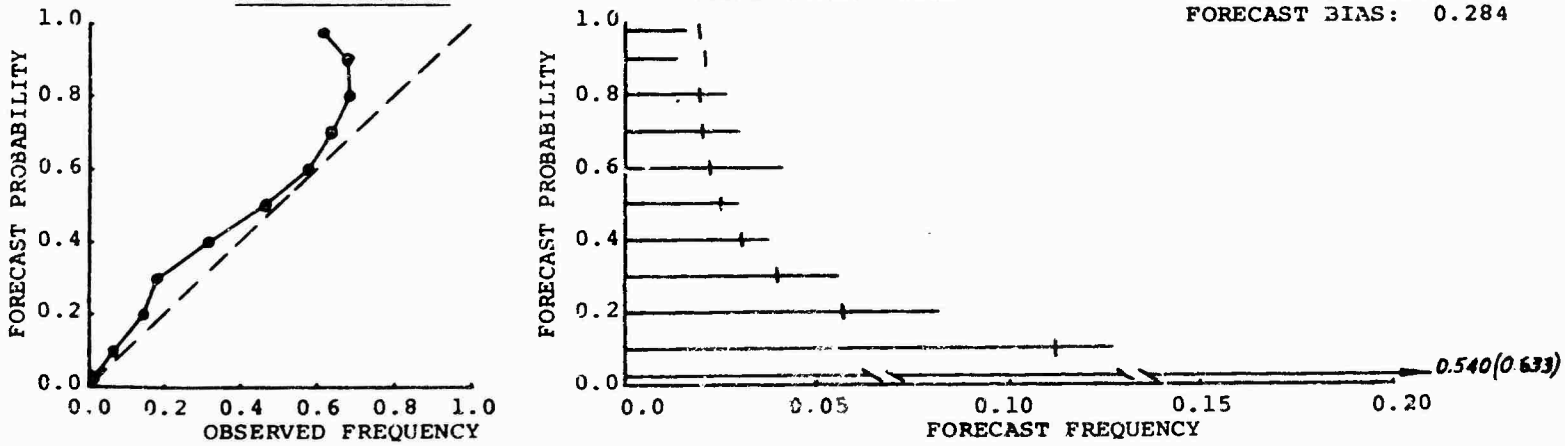




ATCH 17

DETACHMENT CATEGORY C CEILING FEB-MAR





ATCH 18

Comparison

of

Original 23 Units and 7 Units Added in December

Percent Improvement Forecaster Brier Score over Conditional Climatology Brier Score

	3 HOUR		6 HOUR	
	Cig	Vsby	Cig	Vsby
Original 23 (Oct-Mar)	36	13	24	6
Original 23 (Dec-Mar)	35	12	24	10
Added 7 (Dec-Mar)	27	5	22	0
Original 23 (Feb-Mar)	34	17	25	9
Added 7 (Feb-Mar)	30	9	23	5

CONTINGENCY TABLES FOR CIG/VIS AT 3 HR

Persistence

Forecast

OBSERVED	A	B	C	D	T
A	102	67	9	22	200
B	105	797	227	134	1263
C	17	328	870	506	1721
D	30	161	669	7623	8483
T	254	1353	1775	8285	11,667

Maximizes AWS Skill Score

Forecast

OBSERVED	A	B	C	D	T
A	114	61	11	14	200
B	68	946	148	94	1256
C	8	184	1255	264	1711
D	15	76	394	7981	8466
T	205	1267	1808	8353	11,633

Maximizes Log Skill Score

Forecast

OBSERVED	A	B	C	D	T
A	99	79	9	13	200
B	48	988	133	86	1255
C	5	244	1130	332	1711
D	12	92	338	8024	8466
T	164	1403	1610	8455	11,632

Maximizes Gringorten Skill Score

Forecast

OBSERVED	A	B	C	D	T
A	157	25	12	6	200
B	450	646	121	39	1256
C	129	281	1186	114	1710
D	113	164	1264	6922	8463
T	849	1116	2583	7081	11,629

ATCH 20

CONTINGENCY TABLES FOR CIG/VIS AT 6 HR

Persistence

OBSERVED \ Forecast	A	B	C	D	T
A	78	91	19	51	239
B	91	628	281	263	1263
C	27	366	671	642	1706
D	59	269	803	7319	8450
T	255	1354	1774	8275	11,658

Maximizes AWS Skill Score

OBSERVED \ Forecast	A	B	C	D	T
A	82	93	31	33	239
B	68	780	238	166	1252
C	14	232	1025	423	1694
D	11	120	492	7797	8420
T	175	1225	1786	8419	11,605

Maximizes Log Skill Score

OBSERVED \ Forecast	A	B	C	D	T
A	74	111	28	26	239
B	46	847	200	161	1254
C	6	306	882	500	1694
D	7	13	439	7838	8419
T	133	1399	1549	8525	11,606

Maximizes Gringorten Skill Score

OBSERVED \ Forecast	A	B	C	D	T
A	166	35	26	11	238
B	453	516	205	80	1254
C	177	306	1021	190	1694
D	141	243	1506	6530	8420
T	937	1100	2758	6811	11,606

CONTINGENCY TABLES FOR CIG/VIS AT 12 HR

Persistence

OBSERVED \ Forecast	A	B	C	D	T
A	51	91	26	95	263
B	62	422	266	511	1261
C	41	304	477	764	1586
D	117	447	883	7160	8607
T	271	1264	1652	8530	11,717

