NASA Contractor Report NASA/CR-2001-210264

Statistical Short-Range Forecast Guidance for Cloud Ceilings Over the Shuttle Landing Facility

Prepared by:
Winifred C. Lambert
Applied Meteorology Unit

Prepared for:
Kennedy Space Center
Under Contract NAS10-96018

NASA
National Aeronautics and
Space Administration

Office of Management

Scientific and Technical
Information Program
2001

THIS PAGE INTENTIONALLY LEFT BLANK.

Executive Summary

The forecast cloud ceiling over the Shuttle Landing Facility (SLF) at Kennedy Space Center (KSC) is a critical element in determining whether a GO or NO GO should be issued for a Space Shuttle landing. However, the Spaceflight Meteorology Group (SMG) forecasters indicate that the ceiling at the SLF is challenging to forecast. The Applied Meteorology Unit (AMU) was tasked to develop a statistical cloud ceiling forecast technique to aid forecasters in this critical area. Recent studies in the literature have shown success using statistical methods to improve the short-term forecast of the Federal Aviation Administration (FAA) ceiling Flight Rules (FR). These studies provided the basis for the AMU task methodology, which was to create statistical equations to forecast ceilings based on the Shuttle FR.

A 20-year period of record (1978 – 1997) of hourly surface observation data from Daytona Beach (DAB), Orlando (MCO), Patrick Air Force Base (COF), Melbourne (MLB), and the SLF (TTS) was used for the equation development. An analysis of the data revealed that the largest number of ceilings below the Shuttle FR thresholds at the SLF occurred in the cool season (October - March), with very few occurrences in the warm season. Therefore, only the cool season data were used to develop the equations. The data were further stratified into dependent and independent datasets. Equation development was done with the dependent data, and verification done with the independent data. Of the 19 cool seasons in the dataset, 16 were chosen for the dependent dataset, leaving 3 cool seasons for the independent dataset.

Two types of forecast equations were developed: observations-based (OBS) equations that incorporated data from the stations listed previously, and persistence climatology (PCL) equations used as a benchmark against which the OBS equations were tested. Least squares multiple linear regression (MLR) was used as the statistical model for all the equations. Equations were developed for each of three ceiling thresholds at 1-, 2-, and 3-hour lead times, and for each hour of the day. The OBS equation development began with predictor selection using a forward stepwise regression. In most of the OBS equations, the predictor that explained most of the variance was the observation of the predictand at the initial time. This result was consistent with the findings in the literature. Once the predictors were chosen, the OBS forecast equations were developed. The PCL equations were made up of only two predictors: the observation of the predictand at the initial time and the climatology of the predictand at the valid time. The climatology term was a simple mean calculation of the number of ceiling events of each threshold for each hour of every day of the cool season. The PCL equations were developed after the climatologies were calculated.

The OBS and PCL equations were used to make forecasts from all records in the independent data set. The OBS forecast errors were then compared quantitatively to the PCL forecast errors. In order to test the performance of the OBS equations, the probability of detection (POD) and false alarm rate (FAR) were calculated. The quantitative comparison indicated that the OBS equations produced an improvement over the PCL equations. The OBS equation POD values were much larger than the FAR values, which indicated good performance overall. A hypothesis test was used to determine whether the OBS equation improvement was statistically significant. The improvement in skill created by the OBS equations was significant above the 99% confidence level. Therefore, the conclusion of the AMU study is that OBS equations produced more accurate forecasts than the PCL equations, and can be used in operations.

This success was tempered with other findings during the development. The predictors in the OBS equations only accounted for 55-60% of the variance in the data for the 1-hour equations to 35-40% for the 3-hour equations. There are several possible reasons for the unexplained variance. Other data, such as rawinsonde, satellite images or soundings, radar, or input from data assimilation software may be needed to fill the gap that the hourly surface observations could not. Another issue is that several meteorological phenomena could be responsible for the development of ceilings in east-central Florida during the cool season, yet the data were only stratified by season and time of day. A phenomenological stratification of the data would be time-consuming, but may be useful in developing more accurate forecast equations.

The OBS equations developed in this study will be useful since they produce improved forecasts over the PCL equations. They will provide another tool with which to make the ceiling forecasts critical to safe Shuttle landings at KSC. Combined with other observational and model data, as well as forecaster experience, these equations can help to improve the ceiling forecasts at the SLF.

Table of Contents

Executive Summary iii

List of Figures vi

List of Tables vii

1. Introduction 1

  1.1. Experimental Design Background 1

    1.1.1. Vislocky and Fritsch (1997) 2

    1.1.2. Hilliker and Fritsch (1999) 2

  1.2. AMU Study 3

2. Data 4

  2.1. Surface Observation Locations 4

  2.2. Data Pre-processing 5

  2.3. Exploratory Data Analysis 7

  2.4. Data Set Design 9

    2.4.1. Seasonal Stratification 9

    2.4.2. Dependent and Independent Data Sets 10

3. Equation Development 11

  3.1. Initial, Lead, and Valid Times 11

  3.2. Predictands 11

  3.3. Predictors 12

    3.3.1. OBS Predictors 12

    3.3.2. PCL Predictors 13

  3.4. Equations 14

    3.4.1. Preliminary Testing 14

      3.4.1.1. Statistical Models 14

      3.4.1.2. Predictor Selection 15

      3.4.1.3. Test Results 17

    3.4.2. OBS Equation Development 18

    3.4.3. PCL Equation Development 19

    3.4.4. Number of Events 19

4. Tests with Dependent and Independent Datasets 21

  4.1. Equation Comparison 21

  4.2. Hypothesis Testing 24

  4.3. Equation Performance 26

  4.4. Probability Cutoff for Yes/No Forecasts 28

5. Conclusions 33

5.1. Test results 33

5.2. Issues with Test Results 34

5.3. Other Efforts 34

5.4. Operational Use 35

Acknowledgements 35

References 36

List of Acronyms 37

List of Figures

Figure 2.1. A map of the locations of the stations whose data were considered for this task. 4

Figure 2.2. Histogram of the number of observations for all ceiling heights reported in the 20-year POR of hourly surface observations at the Shuttle Landing Facility. 8

Figure 2.3. Contour plot of the percent occurrence of ceilings < 8000 ft at the Shuttle Landing Facility by hour and month for the 20-year POR. 8

Figure 2.4. Graph showing the change in percent occurrence of ceilings < 8000 ft at the Shuttle Landing Facility by hour of day for each month in the 20-year POR. 9

Figure 3.1. A hypothetical example of the difference between Regression Estimation of Event Probabilities (REEP) and logistic regression (LGR). 15

Figure 3.2. R-squared (R^2) values for the 2-hour forecast valid at 1200 UTC for a ceiling of < 10 000 feet at TTS. 17

Figure 3.3. The number of predictand events used in the development of the observations-based equations for top) 1-hour forecasts, middle) 2-hour forecasts, and bottom) 3-hour forecasts. Within each graph, values are given for each predictand category and hour. 20

Figure 4.1. Percent improvement in MSE of the OBS equations over the PCL equations for forecasts of ceilings < 10 000 feet at each lead time and hour of the day for the cool season. 22

Figure 4.2. Percent improvement in MSE of the OBS equations over the PCL equations for forecasts of ceilings < 8000 feet at each lead time and hour of the day for the cool season. 23

Figure 4.3. Percent improvement in MSE of the OBS equations over the PCL equations for forecasts of ceilings < 5000 feet at each lead time and hour of the day for the cool season. 24

Figure 4.4. Distribution of differences between the OBS and PCL MSE values for the 24 hourly equations for 1-hour forecasts of ceilings < 5000 feet. 25

Figure 4.5. Distribution of differences between the OBS and PCL MSE values for the 24 hourly equations for 2-hour forecasts of ceilings < 8000 feet. 25

Figure 4.6. The standard 2 x 2 contingency table used in calculating several measures of the accuracy of binary forecasts. 27

Figure 4.7. The average POD and FAR values for the three lead times at each probability cutoff using the independent dataset. 28

Figure 4.8. The B, HR, KSS, and TS values for the 1-hour lead time. 30

Figure 4.9. The B, HR, KSS, and TS values for the 2-hour lead time. 31

Figure 4.10. The B, HR, KSS, and TS values for the 3-hour lead time. 32

List of Tables

Table 1.1. List of Flight Rules for ceiling thresholds at the Shuttle Landing Facility (SLF). 1

Table 2.1. The values used in the “impossible value” QC routine. 6

Table 3.1. Examples of the predictand binary values used in the equation development. 1 means the observation met the threshold condition, and 0 means the condition was not met...................12

Table 3.2. List of potential predictors and binary thresholds used for the observations-based equation development. A predictor is set equal to 1 if the observation satisfies the binary threshold, otherwise it is set equal to 0. “Continuous” means that the actual value of the variable was used. 13

Table 3.3. A summary of the parameters in the development of the equations used to test which statistical model and predictor cutoff method were most appropriate for the full set of equations. The total number of OBS and PCL equations developed for the testing are shown in the last column. The number of equations includes those for each of the three Shuttle FR binary ceiling categories..18

Table 4.1. Average values of the percent improvement (PI) for the 24 equations, valid 0000 to 2300 UTC, by lead time, dataset, and ceiling category. 21

Table 4.2. The POD and FAR scores of the OBS forecasts using the dependent and independent datasets. The values are averaged for the 24 valid times in each ceiling height/lead time category. 27

1. Introduction

The forecast cloud ceiling over the Shuttle Landing Facility (SLF) at the Kennedy Space Center (KSC) is a critical element in determining whether a GO or NO GO should be issued for a Space Shuttle landing. For a GO decision to be issued, the cloud ceiling constraints defined by the Space Shuttle Flight Rules (FR) (NASA/JSC 1997a) in Table 1.1 must be observed GO at the de-orbit burn decision time and forecast GO for the time of landing. However, the Spaceflight Meteorology Group (SMG) forecasters have found that cloud ceiling is a challenging parameter to forecast, even in the short-term (0-6 hours) when persistence is expected to be a reliable predictor. The Applied Meteorology Unit (AMU) was tasked to develop a statistical cloud ceiling forecast technique to aid forecasters in this critical area.

Table 1.1. List of Flight Rules (NASA/JSC 1997a) for ceiling thresholds at the Shuttle Landing Facility (SLF).

Ceiling Threshold	***Flight Rule***
< 5000 ft	Return to Launch Site (RTLS)
< 8000 ft	End of Mission (EOM)
< 10 000 ft	Navigation Aid Degradation

1.1. Experimental Design Background

The technique described in this report was developed according to the steps outlined by the World Meteorological Organization (1992, hereafter WMO):

- Define the weather element to be predicted, *i.e.* the predictand.
- Determine whether observational data, model output data, or both will be used in the forecast equations. WMO (1992) suggests that observational data alone can be used for short-term forecast equations in the 0-6 hour time period.
- Prepare the datasets by first selecting the predictors and determining which data types are needed, then stratify the data by time of day or year if necessary, and, finally, create the dependent and independent datasets.
- Select the statistical model, *e.g.* linear regression, logistic regression, etc.
- Develop and test the equations using the dependent dataset,
- Test the equations using the independent dataset.

The element to be predicted, or the predictand, was the ceiling observation at the SLF or some categorization thereof, and the short-term nature of the forecasts would imply that observational data only should be used to develop the equations (WMO 1992). Selecting the statistical model and data types to use was a much more daunting task. Several possible statistical models other than those listed above were possible candidates for the equations, from decision trees to neural networks. Many data types were also considered. Observational data available for the SLF included hourly surface observations, wind tower network data, upper air data from rawinsondes and profilers (50- and 915-MHz), Weather Surveillance Radar 1988 Doppler (WSR-88D) and WSR-74C radar data, satellite data, lightning data, and even output from several forecast models. These choices had to be made before the datasets could be prepared and equations developed. Two recent studies presented in the literature provided guidance in making the choices: Vislocky and Fritsch (1997) and Hilliker and Fritsch (1999). Both showed success in improving the short-term cloud ceiling forecast.

1.1.1. Vislocky and Fritsch (1997)

The main purpose of the Vislocky and Fritsch (1997, hereafter VF) study was to determine if an observations-based forecast system could be used to improve short-term forecasts of cloud ceiling and visibility. According to VF, previous studies had used observational data from a single station to develop forecasts of surface conditions at that station in the short-term. The difference in the VF study was that observational data from a network of stations surrounding the station of interest were added as predictors for that station's surface conditions. The dataset used by VF included 13 years of hourly observations from 202 sites in the Eastern United States and Southeastern Canada. Of these 202 stations, 25 were chosen as the stations for which probability forecasts of ceiling and visibility categories would be made.

The hourly surface observations of the standard Federal Aviation Administration (FAA) FR categories valid at 1 hour, 3 hours, and 6 hours after the initial times of 0300 and 1500 UTC were used as predictands. The predictands and most of the predictors were coded as binary in which a "1" signified that an observation satisfied a variable definition and a "0" signified the opposite. A forward stepwise algorithm was used to select the best predictors and least squares multiple linear regression (MLR) (Wilks 1995) was used to develop the predictive equations. The output from the observations-based equations were compared to that of a benchmark called persistence climatology to determine if the new equations offered an improvement. The persistence climatology equations were also developed using MLR and contained the station observation of the predictand and the station climatology of the predictand as predictors. VF called persistence climatology a "formidable benchmark" for short-term forecasts, and any improvement over it would be significant. A model output statistics (MOS)-based system of equations was also produced and compared to persistence climatology.

The equations were developed using 11 years of data and tested with the remaining two years. The observations-based and MOS forecasts were quantitatively compared to those of persistence climatology. The results of the comparisons were averaged over the 25 stations and both initial times for each individual FR category and lead time. They were presented as percent improvements over persistence climatology. Scores for the observations-based ceiling forecasts ranged from 6-20%, depending on the lead time and ceiling category. The lowest ceiling category (< 500 ft) 1-hour forecast had the smallest improvement, and the highest ceiling category (< 6500 ft) 6-hour forecast had the largest improvement. The skill of the observations-based 1-hour and 3-hour equations was superior to that of the MOS-based equations. At the 6-hour forecast, the MOS-based performance was slightly better than the observations-based performance, indicating that model output becomes an important predictor at this point and beyond. These results indicated that for short-term forecasts (< 6 hours), more emphasis should be placed on using observational data from stations surrounding the forecast site instead of model data.

1.1.2. Hilliker and Fritsch (1999)

Following the basic premise of VF, Hilliker and Fritsch (1999, hereafter HF) developed equations to predict hourly probabilities of marine stratus burn-off at San Francisco International Airport (SFO). This study built upon VF by adding upper-air sounding data as possible predictors for the equations. The dataset included 15 years of standard hourly surface data from 14 sites around the San Francisco Bay area and rawinsonde data from Oakland, CA. The data were stratified between warm and cool seasons because of the much larger occurrence of low ceilings during the warm season.

For various reasons as explained in the article, HF chose the ceiling at or below 3000 ft between the hours of 1500 UTC and 2100 UTC, inclusive, as the predictand. This also coincides with one of the standard FAA FR ceiling categories and, as such, was one of the predictors in VF. As in VF, the predictand and most of the predictors were coded as binary. In the experiments, 1- to 6-hour forecast equations were created with 1500 and 1800 UTC as the valid time. As in VF, a forward stepwise algorithm was used to select the best predictors, but logistic regression (LGR) (Wilks 1995) was used to develop the predictive equations instead of MLR. The output of the equations was a probability of occurrence value between 0 and 1 inclusive. Similar to VF, the observations-based equations were compared to the benchmark persistence climatology.

The equations were developed using 12 years of data and tested on the remaining three years. Equations using only surface observations and equations using surface and upper-air data were created to determine if adding rawinsonde data made a significant improvement in the forecast over surface data alone. As in VF, a MOS-based system of equations was produced and also compared to persistence climatology. The results were averaged for each lead time and each season. The observations-based forecasts again showed improvement over persistence climatology from 17-21% in the warm season and 6-10% in the cool season. The addition of rawinsonde data increased the improvement by approximately 3%. As in VF, the observations-based forecasts produced a larger improvement over persistence climatology than did the MOS equations, indicating that observational data is more important for short-term forecasts than is model data.

