P2.17 EVALUATION OF WSR-88D METHODS TO PREDICT WARM SEASON CONVECTIVE WIND EVENTS AT CAPE CANAVERAL AIR FORCE STATION AND KENNEDY SPACE CENTER

James J. Rennie, James P. Koermer* and Thomas R. Boucher
Plymouth State University, Plymouth, New Hampshire

William P. Roeder
45th Weather Squadron, Patrick Air Force Base, Florida

1. INTRODUCTION

Convectively driven wind events are difficult to accurately predict, due to their small spatial resolution, short lifecycles, and large variability. Despite these forecast challenges, there are strong operational requirements for their prediction. For example, these winds create major aviation hazards, because of the strong low level wind shear that can be generated. Another example is processing space launch vehicles and payloads for space launch at the Kennedy Space Center (KSC) and Cape Canaveral Air Force Station (CCAFS) in Florida.

Convective winds are the second most frequent weather warning issued at CCAFS/KSC, after lightning warnings (Dinon *et* al. 2008, Ander *et* al. 2009). Accurate forecasts, as well as proper lead time of these wind gusts, are critical for the safety of over 25,000 people, resource protection for over $20 billion of facilities, avoiding space launch schedule delays, and mission assurance during processing of up to multi-billion dollar payloads or space launch vehicles in the weeks to months before a space launch.

Currently, the 45th Weather Squadron (45 WS) is responsible for predicting convective winds on the CCAFS/KSC complex. The warnings are categorized as either ≥ 35kt or ≥ 50kt from the surface to 300 ft above ground level (AGL). Until recently, there had also been a requirement for ≥ 60kt winds. For winds ≥ 35 kt, the desired lead time in issuing the warning is 30 minutes. On the other hand, winds ≥ 50 kt have a desired lead time of 60 minutes. The previous warning for ≥ 60 kt also had a desired lead time of 60 minutes.

*Corresponding author address: Dr James P. Koermer, MSC #48, 17 High Street, Dept. of Atmospheric Science and Chemistry, Plymouth State University, Plymouth, NH, 03264; e-mail: koermer@plymouth.edu

Over the past few years, much work has been performed by Plymouth State University researchers to aid the 45 WS in improving their forecasting of convective winds. These include creating the climatology of convective wind events based on five minute average peak wind speeds over the wind tower network from 1995-2003 (Loconto *et* al. 2006). This climatology significantly enhanced the 4-year study (1995-1998) by Sanger (1999). More work by Loconto (2006) tested and evaluated current forecasting techniques used by the 45 WS, as well as suggested new methods of forecasting for these winds using both KXMR sounding data and KMLB radar data. Related research to improve downburst prediction using RAOB data is also being conducted (McCue et al. 2010). In 2007, the climatology was reconstructed to include non-warning level convective episodes, and the formal definition of a convective event was introduced (Cummings *et* al. 2007).

In 2007-2008, Level III NEXRAD data for the warning level convective events were added to the climatology and events were classified by cell strength, cell initiation, cell structure, cell group movement, individual cell movement and location of maximum peak wind (Dinon *et* al. 2008). This work was continued in 2008-2009, where non-warning criteria convective events were added to the dataset (Ander *et* al. 2009). In 2009, the climatology was expanded to include data for 2008 and 2009. Currently, the climatology consists of fifteen years of data from 1995 through 2009.

Over recent years, an attempt has been made to create nowcasting techniques using WSR-88D data. Stewart (1996) created an equation to help determine the maximum downdraft speed of a microburst event based on derived WSR-88D products. The equation is the following:

$$w = \sqrt{(20.628571 \cdot VIL) + (-3.125 \times 10^{-6} \cdot ET^2)}$$

where w is the maximum predicted downdraft speed (m s^{-1}) VIL is the cell-based vertically integrated liquid (kg m^{-2}), and ET is the echo top (kft).

In addition, Loconto (2006) set out to create a more effective equation to determine the maximum peak wind gust based on radar variables from KMLB storm structure data. Using a multiple linear regression model, the following equation was generated:

$$GU = (.4138 \cdot VIL) + (.9194 \cdot MaxZ) + (.6253 \cdot Height) - 28.7719$$

where *GU* is the maximum peak wind gust (kt), *VIL* is the cell-based vertically integrated liquid (kg m^{-2}), *MaxZ* is the maximum reflectivity (dBZ), and *Height* is the height of the maximum reflectivity (kft).

Finally, a theory was provided by Loconto (2006) to explain the relationship between the heights of the maximum reflectivity and the peak wind gust. Figure 1 below contains plots of maximum reflectivity versus wind gust and also the difference between the height of the maximum reflectivity and RAOB defined freezing level. Positive height differences indicate the highest reflectivity is *above* the freezing level while a negative height indicates the highest reflectivity is *below* the freezing level.

While there is little correlation of the maximum reflectivity with maximum gust, there appeared to be a strong relationship with the height of the maximum reflectivity above the freezing level. With a positive height difference, there is a higher probability that hail has formed. Hail can provide additional cooling from the latent heat of melting, which provides additional negative buoyancy that helps sustain the depth and momentum of the downdraft to the surface to form the downburst. (Srivastava 1985, Proctor 1989). Figure 1 shows in general that whenever there is a positive height difference the wind gust is greater than the 35 kt threshold.

While these techniques show positive results, the datasets used were small. Since the radar climatology has been updated over the years, there was a need to re-evaluate these results to see if they still prove to be effective or can be improved for forecasting convective wind events. Additionally, the findings were only based on the volume scan at or just before the time of onset. We would like to see if information about a convective event at volume scans prior to the time of onset can be obtained, in order to create a longer lead time to issue the proper warnings.

Finally, additional statistical techniques are applied to the data and tested against an independent dataset. Our goal is to maximize the True Skill Statistic (TSS) to find the optimum compromise between Probability of Detection (POD) and Probability of False Alarm (POFA) for producing convective winds $\geq$ 35kt. The metric POFA is used rather than the equivalent and better known False Alarm Ratio to avoid confusion between False Alarm Rate and False Alarm Ratio (Barnes *et* al 2009).

2. DATA AND METHODOLOGY

Using the Plymouth State convective wind climatology, events from the years 2003 through 2009 were extracted and related information recorded, including the value of the peak gust (in knots), the time of the gust, and the weather tower that reported the gust. While stronger gusts from a downburst event might occur between the weather towers, for the purposes of this research, the peak wind gust that was recorded from one of the 36 weather towers on the CCAFS/KSC complex is considered the true value. It should be noted that null cases, known as events that met the convective event criteria defined by Cummings *et* al (2007), but produced a maximum peak wind gust less than 35 knots, were also included in this database.

In order to compute the predicted version of the peak wind gust for comparison, different predictors were needed. First, storm structure data from the WSR-88D located in Melbourne, FL (KMLB) was obtained from the NCDC radar archives that corresponded with the times of the peak wind gust. Using a great circle distance calculation, the cell that was closest to the tower that reported the peak wind gust was used. Once the cell was located, the following information was taken from the storm structure product:

- vertically integrated liquid (VIL) [kgm^{-2}]
- echo top [kft]
- maximum reflectivity [dBZ]
- height of the maximum reflectivity [kft].

Additionally, VIL Density (VILD) can be derived from storm structure data. It has been shown that VIL Density can be an indicator of the presence of hail (Amburn and Wolf, 1997), and therefore a

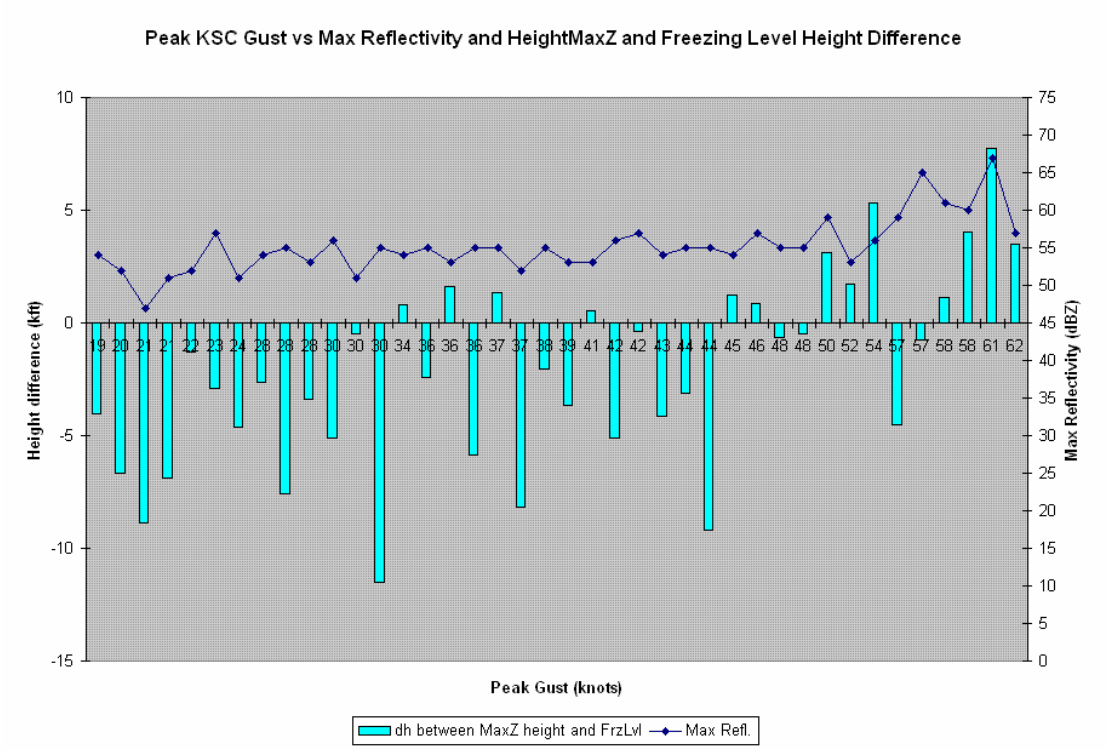

Figure 1. Max reflectivity minus RAOB freezing level height difference (light blue bars), value of max reflectivity versus peak CCAFS/KSC wind gust speed (dark blue line). (from Loconto 2006)

