A data mining and statistical learning method known as a random forest (RF) is employed to generate 2-h forecasts of the likelihood for initiation of mesoscale convective systems (MCS-I). The RF technique uses an ensemble of decision trees to relate a set of predictors [in this case radar reflectivity, satellite imagery, and numerical weather prediction (NWP) model diagnostics] to a predictand (in this case MCS-I). The RF showed a remarkable ability to detect MCS-I events. Over 99% of the 550 observed MCS-I events were detected to within 50 km. However, this high detection rate came with a tendency to issue false alarms either because of premature warning of an MCS-I event or in the continued elevation of RF forecast likelihoods well after an MCS-I event occurred. The skill of the RF forecasts was found to increase with the number of trees and the fraction of positive events used in the training set. The skill of the RF was also highly dependent on the types of predictor fields included in the training set and was notably better when a more recent training period was used. The RF offers advantages over high-resolution NWP because it can be run in a fraction of the time and can account for nonlinearly varying biases in the model data. In addition, as part of the training process, the RF ranks the importance of each predictor, which can be used to assess the utility of new datasets in the prediction of MCS-I.

1. Introduction

Because of their large size, intensity, and longevity, mesoscale convective systems (MCSs) impact society in many ways: public safety (flash flooding); wind farm energy generation, above ground transmission of electricity, and cellular communication towers (severe wind events); agricultural practices (water usage); and safe and efficient air travel (turbulence, wind shear, hail). Better forecasts of MCSs will lead to more advanced public warning of severe weather (Stensrud et al. 2013), improved ability to protect wind farm assets from extreme winds (Mahoney et al. 2012), improved response time for energy and communications infrastructure repairs due to damage caused by MCSs, and improved airline safety and air traffic efficiency by routing aircraft around potential MCS initiation events (Colavito et al. 2011, 2012; Robinson 2014).

Operational high-resolution numerical weather prediction (NWP) models with advanced data assimilation, such as the High Resolution Rapid Refresh (HRRR; Benjamin et al. 2014), are beginning to show promise in providing skillful forecasts of MCSs. Advances in the assimilation of radar reflectivity have improved the initialization of existing MCSs in NWP models, but predicting the timing and location of MCS initiation remains a particularly vexing problem (e.g., Clark et al. 2007, 2014; Pinto et al. 2015; Trier et al. 2014, 2015).
Very short-term predictions of the initiation of an MCS (MCS-I) requires a high-resolution depiction of the evolving stability, shear profile, and potential forcing mechanisms such as surface boundaries or elevated propagating waves (e.g., Jirak and Cotton 2007; Houze 2004). High-resolution models with advanced data assimilation can provide a three-dimensional estimate of the evolving environment, but imperfections in the model and poorly constrained errors in temperature and moisture mean that NWP predictions of MCS-I are still prone to a great deal of uncertainty (Pinto et al. 2015).

Statistical techniques (e.g., linear regression, k-nearest neighbor, analogs, neural networks, random forest, and genetic algorithms) can operate on data much more quickly than a human analyst, enabling the rapid digestion of frequently updating datasets (e.g., surface mesonets, radar, satellite) along with NWP models as often as new data arrive. In this study, we evaluate the utility and predictive skill of a random forest (RF) at predicting MCS-I. The RF technique is still relatively new to most meteorologists yet has shown promise in several other complex weather prediction applications, as described below.

Statistical models have long been a part of weather forecasting. For example, model output statistics (MOS) based on multiple linear regressions are routinely used to compensate for systematic model biases and to generate reliable probabilistic forecasts of precipitation, cloud cover, and other variables (Glahn and Lowry 1972). Analog statistical techniques identify similar past weather patterns and give probabilistic projections based on the observed evolution of those past patterns (Hamill and Whitaker 2006; Delle Monache et al. 2013). The tropical weather community uses statistical models to predict the probability of tropical cyclogenesis, rapid intensification, and eyewall replacement cycles (Rozoff and Kossin 2011; DeMaria and Kaplan 1994). Marzban et al. (2007) used neural networks to predict cloud ceiling and visibility, and Coniglio et al. (2007) used logistic regression to predict MCS maintenance based on vertical wind and stability profiles. More recently, Roebber (2015) used evolutionary programming techniques to generate probabilistic forecasts of minimum surface temperatures.