1.2. AMU Study

These two successful studies provided a model for the AMU to follow in determining the predictands, choosing the statistical model and data types, and narrowing down the possible predictor list for similar observations-based equations. However, the AMU task differed from the studies in the literature in two ways. First, the previous studies used data from the eastern United States (not including Florida) and the San Francisco Bay area where persistent ceilings were known to exist. Such conditions are not the norm in the subtropical environment of east-central Florida. Second, the studies used standard FAA FR cloud ceiling categories as predictands. The predictands in the AMU task were the ceiling thresholds as defined by the Space Shuttle FR.

There are five sections in this report. Section 2 describes all aspects of the data type used including period of record, locations, quality control, exploratory data analysis, and stratification. Section 3 discusses what predictands and predictors were used and the approach used in developing the equations. The statistics showing the performance of the equations are given in Section 4, and the conclusions based on those results are discussed in Section 5.

2. Data

The successful results documented in VF and HF as well as the information from WMO (1992) led to the conclusion that the predictands and predictors should be drawn from observational data exclusively. The equations in VF and HF used the standard hourly surface observations. Therefore, the hourly surface observations from the SLF (3-letter station identifier TTS) and several stations in east-central Florida were collected from the Air Force Combat Climatology Center (AFCCC) to develop the equations in this study. The period of record (POR) for TTS provided by AFCCC covered 20 years from January 1978 to March 1997.

2.1. Surface Observation Locations

The map in Figure 2.1 shows the locations of the stations provided in the AFCCC dataset. The stars and boxes indicate the locations of the hourly surface observations and upper-air soundings, respectively. The triangles are the locations of operational buoys and Coastal Marine Automated Network (CMAN) stations. These datasets were acquired through the National Data Buoy Center (NDBC).

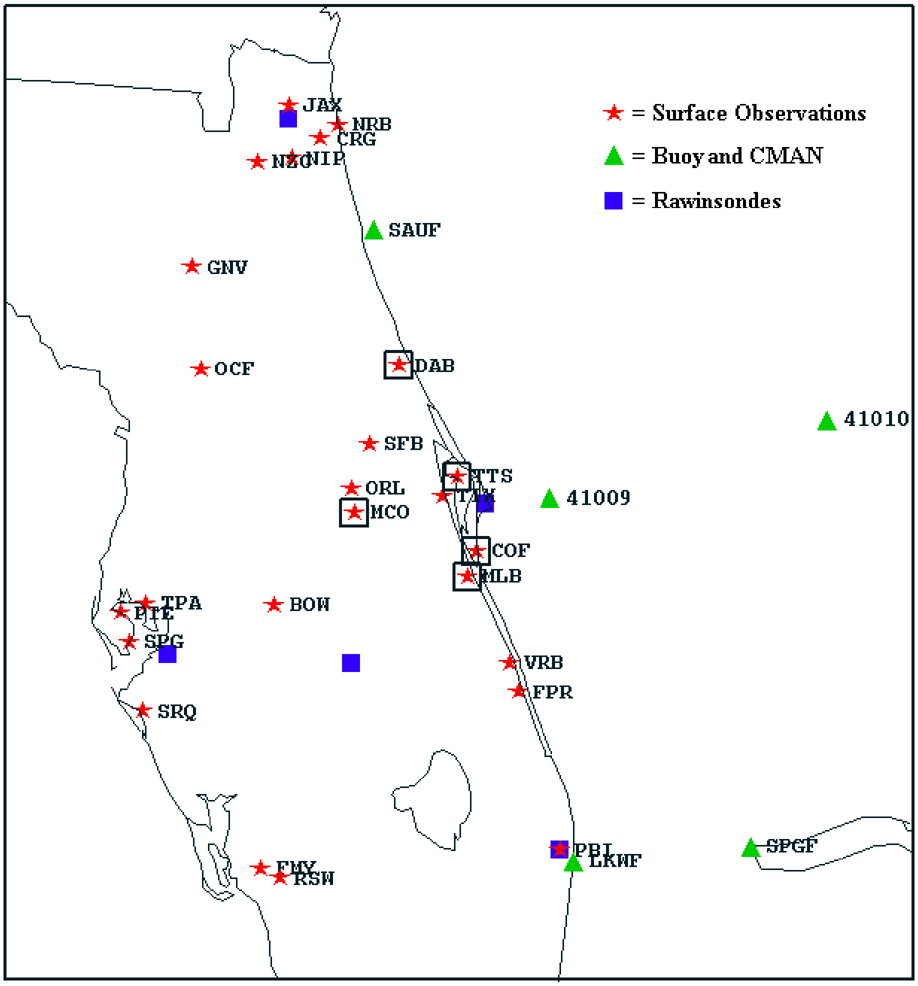


Figure 2.1. A map of the locations of the stations whose data were considered for this task. The stars indicate the hourly surface observation stations, the solid boxes indicate the upper-air sounding stations, and the triangles indicate the CMAN and buoy stations. The surface stations with boxes surrounding them were the only stations used in this study.

The data from all the stations were examined for POR and the amount of missing data. Several of the hourly surface observation stations closed during the 20-year POR, others had long periods of missing data on the order of days to several months at a time. These stations were deemed not usable for the study. The list of possible surface observation stations, other than TTS, was narrowed to the following:

- Jacksonville (JAX),
- Daytona Beach (DAB),
- Orlando (MCO),
- Tampa (TPA),
- Patrick Air Force Base (COF),
- Melbourne (MLB),
- Vero Beach (VRB), and
- West Palm Beach (PBI).

Of these stations, COF was closed during the nighttime hours. However, its data could still be important in daytime forecast equations due to its close proximity to TTS and was kept in the dataset. During equation development (Section 3), data from JAX, TPA, VRB, or PBI were never chosen as potential predictors in the selection process. Therefore, the only hourly surface observations used were from TTS, DAB, MCO, COF, and MLB.

The rawinsonde data were also checked for POR and missing data, and the TTS dataset was deemed usable. However, considering both the amount of time needed to preprocess and quality check the upper-air data and the small gain from using upper-air data found in HF, the AMU decided to develop the equations using only surface observations. If good results were found using the surface data and time permitted, the sounding data would be processed and added to the predictor list. The POR for the buoy data was only 10 years, not long enough to match the POR of the hourly surface and rawinsonde data. As will be seen in Section 3, surface observations of cloud parameters were chosen as the predictors almost exclusively over variables such as temperature or wind speed. The buoy data did not contain cloud reports and would likely not have contributed to the equations. It was, therefore, not used in this study.

2.2. Data Pre-processing

The data from AFCCC were provided unformatted and had not been checked for quality. Therefore, the AMU created software to decode the data and put it in a format that could be imported to a spreadsheet or database. The data were also filtered and checked for quality (QC) during the decoding.

The filter removed all special observations and kept only the regular observations that were made on the hour. Special observations were made only when there is a significant change in the weather within the hour and were not uniform in time. As such, there were relatively few in the dataset and they were not made at consistent times, making it difficult to determine how to use their information in the equation development.

Three QC routines were applied to the hourly data in order to remove erroneous values in the observations. The first QC applied was an “impossible value” check that removed values of certain variables that could not possibly exist, or were very unlikely to exist in central Florida. The values used are shown in Table 2.1 and were based on input from local forecasters.

Table 2.1. The values used in the “impossible value” QC routine.

Wind Direction	< 0°	*or*	> 360°
Wind Speed	< 0 ms	*or*	> 60 ms (116 kts)
Wind Gust	< 0 ms	*or*	> 70 ms (136 kts)
Temperature	< -10C (14°F)	*or*	> 40C (105°F)
Dew Point Temperature	< -18C (0°F)	*or*	> 35C (95°F)
Ceiling Height	< 0 m		
Visibility	< 0 m		
Altimeter	< 960 mb	*or*	> 1040 mb
3 Hour Pressure Change	< -10 mb	*or*	> 10 mb

The second QC removed outliers in the data through a standard deviation check. Mean and standard deviation values for wind speed, temperature, dew point temperature, and altimeter were calculated by hour and month over the 20-year POR. Data were removed if their values were more than 10 standard deviations from the mean. Ceiling, visibility, wind direction, wind gust, and 3-hour pressure change observations were not put through this check. Ceiling and visibility observations encompassed a large range of values and any value above 0 was possible. A wind direction average would likely be meaningless as wind directions changed from day to day and hour to hour. Wind gusts did not occur consistently and could also be discontinuous in their values from day to day, as was also true for the 3-hour pressure change.

The third and final QC was a simple temporal consistency check. This was only done on the temperature, dew point temperature, and altimeter data because these variables tend to be much more continuous from hour to hour than the others. Ceiling and visibility especially were found to be highly discontinuous from hour to hour and not good candidates for this algorithm. As a value (V) was checked, the value observed one hour previous to (V_{-1}) and one hour after (V_{+1}) V were averaged using the equation

$$V_{avg} = (V_{-1} + V_{+1}) / 2.$$

If the absolute value of the difference between V_{avg} and V ($| V_{avg} - V |$) was larger than a certain threshold, then V was removed. The threshold difference values for each of the variables were:

Temperature:	20 C (36°F)
Dew Point Temperature:	20 C (36°F)
Altimeter:	10 mb

If a value was missing before or after the value being checked, the QC was not performed.

The total number of data points removed by all three QC algorithms was very small. This could be due to the fact that the QC algorithms were formulated to remove only the obvious gross errors in the data. Valid observed extremes in all data types, such as very cold temperatures in the cool season or strong winds in any season, were in the 20-year dataset. The goal of the QC was to remove erroneous data while retaining most of the valid observations. It is also possible that the dataset was clean, with very few erroneous values. The percentages of observations removed by the three QC algorithms at each station are:

COF:	0.01%
DAB:	0.04%
MCO:	0.01%
MLB:	0.07%
TTS:	0.01%

Although both VF and HF employed methods of filling in missing data, it was not done with the dataset in this study. In order to fill in missing cloud cover and visibility information, VF and HF used a 'nearest neighbor' algorithm that substituted the values from the nearest station. The SMG and 45th Weather Squadron (45 WS) forecasters indicate that ceilings over the KSC/Cape Canaveral Air Force Station (CCAFS) area have been spatially discontinuous. Given the importance of accuracy in forecasting for Shuttle landings, the risk in using an inaccurate value in place of a missing one was too high. There are also temporal routines that fill in missing data based on past and future values such as linear interpolation or a spline method. A careful examination of the TTS dataset showed that missing data tended to occur in blocks of several hours as opposed to just one missing hour. The examination also showed instances of large differences in ceiling values from hour to hour. These two findings reduce confidence in using interpolated values to replace the missing data.

2.3. Exploratory Data Analysis

Once the data were formatted and checked for quality, an exploratory data analysis (EDA) was conducted to identify characteristics such as temporal trends and climatologies, the frequency of occurrence of the Shuttle FR ceiling categories, and possible relationships between potential predictors and predictands. The images in this section show some of the important results from the EDA.

Figure 2.2 is a histogram showing the number of observed occurrences for all ceiling heights reported in the data for the entire POR. The ceiling heights in this dataset were in discrete 30 m increments up to 1500 m, 300 m increments from 1500 m to 9000 m, and approximately every 1500 m above that. This histogram shows that there are preferred reporting heights above 900 m (~ 3000 ft) and that the data are not evenly distributed, in general. Three of those preferred heights correspond to the Shuttle FR ceiling thresholds at 1500 m (~ 5000 ft), 2400 m (~ 8000 ft), and 3000 m (~ 10 000 ft), as indicated by the arrows on the graph. The observer at the site uses all available data sources to provide a best estimate of the ceiling height. Ceilometer data are available to the observers at the SLF and are treated as a prime source of ceiling height information. However, like the other sources, ceilometer data are used only as a supplement to observer discretion and the final ceiling value is a human estimation. The uneven distribution seen in Figure 2.2 is likely due to human error in estimating ceiling heights and may be part of the reason for the difficulty experienced in forecasting ceilings. It may also be difficult to use such data in a statistical analysis where smooth theoretical distributions are assumed in many of the methods. For comparison, the inset in the upper-left corner of Figure 2.2 shows the histogram of temperatures for the same time period, which shows a smooth distribution that is more amenable to statistical analysis.

Figure 2.3 is a color-fill contour plot of the percent occurrence of ceilings below 8000 ft by month and hour of day (UTC). This height threshold was chosen as an example since the other height thresholds (5000 ft and 10 000 ft) showed similar patterns. The months on the y-axis were ordered from July to June so that the cool season months would be emphasized in the middle of the plot. This produces a graph that clearly shows the maximum occurrences of low ceilings in the cool season. Figure 2.3 shows that the highest percentage of ceilings < 8000 ft occurs during the cool season morning hours that flank the time of local sunrise. For example, 30-35% of the time in December and January between the hours of 1100 and 1400 UTC (0600 – 0900 local time), the ceiling is below 8000 ft.

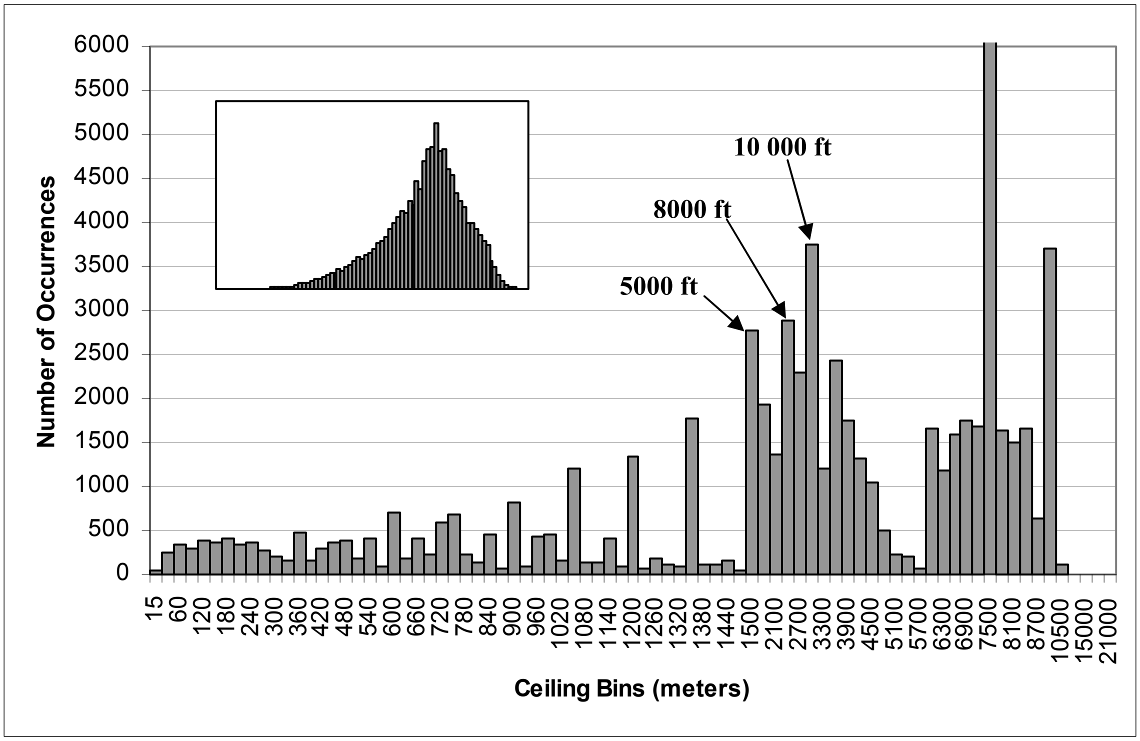


Figure 2.2. Histogram of the number of observations for all ceiling heights reported in the 20-year POR of hourly surface observations at the Shuttle Landing Facility. The total number of observations at 7500 m is 11 582. The scale of the vertical axis was truncated to emphasize the smaller number of observations at other heights. The inset in the upper-left corner is the histogram of temperatures for the same period.