precursor to wet microbursts. VIL Density is defined as the following:

$$VILD = \frac{VIL}{ET}$$

Where VIL is the cell-based vertically integrated liquid (kg m^{-2}) and ET is the echo top (m). Units of VIL Density are in kg m^{-3}. VILD includes both the parameters used in the Stewart (1996) technique to estimate the maximum downburst gust speed a cell may produce. Thus, VILD may be useful not just to predict the occurrence of a downburst, but also its intensity, although Stewarts technique suggests VIL^{-2} may be more useful in predicting the latter. In order to provide lead time to the forecaster, data was taken not only at the onset of the peak wind gust, but also four volumetric scans prior to the occurrence, which is typically 16-25 minutes.

In addition to the storm structure data, previous work showed that low-level boundary interactions might be useful in forecasting downbursts at CCAFS/KSC (Dinon *et* al 2008, Ander *et* al 2009). Data on these boundary interactions were also collected to continue exploring this possible forecast technique. Low level boundary interactions at CCAFS/KSC during the summer are classified into four categories: sea breeze front (SBF), outflow boundary (OFB), both SBF and OFB, or no SBF or OFB.

Due to the general lack of synoptic triggering mechanisms during the warm season, mesoscale interactions play a dominant role in the formation of convective cells at CCAFS/KSC. Figure 2 below displays the percentage of occurrences of both warning (gust speed $\geq$ 35kt) and non-warning (gust speed < 35kt) convective cells based on their cell initiation. For warning cases, a boundary interaction occurred nearly 70 percent of the time. For non-warning cases, there was no mesoscale boundary almost 65 percent of the time. Because of this important distinction between warning and non-warning criteria, this variable is considered as a predictor.

Once the data were collected, the previous methods defined by Stewart (1996) and Loconto (2006) were tested to see if they were still valid with a larger dataset, as well as a longer lead-time. To evaluate the radar gust equations, the root mean square error (RMSE), and mean

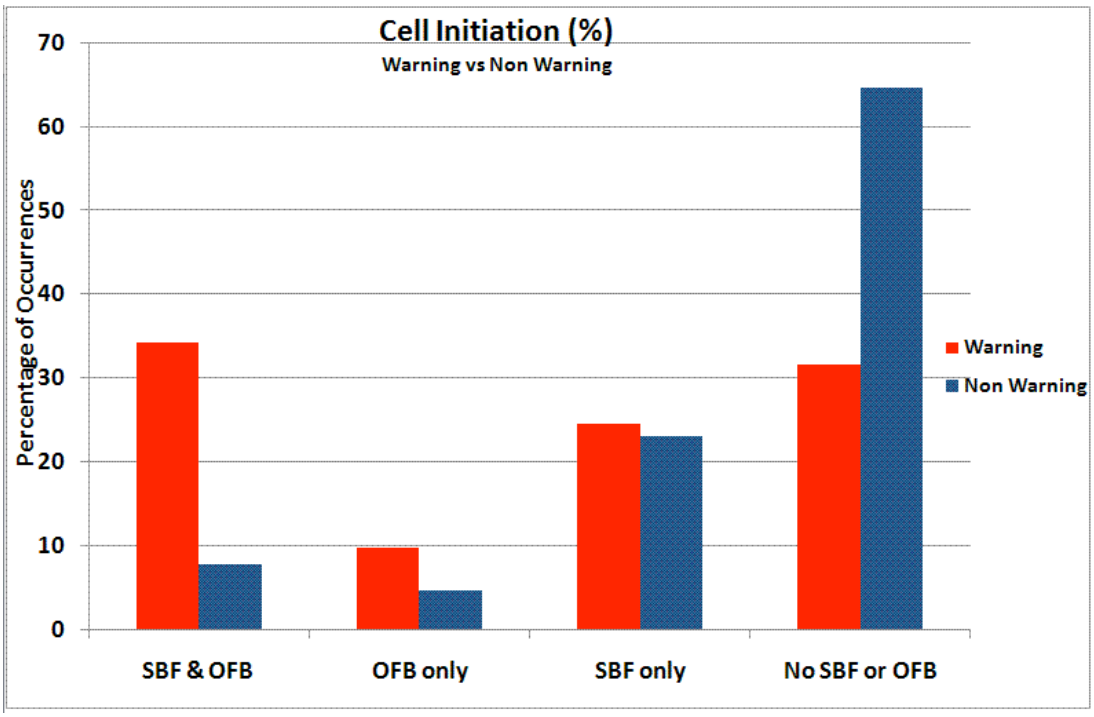

Figure 2. Cell Initiation with mesoscale boundaries with warning level cases in red and non-warning level cases in blue. Initiation is shown on the x-axis and percentage of occurrence on the y-axis. From Ander et al (2009)

absolute error (MAE) are calculated to determine forecast error. The number of hits, defined as a correctly predicted gust with an accuracy of ± 5 kt, is also determined.

In order to test the relationship between the height of the maximum reflectivity and peak wind gust, plots similar to Figure 1 are shown for all five volumetric scans. In addition, a 2X2 contingency table is constructed to determine forecast performance of this technique. Finally, new methods for improving nowcasting of convective winds are introduced using Classification And Regression Trees (CART) (Breiman *et* al 1984). CART can be used to provide objective forecasts without the use of a specific parametric form of the relationship between the response and the predictors. CART uses combinations of cluster analysis and discriminate analysis to optimally stratify the data into objective categories and provide yes/no decision branches to categorize future events into the most likely category. For example, to develop a downburst forecast tool, one might categorize the events into < 35 kt and ≥ 35 kt categories, then let CART select the best order of variables and yes/no decision thresholds to forecast < 35 kt and ≥ 35 kt events in the future.

Using a database with predictors and a response CART divides the sample into homogeneous groups with similar values of the response. This process is also known as "splitting". The samples are split according to splitting rules based on the values of the predictors. Different algorithms will search the dataset and maximize the significance or purity of a response in a group to determine where and when a split will occur. This process continues until each subset of the database attains complete homogeneity, or a stopping criterion is reached.

The advantages of using CART are that it can handle large datasets, does not need variable selection to be run before hand, and does not assume normality of the data or make other parametric assumptions. Additionally, it can handle outliers and nonlinear relationships. Operationally, CART also has the advantages of being easy to implement, easy to train, and easy to automate. A disadvantage of CART is that it tends to over fit the model with too many leaves and end nodes. This can be resolved by applying a process called pruning, where the

model will remove leaves that are deemed insignificant. In addition, the performance metrics associated with CART are often not as easy to interpret as those with more familiar metrics. Operationally, CART has the disadvantage that the reasons for the various decision branch variables and thresholds may not be clear. Studies have shown that weather forecasters are more likely to use techniques they understand, even if another less clear technique can provide better performance.

CART is not a single technique; there are many different approaches to create a CART, each of which can give different solutions for the same data. Therefore, it is good practice to develop CARTs for the same application via different approaches and compare their results, or even combine the approaches in the final forecast tool to create an "ensemble" of CARTs for more robust results. For this research five different tree algorithms are invoked on this dataset. The algorithms, which are implemented through the R Statistical environment (R Development Core Team 2009), are rpart, ctree, bagging, boosting, and random forests. A brief description of each algorithm is noted in the results section.

Before the algorithms are run, the dataset is partitioned into training and test subsets. The CART models are built, using a training set, the data between 2003 and 2007. The data from 2008 and 2009 are used as an independent test set to test the validity of the model. Using this, the probability of detection (POD), probability of false alarm (POFA), and true skill score (TSS) are calculated for each of the five CART models and compared to each other.

POD is the probability that convective wind events occur when forecast to occur. POD varied from 0.0 to 1.0 with 1.0 being best. The POFA is the probability that the convective wind event did not occur when forecast to occur. POFA varied from 0.0 to 1.0, with 0.0 being best. For an ideal forecast, the POD should be high and the POFA should be low. Perfect forecasting has both POD = 1.0 and POFA= 0.0.

TSS determines the model performance relative to random forecasting. The values of TSS can range between -1.0 and 1.0. A value of 1.0 is perfect forecasting, where a 0.0 is the same performance as random forecasting. A TSS less than 0.0 indicates the model performs worse than random forecasting (Wilks 2005). In general, a TSS above 0.3 is necessary for useful forecasting in real-world applications.

3. RESULTS

This section provides an evaluation of the two previous forecast methods and the five new CART methods. A discussion of the results follows in section 4.