In past studies, the skill of the RF has been shown to vary with implementation and application. Prior to its use in meteorology, the RF statistical technique was used successfully in biomedical research to select and classify genes relevant to diseases (e.g., Diaz-Uriarte and de Andrés 2006). More recently, the RF approach was used to diagnose regions of atmospheric turbulence due to convection from radar and satellite observations and NWP model data (Williams et al. 2007, 2008c; McGovern et al. 2011; Williams 2014). Williams et al. (2008a,b) showed how RFs could be used to predict areas where convective storms were likely. Gagne et al. (2009) compared the RF technique to a host of other machine learning algorithms, and found it to be better than all other algorithms at classifying radar-based storm type. Another comparative study described by Lakshmanan et al. (2010) found that RF had a slight edge over competing artificial intelligence learning techniques in classifying storm type. Hall et al. (2011) found that the RF was one of the best algorithms in terms of overall skill metrics for short-term clear-sky forecasts, although its underconfidence (Wilks 2006, p. 288) made it statistically less reliable than other statistical data mining techniques. Recently, Gagne et al. (2014) used RF to add skill to an ensemble of storm-scale precipitation forecasts, while Mecikalski et al. (2015) found RF performed slightly worse than logistic regression in forecasting small-scale convection initiation with NWP and geostationary satellite data.

In this paper, we demonstrate how RFs can be trained to predict the very challenging forecast problem of large-scale convective storm initiation (MCS-I), following an approach similar to that used by Williams (2014) for predicting atmospheric turbulence. Section 2 introduces the input predictor fields and our quantitative definition of MCS-I. Section 2 also describes the predictor selection process and documents the improved skill resulting from expanding the predictor list from a small set of NWP model fields to a combination of smoothed NWP output and observations. The sensitivity of prediction skill to various RF parameters is explored in section 3. Section 4 shows case studies to demonstrate the value that the RF technique offers when compared to the individual constituent data sources. Finally, in section 5, the results are summarized and presented along with a discussion of the strengths and weaknesses of the technique.

2. Methodology

The RF data mining technique requires the definition of a forecast variable of interest, or predictand, and a set of predictor fields that are thought to be related to the predictand. For this study, the predictand is the binary variable representing whether or not MCS-I occurred at a given time and location, and the predictors are derived from radar reflectivity, satellite data, and NWP model output. The RF was trained using data collected during June–August (JJA) of 2011 and evaluated on data from the summer of 2013 to provide a stringent test of its ability to capture MCS-I even when the NWP model changes. The datasets,
predictand, and RF methodology are described in detail below.

a. Datasets

The model diagnostics used to train the RF came from the HRRR (Benjamin et al. 2014). The HRRR is a convection-permitting model run over the entire continental United States (CONUS) with hourly cycling and 3-km grid spacing. The 2011 version of the HRRR gained information on the location of existing storms indirectly via the three-dimensional variational data assimilation (3DVAR) of radar reflectivity into the 13-km Rapid Refresh (RAP) model, which was used to initialize the HRRR forecasts. In 2013, the HRRR was updated to include direct assimilation of radar reflectivity into its 3-km grid (Benjamin et al. 2014). This change notably improved the performance of the HRRR, particularly its ability to capture existing MCSs (Pinto et al. 2015). As a result, training the RF on 2011 HRRR data and testing it on 2013 HRRR data demonstrates whether or not the RF is robust for use with different NWP model analysis systems.

Extrapolated radar observations started with composite reflectivity provided by the National Mosaic and Multi-Sensor Quantitative Precipitation Estimation (NMQ) system from the National Severe Storms Laboratory (Zhang et al. 2011). This product merges multiple radar volumes into a 3D grid with 1-km spacing in the horizontal and 0.5-km spacing in the vertical and then derives 2D fields such as composite reflectivity.