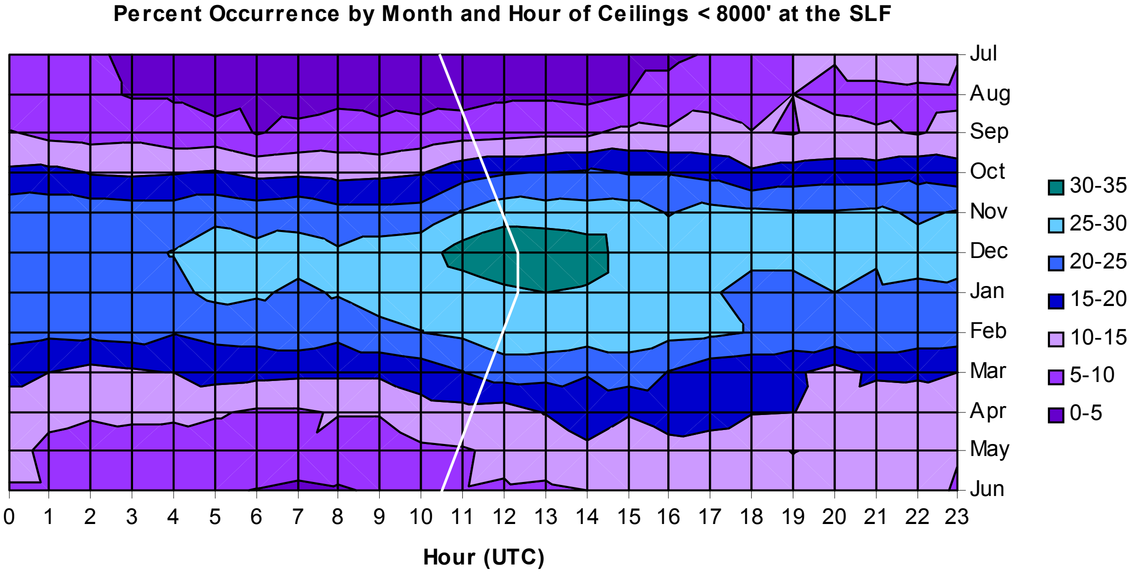


Figure 2.3. Contour plot of the percent occurrence of ceilings < 8000 ft at the Shuttle Landing Facility by hour and month for the 20-year POR. The legend at right shows the association of the colors with percentage ranges in intervals of 5%. The white line indicates the approximate time of local sunrise at KSC.

Figure 2.4 is another way of looking at the data in Figure 2.3, with each line showing the values of the percent of occurrence of ceilings < 8000 ft for a specific month. As expected, the cool months (October – March) show a higher percent of occurrence at all hours with peaks in the morning (~ 1000 – 1500 UTC) with a maximum of close to 35% in December. There is a much lower percent of occurrence during the warm season (April – September). Also, the warm season peaks tend to be less than 15% (April reaches 17%), which is only slightly higher than the lowest cool season values in October and March of approximately 14%. The large disparity in the number of events between the warm and cool seasons as shown in Figures 2.3 and 2.4 suggests that the data should be stratified by these seasons.

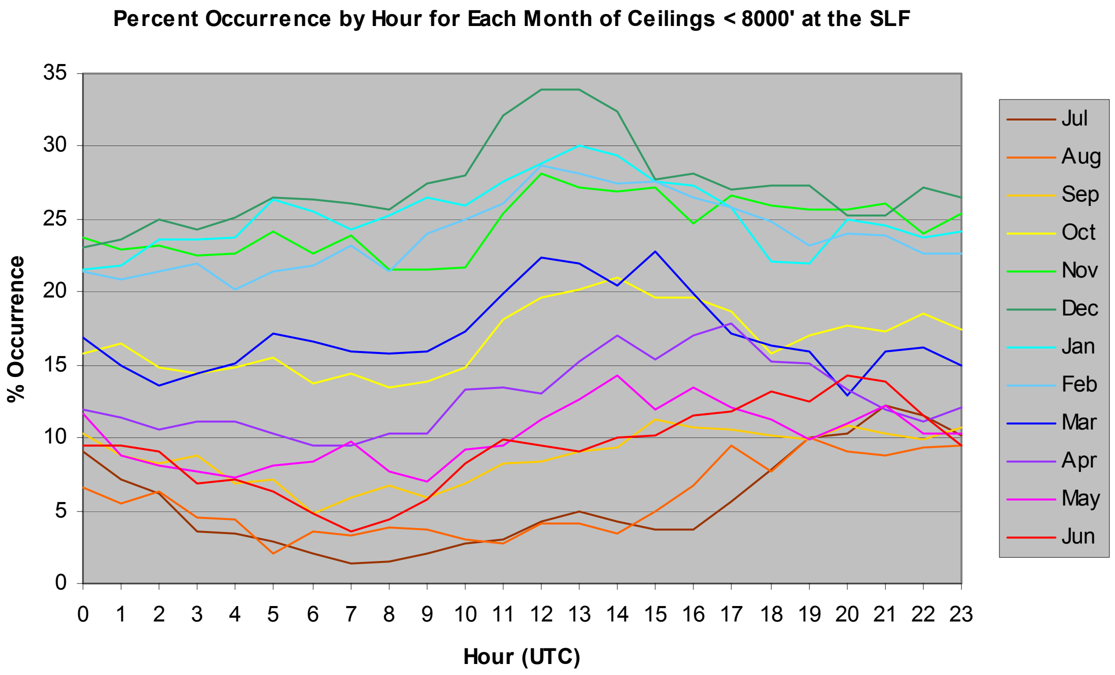


Figure 2.4. Graph showing the change in percent occurrence of ceilings < 8000 ft at the Shuttle Landing Facility by hour of day for each month in the 20-year POR. The legend at right shows the association of the colors with each month in the year.

2.4. Data Set Design

After the EDA, the datasets to be used for the equation development were prepared. This involved stratification by season, then separation into development (dependent) and testing (independent) datasets.

2.4.1. Seasonal Stratification

The EDA revealed large differences in the occurrence of low ceilings between the warm and cool seasons. This could mean that different climatological processes were responsible for the creation of low ceilings and, therefore, the forecast equations for one season may be markedly different from those for the other season. In order to obtain the greatest forecast accuracy, the data from all stations were stratified into cool (October-March) and warm (April-September) season datasets prior to equation development. Because of the relative dearth of FR threshold ceiling events in the warm season (see Figures 2.3 and 2.4), there were not enough events from which to develop robust relationships between the TTS ceiling observations and other data types. Therefore, only the cool season data were used in the creation of the forecast equations.

2.4.2. Dependent and Independent Data Sets

The cool season data was then separated into two more datasets known as the dependent and independent datasets. The dependent dataset was used to develop the equations, and the independent dataset was used to test the equations. According to WMO (1992), the dependent dataset should contain enough samples so that the resulting set of equations is stable, i.e. the equations maintain consistent forecast accuracy on different datasets. A small dataset may not contain a representative set of events, and the equations developed from such a small set may show wide variations in accuracy on different datasets and forecasters will not have confidence in their results. It is also desirable, in a statistical sense, that the events in the dependent dataset have no correlation between them. This is not usually the case with meteorological data since the datasets consist of a time series of events that are serially correlated. In general, according to WMO (1992), the greater the correlation between events, the smaller the amount of new information contributed by each event, and the larger the dependent dataset must be to create equations representative of the relationship between the predictors and the predictand.

The independent dataset was needed for equation testing in order to have a more realistic view of how the equations would perform in operations. It was expected that the equations would not perform as well on the independent data as they would with the data from which they were developed. However, if performance were a great deal worse with the independent data, this would indicate that either too many predictors were chosen and the equations were fit too strongly to the dependent data, or the dependent dataset was too small.

In order to develop stable equations, a large portion of the cool season dataset was set aside as the dependent dataset. There were 19 full cool seasons in the 20-year POR, 16 of which comprised the dependent dataset with the other three making up the independent dataset. Instead of choosing a chronological sequence of years for the independent dataset, statisticians at ENSCO, Inc. recommended random years be used. This would ensure that the results with the independent data would not be contaminated by any climate trends, but would reflect how the equations would perform in any time period. The independent cool season years were chosen using the random number generator function in Microsoft Excel©. In the range of cool seasons from 1978/79 to 1996/97, the cool seasons chosen for the independent dataset were 1979/80, 1987/88, and 1995/96, with the rest making up the dependent dataset.

3. Equation Development

Equation development began after the datasets were prepared. Equations for both the observations-based (OBS) and persistence climatology (PCL) methods were created. As in VF and HF, the performance of the OBS equations was compared against the benchmark PCL equations to determine if they improved the short-term ceiling forecast. The development was done with a commercial-off-the-shelf statistical software package called S-PLUS® (Insightful Corporation 1999). This software package contained all of the functions necessary to do the entire development and testing on a PC. Use of S-PLUS® also eliminated the need to develop and test statistics code from scratch, saving valuable time.

Before the equation development could take place, it was important to establish the initial, lead, and valid time for each equation, and determine the number of predictands. These parameters determined how many equations were developed. Then the predictors were chosen, and the predictands and predictors prepared for input into the S-PLUS® algorithms that created the forecast equations. This section will describe how these parameters were chosen along with the details of the mathematical process used in the creation of the equations.

3.1. Initial, Lead, and Valid Times

Shuttle landings occur at various times around the clock. Therefore, the forecast equation valid times needed to cover the entire 24-hour period at the highest temporal resolution possible. Standard surface observations were reported every hour on the hour at most stations in the datasets, with an occasional special report given off the hour if there was a significant change in the weather. The special reports in the dataset were too few and too sporadic to allow development of equations with valid times off the hour. Therefore, the forecast equations developed in this study were valid for each hour of the day, given that the highest consistent temporal resolution in the dataset was hourly.

The initial time for each equation was determined by the lead time. The customers requested that the equations have lead times of 1-, 2-, and 3-hours. Therefore, for each valid time there were three initial times: one hour, two hours, and three hours prior to the valid time.

3.2. Predictands

The predictands, or weather elements to be predicted, could be either continuous in which an exact value was predicted, or categorical in which a prediction for a specific range of ceiling heights was made. The difficulty in forecasting an exact ceiling value was underscored in Figure 2.2. The distribution of ceiling height observations in the historical dataset was not smooth but showed a tendency to have observations consolidated at specific heights. The values were estimated through human observations, not measured. This may have introduced errors such as a bias to report only certain heights as evidenced in Figure 2.2. Since an accurate measured value of the ceiling height was not reported in the historical dataset, it was impossible to create equations that would forecast the continuous ceiling height values accurately.

The forecasters at SMG are interested in forecasting the probability of whether a ceiling threshold defined in the Shuttle FR (Table 1.1) will be violated. This requirement and the data distribution issue seen in Figure 2.2 were the deciding factors for using categorical predictands of the three ceiling thresholds in the Shuttle FR for the SLF: < 10 000 ft, < 8000 ft, and < 5000 ft. The specific parameter forecast would be the probability of a ceiling occurring within each of the FR ceiling thresholds at TTS. In the equation development, the predictands were coded as binary variables in which a 1 was used if an observation satisfied the predictand definition and a 0 otherwise. For example, for a ceiling observation of 7000 ft, the predictands < 10 000 ft and < 8000 ft would be set to 1, and the predictand < 5000 ft would be set to 0. Other examples are shown in Table 3.1.

Table 3.1. Examples of the predictand binary values used in the equation development. A value of 1 means the observation met the threshold condition, whereas a value of 0 means the condition was not met.

Ceiling Height Ob	***Predictand Category***		
	< 10 000 FT	**< 8000 FT**	**< 5000 FT**
11 000 ft	0	0	0
10 000 ft	0	0	0
9000 ft	1	0	0
8000 ft	1	0	0
7000 ft	1	1	0
5000 ft	1	1	0
4000 ft	1	1	1

3.3. Predictors

The OBS and PCL methods each had their own set of predictors. For the OBS method, several observed parameters from all the stations at the forecast initial time were considered as predictors. For the PCL method, the predictors were the observation of the ceiling at the initial time and a climatological value for the valid time.

3.3.1. OBS Predictors

A vast number of predictors could have been derived from the observed data. The predictor lists in VF and HF were used as starting points in this study to reduce the size of the potentially large list of possible predictors. They conducted preliminary studies in order to reduce the large number of potential predictors while retaining the predictors best suited for ceiling and visibility forecasts. Their final lists were very similar and consisted of binary thresholds for visibility, wind direction, and several cloud observations including ceiling height and total cloud cover, and continuous values of wind speed, temperature, dew point temperature, and pressure.

The predictor lists were modified for this study by using knowledge of local meteorology and conducting tests to create a list of physically reasonable predictors for categorical ceiling forecasts at the SLF. Previous AMU work showed that the visibility thresholds were not good predictors for cloud ceiling in east central Florida and were, therefore, removed from consideration. Cloud deck height thresholds were added to the list. Heights for up to four cloud decks were reported in the standard hourly observations, and were categorized according to the Shuttle FR. The basic idea was that if any amount of clouds within a FR height category existed at the initial time, that cloud deck observation might be a good predictor for that ceiling threshold. The wind direction predictors were modified to reflect the locations of stations relative to TTS: north (DAB), west (MCO), south (MLB and COF), and east (no stations). The continuous variables were the same as in VF.

The resulting predictor list used in this study is shown in Table 3.2. As with the predictands, most of the predictors were coded as binary in which a 1 was used if an observation satisfied the binary threshold and a 0 otherwise. Where the word “continuous” is seen in the table, the actual value of the variable was used in the equation development. The first three predictors in the table correspond to the predictand categories. One of the main conclusions in both VF and HF was that the most important predictor was the observation of the predictand at the initial time.

Table 3.2. List of potential predictors and binary thresholds used for the observations-based equation development. A predictor is set equal to 1 if the observation satisfies the binary threshold, otherwise it is set equal to 0. "Continuous" means that the actual value of the variable was used.

Variable	*Binary Threshold*	*Variable*	*Binary Threshold*
Ceiling Height	< 5000 ft	Second Cloud Deck Base	< 5000 ft
Ceiling Height	< 8000 ft	Second Cloud Deck Base	< 8000 ft
Ceiling Height	< 10 000 ft	Second Cloud Deck Base	< 10 000 ft
Total Cloud Cover	> 1/10	Third Cloud Deck Base	< 5000 ft
Total Cloud Cover	> 5/10	Third Cloud Deck Base	< 8000 ft
Total Cloud Cover	> 9/10	Third Cloud Deck Base	< 10 000 ft
Wind Direction	> 315 and < 45	Fourth Cloud Deck Base	< 5000 ft
Wind Direction	> 45 and < 135	Fourth Cloud Deck Base	< 8000 ft
Wind Direction	> 135 and < 225	Fourth Cloud Deck Base	< 10 000 ft
Wind Direction	> 225 and < 315	Wind Speed	Continuous
Precipitation	Yes	Temperature	Continuous
First Cloud Deck Base	< 5000 ft	Dew Point Temperature	Continuous
First Cloud Deck Base	< 8000 ft	Dew Point Depression	Continuous
First Cloud Deck Base	< 10 000 ft		

3.3.2. PCL Predictors

Forecasts from the PCL equations were used as the benchmark against which the forecasts from the OBS equations were tested. Development of the PCL equations was based on the method described in VF. There were only two predictors for the PCL equations, which were the observation of the predictand at the initial time and a term representing the predictand climatology at the valid time. The observation of the predictand at the initial time represented the persistence part of the equation. A persistence forecast is one in which the assumption is that the ceiling will be the same at the valid time as it was at the initial time. The climatology term was represented by a fractional value indicating the frequency of occurrence of the predictand at a certain hour on a certain day. While the persistence term was readily available from the observations, the climatology term had to be calculated.

Calculation of the climatology term, called the event relative frequency (ERF), followed the procedure described in VF. An ERF was calculated for each hour of the day and each day of the year, a total of 8760 values, using the entire 20-year POR. ERFs were calculated for each day of the cool season, each hour of the day, and each ceiling category threshold. In other words, there was one ERF value per ceiling category for each hour of the day in each day of the cool season.