3.1 Number of Cases

Table 1 displays the number of events that contained available WSR-88D data for each volumetric scan for the years 2003 through 2009. This includes events that produced winds < 35 kt as well as ≥ 35 kt. Overall, there were 377 cases used in this study. Note that the number of available data for earlier volumetric scans decreases. This occurred for one of two possible reasons. First, the cell might not have existed yet. The second and more probable reason could have been due to the corruption of the data. Some files obtained from NCDC were corrupted, and no data were present. There were times when the cell existed on all five volumetric scans, but did not have complete data.

Table 1. Number of cases where available data were present for all scans

Onset	377
Scan1	322
Scan2	280
Scan3	264
Scan4	250

3.2 Evaluation of Previous Methods

This section provides an evaluation of the previous two forecast methods: 1) radar gust equations, and 2) maximum reflectivity above freezing level.

3.2.1 Radar Gust Equations

Tables 2 and 3 show the RMSE, MAE, and number of hits for the ET/VIL relationship generated by Stewart (1996) and the Radar Gust Equation developed by Loconto (2006),

respectively. These values were calculated using the larger dataset from 2003-2009, as well as four volumetric scans prior to the onset of the peak wind gust.

The equation developed by Loconto appears to provide smaller error values than the ET/VIL relationship. Loconto's relationship has errors that are generally two knots lower than Stewart's. Additionally, Loconto's equation appears to be more accurate within five knots, providing a higher hit rate.

Even though Loconto's equation proves to have better performance, both methods perform poorly. RMSE values vary between 12 and 14 knots for all volumetric scans, and MAE values range between 9 and 11 knots. This large error range could make it difficult to distinguish between warning and non-warning cells. Additionally, hit rates for both methods at all volume scans are less than 34 percent. Therefore, the validity of these equations comes into question, and they may not be as useful as once thought.

Table 2. Forecast errors for ET/VIL relationship (Stewart 1996) for all five volumetric scans.

	RMSE	MAE	HITS	% HITS
Onset	14.03	11.23	97	26
Scan1	14.22	11.40	82	25
Scan2	13.72	10.86	80	29
Scan3	13.94	11.14	71	27
Scan4	13.89	11.22	68	27

Table 3. Same as Table 2, but with Loconto Radar Gust Equation (Loconto 2006).

	RMSE	MAE	HITS	% HITS
Onset	12.72	10.03	125	33
Scan1	12.73	10.03	95	30
Scan2	12.75	9.93	85	30
Scan3	12.12	9.60	89	34
Scan4	12.11	9.52	83	33

3.2.2 Maximum Reflectivity above Freezing Level

Figures 3 to 7 below indicate the averaged height difference and averaged maximum reflectivity versus peak wind gust. Whenever there is a positive height difference, the maximum reflectivity is above the freezing level. Two major points can be made. First, whenever there is a positive height difference, the wind gust tends to be greater than 35 knots. It needs to be noted that the relationship is not perfect, since some wind gust values have an averaged height difference less than zero, yet are greater than 35 knots.

Another point can be made here that was not apparent from Loconto's previous analysis. As the wind gust increases, the maximum reflectivity increases as well, whereas no relationship could be made from Figure 1. The coefficients of determination confirm that there may be a linear relationship between the two.

It appears that whenever there is a positive height difference, a warning should be issued. However, it is important to note that these are averaged height differences. The averages may be skewed by outliers and can severely misrepresent the results. In order to account for this, contingency tables were constructed, and the POD, POFA, and TSS were calculated.

The results, which are shown in Table 4, appear to be poor. The probability of detection is very low, ranging between 24 and 41 percent. While the probability of false alarm is also relatively small, they increase dramatically with each volume scan. In addition some scans, for example scan1, have higher POFA than POD. The TSS values range between 0.10 and 0.20, which means that this method of forecasting for peak wind gusts is only slightly better than random forecasting. A TSS > 0.30 is usually considered to be necessary for useful operational forecasting. Because of the poor performance metrics, the assumption that a positive height difference will produce a gust greater than 35 kt is no longer a certainty.

Table 4. POD, POFA, and TSS calculated for all volumetric scans of the freezing level relationship

	POD	POFA	TSS
Onset	0.24	0.26	0.17
Scan1	0.30	0.32	0.18
Scan2	0.41	0.36	0.20
Scan3	0.38	0.39	0.17
Scan4	0.33	0.45	0.10

3.3 Introduction of CART Methods

This section provides an evaluation of the five CART methods used to create forecast models

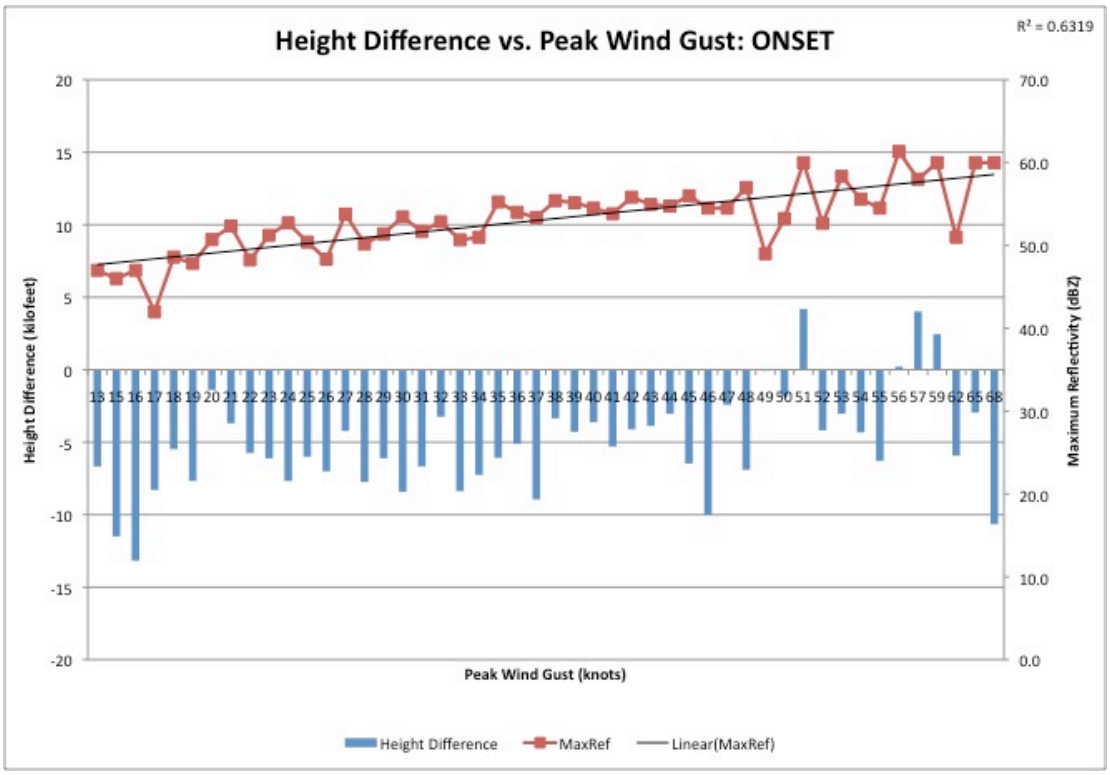

Figure 3. Max reflectivity minus RAOB freezing level height difference, value of max reflectivity versus peak CCAFS/ KSC wind gust speed using updated dataset from 2003-2009 for onset of maximum peak wind.

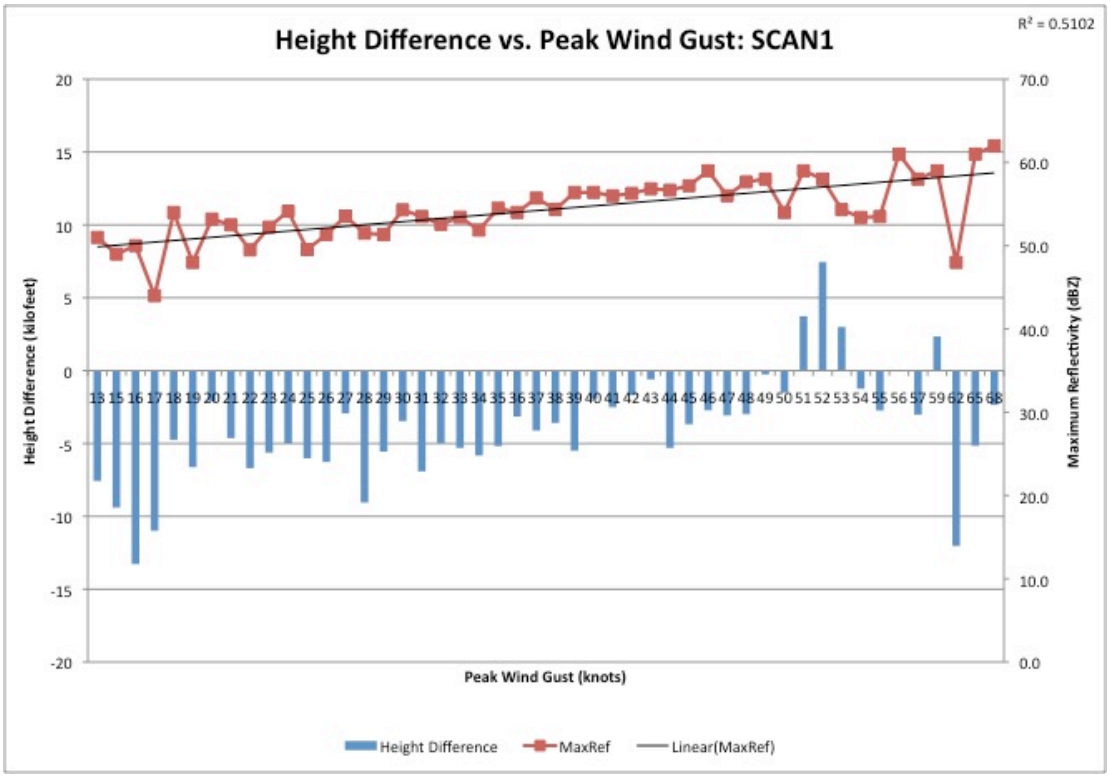

Figure 4. Same as Figure 3, but for first volumetric scan prior to onset.