Satellite observations came from the Geostationary Operational Environmental Satellite system (GOES), operated by the National Environmental Satellite, Data, and Information Service (NESDIS). Brightness temperature in the longwave IR channel (10.7 μm) was subtracted from the CO₂ channel (13.3 μm) to yield a satellite brightness temperature difference (SBTD) field. The SBTD has been shown to distinguish between growing cumulonimbi and low cumulus or thin cirrus (Mecikalski and Bedka 2006), so it is useful to delineate areas of growing cumuli that may consolidate into an MCS. Thin cirrus and shallow cumulus have brightness temperature differences of less than −25°C, while developing storms exhibit values between −25° and −5°C, and difference values near zero indicate mature cumulonimbi. Mecikalski and Bedka (2006) and Mecikalski et al. (2008) included SBTD as one of the components of their satellite-based convection initiation algorithm.

The radar reflectivity and SBTD were interpolated onto the HRRR 3-km grid using bilinear interpolation. These fields were then advected to their expected downstream locations at later times based on the motions detected in the vertically integrated liquid water (VIL) field from the Corridor Integrated Weather System (CIWS; Evans and Ducot 2006; Dupree et al. 2009).

b. Definition of MCS initiation

National composites of VIL that are available from CIWS were used to identify MCSs following the method described in Pinto et al. (2015). As in Pinto et al. (2015), in this study we define MCSs as consisting of an area of VIL exceeding 3.5 kg m⁻² with a horizontal extent of at least 100 km (allowing gaps of up to 10 km). These conditions must be met for at least two consecutive tops of the hour. While not essential to the conclusions of the paper, the criteria that have been adopted to classify MCSs are similar but not identical to those used in many prior studies (e.g., Geerts 1998; Houze 2004; Coniglio et al. 2010). The lifetime threshold was set relatively low to ensure an adequate sample size for developing the training dataset using data obtained for a limited time period. Larger-sized storms of longer duration are much less frequently occurring (e.g., Davis et al. 2006) and, therefore, would require a longer period from which to draw an adequate number of representative cases.

Once the MCS definition is satisfied, the area spanned by the core area of high VIL is dilated by 125 km as shown in Fig. 1 to define the MCS region. VIL is used to detect MCSs instead of radar reflectivity because it is relatively insensitive to brightband contamination and anomalous propagation artifacts (e.g., Smalley and Bennett 2002). VIL also includes the integrated effect of hydrometeors at all vertical levels, making its intensity more closely related to convective vigor than a single level of radar reflectivity.

After MCSs are identified for each time, they are checked to see if they qualify as an initiation event (MCS-I). To qualify, an MCS must be at least 125 km removed from any previously existing MCS that was present during the previous 2 h, and it must persist for at least 1 h. MCS-I is evaluated only at the top of each hour, when HRRR model forecasts are valid, and an MCS-I event occurs only in the first hour that a temporally and spatially isolated MCS is identified. A detailed description of the MCS-I identification algorithm is given in Pinto et al. (2015). The data points around the MCS-I are used in the RF training set as positive events, while all others are nonevents. The expansion of the MCS region accounts for potential offsets or timing errors between the observed MCS-I and the environmental conditions as represented by the model. This increases the number of training data points that go into the RF and allows for some positional error in a forecast.

c. Random forest algorithm

A decision tree is a common tool in machine learning (Breiman et al. 1984; De’ath and Fabricius 2000;
Dattatreya 2009), and an RF is an ensemble of weakly correlated decision trees (Breiman 2001). Collectively, the trees function as an ensemble of “experts,” each casting a vote for whether MCS-I will occur (e.g., Fig. 2). All of the nodes of a decision tree can be reduced to simple rules of the form: if predictor $P$ is $x$ or less (where $x$ is any number), then follow branch A; otherwise, follow branch B. A predictor may be used at multiple nodes in the same tree. Each branch will either lead to another node or terminate with a “yes” or “no” vote. When a decision tree is being...
trained, the algorithm finds a predictor and a threshold that “splits” the training data instances that reach a node into two subsets in a way that maximizes the homogeneity of the subsets with respect to the predictand, for example, by minimizing the “Gini impurity” (Breiman 1996).