Using ceilings < 5000 feet occurring at 1200 UTC on 1 January as a specific example, the following procedures were used to calculate the ERFs:

- The equation used to calculate the ERF was a simple average given by

$$\mathrm{ERF} = \frac{1}{n}\sum_{i=1}^{n}(o_i)\,,$$

  where n is the number of possible observations and o_i was 1 if the ceiling category was observed, 0 if it was not.

- There were a total of n=20 1200 UTC 1 January observations of ceilings < 5000 feet in the 20-year POR. According to VF, this number is too small to calculate a reliable climatological average.

- To increase n, all observations of ceilings < 5000 feet at 1200 UTC from 15 days before and 15 days after 1 January, inclusive, were used for a total of 31 days. The new value for n, then, is 620 (20 years x 31 days).
- The 620 binary observation values were added together and divided by n, according to the equation above.
- This procedure was repeated for every hour of every day in the cool season.
- With 182 days in the cool season and 24 hours/day, there were 4368 ERF values per category used in the development of the PCL equations. For leap years, the 29 February values were calculated by averaging the 28 February and 1 March values.
- The ERF values ranged from 0.06 at 0900 UTC in October to 0.40 at 1400 UTC in December.

The result was 8760 values that could be matched with each day/hour combination in every year.

3.4. Equations

Equations were developed for each of the predictands at each of the lead times for each hour of the day (0000 to 2300 UTC) using cool season data only. Three predictands at three lead times over 24 hours yielded

*3 Predictands x 3 Lead Times x 24 Hours = **216 Equations**.*

A full set of equations was developed for each forecast method, OBS and PCL. This yielded

*2 Forecast Methods x 216 Equations/Method = **432 Total Equations**.*

These 432 equations would be used for every day in the cool season. In addition, predictors for the OBS equations must be chosen from all predictors from all stations:

*5 Stations x 27 Predictors/Station = **135 Possible Predictors**.*

It was clear that, from the large number of equations to create and the number of possible predictors from which to create them, a standardized automatic method of development and testing was needed. The time allotted for the task did not permit development, testing, and refinement of each individual equation. Therefore, all equations would use the same model formulation and all OBS predictors would be selected using the same criteria for each individual equation.

3.4.1. Preliminary Testing

Tests were conducted on portions of the data to determine (1) the appropriate statistical model for both forecast methods and (2) the predictor selection criteria for the OBS equations. Equations for 1- and 2-hour forecasts valid at 1200 and 1500 UTC were used to test the different statistical models and predictor selection methods and determine which were best suited for TTS ceiling forecasts. These times were chosen because of the higher frequency of occurrence of low ceiling events during the cool season morning hours (see Figure 2.3). Testing equations built from a subset of the data containing the most events would likely produce the most accurate results and provide the best information from which to choose the appropriate methods to apply to the rest of the data.

3.4.1.1. Statistical Models

The candidate statistical models were multiple linear regression (MLR), as used in VF, and logistic regression (LGR), as used in HF. The equations were to calculate the probability of occurrence of the predictands using binary values for the predictands and most of the predictors. Wilks (1995) identifies two regression methods as viable options. The first is called Regression Estimation of Event Probabilities (REEP). REEP is simply MLR that creates a line defining the relationship between the binary predictors and predictand. It is in the form

$$P = C_o + C_1 x_1 + C_2 x_2 + \ldots C_n x_n,$$

where P is the forecast probability, C_n are the coefficients, and x_n are the predictors. This method produces the smallest error between the observations and the line for the predictions given a set of predictors. It is the least computationally demanding of the two methods, but in theory it can produce impossible probabilities that are less than 0 or greater than 1. A more desirable model for probability forecasts of binary predictands is LGR, which uses a nonlinear equation of the form

$$P = 1 \,/\, (\, 1 + \exp[C_o + C_1 x_1 + C_2 x_2 + \ldots C_n x_n]\,)$$

to calculate a curve. Like REEP, it also minimizes the error between the curve and the observations. Unlike REEP, it cannot calculate probabilities less than 0 or greater than 1. Wilks states that LGR is the theoretically preferred method over REEP when the predictands are binary, such as in this study. Figure 3.1 shows a hypothetical example of the difference between REEP and LGR. The curved fit appears to reduce the error further from that of the linear fit. This indicates that LGR might produce a more accurate forecast than REEP.

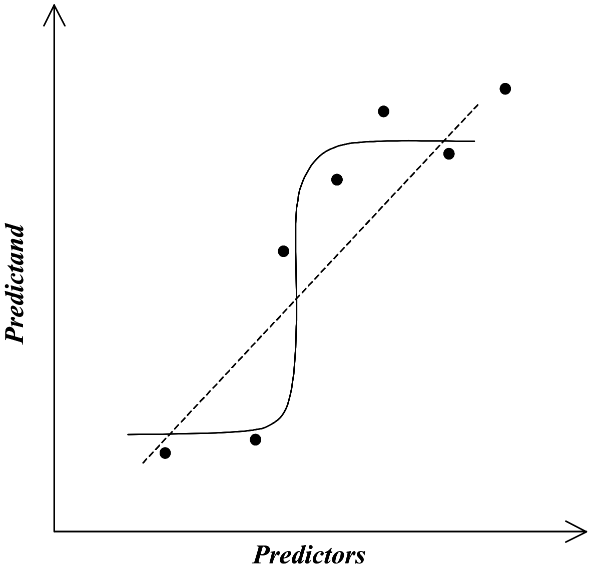


Figure 3.1. A hypothetical example of the difference between Regression Estimation of Event Probabilities (REEP) and logistic regression (LGR). The dashed line represents an estimate of the best fit from the REEP and the curve is an estimate of the fit from the LGR.

3.4.1.2. Predictor Selection

A forward stepwise regression technique was used with the dependent dataset to choose the best predictors for the OBS equations. This procedure started by choosing the predictor that accounted for the largest percentage of variance between the predictor and predictand. In subsequent steps, it tested the remaining predictors and chose them one at a time based on which one accounted for the most variance when combined with the previously chosen predictors. If left unchecked, the stepwise technique would have found a relationship between the predictand and every predictor in the set in an effort to explain as much of the variance as possible. This would have caused two problems. The obvious problem was that it would be cumbersome and impractical to enter 135 values for the predictors in an operational setting. The most critical problem was that, with so many predictors, an excellent fit could be achieved on the dependent data, but the relationship would not be optimal for the independent or any other dataset and would not produce the best-performing forecasts in an operational setting. This is known as over-fitting the equation to the data. Therefore, a balance had to be achieved in which enough physically meaningful predictors were chosen to define a proper relationship between the predictand and predictors, yet not so many that the relationship would only be valid for the dependent dataset.

The S-PLUS® software allows the designation of an F-value (Wilks 1995) to stop adding more predictors in the forward stepwise regression. The F-value is a ratio given by the equation

$$F = MSR/MSE.$$

MSR is the mean of the regression sum of squares given by

$$MSR = \frac{1}{df}\sum_{i=1}^{n}\left(f_i - \overline{o}\right)^2 ,$$

where df is the degrees of freedom, n is the number of observations in the dataset, f_i is the prediction, and $\overline{o}$ is the average of the predictand observations. The MSE is given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(o_i - f_i)^2 ,$$

where o_i is the observation corresponding to the f_i prediction. As predictors are added to the equation, more of the variance is explained causing MSR to increase. The value of df is the difference between n and the number of predictors in the equations. Therefore it decreases as the number of predictors increases. As MSR increases the forecast error decreases causing MSE to decrease. This combination causes F to increase as predictors are added to the equation. The F-value designates the maximum F to be achieved as predictors are added to the equation. If a predictor causes F to exceed the predetermined F-value, it is not added to the equation and the forward stepwise procedure is stopped. Since HF used the same software, their value of 10 for F was used. This value was sufficient to narrow down the list of possible predictors from 135 to at most 20 physically relevant variables per equation that were ranked from the most to the least important in terms of explaining the linear relationship.

A quantitative examination of the predictors chosen in the stepwise procedure indicated the possibility that some of the less relevant predictors chosen later in the algorithm may not have been physically reasonable for the prediction and might not add much, if any, predictive value to the equation. Therefore, another technique to reduce the predictor list further was employed. One of the values output from the regression algorithm was the percentage of the variance in the predictor data accounted for by the regression, R^2, also known as the coefficient of determination. In a perfect prediction, this value would be 100%. Wilks (1995) suggested a practical cutoff would be when none of the remaining predictors increase the R^2 value by more than a specified amount. This method was tested by adding the predictors chosen by stepwise one at a time until the increase in R^2 was less than 1%, 0.5%, and 0.05%.

The graph in Figure 3.2 shows the R^2 calculations for the predictors for the 2-hour forecast valid at 1200 UTC of ceilings < 10 000 feet. The seven predictors along the axis are in the order chosen by the stepwise regression using F = 10 for this particular forecast. The first three letters in each predictor name represent the station from which the observation was taken. The XXX.C.10000 predictors represent the binary observations of ceilings < 10 000 feet at the specified station, DAB.C.5000 is the same for ceilings of < 5000 feet at DAB, TTS.CD1.10000 is the binary predictor for whether the first cloud deck was < 10 000 feet at TTS, and TTS.CV.9.10 is the binary predictor for 9/10 or greater cloud cover at TTS. The left y-axis and the solid line represent the R^2 values as each predictor was added, and the right y-axis and dashed line represents the amount R^2 was increased by adding the associated predictor. This graph shows that if the 1% cutoff was used, only the first 3 predictors would be in the final equation. If 0.5% was used, the last two predictors would have been left out, and if 0.05% was used all predictors would be part of the equation. It should be noted that this is a specific example and not completely representative of all the equations. There were variations in the number of predictors chosen and their associated R^2 values and differences.

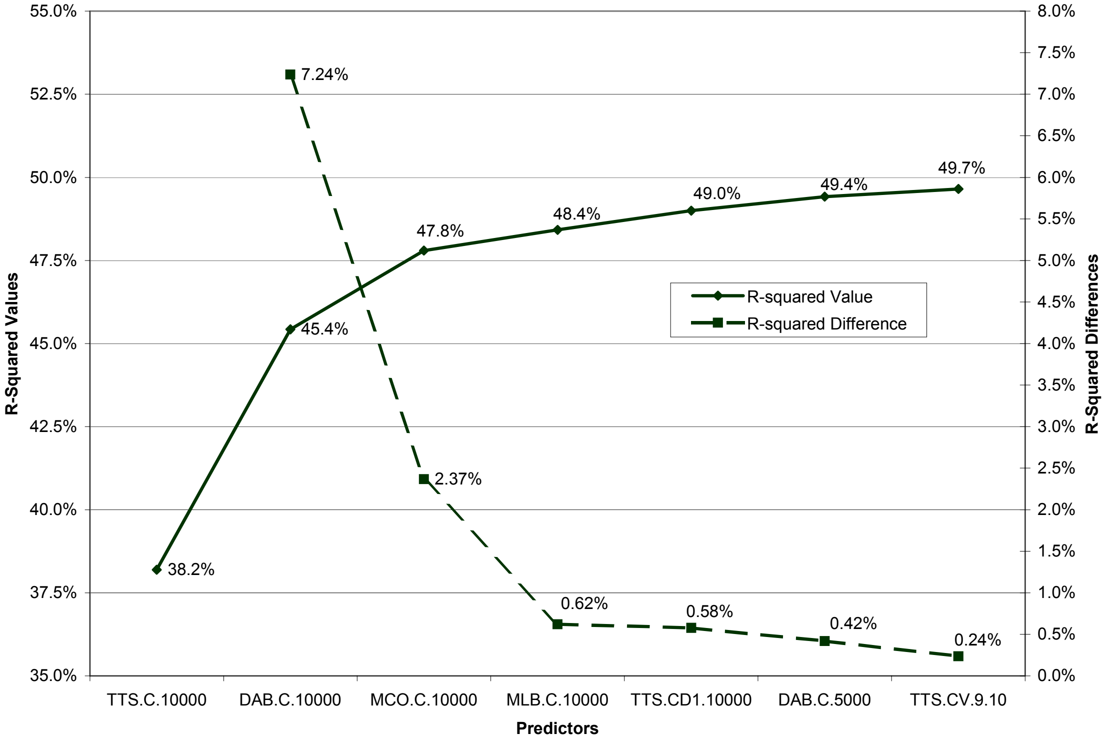


Figure 3.2. R-squared (R^2) values for the 2-hour forecast valid at 1200 UTC for a ceiling of < 10 000 feet at TTS. The predictors along the x-axis are in order according to the amount of variance explained as each predictor was added to the equation. The left y-axis is the value of the variance, R^2, as each predictor is added to the equation, and the right y-axis is the amount of increase in R^2 caused by adding the associated predictor.

3.4.1.3. Test Results

As stated earlier, equations for 1- and 2-hour forecasts valid at 1200 and 1500 UTC for each of the ceiling categories were developed to conduct the tests. The goal was to find the best combination of statistical model and predictor cutoff method in order to decrease the mean square error (MSE) of the forecasts. The MSE is given by the equation

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(o_i - f_i)^2 ,$$

where n is the number of forecasts, o_i is the binary ceiling category observation (0 or 1), and f_i is the probability forecast ($0 \le f \le 1$) given by the equation. If MSE = 0, the forecast is considered “perfect”.

First, F = 10 was used to narrow down the list and rank the predictors from the most to least important, then the three R^2 cutoff criteria were imposed to create three predictor sets. Then, REEP and LGR forecast equations for each of the Shuttle FR ceiling categories were developed using these predictor sets for the OBS method and the PCL method. Table 3.3 shows a summary of the equations that were developed for testing. An N/A appears in the third column for PCL because that method always used the same two predictors: the binary ceiling observation at the initial time and a term for the ceiling climatology at the valid time. Predictions were made with both methods using observations from the dependent, then independent datasets. They were tested on the dependent data first to ensure adequate performance on the dataset from which they were developed. If they performed poorly on the dependent dataset, they would not likely perform well on any data.

Table 3.3. A summary of the parameters in the development of the equations used to test which statistical model and predictor cutoff method were most appropriate for the full set of equations. The total number of OBS and PCL equations developed for the testing are shown in the last column. The number of equations includes those for each of the three Shuttle FR binary ceiling categories.

Forecast Method	*Initial Time: Valid Time (UTC)*	*R^2 Cutoffs (F = 10)*	*Statistical Models*	*Total # Equations*
OBS	1000, 1100: 1200 1300, 1400: 1500	1%, 0.5%, 0.05%	REEP, LGR	72
PCL	1000, 1100: 1200 1300, 1400: 1500	N/A	REEP, LGR	24

After all the predictions were made, the MSE value resulting from each of the equations was calculated. None of the MSEs were 0, but were all on the order of 0.1. The lowest MSE values calculated from forecasts made using the dependent data were from the REEP model using 0.5% and 0.05% as the R^2 cutoffs, but their values were very similar at approximately 0.08. It appeared from this result that the extra predictors allowed by the smaller cutoff did not add any predictive value to the equations. This was confirmed by the results using the independent data. The lowest MSEs from the independent data were from the REEP model using 0.5% as the R^2 cutoff. The extra predictors allowed by the 0.05% cutoff appeared to cause some over-fitting of the dependent data.

The LGR model was expected to produce the lowest MSEs because it is able to fit the predictors to the predictand in a non-linear fashion (Section 3.4.1.1, Figure 3.1). This was not the case in these preliminary tests as the MSEs from the LGR equations were higher in every case when compared to the MSEs from the corresponding REEP equations. Another reason to use LGR was to avoid probability predictions that were <0 or >1, as can be produced by REEP. No such probabilities were produced by the REEP equations in the preliminary testing. Wilks states that REEP rarely produces such erroneous probabilities in an operational setting when there are enough predictors in the forecast equation. Based on these preliminary tests, the model used for the OBS and PCL equations was REEP and the predictors were chosen for the OBS equations using the 0.5% R^2 cutoff.