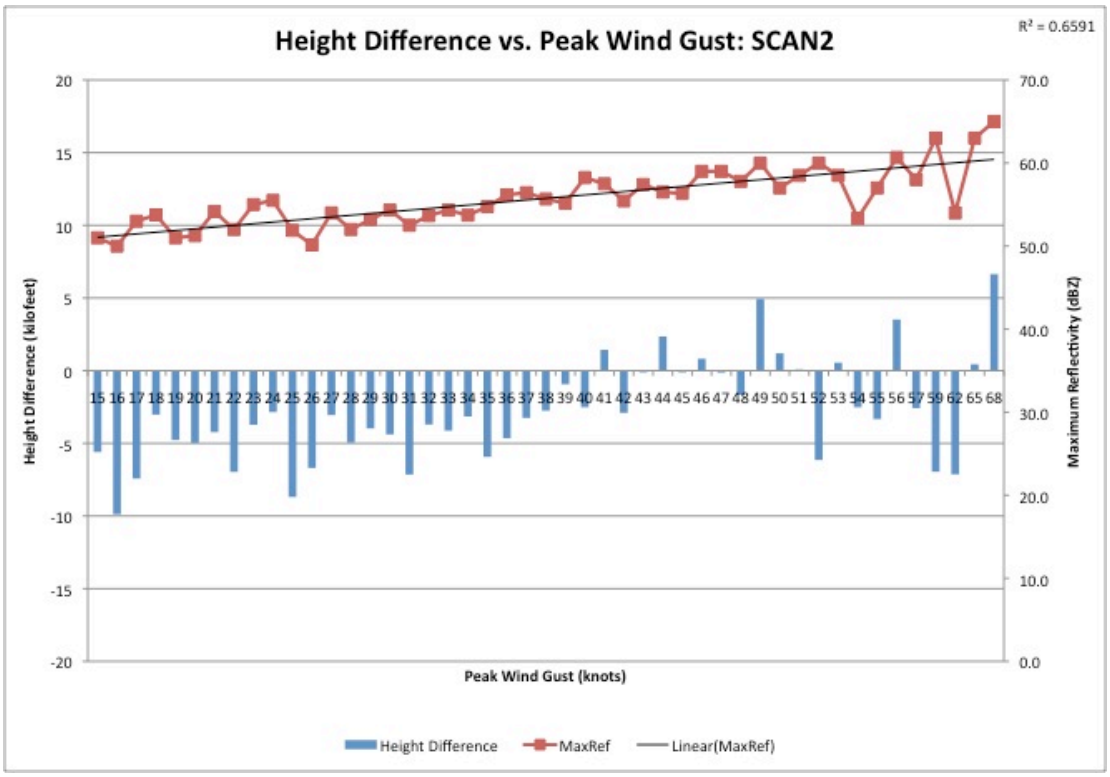

Figure 5. Same as Figure 3, but for second volumetric scan prior to onset.

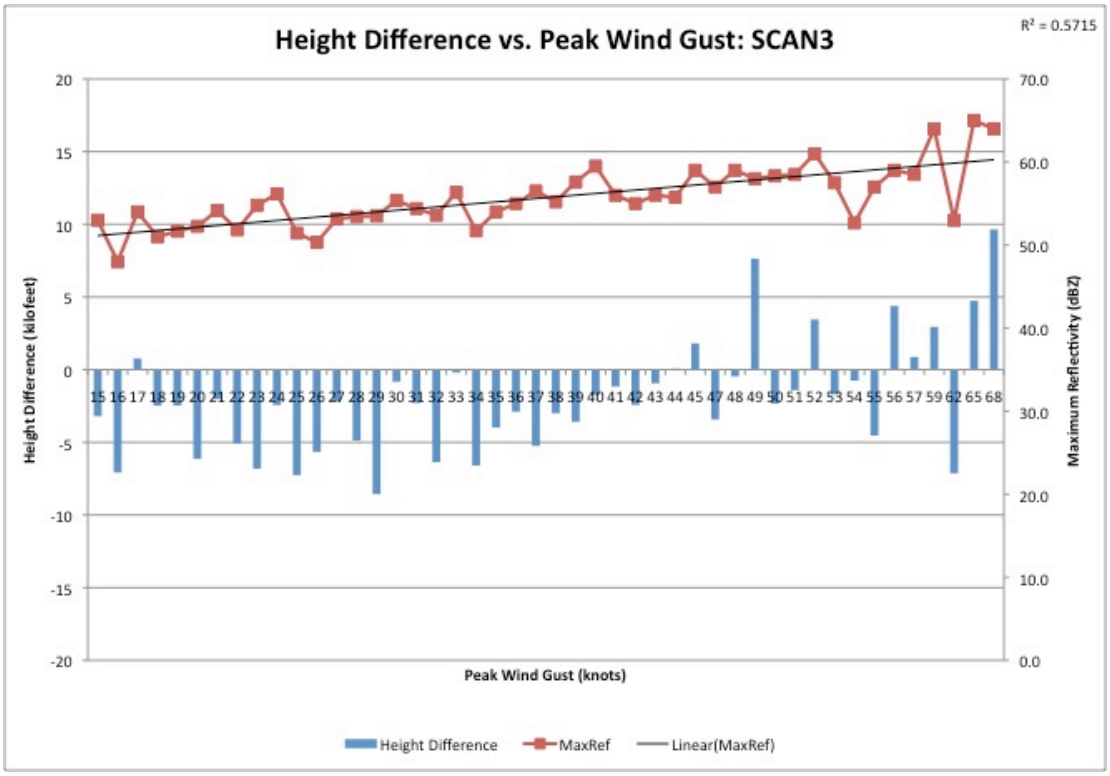

Figure 6. Same as Figure 3, but for third volumetric scan prior to onset.

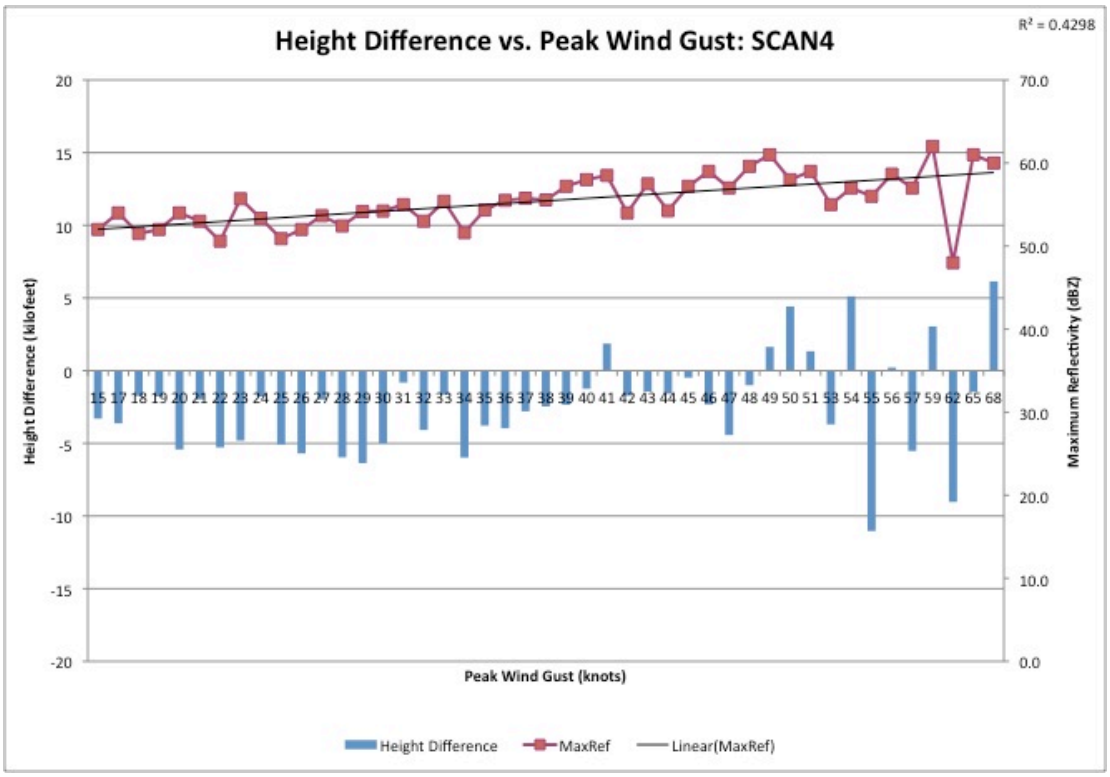

Figure 7. Same as Figure 3, but for fourth volumetric scan prior to onset.

of warning/non-warning convective winds. The five methods are 1) recursive binary partitioning and regression trees, 2) conditional inference trees, 3) bootstrap aggregation, 4) boosting, and 5) random forests.