RF trees differ from conventional decision trees in that each RF tree is trained on a bootstrapped sample of the training cases (illustrated by Fig. 3). Additionally, at each node of the tree, only a limited randomly selected subset of predictors is chosen as candidates for splitting, whereas standard decision trees consider all predictors as candidates. (The predictor candidates are selected randomly with replacement, so that any predictor may be a candidate at any node.) An implication of bootstrapping is that roughly one-third of the training cases are not used for any given tree, and these “out of bag” cases are used as test cases to quantify the importance of each predictor field. Bootstrapping the training cases and ignoring some predictors at each node make individual trees weaker, but these steps also ensure the trees are not strongly correlated with each other. Thus, the forest is less susceptible to overfitting the peculiarities of the training set, and can provide probabilistic information. The number of trees and number of predictors chosen as candidates for splitting at each node are tunable parameters to which predictive performance sensitivity is tested below.

The RF has several advantages over other data mining techniques. For one, the empirical model created from the RF ensemble does not require the predictors to be monotonically related to the predictand, meaning that it can represent a variety of functional relationships. Alternative techniques like logistic regression are inherently linear. The decision trees in the RF are also human readable, such that relationships between data and how they were used to predict can be explored. In addition to their predictive capabilities, RFs can rank the importance of individual predictors (Breiman 2001; Topi´c et al. 2014). A predictor’s importance is quantified by scrambling its values in the out-of-bag training cases for each tree and seeing how much the classification accuracy of the RF goes down. For example, the expected importance of a random variable is zero. The importance value is often scaled by dividing it by a quantity akin to its standard error (e.g., see supplemental material for Díaz-Uriarte and de Andrés 2006). Importance scores provide a helpful starting point for comparing the potential contributions of different variables and selecting a small but skillful subset of predictors.

1) TRAINING

The RF is trained to use predictors available at a given time to forecast the occurrence of MCS-I 2 h in the future. To create the RF training suite, predictor values and associated MCS-I “truth” values from 2 h later were interpolated onto a cylindrical grid covering 25°–48°N and 125°–67°W with horizontal spacing of about 4 km (0.036° latitude and 0.038° longitude). The geographical coverage of data points used in the training suite is shown in Fig. 1. Points over the Atlantic Ocean and parts of Canada and Mexico that are beyond the WSR-88D radar network coverage were not included. The analysis was done using data available at the top of each hour. Over the 3-month period, from June through August 2011, there were over 200 million potential data points. Even though there were many cases to choose from, most of them were null events (no MCS-I). Even in the most MCS-I-prone geographical regions in the United States, MCS-I events occur only 3% of the time (Pinto et al. 2015). The average MCS-I frequency for the entire domain is only 0.3%. This rarity makes MCS-I a difficult
predictand for any statistical forecast algorithm to handle. One can achieve 99.7% accuracy by always predicting a null event at every grid point. To help the RF algorithm discriminate between MCS-I and non-MCS-I cases, the MCS-I cases are oversampled such that they make up 30% of the training set. This artificial increase in the proportion of events in the training set can be accounted for in the RF vote calibration phase.

The RF parameter sensitivity tests and predictor importance analyses were conducted using 10 disjoint training sets: 5 sets of 18,000 cases each were drawn randomly without replacement from odd days, and another 5 sets were drawn similarly from even days. The standard deviation of skill over these 10 training sets provides a means for assessing the relative significance of differences in mean skill score when the RF parameters are changed. While one standard deviation is not a particularly stringent requirement, there is only a 22% chance that the mean of 10 samples will differ by more than one standard deviation from the mean of another 10 samples drawn from the same population. Selecting the sets from even or odd days also permits testing a model trained on even days against independent data from odd days, and vice versa.

In general, one wants as many training cases as possible to fully sample the general population of weather scenarios. On the other hand, given finite resources, one must limit the number of cases. We found that 18,000 cases allowed for efficient training of the RF while fully sampling the parameter space. This number of cases is actually quite large compared to other recent studies. For example, McGovern et al. (2011) successfully trained an RF to predict atmospheric turbulence with only 2055 cases, and Mecikalski et al. (2015) predicted the onset of radar reflectivity $Z_{>35}$ at the $-10^\circ$C level (i.e., a formal definition for convective initiation) with only 9015 cases. While the number of cases is relatively large, it is important to note that a higher number of training examples is often needed when the event is rare, so that both verifying classes are sufficiently sampled.