3.4.2. OBS Equation Development

The development of the 216 OBS equations was done using an automated procedure. Nonetheless, output from the procedure was monitored closely to ensure that reasonable predictors were chosen for each equation. Important characteristics of the equations were:

- The number of predictors in each equation ranged from 1 to 9, with an average of 4 to 5 predictors per equation.
- In general, the number of predictors per equation increased with increasing lead time.
- In 212 of the 216, the predictor that explained most of the variance in the data was the observation of the predictand at the initial time, indicating that persistence was an important predictor. For example, for the 1-hour forecast of TTS ceilings < 8000 feet valid at 1200 UTC, the 1100 UTC observation of TTS ceilings < 8000 feet explained most of the variance in the data. This finding was consistent with VF and HF.
- The binary ceiling observations at DAB were most often chosen as predictors after the observation of the predictand at the initial time, followed by the binary observations of ceiling from the other stations.
- The binary total cloud cover and cloud deck predictors at TTS and the other stations were

next in importance after the ceiling observations.

- Precipitation occurrence, wind direction, and the rest of the continuous predictors were rarely chosen, and were of low importance in the equation when they were.
- The amount of variance explained by the predictors (R^2), was 55-60% for the 1-hour forecasts, 45-50% for the 2-hour forecasts, and 35-40% for the 3-hour forecasts. Within each lead-time category, the equations for ceilings < 5000 feet had the lowest R^2 values, and the equations for ceilings < 10 000 feet had the highest values.
- A large percentage of R^2 was explained by the observation of the predictand at the initial time, with the other predictors only explaining an additional 5-10%, indicating further the importance of persistence.

3.4.3. PCL Equation Development

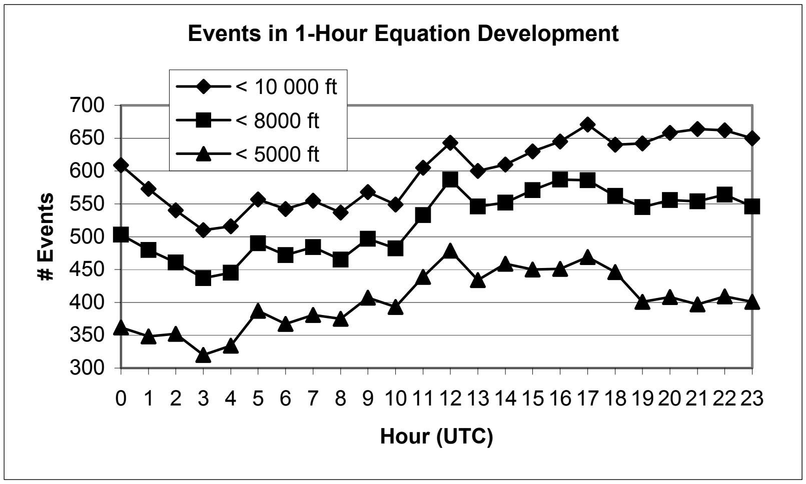
The development of the PCL equations was less complicated than the OBS equations because there were only two predictors: the binary observation of the predictand at the initial time and a climatological term at the valid time. The ERFs were incorporated into the dependent dataset, and PCL equations were developed using the REEP statistical model. Important characteristics of the equations were:

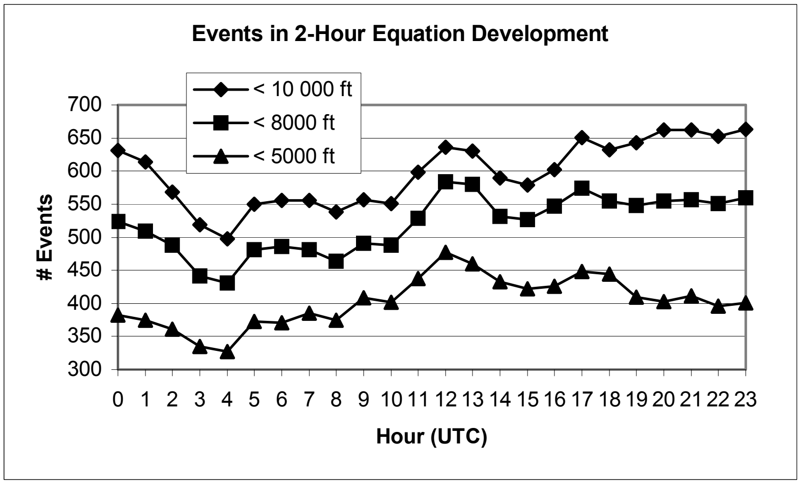
- In all of the 1-hour equations, the predictor that explained most of the variance in the data was the observation of the predictand at the initial time, similar to that of the OBS equations.
- In the 2-hour equations, the ERF value at the valid time became increasingly important and was occasionally chosen as the predictor that explained most of the variance, likely indicating the decreasing contribution of persistence as lead time increases.
- In more than half of the 3-hour equations, the predictor that explained most of the variance in the data was the ERF value, again likely indicating the decreasing contribution of persistence at longer lead times.
- The amount of variance explained by the two predictors (R^2), was 5-10% lower than that of the corresponding OBS equations. This was the first indication that the OBS equations would produce more accurate forecasts since they explained more of the variance.

3.4.4. Number of Events

One other important characteristic of the OBS and PCL equations had to be considered before proceeding to testing with dependent and independent data. The WMO (1992) recommends that at least 250 events of the predictand should be in the dependent dataset in order for stable relationships to be defined by the equations. A predictand event occurs when the binary value is equal to 1. All of the predictand events from the dependent dataset were counted and stratified according to lead time, Shuttle FR, and hour. The results are shown in the graphs of Figure 3.3.

A cursory examination of all three graphs shows that each equation was developed from a dataset that had more than 300 predictand events, over 50 events more than recommended by WMO (1992). The number of events increases with ceiling height threshold. This is because all ceilings below a threshold were used in the equation development. For example, for the FR category < 10 000 feet, all ceilings less than 10 000 feet were used in equation development, even those below 8000 feet and 5000 feet. There is also a temporal trend with a minimum in the late evening/early morning hours (0300 – 0400 UTC), rising gradually to ~1000 UTC, then a large increase to 1200 UTC near sunrise, slight decrease through the morning hours then rising through the afternoon through evening. Regardless of the variations, all equations were developed using a sufficient number of events to establish a stable relationship between the predictand and the predictors.





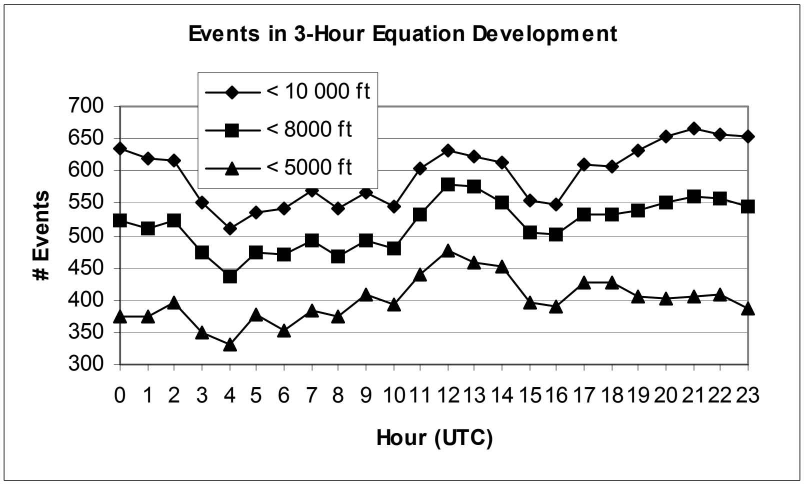


Figure 3.3. The number of predictand events used in the development of the observations-based equations for top) 1-hour forecasts, middle) 2-hour forecasts, and bottom) 3-hour forecasts. Within each graph, values are given for each predictand category and hour.

4. Tests with Dependent and Independent Datasets

After the equations were developed, they were used to make probability forecasts with every record in the dependent and independent datasets. Comparison testing between the OBS and PCL equation forecasts was done to determine whether the OBS method produced more accurate forecasts than the PCL method. The tests with the independent data would most likely reveal the operational performance of the equations, which was the main goal of the task, but tests with the dependent dataset were also done for two reasons. First, if the equations performed poorly on the dataset from which they were developed, they could not be relied upon to produce accurate forecasts with the independent or any other dataset. Second, the dependent test results were compared to those using the independent dataset to ensure comparable equation performance. It was expected that the equations would perform better on the dependent data than on the independent data (Wilks 1995), but a large degradation in performance with the independent data could mean that the equations had been overfit.

4.1. Equation Comparison

The first test done was to determine the relative skill of the OBS equations using the PCL forecasts as the benchmark. The MSE values for the OBS and PCL predictions were calculated using the equation in Section 3.4.1.3 and were used in the following equation:

$$\mathrm{PI} = \frac{(\mathrm{MSE}_{obs} - \mathrm{MSE}_{pcl})}{(\mathrm{MSE}_{perfect} - \mathrm{MSE}_{pcl})} \times 100 ,$$

where PI is the percent improvement of the OBS equations over the PCL equations, MSE_{obs} is the MSE for the OBS equations, MSE_{pcl} is the MSE for the PCL equations, and $\mathrm{MSE}_{perfect}$ is the MSE for a perfect forecast, which is 0. If PI is positive, it indicates an improvement of the OBS method over the PCL method. A negative PI would indicate that the PCL method produces a more accurate forecast. The PI values were calculated for all 216 equations from both methods.

The PI values for both dependent and independent datasets were examined and compared in detail. For brevity, the mean of the PI values from the 24 equations valid from 0000 to 2300 UTC were averaged by lead time, ceiling category, and dataset, and are shown in Table 4.1. The results from using the dependent and independent datasets are side-by-side for easy comparison. The most important result is that all PI values were positive from both datasets, indicating that the OBS method produces improved probability forecasts over the PCL method. Another important result is that the values from using the independent dataset were similar, albeit slightly smaller, than those from using the dependent dataset. This result indicates that the dependent dataset was large enough to develop stable equations and that the equations were not overfit to the dependent data. In general, the PI values increase with increasing lead time. This may indicate the decreasing importance of persistence and an increasing importance of observations from the surrounding stations as the lead time increases. The PI values also decrease with decreasing ceiling height category.

Table 4.1. Average values of the percent improvement (PI) of the OBS method over the PCL method for the 24 equations, valid 0000 to 2300 UTC, by lead time, dataset, and ceiling category.

Lead Time	*< 10 000 feet*		*< 8000 feet*		*< 5000 feet*	
	DEP	**INDEP**	**DEP**	**INDEP**	**DEP**	**INDEP**
1-Hour	12.05	11.89	11.04	10.03	10.15	8.75
2-Hour	15.99	15.27	14.27	13.45	13.31	11.94
3-Hour	15.95	14.88	14.21	13.62	13.75	13.44

Since the results from testing with the independent data would be most indicative of the expected operational performance, they are shown in detail in Figures 4.1 – 4.3. Over 500 observations were available in the independent dataset to calculate the PI values. The hour-to-hour variance seen in Figures 4.1 – 4.3 may be an artifact of the relatively small sample size used in the calculations. The variance in all three traces is smaller for the dependent dataset. This could be due to the fact that the dependent dataset has almost 3000 observations per equation, or due to the better fit achieved on the dependent dataset. As more data are collected over time, this variance may decrease.

The PI values for each of the equations that produced probability forecasts for ceilings < 10 000 feet are shown in Figure 4.1. The most important aspect of this graph, as with Table 4.1, is that all the values were positive, indicating that the OBS equations produced more accurate forecasts than the PCL equations. The values tended to increase with lead time, and the 3-hour lead time appeared to produce a larger variance in skill over 24 hours. No diurnal trend was apparent for any of the equations.

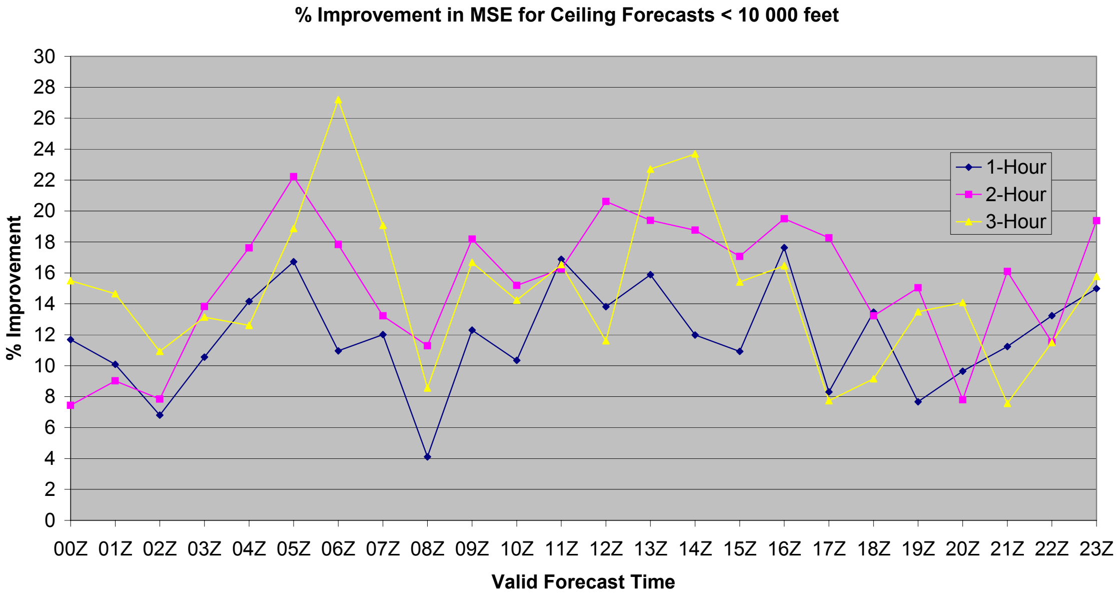


Figure 4.1. Percent improvement in MSE of the OBS equations over the PCL equations for forecasts of ceilings < 10 000 feet at each lead time and hour of the day for the cool season. Each line represents the PI values for the three lead times: blue) 1-hour forecasts, pink) 2-hour forecasts, and yellow) 3-hour forecasts.

Figure 4.2 is analogous to Figure 4.1 except that it shows the PI values for the equations that produced forecasts for ceilings < 8000 feet. Again, all the values were positive, indicating improved performance by the OBS equations, they tended to increase with lead time, and no diurnal trend was apparent.

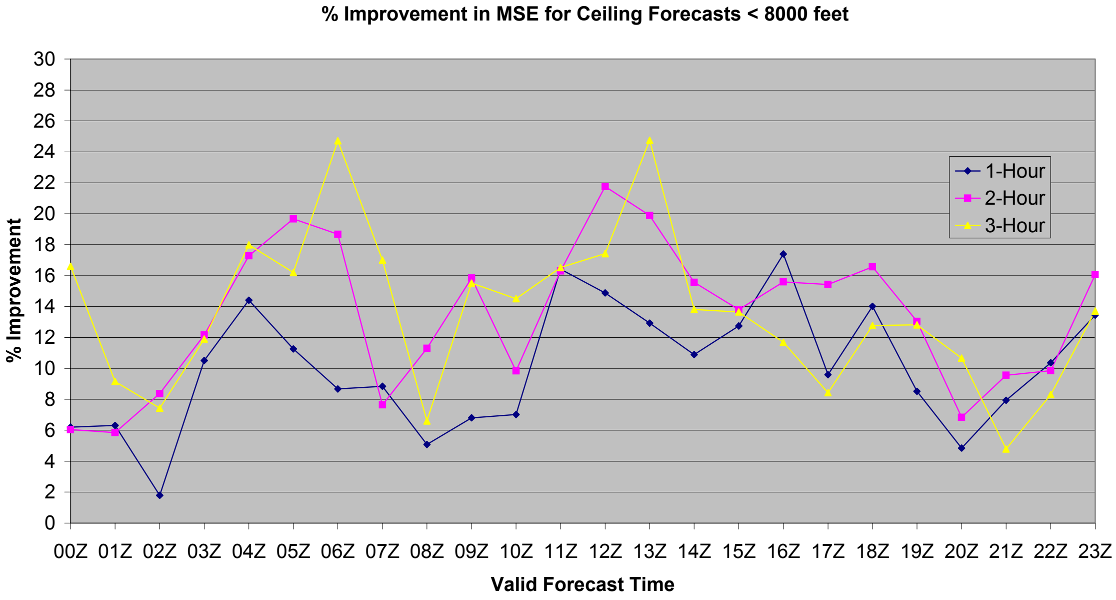


Figure 4.2. Percent improvement in MSE of the OBS equations over the PCL equations for forecasts of ceilings < 8000 feet at each lead time and hour of the day for the cool season. Each line represents the PI values for the three lead times: blue) 1-hour forecasts, pink) 2-hour forecasts, and yellow) 3-hour forecasts.