3.3.1 Recursive Binary Partitioning and Regression Trees (rpart)

The rpart algorithm (Therneau and Atkinson 1997) is the basic CART producer in the R statistical environment. The algorithm recursively partitions (splits) the sample into smaller subgroups that are more homogeneous with respect to the response (purity). Splitting rules based on values of the predictors are used to assign observations to subgroups. At the first step the sample is split into two subgroups. On the next step the process is repeated on each of the subgroups resulting from the previous step. At each step possible splits are evaluated by calculating the purity of the response in the subgroups determined by that candidate split. The split chosen is the one that yields the greatest increase in purity of the response. This recursive partitioning of the data set continues until a stopping criterion is met. The final model is called a tree due to the fact that a graphical depiction of this process has a branching form similar to a tree. Settings can be adjusted by the user, including the number of possible splits, the minimum number of observations in a subgroup, as well as the depth of the tree. In addition, the rpart algorithm provides a cost complexity parameter (cp). This parameter is important for tree pruning, as CART models tend to become too large for operational use. The cost complexity parameter combines a measure of how well the tree fits the data with a penalty for over fitting. Minimizing this value will prevent the classification tree from growing too large.

Using the training set of data from 2003-2007, trees were generated using the default settings in the rpart algorithm. However, the final tree was determined by cross validation through the cp parameter. Table 5 below compares the POD, POFA, and TSS for all five volumetric scans together for comparison. These values were calculated using the independent test set of data from 2008-2009.

With the exception of the fourth volumetric scan, all scans have produced a tree has a POD $\geq$ 50 percent, and even the fourth volume scan had POD close to that value (48 percent). The POFA is also less than 50 percent, ranging

between 31 percent and 42 percent. Values of the true skill score range between 26 and 37 percent. This indicates that all of the models are better than random forecasting, but not by much and are just barely useful for operational forecasting.

Table 5. Performance metrics for rpart algorithm.

	POD	POFA	TSS
Onset	0.69	0.35	0.37
Scan1	0.53	0.32	0.29
Scan2	0.56	0.36	0.26
Scan3	0.76	0.42	0.26
Scan4	0.48	0.31	0.32

In terms of the optimal model to use, the third volumetric scan seems best. It is able to correctly identify events 76 percent of the time, while holding a modest false alarm ratio and true skill score. Scan2 has the same TSS as Scan3, but Scan3 provides more lead time. An argument could be made for onset, with a relatively high POD and better POFA and TSS than scan3. This also makes better physical sense, since forecasts should improve as the event draws closer in time. However, warning at onset is not operationally useful, since it does not allow enough time to issue a warning, because a peak wind gust will have occurred or will be occurring at that time. With the third volumetric scan, a lead time of approximately 15 minutes could be achieved. However, all this is moot for development of an operational forecast technique since even the best of these TSS with positive lead time (Scan1 or earlier) is < 0.3 and not operationally useful or at best only marginally useful.

Figure 8 displays the tree generated from the rpart algorithm for the third volumetric scan. It tests the VIL, VIL density, and boundary interactions of the cell starting with the vertically integrated liquid. For each split, if it passes the criteria, it will continue down the left side of the node. So if the cell has a VIL higher than 23.5 kgm^{-2}, then a convective event is predicted to occur. If not, it checks if a cell was initiated with any kind of boundary interaction (either sea breeze front, outflow boundary, or both). If it has, then it checks the VIL density of the cell, and if not, a peak wind gust will not occur.

3.3.2 Conditional Inference Trees (ctree)

The ctree algorithm (Hothorn *et* al 2006) uses conditional inference in an attempt to prevent over fitting of the tree. While the rpart algorithm does not use hypothesis testing to build the tree, the ctree algorithm does. For each split in question, p-values are generated for each candidate predictor. If the lowest p-value is below the standard statistically significant threshold of 0.05, then that candidate predictor has the strongest association. The null can then be rejected in favor of the alternative, and a split is performed. The tree is grown until no more nodes have a statistically significant relationship.

Using the ctree algorithm, trees are generated for all volume scans using the default settings. Table 6 displays performance of each model when the tree is tested against the independent dataset. It appears that the models at later volume scans (scan3 and scan4) have poor performance, with very low probability of detections, high probability of false alarms, and low true skill scores. The most recent three scans on the other hand, do provide higher performance. The setback is that they are closer to the onset of peak wind gust on the CCAFS/KSC complex and so provide less lead time. However the second volumetric scan, approximately 10 minutes prior to onset, can predict an occurrence approximately 75 percent of the time, while still holding a moderate probability of false alarm and true skill score (35 and 36 percent, respectively).

Table 6. Performance metrics for ctree algorithm.

	POD	POFA	TSS
Onset	0.67	0.32	0.40
Scan1	0.63	0.31	0.36
Scan2	0.75	0.35	0.36
Scan3	0.31	0.43	0.13
Scan4	0.30	0.22	0.24

The tree built for the second volumetric scan is shown in Figure 9. For each node, the p-value is given to determine the significance of the split. Additionally, once the end node is reached, the probability of the wind gust occurrence is displayed, as well as the number of cases that fell into the criteria when the model was constructed.

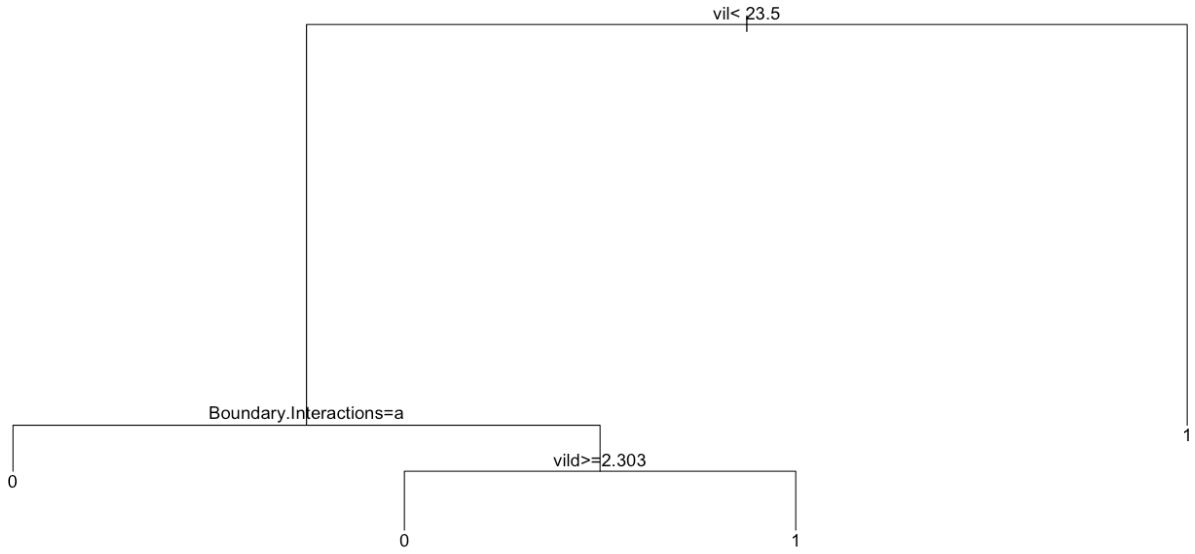

Figure 8. Tree generated from rpart algorithm for third volumetric scan (approximately 12-15 minutes prior to the onset of peak wind gust). a = cell initiated by neither SBF nor OFB.

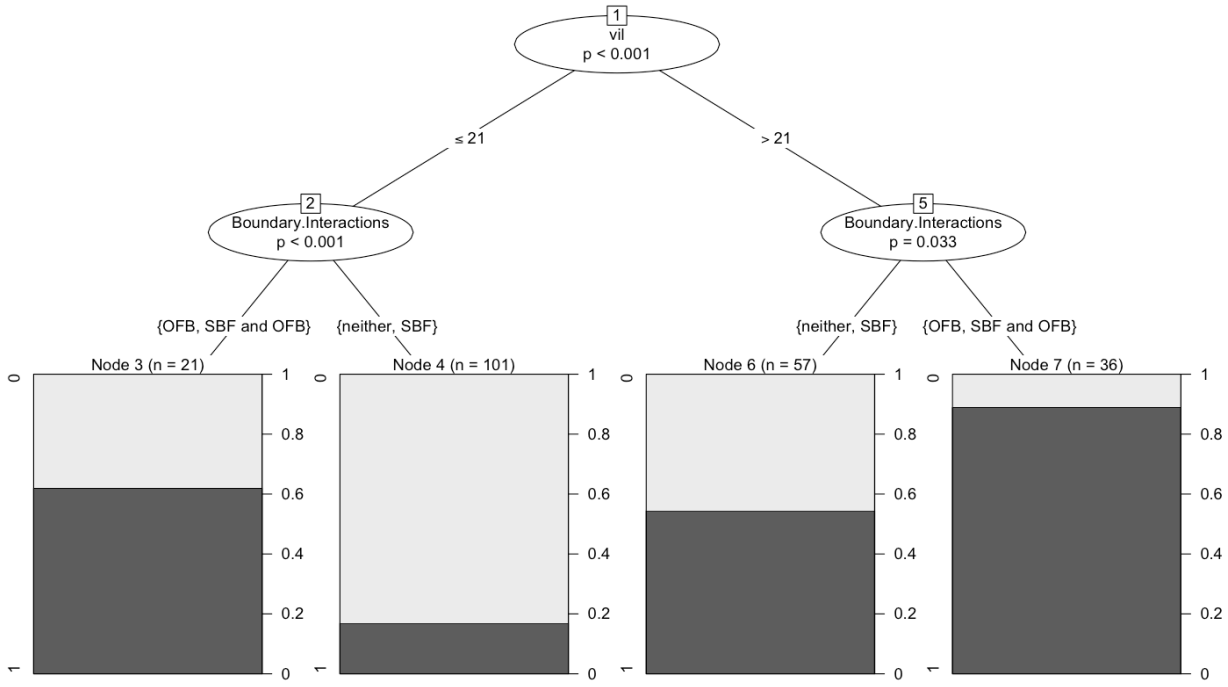

Figure 9. Tree generated from ctree algorithm for second volumetric scan prior to onset. P-values for the significance of split is displyed for each node, as well as the probabilities of strong wind gust occurrence at each end node.