### 2) PREDICTOR SELECTION

As a proof of concept, the RF was first trained with a select group of diagnostic fields obtained from the HRRR. Later, the value of adding observational predictor fields was explored. Diagnostic output from the HRRR included 17 two-dimensional fields deemed to be relevant for the prediction of MCS-I (Table 1). Environmental factors that contribute to the development of MCSs are discussed in Houze (2004). Undoubtedly, there are other fields that may be derived from the full three-dimensional HRRR dataset that would have potential for adding value to the prediction of MCS-I (e.g., vertical wind shear), but for simplicity we limited our training sets to fields available within the HRRR two-dimensional data stream. In addition, local solar time was added as a predictor field as a simple way to account for differing mechanisms responsible for daytime and nocturnal MCS-I.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Description</th>
<th>Unit</th>
<th>Mean selection count</th>
</tr>
</thead>
<tbody>
<tr>
<td>PWAT_EATM</td>
<td>Precipitable water in model column</td>
<td>kg m$^{-2}$</td>
<td>50.4</td>
</tr>
<tr>
<td>PRES_SFC</td>
<td>Surface pressure (a proxy for terrain height)</td>
<td>hPa</td>
<td>49.0</td>
</tr>
<tr>
<td>Local solar time</td>
<td>UTC hour + ('E)/15 h$^{-1}$</td>
<td>h</td>
<td>45.9</td>
</tr>
<tr>
<td>REFC_EATM</td>
<td>Max reflectivity in model column</td>
<td>dBZ</td>
<td>40.8</td>
</tr>
<tr>
<td>TSOIL_SFC</td>
<td>Soil temperature at the surface</td>
<td>K</td>
<td>38.3</td>
</tr>
<tr>
<td>CAPE_SFC</td>
<td>CAPE of surface parcel</td>
<td>J kg$^{-1}$</td>
<td>34.8</td>
</tr>
<tr>
<td>HPBL_SFC</td>
<td>Height of planetary boundary layer</td>
<td>m</td>
<td>28.3</td>
</tr>
<tr>
<td>RH_HTGL</td>
<td>2-m relative humidity</td>
<td>%</td>
<td>28.0</td>
</tr>
<tr>
<td>DZDTH1H_SIGL</td>
<td>Vertical velocity in 900–700-hPa layer</td>
<td>m s$^{-1}$</td>
<td>26.0</td>
</tr>
<tr>
<td>ACP1Hr_SFC</td>
<td>Accumulated model precipitation</td>
<td>mm</td>
<td>24.6</td>
</tr>
<tr>
<td>DSWRF_SFC</td>
<td>Downward shortwave radiation flux at surface</td>
<td>W m$^{-2}$</td>
<td>23.2</td>
</tr>
<tr>
<td>ULWRF_NTAT</td>
<td>Upward longwave radiation flux at top of atmosphere</td>
<td>W m$^{-2}$</td>
<td>22.0</td>
</tr>
<tr>
<td>SHTFL_SFC</td>
<td>Sensible heat flux at surface</td>
<td>W m$^{-2}$</td>
<td>20.9</td>
</tr>
<tr>
<td>LHTFL_SFC</td>
<td>Latent heat flux at surface</td>
<td>W m$^{-2}$</td>
<td>20.0</td>
</tr>
<tr>
<td>DPT_HTGL</td>
<td>2-m dewpoint</td>
<td>K</td>
<td>18.4</td>
</tr>
<tr>
<td>×4LFTX_SPDLa</td>
<td>Best (four layer) lifted index</td>
<td>K</td>
<td>15.9</td>
</tr>
<tr>
<td>SPFH_HTGLa</td>
<td>Specific humidity</td>
<td>g kg$^{-1}$</td>
<td>15.2</td>
</tr>
<tr>
<td>CIN_SFC</td>
<td>CIN of surface parcel</td>
<td>J kg$^{-1}$</td>
<td>9.3</td>
</tr>
</tbody>
</table>

aThese two predictors were removed after the mean selection count analysis for reasons described in the text.
2D data stream that we thought might be of value in the prediction of MCS-I was evaluated. Hall et al. (2011) also note that while the RF was designed to effectively utilize large numbers of predictors, it can be susceptible to noise from extraneous or redundant features. To reduce the number of predictors used in the training sets, we implemented a method that systematically determines which predictor fields should be retained, allowing some of the correlated predictor fields to be eliminated. As will be discussed below, choosing which predictors to retain depends on the entire set of predictor fields under evaluation. This is particularly true for RFs since by utilizing decision trees that split on multiple predictors in succession, an RF captures and exploits relationships between the predictors.