Figure 4.3 shows the PI values for each of the equations that produced forecasts for ceilings < 5000 feet. All except one of the values were positive for this set of equations, and another was close to 0. The 1-hour OBS equations valid at 0200 and 2000 UTC produced PI values of 0.2 and –0.6, respectively. Although the second was negative, both values were close to 0 indicating that the OBS and PCL equations for those times performed almost equally. Otherwise, the values show traits similar to those of the other two graphs.

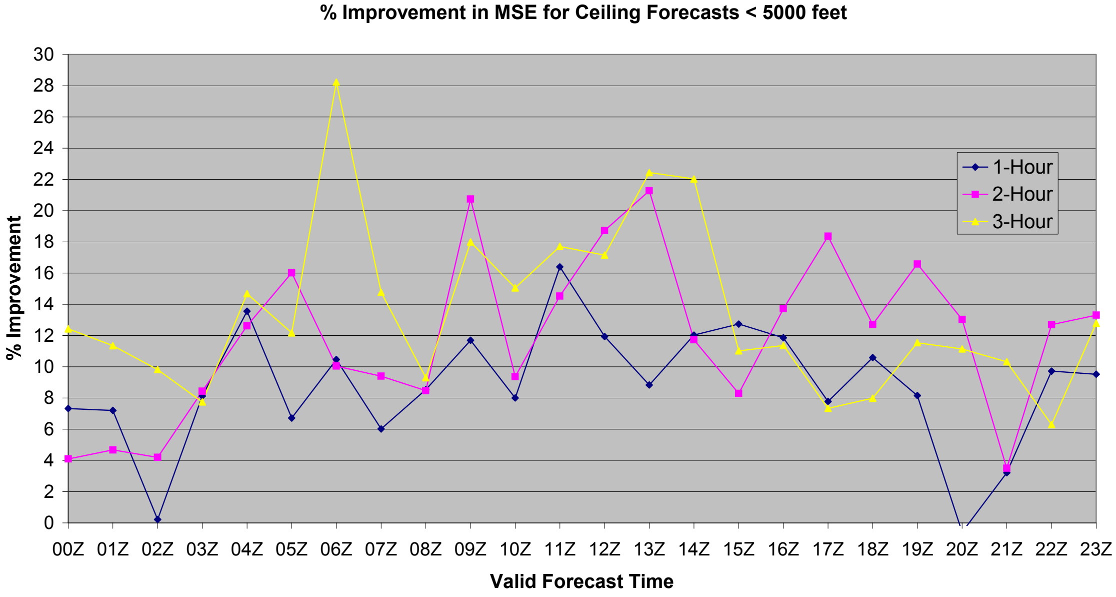


Figure 4.3. Percent improvement in MSE of the OBS equations over the PCL equations for forecasts of ceilings < 5000 feet at each lead time and hour of the day for the cool season. Each line represents the PI values for the three lead times: blue) 1-hour forecasts, pink) 2-hour forecasts, and yellow) 3-hour forecasts.

The corresponding graphs for the dependent data (not shown) show less variability from hour to hour and no negative values or values very close to 0. It is important to note that the MSE and subsequent PI values for the independent dataset were calculated with 3 cool seasons compared to 16 cool seasons in the dependent dataset. As the equations are used in cool-season operations and more data on equation performance is collected, use of a larger dataset in calculating MSEs may result in more smooth graphs of hourly PI values.

4.2. Hypothesis Testing

While it was important to show that the OBS equations produce improved forecasts over that of the PCL equations, it was also important to know whether that improvement was statistically significant. Statistical significance was determined through the use of hypothesis testing. A null hypothesis was defined first, then the testing determined whether that hypothesis could be rejected. For this study, the null hypothesis was that the mean of the differences between the OBS and PCL MSE values was 0, which would indicate that the OBS equation improvement was not significant and that use of the PCL equations would produce forecasts as accurate as the OBS equations.

Hypothesis tests are broadly classified into parametric and nonparametric tests. A parametric test is used when the data used in the test are represented by a known theoretical distribution, such as Gaussian. A nonparametric test does not require the data to be in any particular distribution. In order to choose an appropriate test, the MSE differences were calculated by subtracting PCL MSEs from their corresponding OBS MSEs and the distributions of the differences were examined by lead time and ceiling category. Figure 4.4 shows the distribution of the 24 MSE differences for 1-hour forecasts of ceilings < 5000 feet, and Figure 4.5 shows the distribution of the 24 MSE differences for 2-hour forecasts of ceilings < 8000 feet. Very few of the distributions were theoretical like the Gaussian distribution in Figure 4.4. Distributions that did not bear similarity to any theoretical distribution like that in Figure 4.5 were more the norm.

It is important to note here that most of the differences were negative, indicating that the OBS MSEs were smaller than the PCL MSEs. This re-confirms the previous finding that the OBS equations produce more accurate forecasts than the PCL equations.

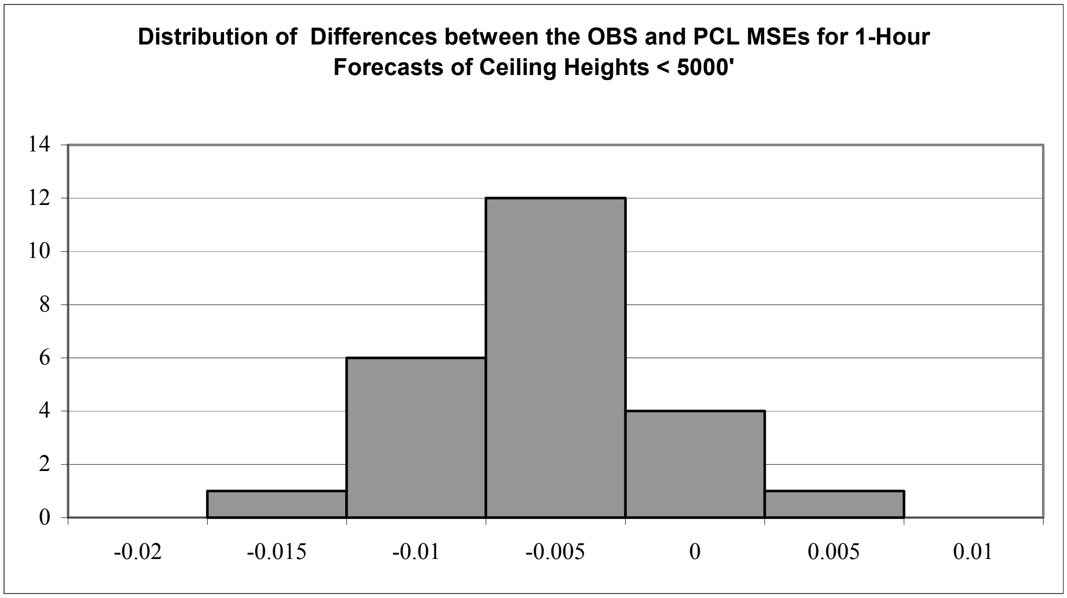


Figure 4.4. Distribution of differences between the OBS and PCL MSE values for the 24 hourly equations for 1-hour forecasts of ceilings < 5000 feet. The value under each bin represents the upper bound of that bin’s range.

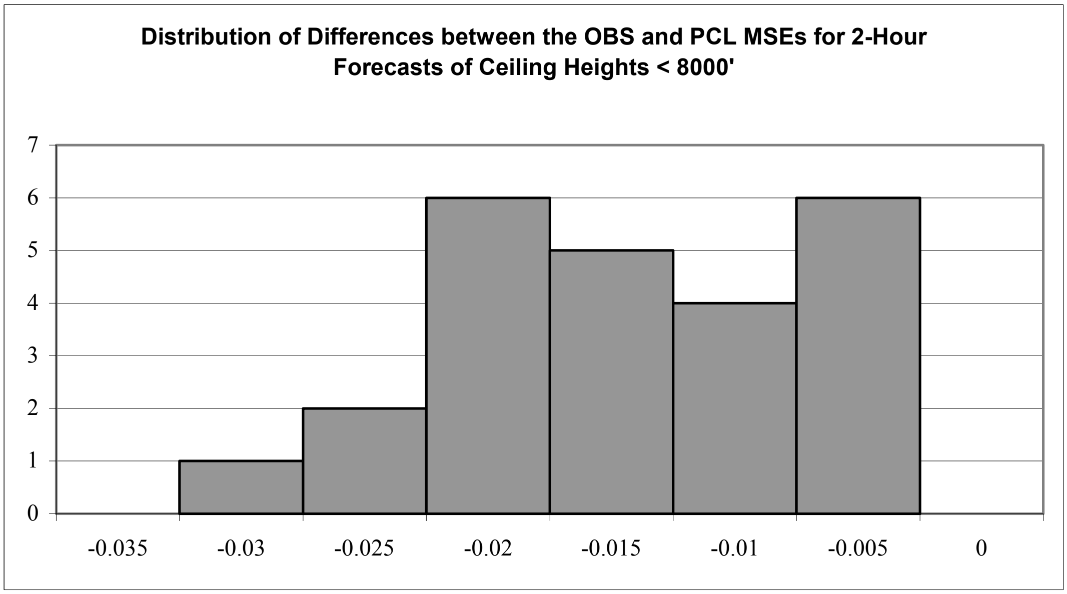


Figure 4.5. Distribution of differences between the OBS and PCL MSE values for the 24 hourly equations for 2-hour forecasts of ceilings < 8000 feet. The value under each bin represents the upper bound of that bin’s range.

Given the inconsistencies in the distributions and their non-similarity to any theoretical distributions, a nonparametric test called the Wilcoxon Signed Rank test (Wilks 1995, Insightful Corp. 1999) was chosen to determine statistical significance. This test calculated the difference between corresponding pairs of OBS and PCL MSE values and used these differences to determine if the mean of the differences was not equal to 0 (Wilks 1995). The critical output value from the test was the p-value. The p-value represents the probability of error involved in accepting that the difference between the two MSEs is valid (Statsoft, Inc. 2001). Put another way, it represents the likelihood that the difference is due to chance (Sheskin 1997). The smaller the p-value, the less likely the difference is due to chance and the more probable that the difference is significant. The common convention is to use a p-value of 0.05 as the threshold value to accept or reject the null hypothesis. This is interpreted as having 95% confidence that the mean of the differences is not equal to 0. One-tailed tests were performed that would determine whether or not the OBS MSEs were significantly less than the PCL MSEs. A two-tailed test would only determine if the MSEs were significantly different ignoring whether one set of MSEs were greater or lesser than the other. Since it was important to show that the OBS MSEs were less than the PCL MSEs, this one-tailed test was employed.

A total of nine tests were conducted using the MSE values generated from forecasts with the independent dataset. The corresponding OBS and PCL MSE values from the 24 valid times in each lead time/ceiling category group were input to the algorithm. All nine p-values were only slightly larger than 0, on the order of between 1e-7 to 1e-8. The smallness of the p-value suggested that the null hypothesis, which was that the mean of the differences between the OBS and PCL MSEs was 0, could be rejected with greater than 99% confidence. The alternative hypothesis, that the OBS MSEs were significantly less than the PCL MSEs, was accepted with the same confidence. Therefore, the improvement in forecast accuracy by the OBS equations over the PCL equations was statistically significant.

4.3. Equation Performance

The information in Sections 4.1 and 4.2 showed that the OBS equations produced an improvement over the PCL method and that the improvement was statistically significant. This answered the important question of whether the addition of data from surrounding stations in the development of the equations improves the short-term ceiling forecast at TTS. However, PI is a relative measure of performance and does not indicate the absolute performance of the OBS equations. If the PCL method produced highly inaccurate forecasts to begin with, the OBS equation improvement may not be meaningful. Therefore, another important question to answer was whether the OBS equations actually produced useful forecasts.

Two scores were used to help determine the absolute performance of the OBS equations. The first was the probability of detection (POD), which is the fraction of times an event occurred when it was forecast to occur. The second was the false alarm rate (FAR), which is the fraction of times an event did not occur when it was forecast to occur. The POD and FAR for perfect forecasts is 1 and 0, respectively, and were calculated using the equations in Figure 4.6. They were computed using the values from a standard contingency table, as shown in Figure 4.6 (Wilks 1995). The observations were binary, 1 for Yes and 0 for No and could be incorporated easily into the table. However, the forecasts were probability values between 0 and 1, inclusive, not binary like the observations. In order to input the forecasts into the contingency table, a forecast probability value of 0.5 was chosen as the boundary between Yes and No forecasts. Values ≥ 0.5 were considered Yes forecasts and those < 0.5 were No forecasts. The contingency table was filled out and the POD and FAR calculated for all 216 OBS forecasts created with the dependent and independent datasets.

Forecast	*Observed* Yes	*Observed* No
Yes	a	b
No	c	d

$$\text{POD} = \frac{a}{a+c}$$

$$\text{FAR} = \frac{b}{a+b}$$

Figure 4.6. The standard 2 x 2 contingency table used in calculating several measures of the accuracy of binary forecasts. Each letter represents the number of times a) an event was forecast and observed, b) an event was forecast but not observed, c) an event was not forecast but observed, and d) an event was not forecast and not observed. The equations for the two accuracy measures, probability of detection (POD) and false alarm rate (FAR), are shown.

The POD and FAR values for the 24 equations in each lead time/ceiling category were averaged (as in Table 4.1) and are shown in Table 4.2. More accurate forecasts would maximize POD and minimize FAR. In general, the POD values were much larger than the FAR values, which was consistent with good performance. The values for the dependent and independent dataset were similar, although slightly better for the independent data. This re-confirms that the equations were not overfit to the dependent dataset from which they were developed. The POD values decreased rapidly with lead time in each ceiling category and dataset and decreased slightly with decreasing ceiling height category value. Conversely, the FAR values increased rapidly with lead time in each ceiling category and dataset, and increased with decreasing ceiling height category. More specifically, the 1-hour equations produced PODs above 0.8 and FARs below 0.2 using the independent data, which indicated that equations with this lead time were relatively accurate. The values for the other two lead times deteriorated substantially. The POD values were smaller and the FAR values were larger.

Table 4.2. The POD and FAR scores of the OBS forecasts using the dependent and independent datasets. The values are averaged for the 24 valid times in each ceiling height/lead time category.

Lead Time by Score	*< 10 000 feet*		*< 8000 feet*		*< 5000 feet*	
	DEP	INDEP	DEP	INDEP	DEP	INDEP
POD						
1-Hour	0.81	0.83	0.80	0.83	0.76	0.80
2-Hour	0.69	0.73	0.66	0.70	0.61	0.65
3-Hour	0.62	0.67	0.57	0.63	0.49	0.54
FAR						
1-Hour	0.18	0.16	0.20	0.17	0.23	0.18
2-Hour	0.23	0.21	0.25	0.23	0.28	0.24
3-Hour	0.27	0.25	0.29	0.27	0.30	0.27

The all-around degradation in performance with increasing lead time and decreasing ceiling height may be caused by other meteorological phenomena not contained in the datasets. As lead time increases, persistence becomes less important and it appears that the observations from the surrounding stations did not completely make up the difference in performance. The degradation in performance with decreasing ceiling height was very small but noticeable. It is possible that processes causing lower ceiling heights were not accounted for in the surface METAR data, such as an elevated cool and/or moist layer that could only be sensed by a rawinsonde.