For this model, the tree begins by checking the VIL, and then the boundary interactions. If the VIL is high enough, there will be a good chance a strong wind gust will occur on the CCAFS/KSC complex. The probabilities increase when the cell was initiated by a strong boundary interaction. On the other hand, if the VIL is not high enough, a strong wind gust can only occur if the boundary interactions are strong enough. This all makes good physical sense. It is interesting to note that both rpart and ctree picked VIL as the first predictor with very similar thresholds of 23.5 kgm^{-3} and 21 kgm^{-3}, respectively. The fact that two somewhat independent techniques yielded very similar results increases confidence in those first branches.

3.3.3 Bootstrap Aggregation (bagging)

The next three algorithms use a technique called bootstrapping (Wilks 2005). Bootstrapping, in its simplest terms, provides multiple re-samples of data. Using the original dataset, random sampling with replacement is performed to create a subset of the data. It is important to note that some data is repeated in the different subsets, so the results are not fully independent of each other. The statistical method is then applied to each dataset, and multiple values are then produced for statistical inference. Bootstrapping has the ability to reduce the variance of predictions made by models for the original dataset, but can be computationally intensive, especially with large datasets.

Bootstrap aggregation, or bagging (Breiman 1996), uses the above technique to create multiple re-samples and produce multiple classification trees, one for each bootstrap sample. While different trees can provide different classifications, the final classification is determined by popular vote. For example, if 100 trees are generated, and the end node of 90 trees is 0, while 10 are 1, then popular vote dictates the final classification will be 0. By creating an ensemble of classification tree forecasts, presumably the final forecast is more robust than any single tree.

By default, the bagging algorithm creates 100 trees using rpart. The settings of the bagging algorithm, as well as the rpart algorithm which makes the trees, are unchanged. Using the independent dataset, classifications of 0 and 1 are created by running through the trees and then popular vote determines the final classification. The POD, POFA, and TSS can then be calculated and is seen in Table 7.

Overall the results indicate marginal utility for operational forecasting. The exception to this is the third volumetric scan, with a POD of 41 percent, a POFA of 29 percent and a TSS of 26 percent. The other scans however do have TSS > 0.3, POD values above 60 percent, and POFA values between 22 and 35 percent.

The fourth volumetric scan appears to provide one of the better results. Using this model, convective winds above 35 knots will be predicted 61 percent of the time, with a probability of false alarm of only 22 percent. With a 48 percent true skill score, this method is much better than random forecasting. On top of the promising performance, this model can provide approximately 20 minutes of lead time to the forecaster, which is very important.

Table 7. Performance metrics for bagging algorithm.

	POD	POFA	TSS
Onset	0.61	0.35	0.32
Scan1	0.63	0.32	0.34
Scan2	0.69	0.35	0.32
Scan3	0.41	0.29	0.26
Scan4	0.61	0.22	0.48

The skill increasing for longer lead-times may seem counter-intuitive since skill usually increases closer in time to the event. Note that the same pattern occurs in some of the subsequent CART models increasing confidence that the pattern is valid. However, it is important to define the event rigorously. In this case, the event is the time of maximum gust at or near the surface. But the radar is measuring conditions in the convective cells, where the downdraft begins, not the downburst that forms once the downdraft reaches the surface. Allowing time for the downdraft to form, reach the surface, and reach the tower where the maximum gust speed is recorded, lead-times of 15-20 minutes may be reasonable. The time to reach the ground and the time for the downdraft to form, may be estimated as follows. Using a typical downburst producing downdraft height of 575 hPa, as implied by the Microburst-Day Potential Index

Table 8. Variable importance for all six predictors in the bagging algorithm for all volumetric scans. Variable of highest importance noted in red and second highest importance noted in blue.

	TOP	VIL	MAXREF	HEIGHT	VILD	BOUNDARY
Onset	**26.13**	15.77	12.61	4.50	4.50	**36.49**
Scan1	16.14	**27.80**	11.66	6.28	6.28	**31.84**
Scan2	15.27	**30.54**	4.93	3.94	12.32	**33.00**
Scan3	21.62	20.72	6.31	2.25	**22.97**	**26.13**
Scan4	**31.76**	9.87	13.30	7.73	7.73	**29.61**

(Wheeler and Roeder, 1996), which has an average height of 4,750 m in July (Range Reference Atmosphere, 2006), the middle of the local convection season, and a typical downburst speed of around 28 kt (14 m/s) from the CCAFS/KSC downburst climatology (Ander et al. 2009, Dinon et al. 2008, and Loconto et al 2006), a typical downdraft takes 5.7 minutes to reach the ground. Given the typical size of a downburst and the typical spacing of the weather towers at CCAFS/KSC, the tower with the maximum gust speed will typically be 1,000-2,000 m from where the downdraft hits the ground, which implies a time of 1-2 minutes. It is reasonable to assume that it takes several minutes for the downdraft to form after the convective cell reaches its maximum intensity. Altogether radar lead-times of 15-20 minutes may be reasonable for downbursts at CCAFS/KSC. It would be interesting to extend the analysis to even earlier volume-scans to see if the TSS begins to drop, as expected.

Since bagging produces an ensemble of 100 different classification tees, showing a figure of the tree and discussing the variables and decision branches, as done for the first two classification trees, is not practical. However the algorithm can provide variable importance. This value dictates the relative importance of each variable in the classification task. In other words it takes into account how many times the predictor is used for splitting. The higher the value is, the more important that predictor is to the model.

Table 8 provides the variable importance for the bagging algorithm. The variable with the highest importance is noted in red and second highest in blue. It can be shown that the boundary interactions have the highest importance in most of the models. The exception is the fourth volumetric scan, with the echo top only having a slightly higher importance. Even then, the boundary variable was a close second and the difference between the two variables may not have been statistically significant. The height parameter appears to have the weakest importance in all five models generated.

3.3.4 Boosting

The boosting algorithm (Freund and Schapire 1996) will create multiple classification trees based on re-sampled datasets, however instead of randomly sampling the data, the original dataset is weighted. These weights, which are adjusted after each iteration, will become a factor when classes are identified. Classes that were correctly identified during the previous step are given a lower weight, and ones that are incorrectly classified are given higher weight. Final classifications are then determined not by popular vote, but rather a weighted vote of the iteratively produced classifiers.

Similar to the bagging algorithm, 100 trees are generated (using rpart defaults) and tested against the independent dataset. The results can be seen in Table 9. In general the boosting algorithm provides positive performance. Probability of detection values range between 58 and 66 percent, and all of the probability of false alarms are lower than one third. True skill scores vary between 37 and 48 percent, and are higher at earlier volumetric scans.

Table 9. Performance metrics for boosting algorithm.

	POD	POFA	TSS
Onset	0.61	0.31	0.37
Scan1	0.58	0.23	0.41
Scan2	0.63	0.29	0.38
Scan3	0.66	0.24	0.47
Scan4	0.61	0.22	0.48

As for which scan provides the best performance, it appears that both the third and

Table 10. Variable importance for all six predictors in the boosting algorithm for all volumetric scans. Variable of highest importance noted in red and second highest importance noted in blue.

	TOP	VIL	MAXREF	HEIGHT	VILD	BOUNDARY
Onset	**24.75**	11.17	16.50	11.17	11.17	**25.24**
Scan1	**27.67**	15.05	14.08	10.68	11.65	**20.87**
Scan2	16.97	**20.64**	5.05	17.89	15.60	**23.85**
Scan3	**33.63**	10.18	3.10	8.85	**23.89**	20.35
Scan4	**27.83**	7.55	13.20	14.62	**18.40**	**18.40**

fourth volumetric scans are adequate. For both models, probability of detection values are above 60 percent, probability of false alarms are less than 25 percent, and true scores are nearly 50 percent. Similar to the bagging algorithm, better performance appears at earlier volumetric scans, which can provide longer lead times to the forecaster.

As with bagging, boosting produces an ensemble of 100 different classification trees, so showing a figure of the tree and discussing the variables and decision branches, as done for the first two classification trees, is not practical. The variable importance of each predictor in the boosting algorithm can be seen in Table 10. Unlike the bagging algorithm, the boundary interactions do not pay the most significant role in the generation of these models. In fact, most of the time, the echo top has the dominant importance.