A predictor selection trial was performed using a series of two forward selection steps and one backward elimination step. At each forward selection step, all unselected predictors were tested individually as candidates for retention by joining them to the predictors already selected and evaluating the resulting RF’s predictive skill on an independent testing set. The predictor whose inclusion made the RF most skillful was retained for the next step. After two forward selection steps, all the retained variables were tested to see which one’s removal caused the smallest drop in the RF’s skill (backward elimination). The predictor associated with the smallest drop in skill was then removed from the retained variable group and added back into the group of unselected predictor fields. This process was repeated until all 18 variables were retained. Each predictor selection trial was repeated 10 times with the different training sets, with training on odd days and testing on even days and then vice versa. Figure 4 summarizes the results obtained using 10 trials. After step 1, model reflectivity (REFC_EATM) was retained for 9 out of 10 trials, and model precipitable water (PWAT_EATM) was retained once. After step 2 the most frequently retained variables were model reflectivity (9 out of 10 trials) and model precipitable water (9 out of 10 trials), but model surface pressure (PRES_SFC, which is indicative of terrain height) and model lifted index (34LFTX_SPDL) were also retained once. The average number of steps per trial for which a predictor was retained is given in Table 1. The results suggest that the presence of a deep column of water vapor is important for MCS-I given that model precipitable water was the most frequently retained predictor (50.4 steps per trial). Fixed parameters such as solar time and surface pressure, which is a proxy for terrain height, were also retained quite often, indicating the importance of temporal and geographic regimes. On the other hand,
model convective inhibition (CIN_SFC) was retained least often (9.3 steps per trial). It is unclear as to why CIN_SFC seems to have lower importance in the training set; however, it is retained owing to previous reports of its importance in the prediction of MCS-I (e.g., Jirak and Cotton 2007).

3) SCORING AND EVALUATION

The primary objective measure used to assess the performance of the probabilistic RF 2-h MCS-I forecasts was the area under the receiver operating characteristic (ROC) curve (AUC; Marzban 2004). ROC curves (Fig. 5) are obtained by finding the relationship between the hit rate [Hits/(Hits + Misses)] and false alarm rate [False_Positive/(False_Positive + Correct_Null)] for a range of RF vote thresholds. AUC has a long history in evaluating machine learning algorithms. The ROC curve maps hit rate as a function of false alarm (FA) rate across a range of thresholds available within the prediction (e.g., RF vote counts or likelihood values). An AUC value of one is indicative of a perfect forecast while an AUC value of 0.5 is indicative of a purely random forecast. We also used the Gilbert skill score, commonly known as the equitable threat score (ETS), as a second metric to evaluate the RF forecasts. In this case, we took the maximum ETS value over all RF vote thresholds. Both AUC and maximum ETS can be used to compare RF performance to that of other forecasts, even if they have different units or are calibrated differently (Wilks 2006). Finally, we used the symmetric extreme dependency score (SEDS) to evaluate and intercompare the performance of the RF and other short-term forecasts in the real-time prediction of MCS-I observed during a 5-week period in 2013. The SEDS score, which is described by Hogan et al. (2009), is an equitable skill score designed to more effectively evaluate the performance of forecasts of infrequently occurring events such as MCS-I.

d. Predictor field optimization

In the first optimization step, redundant predictors were removed in order to reduce the amount of unnecessary information going into the training set. Highly correlated or redundant predictors like CAPE and lifted index or 2-m dewpoint and 2-m specific humidity were compared. It was found that the better variable to use depended on the number of variables in the training set. Lifted index was selected more often than CAPE when the suite was limited to five or fewer predictors (before step 15 in Fig. 4), but for more than five predictors, CAPE was selected more often, meaning that it was more valuable in combination with the other predictors in the larger set. Likewise, specific humidity was preferred when the number of predictor variables was small, but dewpoint worked better when a greater number of predictors were used. Since our final predictor suite has a larger number of predictors, CAPE and dewpoint were retained.