4.4. Probability Cutoff for Yes/No Forecasts

The use of 0.5 as the probability cutoff between Yes/No forecasts produced POD and FAR scores that indicated good forecast accuracy. The appropriate value for operations, however, would depend on whether the user wished to maximize POD or minimize FAR. To assist in this decision, contingency tables were filled out for 11 cutoff values from 0 to 1 in 0.1 increments, inclusive. The POD and FAR scores were calculated from these tables and plotted in Figure 4.7. These were calculated from forecasts using the independent dataset only since equation performance between the two datasets was similar. The values plotted are averaged over valid time and ceiling height category, resulting in a mean value for each lead time. All POD and FAR values decreased with increasing probability cutoff value. As stated earlier, the 1-hour forecasts produced the highest POD and lowest FAR values. The POD values decreased slowly from 1 at the 0.0 probability cutoff to 0.83 at the 0.4 cutoff, remained constant through 0.5 and dropped rapidly after the 0.6 cutoff. The FAR values also decreased, but did so more rapidly from the 0.0 probability cutoff through the 0.4 cutoff. They remained somewhat constant through the 0.6 cutoff and decreased slowly thereafter. The values for the other two lead time categories decreased similarly, although in a more linear fashion.

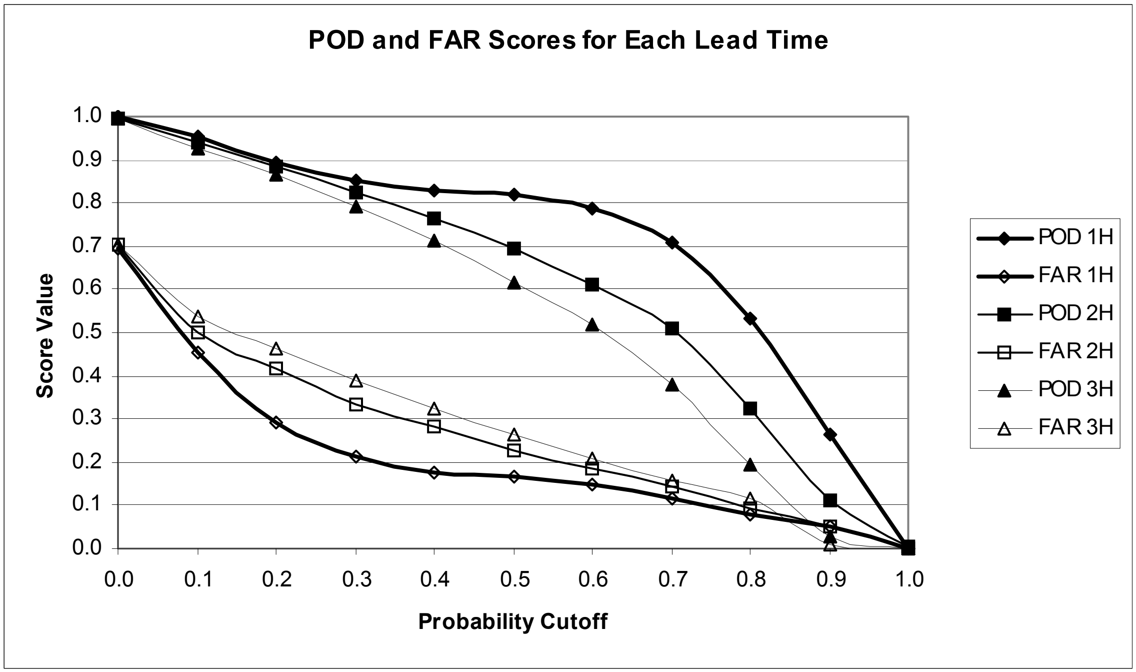


Figure 4.7. The average POD and FAR values for the three lead times at each probability cutoff using the independent dataset. The bold lines represent the 1-hour forecast values, the medium-weight lines represent the 2-hour forecast values, and the light-weight lines represent the 3-hour forecast values. The POD values are denoted by filled markers and the FAR values are denoted by unfilled markers.

While the POD and FAR values in Figure 4.7 were useful in determining equation performance for different probability cutoff values, other measures of accuracy and skill were needed to determine an appropriate cutoff value or range of values. Therefore, four other measures were calculated using the contingency table described in Figure 4.6. Wilks (1995) recommended using the threat score (TS) and bias (B) ratio as good indicators of the appropriate cutoff value. The TS is the correct number of Yes forecasts divided by the total number of times the event was either forecast and/or observed, which is the percent of correct Yes forecasts. The best TS is 1 and the worst is 0. The B ratio is the ratio of the number of Yes forecasts to the number of Yes observations. An unbiased forecast has B=1. If B > 1 the ceiling category was forecast more often than observed, which is overforecasting. If B < 1, the ceiling category was forecast less often than observed, which is underforecasting. When choosing a probability cutoff value, Wilks (1995) suggests choosing the probability where TS is maximized and/or where B=1. These two scores are calculated as (see Figure 4.6 for the definitions of a, b, and c):

$$\mathrm{TS} = \frac{a}{a+b+c}, \text{ and } \mathrm{B} = \frac{a+b}{a+c}.$$

The other two measures employed to help determine the best probability cutoff were the Hit Rate (HR) and Kuipers Skill Score (KSS) (Wilks 1995). They were calculated to help narrow down or confirm the probability cutoff values chosen by the TS and B scores. Where the TS is the percent of correct Yes forecasts, the HR is the percent of all correct forecasts, Yes and No. The best HR is 1 and the worst is 0. The KSS indicates whether the forecast is better than, equal to, or worse than random forecasting. A perfect forecast is indicated by KSS = 1, a forecast similar to a random forecast is indicated by KSS = 0, and KSS < 0 indicates a forecast that is worse than a random forecast. These two scores are calculated as (see Figure 4.6 for the definitions of a, b, c, and d):

$$\mathrm{HR} = \frac{a+d}{a+b+c+d}, \text{ and } \mathrm{KSS} = \frac{ad-bc}{(a+c)(b+d)}.$$

Figures 4.8 – 4.10 show the values of these scores for each lead time using the independent dataset. The values at each valid time in each ceiling category were similar, but were not similar among lead times. Therefore, the values at each valid time in each ceiling category were averaged, resulting in a mean value for each lead time. The gridline at Score Value = 1 is bold in each chart to emphasize the location where B = 1. To help analyze the Wilks (1995) method of determining the probability cutoff, the B and TS curves are bold.

Figure 4.8 shows the average values of the four measures for the 1-hour lead time. The TS was maximized between the probability values of 0.4 and 0.5, and B =1 at the same location. The value of TS at the maximum was 0.7, indicating 70% correct Yes forecasts when using a probability cutoff between 0.4 and 0.5. The co-location of B = 1 and the TS maximum indicated that using a probability value between 0.4 and 0.5 as the cutoff would produce the most accurate forecasts. This was confirmed by the HR and KSS scores. The HR maximum was 0.9 at the 0.5 probability value indicating that 90% of the forecasts were correct, and the KSS maximum was 0.76 at the 0.4 probability value. A positive KSS value close to 1 indicated that these forecasts were much more accurate than random forecasts.

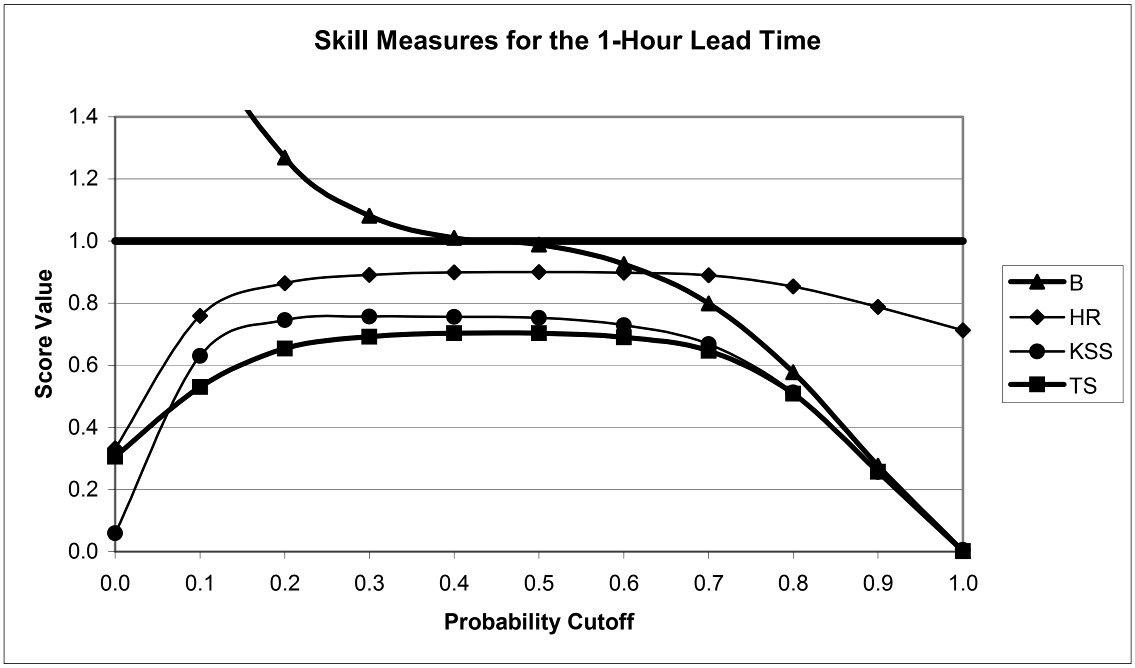


Figure 4.8. The B, HR, KSS, and TS values for the 1-hour lead time. The gridline for Score Value = 1 is bold to emphasize the location where B = 1, and the B and TS curves are bold since they are used to determine the appropriate cutoff value according to Wilks (1995). The scale of the vertical axis was truncated to emphasize the smaller values of the other skill and accuracy measures. The maximum B value was 3.4 at probability cutoff 0.0.

Figure 4.9 shows the average values of the four measures for the 2-hour lead time. The TS was maximized at 0.4 and B = 1 between 0.4 and 0.5, but closer to 0.4. The value of TS at the maximum was 0.58, indicating 58% correct Yes forecasts, less than that for the 1-hour forecasts. The HR maximum was 0.85 at the 0.5 probability value indicating that 85% of the forecasts were correct, and the KSS maximum was 0.65 at the 0.3 probability value. The positive KSS value indicated that these forecasts were much more accurate than random forecasts. The best probability cutoff value appeared to be less for the 2-hour lead time than the 1-hour lead time, likely close to 0.4.

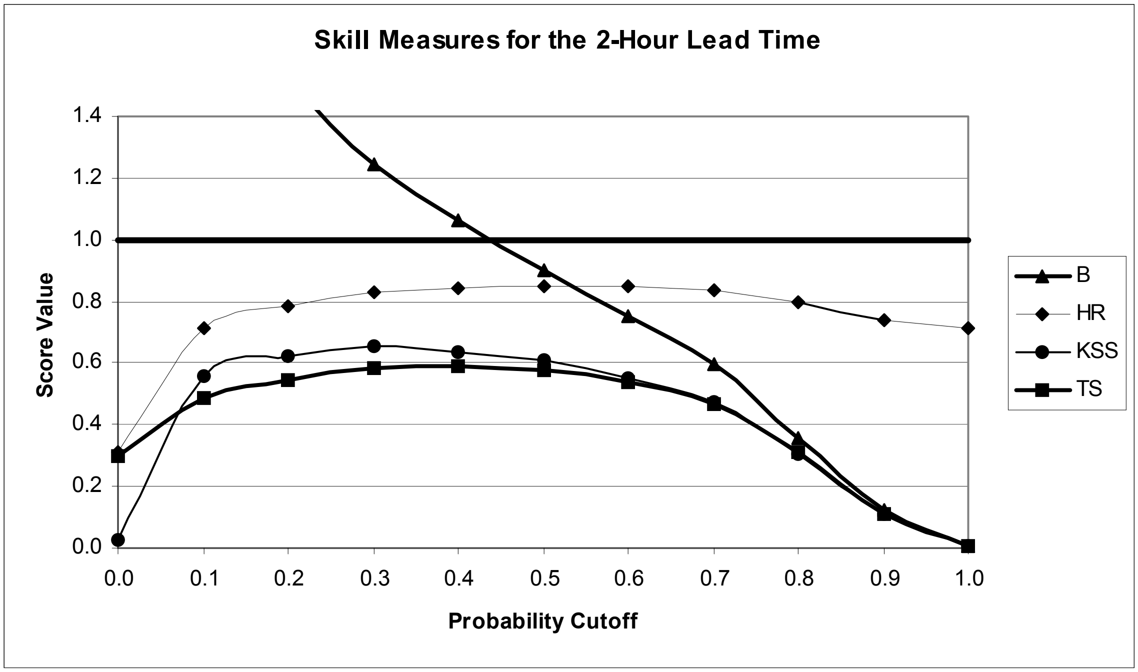


Figure 4.9. The B, HR, KSS, and TS values for the 2-hour lead time. The gridline at Score Value = 1 is bold to emphasize the location where B = 1, and the B and TS curves are bold since they are used to determine the appropriate cutoff value according to Wilks (1995). The scale of the vertical axis was truncated to emphasize the smaller values of the other skill and accuracy measures. The maximum B value was 3.5 at probability cutoff 0.0.

Figure 4.10 shows the average values of the four measures for the 3-hour lead time. As for the 2-hour forecasts, the TS was maximized at 0.4 and B = 1 between 0.4 and 0.5, but closer to 0.4. The value of TS at the maximum was 0.53, indicating 53% correct Yes forecasts, less than that for the 1- and 2-hour forecasts. The HR maximum was 0.82 at the 0.5 probability value indicating that 82% of the forecasts were correct, and the KSS maximum was 0.58 at the 0.3 probability value. The positive KSS value indicated that these forecasts were more accurate than random forecasts. The best probability cutoff value appeared to be similar to, although slightly smaller than, the 2-hour lead time at 0.4.

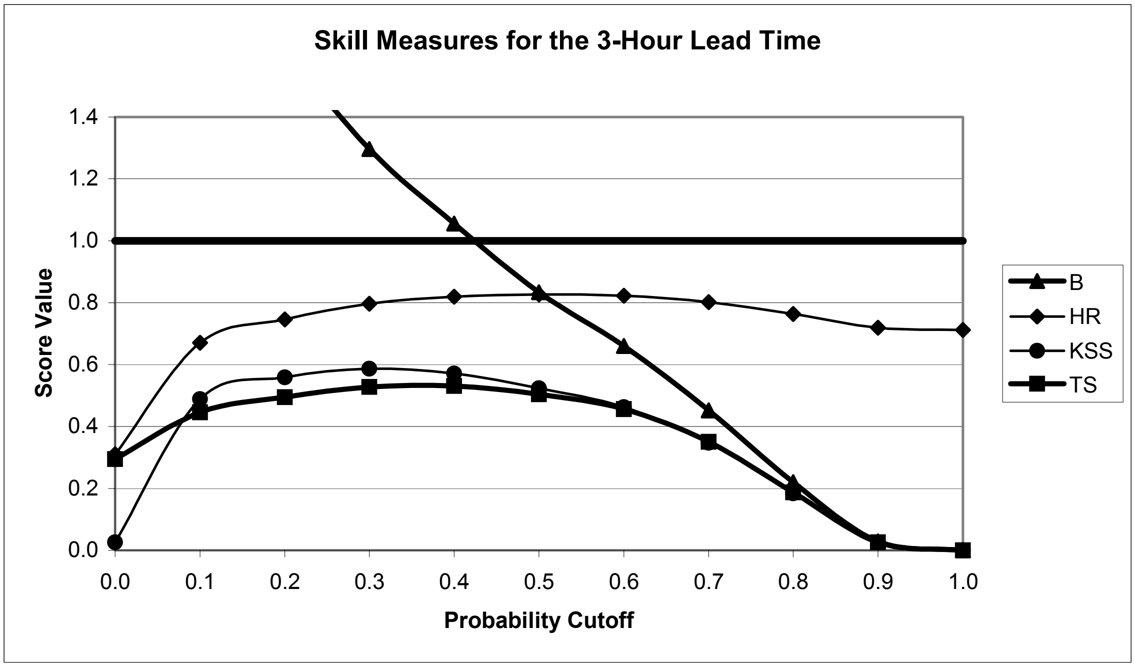


Figure 4.10. The B, HR, KSS, and TS values for the 3-hour lead time. The gridline at Score Value = 1 is bold to emphasize the location where B = 1, and the B and TS curves are bold since they are used to determine the appropriate cutoff value according to Wilks (1995). The scale of the vertical axis was truncated to emphasize the smaller values of the other skill and accuracy measures. The maximum B value was 3.5 at probability cutoff 0.0.