3.3.5 Random Forests

One of the issues of the bagging and boosting algorithms is that the generated trees are correlated with each other. This may cause the unnecessary use of extra trees. The random forest algorithm (Breiman 2001) will create trees that have less correlation with each other, with hopes of a greater reduction in prediction variance. The process is similar to bagging and boosting in that it then determines classification by a popular vote over the ensemble of trees. It is different from the bagging and boosting algorithms in that 500 rpart trees are the default instead of 100. Similar to the previous tree algorithms, the default settings are not changed.

Table 11 displays the model performance when tested against the independent dataset. Like all of the other models, the results show some utility for operational forecasting. Probability of detection values are fair, ranging between 56 and 60 percent, while at the same time maintaining small probability of false alarms, which range between 28 and 35 percent. True skill scores range between 34 and 39 percent, meaning that they are about one third better than simple random guessing.

Table 11. Performance metrics for random forest algorithm.

	POD	POFA	TSS
Onset	0.59	0.29	0.39
Scan1	0.60	0.31	0.34
Scan2	0.56	0.28	0.35
Scan3	0.59	0.32	0.34
Scan4	0.57	0.35	0.34

Because of the small variability in the performance metrics for all volumetric scans (4 percent for POD, 7 percent for POFA, and 5 percent for TSS), any of the random forest models could be used to forecast convective winds on the CCAFS/KSC complex. However since the idea is to provide adequate lead times to the forecaster, logic dictates that the fourth volumetric scan will provide the best model. Approximately twenty minutes out, a wind gust greater than 35 knots will be accurately detected 57 percent of the time, while holding a POFA of 35 percent and a true skill score of 34 percent.

As with bagging and boosting, random forests produce an ensemble of different classification trees, so showing a figure of the tree and discussing the variables and decision branches, as done for the first two classification trees, is not practical. It can be noted that the variable importance values (displayed in Table 12) are evenly distributed throughout all five models. The boundary interactions, which were highly important in the bagging and boosting algorithms, do not appear to have any importance here. In fact, the height parameter, which is often disregarded in most of the previous models, has a higher importance than

Table 12. Variable importance for all six predictors in the random forest algorithm for all volumetric scans. Variable of highest importance noted in red and second highest importance noted in blue.

	TOP	VIL	MAXREF	HEIGHT	VILD	BOUNDARY
Onset	**28.58**	22.51	18.44	18.84	**23.02**	17.78
Scan1	**23.78**	**22.39**	17.66	16.76	21.11	14.57
Scan2	**21.16**	**21.39**	12.42	16.12	18.46	14.97
Scan3	**22.39**	**18.54**	12.96	13.26	18.00	12.65
Scan4	**20.31**	15.35	13.83	16.12	**17.12**	12.16

the boundary interactions. It appears that for all models, the echo top has the most significance, however the cells VIL comes in a close second.

4. DISCUSSION

After analyzing previous methods to forecast convective winds using WSR-88D technology with a larger dataset and longer lead time, it appears that they do not perform as well as initial results indicated. For both the ET/VIL relationship and Loconto's Radar Gust Equation, error values were no lower than 9 knots and the hit rates were no higher than 35 percent. The ability to differentiate between downbursts less than 35 kt and greater than 35 kt has been identified as important to 45 WS operations based on the previous climatology of distribution of convective wind speeds. (Figure 10). This is close to the peak in frequency distribution of downburst speeds and so the decision between issuing a warning for ≥ 35 kt or not issuing a warning is the most frequent convective wind warning faced by 45 WS. The warning decision between ≥ 50 kt and just ≥ 35 kt occurs much less frequently. It is very rare for a ≥ 50 Kt warning or no warning decision to occur. An error of 9 knots could inaccurately place the wind above or below 35kt or 50kt thresholds, which could generate not only false alarms, but also warnings that weren't issued which should have been.

In addition, it appears that the relationship between the peak wind gust and the height of the maximum reflectivity has less validity. While the updated plots look highly similar to those constructed in previous analysis, POD values were too low and the POFA values too high to warrant any confidence. There is some potential for the reflectivity relationship, as there appears to be a strong linear correlation with the peak wind gust, but it was not looked into any further in this study.

The CART based statistical techniques introduced in this paper all appear to provide promise to better nowcast convective winds on the CCAFS/KSC complex. Almost all of the methods at all volumetric scans have POD values greater than 50 percent and POFA values less than 50 percent. TSS values varied, but because they have positive values instead of negative, they are all better than simple random forecasting. Most had TSS around 30% or less, indicating only marginal use in operational forecasting. However, one of the CART techniques, boosting (see section 3.3.4), had TSS approaching 50%, near the threshold for good utility in operational forecasting.

It is important to note that each CART algorithm is unique and handles datasets differently. Some methods are more efficient and can be easily understood by the user. Others are computationally intensive yet provide better performance in terms of model fit and predictive accuracy. Forecasters may prefer simplicity to slightly higher accuracy. However, through technological advances, these models can be automated quickly and efficiently.

It may not be appropriate at this time to choose which model is the best fit to be put into operation. While overall performance appears to be promising, these models should be tested during warm season forecasting operations to determine in "real time" how well they actually perform.

5. FUTURE WORK

It was noted in section 3.2.2 that there might be a linear correlation between the maximum reflectivity of the cell to the maximum peak gust. This suggests creating a new reflectivity threshold that would distinguish cells between warning and non-warning criteria. For example, a modified method might be to forecast ≥ 35kt if the maximum reflectivity is above the freezing

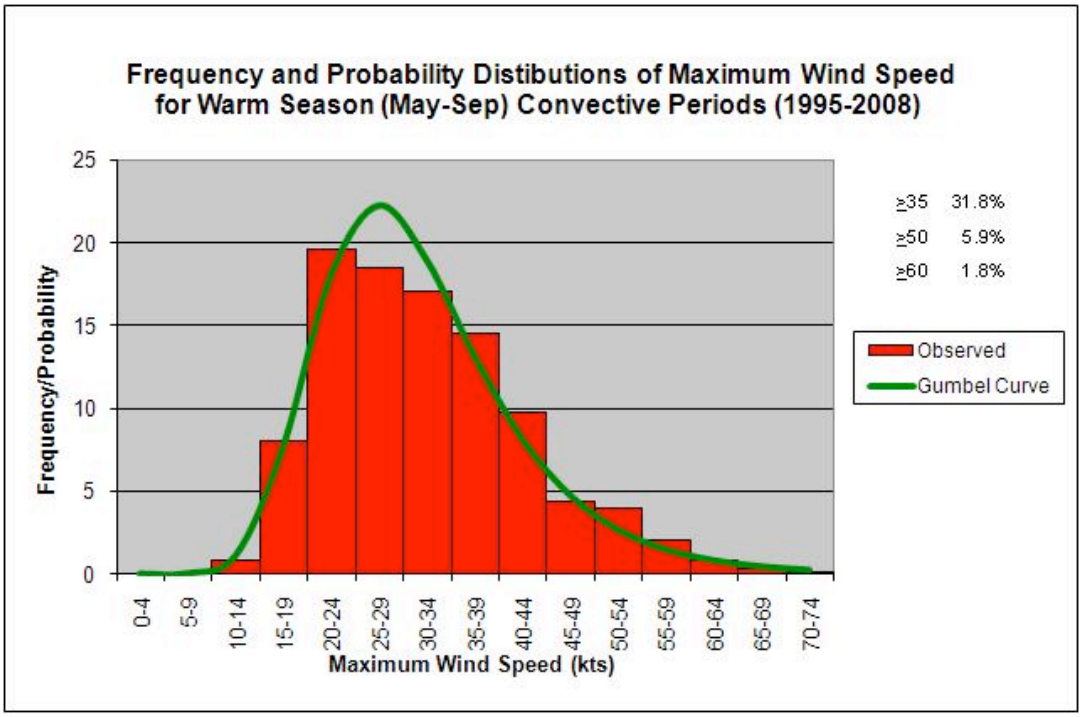

Figure 10. Frequency distribution of maximum convective peak wind observations by 5-knot increments, with a Gumbel probability curve fit to the observed data for the 924 convective periods for the warm-season (May-Sep) months in the 14-year (1995-2008) study period.

level and the maximum reflectivity is ≥ X dBZ, where 'X' would be tuned empirically to maximize TSS. Also, if the threshold between the height of the maximum reflectivity and freezing level were changed from 35 to 45 kt, there may be an improvement in model performance. Another approach might be to apply binary logistic regression combining if maximum reflectivity is above a to-be-determined threshold and above the freezing level to produce a probability forecast. The performance of the two approaches could be compared to pick the best approach. In addition, the amount of height of maximum reflectivity above freezing level could also be analyzed, as apposed to a yes/no threshold that it is above the freezing level.

It should be emphasized that all five tree algorithms in the R statistical environment were run using default parameters. The exception to this is the cost complexity parameter in the rpart function. It might be helpful to adjust these settings to see if model performance can be improved. Some of these settings include maximum number of nodes, minimum number of observations within a node, and number of bootstraps that are applied to a dataset.

For the bootstrapping algorithms (bagging, boosting, random forests) an ensemble of trees are generated, and a final classification is determined by popular vote. It might be useful to create probability forecasts from these classifications. For example, if 90 classifications out of 100 are placed under the 'yes' category, then there might be a 90 percent probability of occurrence. The percentage of votes in favor of a warning may not correspond directly to the probability forecast, so a conversion technique may need to be developed. In addition, the voting threshold for yes forecasts could be restructured to improve model performance. Instead of assuming that 50 percent of 'yes' classifications determine the final classification, this percentage could be increased or decreased to maximize TSS, in essence applying a bias correction.