In the second set of predictor optimization experiments, the impact of predictor field smoothing was explored. Each of the remaining HRRR forecast fields was smoothed with circular filters with radii ranging from 10 to 80 km. It was found that using a 40-km circular smoothing filter resulted in the best skill scores. Figure 5 shows ROC curves obtained for RF predictions that were based on HRRR data only at raw resolution versus those obtained using a 40-km circular smoothing filter. There is no overlap between the 10 curves obtained using raw resolution and those obtained using a 40-km circular smoothing filter. The average AUC increased from 0.84 to 0.86, and the maximum ETS increased from 0.33 to 0.37 (Table 2). Both increases were large relative to the standard deviation across the 10 training sets, further indicating the significance of this result.

The final predictor optimization step was designed to assess the impact of observation-based variables to the skill of the RF forecasts. The value of adding radar reflectivity and 13.3–10-μm SBTD was assessed both
individually and in combination. These two fields were added to include information regarding the location and spacing of cloud and precipitation areas that have yet to reach the size threshold required to be classified as an MCS. For consistency, these fields were also smoothed using a 40-km circular filter. To account for storm motion, these fields were extrapolated to their expected locations 2 h later to be consistent with the corresponding HRRR forecast fields. Adding smoothed SBTD had little impact on the skill of the RF forecasts whether added alone or in combination with radar reflectivity, while adding radar reflectivity resulted in a significant increase in skill (Table 2). This increase in skill associated with including radar reflectivity as a predictor field is comparable to that obtained by smoothing the model data. Despite the failings of the SBTD, this field was retained because of the value found in other studies (e.g., Mecikalski and Bedka 2006; Mecikalski et al. 2015).

3. Sensitivity to RF parameters

We tested the sensitivity of the RF performance to several parameters that control aspects of the training. These parameters include the size of the forest (the number of trees), the percentage of positive events in the training set, and the number of candidate variables to use for splitting at each tree node. A forest with more trees will generally be more skillful than one with fewer trees, because it can accommodate more of the nuances of the training set. However, there comes a point when the rate of improvement with more trees is negligible. Using the datasets described above, forests were trained with sizes ranging from 4 to 500 trees. The AUC and maximum ETS affirm that more trees lead to better scores (Fig. 6). However, the improvement slows greatly after about 50 trees. The mean scores for the 50-tree forests were within one standard deviation of the mean scores for the 500-tree forests. This pattern of diminishing returns with greater number of trees is similar to that found by McGovern et al. (2011). Henceforth, 200 trees are used in all the forests.

The RF skill was found to be fairly insensitive to the number of candidate predictors used for splitting at each node. By default, the Topić et al. (2014) software uses the integer value of the square root of the total number of predictors for this parameter. With 19 total predictors, 4 would be the default. Our analysis reveals that using fewer predictors was slightly better (Fig. 7). The best AUC was achieved with two predictors and the best maximum ETS was with three predictors (Fig. 7). Most of the ±1 standard deviation ranges overlap, so in any case, the results are not overly sensitive to this RF parameter. For the rest of our experiments, splitting of the RF at nodes is done using two predictors.

The AUC and maximum ETS of the RFs were most sensitive to the ratio of events to nonevents in the training set. Williams (2014) alluded to the importance of rebalancing the proportion of events to nonevents in the training set when trying to predict very rare events. Results of our sensitivity analysis indicate that the best ratio was between 20% and 40%. The best AUC was achieved with 40% events, and the best maximum ETS was achieved with 30% events (Fig. 8). It is clear that using an event ratio of 5%, which is closest to the climatological frequency of occurrence of MCS-I over the entire domain (0.3%), resulted in the worst performance.