Maximizes AWS Skill Score

OBSERVED \ Forecast	A	B	C	D	T
A	24	95	37	85	241
B	29	563	260	390	1242
C	6	313	489	766	1574
D	14	297	585	7661	8557
T	73	1268	1371	8902	11,614

Maximizes Log Skill Score

OBSERVED \ Forecast	A	B	C	D	T
A	10	141	38	70	259
B	8	632	287	325	1252
C	2	373	460	740	1575
D	6	348	685	7518	8557
T	26	1494	1470	8653	11,643

Maximizes Gringorten Skill Score

OBSERVED \ Forecast	A	B	C	D	T
A	140	56	32	31	259
B	577	317	239	119	1252
C	426	319	574	256	1575
D	922	692	1905	5038	8557
T	2065	1384	2750	5444	11,643

CONTINGENCY TABLES FOR CIG/VIS AT 24 HR

Persistence

OBSERVED \ Forecast	A	B	C	D	T
A	41	76	48	96	261
B	59	336	249	683	1327
C	47	344	349	862	1602
D	121	496	1014	6890	8521
T	268	1252	1660	8531	11,711

Maximizes AWS Skill Score

OBSERVED \ Forecast	A	B	C	D	T
A	8	85	37	130	260
B	14	437	236	626	1313
C	14	276	341	966	1597
D	8	251	552	7669	8480
T	44	1049	1166	9391	11,650

Maximizes Log Skill Score

OBSERVED \ Forecast	A	B	C	D	T
A	3	93	58	106	260
B	3	474	304	532	1313
C	0	296	398	893	1587
D	2	298	705	7475	8480
T	8	1161	1465	9006	11,640

Maximizes Gringorten Skill Score

OBSERVED \ Forecast	A	B	C	D	T
A	97	66	49	48	260
B	534	316	266	197	1313
C	416	334	486	365	1601
D	876	901	1869	4824	8470
T	1923	1617	2670	5434	11,644

Overall Percent of Correct, Optimistic, and Pessimistic Forecasts *

3 HOUR

	Pers	AWS	Log	Gring
% of Correct Forecasts	80.5	88.5	88.0	76.6
% of Optimistic Forecasts	8.3	5.1	5.6	2.7
% of Pessimistic Forecasts	11.2	6.4	6.4	20.7

6 HOUR

	Pers	AWS	Log	Gring
% of Correct Forecasts	74.6	83.4	83.1	70.9
% of Optimistic Forecasts	11.6	8.5	8.8	4.7
% of Pessimistic Forecasts	13.8	8.1	8.1	24.4

12 HOUR

	Pers	AWS	Log	Gring
% of Correct Forecasts	69.2	75.2	74.0	52.1
% of Optimistic Forecasts	15.0	14.1	13.8	6.3
% of Pessimistic Forecasts	15.8	10.7	12.2	41.6

24 HOUR

	Pers	AWS	Log	Gring
% of Correct Forecasts	65.0	72.6	71.7	49.2
% of Optimistic Forecasts	17.2	17.9	17.1	8.5
% of Pessimistic Forecasts	17.8	9.5	11.2	42.3

*Made of 22 stations for CIG/VSBY combined.

Number of Hits and Busts*

3 HOUR

	Pers	AWS	Log	Gring
# of Hits	9392	10,296	10,241	8911
# of 1 Cat Busts	1902	1,119	1,174	2255
# of 2 Cat Busts	321	189	192	344
# of 3 Cat Busts	52	29	25	119

6 HOUR

	Pers	AWS	Log	Gring
# of Hits	8696	9684	9641	8233
# of 1 Cat Busts	2274	1546	1602	2695
# of 2 Cat Busts	578	331	330	526
# of 3 Cat Busts	110	44	33	152

12 HOUR

	Pers	AWS	Log	Gring
# of Hits	8110	8737	8620	6069
# of 1 Cat Busts	2370	2048	2234	3352
# of 2 Cat Busts	1025	730	713	1269
# of 3 Cat Busts	212	99	76	953

24 HOUR

	Pers	AWS	Log	Gring
# of Hits	7616	8455	8350	5723
# of 1 Cat Busts	2604	2129	2294	3434
# of 2 Cat Busts	1274	928	888	1563
# of 3 Cat Busts	217	138	108	924

*For 22 stations for CIG/VSBY combined.

Ratio of Forecasts* to Observations Made by Category

3 HOUR

Category	Pers	AWS	Log	Gring
A	1.270	1.025	.820	4.245
B	1.071	1.009	1.118	.889
C	1.031	1.057	.941	1.511
D	.977	.987	.999	.837

6 HOUR

Category	Pers	AWS	Log	Gring
A	1.067	.732	.556	3.937
B	1.072	.978	1.116	.877
C	1.040	1.054	.914	1.628
D	.979	1.000	1.013	.809

12 HOUR

Category	Pers	AWS	Log	Gring
A	1.030	.303	.100	7.973
B	1.002	1.021	1.193	1.105
C	1.042	.871	.933	1.746
D	.991	1.040	1.011	.636

24 HOUR

Category	Pers	AWS	Log	Gring
A	1.027	.169	.031	7.396
B	.943	.799	.884	1.232
C	1.036	.730	.923	1.668
D	1.001	1.107	1.062	.642

*Made for 22 stations for CIG/VSBY combined.