The 45 WS has recently installed a new polarimetric Doppler system approximately 23 nautical miles southwest of the CCAFS/KSC complex (Roeder et al. 2009). It is currently being tested with operational implementation projected for February 2010. Polarimetric radars provide both vertically and horizontally polarized wavelengths to infer the size, species, and

quantity of the hydrometeors. These data should provide valuable new information for predicting downbursts and may provide a large leap forward in this important forecast challenge. If polarimetric radars provide a higher forecast performance for downbursts, then forecast tools would need to be constructed. Variables such as the reflectivity depolarization ratio (Z_{DR}) and the differential propagation phase (ϕ_{DP}) could be used as predictors along with other polarimetric and traditional radar signatures, and statistical models, such as the ones described in this paper, could be developed.

Also, most of this work only considered the reflectivity-derived parameters. It would be useful to look at the radial velocity parameters as well. It has been shown that mid level radial convergence leads to strong convective wind speeds (Roberts and Wilson 1989) and was used as one of the forecast parameters in the Damaging Downburst Prediction and Detection Algorithm developed for the NEXRAD WSR-88D radars (Smith *et* al 2004). It would be worthwhile to determine if this signature is observed not only on the CCAFS/KSC complex, but also a few volumetric scans prior to the onset of peak winds.

A tool used to predict the onset of downbursts in east central Florida using timelines of WSR-88D cell trends was developed by the Applied Meteorology Unit (Wheeler, 1998). While that technique showed promise it was developed on a relatively small sample size. Further testing and fine tuning of that technique may be worthwhile.

Finally, previous research has indicated that Storm Top (Sullivan, 1999) is better than Echo Top in the Echo Top/VIL Maximum Gust technique (Stewart, 1996). Storm Top should be considered as a possible variable in future radar downburst gust equations.

6. SUMMARY

This paper attempted to evaluate previous WSR-88D to forecast convective winds on the CCAFS/KSC complex. Using an updated dataset with earlier lead times, it was shown that these methods do not provide adequate results. RMSE and MAE for the two predictive gust equations ranged between 11-13 knots and 9-11 knots, respectively. In addition, the relationship between the peak wind gust and the height of the maximum reflectivity has been rendered ineffective.

Five CART algorithms were developed on test data and verified on independent test data. Results showed that all of the models were better than random forecasting and could serve to provide beneficial forecast information to operational forecasters. While some are more computationally efficient than others, automation could be used to quickly generate guidance from the most complicated methods.

Much more detailed data, analyses, and many of these and additional references for these studies are available online at the following URL:

http://vortex.plymouth.edu/conv_winds/

7. ACKNOWLEDGEMENTS

This work was supported through the NASA Space Grant Program under University of New Hampshire Subcontract #PZ05008. We would also like to thank NCDC, the Applied Meteorology Unit (AMU), and Computer Sciences Raytheon (CSR), who provided valuable archived data for our work. The Plymouth State University authors would also like to thank all personnel from the AMU and the 45 WS, who provided invaluable guidance while working at CCAFS during the summer of 2009.

8. REFERENCES

Amburn, S. A., and P. W. Wolf, 1997: VIL density as a hail indicator. *Weather and Forecasting,* **12,** 473–478.

Ander C. J., A. J. Frumkin, J. P. Koermer and W. P. Roeder, 2009: Study of sea-breeze interactions which can produce strong warm-season convective winds in the Cape Canaveral area. *16th Conf. on Air-Sea Interaction/8th Conf. on Coastal Atmospheric and Oceanic Prediction and Processes*, 12-15 Jan, Phoenix, AZ.

Barnes, L. R., D. M. Schultz, E. V. Gruntfest, M. H. Hayden, C. C. Benight, 2009: False Alarm Rate or False Alarm Ratio? *Weather And Forecasting*, **24**, 1452-1454.

Breiman, L., 1996: Bagging predictors. *Machine Learning*, **24**, 123-140.

---, 2001: Random forests. *Machine Learning*, **45**, 5-32.

---, Friedman, J., Olshen, R.A., and Stone, C.J., 1984: *Classification and Regression Trees.* Wadsworth, 368 pp.

Cummings, K. A., E. J. Dupont, A. N. Loconto, J. P. Koermer and W. P. Roeder, 2007. An updated warm-season convective wind climatology for the Florida space coast, *16th Conf. of Applied Climatology*, 14-18 Jan 2007, San Antonio, TX.

Dinon, H. A., M. J. Morin, J. P. Koermer, and W. P. Roeder, 2008. Convective winds at the Florida Spaceport: year-3 of Plymouth State research, *13th Conf. on Aviation, Range, and Aerospace Meteorology,* 21-24 Jan 2008, New Orleans, LA.

Freund, Y., and Schapire, R.E., 1996: Experiments with a new boosting algorithm. *Proc.13th International Conference on Machine Learning*, Bari, Italy, 148-156.

Hothorn, T., Hornik, K., and Zeileis, A., 2006: Unbiased recursive partitioning: a conditional inference framework. *Journal of Computational and Graphical Statistics,* **15**, 651-674.

Loconto, A. N., J. P. Koermer, and W. P. Roeder, 2006: An updated warm-season convective wind climatology for Cape Canaveral Air Force Station/Kennedy Space Center, *12th Conference on Aviation Range and Aerospace Meteorology*, 30 Jan–2 Feb 2006, Atlanta, GA.

---, 2006. Improvements of Warm-Season Convective Wind Forecasts at the Kennedy Space Center and Cape Canaveral Air Force Station. M.S. Thesis, Dept. of Chemical, Earth, Atmospheric and Physical Sciences, Plymouth State University, Plymouth, NH

McCue, M. H., J. P. Koermer, T. R. Boucher, and W. P. Roeder, 2010: Validations and Development of Existing and new RAOB-based Warm-season Convective Wind Forecasting Tools for Cape Canaveral Air Force Station and Kennedy Space Center, 22nd Conference on Climate Variability and Change, 18-21 Jan 10, Paper P2.16.

Proctor, F.H., 1989. Numerical simulations of an isolated microburst. Part II: Sensitivity experiments. *Journal of the Atmospheric Sciences,* **46,** 2143-2164.

R Development Core Team, 2009: R: A language and environment for statistical computing: version 2.10. R Foundation for Statistical Computing, Vienna, Austria.

Range Reference Atmosphere, 2006: Radiosonde climatology for Cape Canaveral Air Force Station Jan 1990-Jan 2002 (listed as Kennedy Space Center, EXCEL spreadsheet, *14th Weather Squadron*, 14 WS, 250 Patton Ave., Rm 120, Asheville, NC 28801-5002, https://notus2.afccc.af.mil, 30 Aug 06

Roberts, R.D., and J.W. Wilson, 1989. A proposed microburst nowcasting procedure using single-Doppler radar. *Journal of Applied Meteorology,* **28,** 285-303.

Roeder, W. P., T. M. McNamara, B. F. Boyd, and F. J. Merceret, 2009: The new weather radar for America's space program in Florida: an overview, *34th Conference on Radar Meteorology*, 5-9 Oct 09, Wiiliamsburg, VA, Paper 10B.6, 9 pp.

Sanger, N. T., 1999: A Four-Year Summertime Microburst Climatology And Relationship Between Microbursts And Cloud-To Ground Lightning Flash Rate For The NASA Kennedy Space Center, Florida: 1995-1998, M.S. *Thesis*, Texas A&M University, Aug 99, pp 116

Smith, T.M., K.L. Elmore, and S.A. Dulin, 2004. A Damaging Downburst Prediction and Detection Algorithm for the WSR-88D. *Weather and Forecasting,* **19**, 240-250.

Srivastava, R.C., 1985. A simple model of evaporatively-driven downdraft: Application to microburst downdraft. *Journal of the Atmospheric Sciences,* **42,** 1004-1023.

Stewart, S.R., 1996. Wet Microbursts—Predicting Peak Wind Gusts Associated with Summertime Pulse-Type Thunderstorms. Preprints, 15th Conf. on Weather and Forecasting. Amer. Meteor. Soc., 324-327.

Sullivan, G. D., 1999: Using the WSR-88D to Forecast Downburst Winds at Cape Canaveral Air Station and the Kennedy Space Center, M. S. Thesis, Air Force Institute of Technology, AFIT/GM/ENP/99M-12, Mar 99, 96 pp.

Therneau, T.M., and Atkinson, E.J., cited 2009: An Introduction to Recursive Partitioning Using the rpart Routine [available online at http://www.mayo.edu/hsr/techrpt/61.pdf].

Wheeler, M. M., 1998: WSR-88D cell trends final report, NASA Contractor Report CR-207904, prepared by Applied Meteorology Unit, http://science.ksc.nasa.gov/amu/final-reports/wsr88d-cell-trends.pdf, 42 pp.

Wheeler, M. M., and W. P. Roeder, 1996: Forecasting wet microbursts on the central Florida Atlantic coast in support of the United States space program, 18th Conference on Severe Local Storms, 19-23 Feb 96, 654-658

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences, Volume 91, Second Edition*. Academic Press, 648 pp.