An analysis of the values in Figures 4.8 – 4.10 indicated that the appropriate probability cutoff values decreased with lead time, but tended to be between 0.4 and 0.5. Since these values were calculated from forecasts using the independent data, they are good indicator of the appropriate probability cutoff values to be used in operations. Ultimately, operational use of the equations will determine the most appropriate cutoff values, but these tests provide a starting point.

5. Conclusions

Following procedures outlined in the literature, the AMU developed observations-based (OBS) short-term statistical forecast equations for cloud ceiling at the SLF that outperform persistence climatology (PCL). Hourly surface observations from TTS and surrounding stations during the cool season in east-central Florida were used to develop and test the equations. There were 16 cool seasons in the dependent dataset from which the equations were developed, and 3 cool seasons in the independent data set on which the equations were tested. The equations calculated probability forecasts for the three ceiling categories in the Shuttle FR at 1-, 2-, and 3-hour lead times for each hour of the day in the cool season (October – March). The same equations are valid for every day of the cool season, e.g. the same equation used to make a 1-hour forecast of ceilings < 8000 feet valid at 1200 UTC is applicable to every day in the cool season.

5.1. Test results

Four tests were conducted to determine OBS equation performance. First, the OBS equation MSE values were compared to PCL equation MSE values to quantify a percent improvement or degradation in performance realized by the OBS equations over the PCL equations. Second, hypothesis testing was conducted to determine the statistical significance of the difference in performance. The actual performance of the OBS equations was determined by calculating the PODs and FARs for each lead time/ceiling height combination. These tests were conducted using the records in both the dependent and independent datasets. Finally, the TS, B, HR, and KSS scores were calculated to determine the most appropriate probability cutoff value to use in operations when determining whether a Yes or No forecast should be issued. These are the major conclusions from the tests:

- The comparison tests revealed that the OBS equations produced an improvement in the probability forecasts over the PCL forecasts from 9% to 15% using the independent data. These values were only slightly less than found from using the dependent data, indicating the equations were not overfit. They are also consistent with the PI values found in the literature.
- The PI values are smaller for the 1-hour lead times and increase through the 2- and 3-hour lead times. They also decrease with decreasing height category.
- The hypothesis testing showed that the improvement realized by the OBS equations over the PCL equations was statistically significant beyond the 99% confidence level.
- The POD values were significantly higher than the FAR values for the OBS independent data forecasts, indicating favorable equation performance. The highest PODs and lowest FARs were generated by the 1-hour forecasts. The performance degraded appreciably with lead time, but the PODs still remained higher than the FARs.
- The TS, B, HR, and KSS values calculated to determine the appropriate probability cutoff all indicated that the equations predicted a large percentage of correct forecasts (TS and HR), that they were unbiased (B = 1 at specific probability values), and they produce forecasts superior to that of random forecasts (KSS > 0).

These major conclusions led to the final conclusion that the OBS equations performed well on the independent dataset, indicating they will likely perform well in operations and produce more accurate forecasts than the PCL method.

5.2. Issues with Test Results

The positive result described in the previous section must be tempered with some peripheral findings during the equation development. While analyzing R^2 values to determine an appropriate predictor cutoff threshold (Section 3.4.1.2), it was noted that the predictors accounted for 55-60% of the variance in the 1-hour equations, 45-55% in the 2-hour equations, and 35-40% in the 3-hour equations. The lowest R^2 values in each lead time category were associated with the forecasts for ceilings < 5000 feet, and the highest values were associated with forecasts for ceilings < 10 000 feet. According to VF, their predictors accounted for 85-90% of the variance in most of their equations, regardless of lead time or ceiling height category. The difference in R^2 values indicated the possibility that the OBS equations in this study were not developed with sufficient data to capture the phenomena necessary for low ceiling formation. There are three possible explanations for the 40-65% unaccounted variance.

First, only hourly surface observations were used to develop the equations. While this data type worked well for VF and HF, it may not be enough to predict ceilings in the sub-tropical environment of east-central Florida. Rawinsonde data were shown to improve the forecast in HF. Although the improvement was small and only for the hours immediately after the sounding was taken, the use of upper-air data in the equations may explain more of the variance. Other upper-air data such as that from satellite, radar, 50- and 915 MHz profilers, or model output may also provide predictors that would be helpful in explaining part of the variance.

The second issue is that the surface observation data were grouped into a cool season dataset, stratified only by time of day. This means that the same equation used to make a 1-hour forecast of ceilings < 8000 ft at 1500 UTC will be applied every day from the beginning of October to the end of March. This stratification was necessary to ensure that the number of ceiling events in each Shuttle FR category was large enough to develop robust equations valid for each hour of the day. In the period from October to March, several meteorological phenomena could be responsible for the development of ceilings in east-central Florida. They can range in size from a synoptic cold front to the local sea breeze, and from tropical to extra-tropical in nature, such as a passing tropical disturbance or a continental air mass. Most of the low ceiling cases in the cool season likely involve events such as fog and frontal systems. A phenomenological stratification of the data would be time-consuming and more data would have to be collected to ensure a sufficient number of events were available for the development of stable equations, but it may be useful in developing more accurate forecast equations.

Finally, there was also a question of which statistical model was most appropriate for the equations. Although the tests with the data indicated otherwise (Section 3.4.1.1), HF and Wilks state that LGR is a more appropriate model than REEP to apply in situations where a probability forecast is to be created from binary predictands and predictors. Other possible models that were not tested include multiple discriminant analysis, decision trees, and neural networks. A fuzzy logic system to forecast ceilings was developed by Hicks (1997) for the aviation community and has produced large improvements over persistence. It is possible that REEP is not the most appropriate model and that some other model may help explain more of the variance.

5.3. Other Efforts

The issue of accurate cloud ceiling forecasts is also highly important to the aviation community. Delays to passenger and freight air flights are costly, and many of the delays are caused by low ceilings and visibilities (Wilson and Clark 2000; Gurka and Mosher 2000). Yet, forecasters for aviation still find that the prediction of FAA FR ceiling categories is a challenging task. Several studies funded by the FAA, the Department of Defense, and other groups are underway to help improve this forecast. Wilson and Clark (2000) describes a project at the Massachusetts Institute of Technology, Lincoln Laboratory that assimilates hourly surface observation, rawinsonde, satellite, sodar, and pyranometer data into algorithms that forecast the time of marine stratus burn-off at SFO. Petty et. al. (2000) describes another project at the National Center for Atmospheric Research (NCAR) in which a prototype fuzzy logic method utilizes surface observations, satellite data, numerical model output, climatology, pilot reports, and radar data to make short-term ceiling and visibility forecasts. Geiszler et. al. (2000) showed preliminary results of comparisons between the Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS) predictions of ceiling and visibility to those from two NCAR algorithms that use Penn State/NCAR 5th Generation Mesoscale Model (MM5) output. They found large differences between the forecasts and the observations and advocated improving the NCAR algorithms and the COAMPS model in order to create reliable forecasts of these phenomena.

It is clear that the issue of ceiling and visibility forecasting is receiving much attention by atmospheric researchers. The previously mentioned studies involved participation by several scientists and cooperation between agencies. Several data types were used in addition to hourly surface observations as input to algorithms and numerical weather prediction models. The use of several data types in the algorithms developed in these studies may help account for the missing variance found in the OBS equations developed by the AMU. However, all of the studies mentioned are ongoing and it is unknown when a finished product, if any, will be available. Although they are mainly concerned with forecasting the FAA FR, results from these studies should be monitored closely to determine if any of the methods could be applied to Shuttle FR ceiling forecasts.

5.4. Operational Use

There are 216 OBS equations and 216 PCL equations valid every hour of the day for each of 9 ceiling height/lead time categories. These equations were developed specifically for use only for days in the cool season from October to March and only at the SLF. All of the equations require input that is readily available from the standard hourly surface observation METAR code, except for the ERF in the PCL equations. The ERF values will be provided to customers interested in using the PCL equations.

These equations were developed specifically for use by SMG when making cool-season ceiling forecasts for Shuttle landings at the SLF. However, the 45 WS may also have a need for some of these forecast equations. The 45 WS launch weather officers provide estimations and forecasts of ceiling to the Eastern Range (ER) Safety group. ER Safety is required to visually track the solid rocket boosters through 8000 feet using a combination of ground and airborne observers. Since one of the ceiling height categories is for ceilings < 8000 feet, the 45 WS may be able to use these equations in helping to make ceiling forecasts near the Shuttle launch complexes. Caution should be exercised if the equations are used for this purpose since the equations were developed for the SLF and, as stated earlier, ceilings over the entire KSC/CCAFS area can be discontinuous.

Even use of the equations as intended must be done with care. The fact that they do not account for all phenomena that create low ceilings may cause them to calculate incorrect probabilities at times. Nonetheless, the OBS equations developed in this study are still useful in that they are an improvement over the PCL method. They provide the forecasters at SMG another tool with which to make the ceiling forecasts critical to safe Shuttle landings at KSC. Combined with other observational and model data, as well as forecaster experience, these equations will likely help to improve the ceiling forecasts at the SLF.

Acknowledgements

Several individuals contributed to the success of this task. Mr. William Roeder of the 45 WS contributed a great deal of knowledge on statistical methods and forecast verification through numerous discussions with the author. Mr. John Shuford of ENSCO, Inc. also provided detailed information on data stratification, statistical models, and hypothesis testing. Dr. Robert Vislocky and Mr. Joby Hilliker, the primary authors of the articles that provided the basis for this task, helped explain some of the exact procedures to be followed in doing the work. Finally, innumerable interactions between AMU team members and the customers helped focus the work.

References

Geiszler, D. A., J. Cook, P. Tag, W. Thompson, R. Bankert, and J. Schmidt, 2000: Evaluation of ceiling and visibility prediction: Preliminary results over California using the Navy's Coupled Ocean / Atmosphere Mesoscale Prediction System (COAMPS). Preprints, *9th Conference on Aviation, Range, and Aerospace Meteorology*, Orlando, FL, 11-15 September, Amer. Meteor. Soc., 334-338.

Gurka, J. J., and F. R. Mosher, 2000: Steps to improve ceiling and visibility forecasts for aviation. Preprints, *9th Conference on Aviation, Range, and Aerospace Meteorology*, Orlando, FL, 11-15 September, Amer. Meteor. Soc., 327-330.

Hicks, T., 1997: FuzzyMOS: A fuzzy logic system for objective aviation forecasting. NOAA Tech. Memo. NWS SR-194, 20 pp. [Available from the U.S. Dep. of Commerce, Washington D.C. 20235.]

Hilliker, J. L., and J. M. Fritsch, 1999: An observations-based statistical system for warm-season hourly probabilistic forecasts of low ceiling at the San Francisco International Airport. *J. Appl. Meteor.*, **38**, 1692-1705.

Insightful Corporation, 1999: *S-PLUS® 2000 User's Guide*, Insightful Corp., Seattle, WA, 558 pp.

NASA/JSC, 1997a: NASA Operational Shuttle Flight Rules (NSTS 12820), Final June 6, 1996, PCN-11 December 7, 2000, Vol A, Section 2.1.1-6. NASA/Johnson Space Center, 2-11 – 2-33. [Available from JSC/DA8, Houston, TX 77058.]

Petty, K. R., A. B. Carmichael, G. M. Wiener, M. A. Petty, and M. N. Limber, 2000: A fuzzy logic system for the analysis and prediction of cloud ceiling and visibility. Preprints, *9th Conference on Aviation, Range, and Aerospace Meteorology*, Orlando, FL, 11-15 September, Amer. Meteor. Soc., 331-333.

Sheskin, D. J., 1997: *The Handbook of Parametric and Nonparametric Statistical Procedures*. CRC Press LLC, Boca Raton, FL, 719 pp.

StatSoft, Inc., 2001: *Electronic Statistics Textbook.* http://www.statsoft.com/textbook/stathome.html, StatSoft, Tulsa, OK.

Vislocky, R. L., and J. M. Fritsch, 1997: An automated, observations-based system for short-term prediction of ceiling and visibility. *Wea. Forecasting*, **12**, 31-43.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*. Academic Press, Inc., San Diego, CA, 467 pp.

Wilson, F. W., and D. A. Clark, 2000: Forecast aids to lessen the impact of marine stratus on San Francisco International Airport. Preprints, *9th Conference on Aviation, Range, and Aerospace Meteorology*, Orlando, FL, 11-15 September, Amer. Meteor. Soc., 317-322.

World Meteorological Association (WMO), 1992: *Methods of Interpreting Numerical Weather Prediction Output for Aeronautical Meteorology.* Technical Note No. 195, ISBN 92-63-10770-X, 89 pp.

List of Acronyms

45 WS	45th Weather Squadron
AFCCC	Air Force Combat Climatology Center
AMU	Applied Meteorology Unit
B	Bias Ratio
CCAFS	Cape Canaveral Air Force Station
CMAN	Coastal Marine Automated Network
COAMPS	Coupled Ocean/Atmosphere Mesoscale Prediction System
COF	Patrick Air Force Base, FL 3-letter identifier
DAB	Daytona Beach, FL 3-letter identifier
EDA	Exploratory Data Analysis
EOM	End of Mission
ER	Eastern Range
ERF	Event Relative Frequency
FAA	Federal Aviation Administration
FAR	False Alarm Rate
FR	Flight Rules
HF	Hilliker and Fritsch 1999
HR	Hit Rate
JAX	Jacksonville, FL 3-letter identifier
KSC	Kennedy Space Center
KSS	Kuipers Skill Score
LGR	Logistic Regression
MCO	Orlando, FL 3-letter identifier
MLB	Melbourne, FL 3-letter identifier
MLR	Multiple Linear Regression
MM5	Penn State/NCAR 5th Generation Mesoscale Model
MOS	Model Output Statistics
MSE	Mean Squared Error
NCAR	National Center for Atmospheric Research
NDBC	National Data Buoy Center
NWS	National Weather Service
OBS	Observations-based forecast method
PBI	West Palm Beach, FL 3-letter identifier
PCL	Persistence Climatology forecast method
PI	Percent Improvement
POD	Probability of Detection
POR	Period of Record
QC	Quality Control
REEP	Regression Estimation of Event Probability
RTLS	Return to Launch Site
SFO	San Francisco International Airport, CA 3-letter identifier
SLF	Shuttle Landing Facility
SMG	Spaceflight Meteorology Group
TPA	Tampa, FL 3-letter identifier
TS	Threat Score
TTS	Shuttle Landing Facility, FL 3-letter identifier
VF	Vislocky and Fritsch 1997
VRB	Vero Beach, FL 3-letter identifier
WMO	World Meteorological Organization 1992
WSR-74C	Weather Surveillance Radar, model 74C
WSR-88D	Weather Surveillance Radar 1988 Doppler

NOTICE

Mention of a copyrighted, trademarked or proprietary product, service, or document does not constitute endorsement thereof by the author, ENSCO, Inc., the AMU, the National Aeronautics and Space Administration, or the United States Government. Any such mention is solely to inform the reader of the resources used to conduct the work reported herein.