4. Evaluation and case studies

Based on these sensitivity experiments, we used a training set that consisted of 30% MCS-I events selected from the JJA period in 2011 to train a 200-tree RF to make 2-h forecasts of MCS-I in real time. A new RF-based MCS-I forecast was issued every hour. The predictive skill of these forecasts was evaluated over the period 11 June–5 August 2013, inclusive. Vote counts were converted to interest or likelihood values using a simple linear transform: \( p = V/200 \), where \( p \) is the likelihood of an MCS-I event and \( V \) is the vote count. Only forecasts for which a complete set of predictors was available were evaluated, resulting in a total of 654 forecasts during the evaluation period.

\[ \text{Table 2. Skill scores for predictor optimization experiments.} \]

<table>
<thead>
<tr>
<th>Predictors</th>
<th>AUC</th>
<th>Std dev of AUC</th>
<th>Max ETS</th>
<th>Std dev of ETS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw HRRR (17) + local solar time (LST)</td>
<td>0.84</td>
<td>0.004</td>
<td>0.33</td>
<td>0.008</td>
</tr>
<tr>
<td>Smoothed HRRR + LST</td>
<td>0.86</td>
<td>0.003</td>
<td>0.37</td>
<td>0.011</td>
</tr>
<tr>
<td>Smoothed HRRR + LST + SBTD</td>
<td>0.87</td>
<td>0.003</td>
<td>0.38</td>
<td>0.010</td>
</tr>
<tr>
<td>Smoothed HRRR + LST + RadarREFC</td>
<td>0.88</td>
<td>0.003</td>
<td>0.40</td>
<td>0.007</td>
</tr>
<tr>
<td>Smoothed HRRR + LST + RadarREFC + SBTD</td>
<td>0.88</td>
<td>0.004</td>
<td>0.40</td>
<td>0.010</td>
</tr>
</tbody>
</table>

\[ ^1 \text{Note that the predictions are not actually probabilities since no attempt has been made to calibrate the predicted likelihood values.} \]
Because of the uncertainties associated with the prediction of convection initiation and the processes responsible for the upscale growth of convective storms into an MCS, the probabilistic nature of the RF forecasts is advantageous compared to deterministic forecasts, such as those provided by either extrapolated reflectivity or a single HRRR forecast. While the likelihood values obtained using RF are not inherently calibrated, the values can still be used in a relative sense. No attempt has been made to calibrate the RF forecasts since the relative variations in the RF likelihood field alone were found to be highly useful; however, this could be done using the approach described in Williams (2014).

The relative performance of four different forecast techniques was assessed using the ROC diagram, AUC, and the SEDS. Skill scores were accumulated from 34 days during the evaluation period for which all forecast datasets (extrapolated reflectivity, HRRR composite reflectivity forecasts, and RF-based MCS-I likelihood forecasts) were available.

\footnote{Note that 4-h forecasts of composite reflectivity from the HRRR are evaluated in this study to account for a 2-h latency of the HRRR forecast products used in the RF predictions.}
The ROC curves shown in Fig. 9 indicate the ability of forecasts to discriminate between events and nonevents. To generate ROC curves from the deterministic HRRR reflectivity and extrapolated reflectivity, the hit rate and FA rate were obtained at 5-dBZ intervals for thresholds ranging from 0 to 65 dBZ. The skill of each method is compared with that obtained using an “informed” climatology as the forecast. The informed climatology was obtained by grouping MCS-I occurrences observed across the eastern two-thirds of the United States during 2011 into two periods [day hours (1200–2300 UTC) and night hours (0000–1100 UTC)]. This aggregation was necessary to build a useful regional climatology from a single summer of data. ROC curves where obtained from the MCS-I climatology using MCS-I frequencies ranging from 0 to 0.05 with an interval of 0.005.

As can be seen in Fig. 9 and Table 3, the RF outperforms the other forecast methods. While the RF AUC values are much lower than those obtained when the training and verification truth are both from the same year (0.76 versus 0.88 from Table 3 and Table 2, respectively), the AUC values obtained for the RF MCS-I forecasts made in 2013 are much higher than those obtained with the other forecast methods. The RF forecasts also have the highest SEDS and the highest hit rate for all FA rates of 0.1 or greater. The RF forecasts also have the highest SEDS and the highest hit rate for all FA rates of 0.1 or greater. The RF forecasts are also clearly more skillful than using an informed climatology; however, the relative pickup in skill over the informed climatology is seen to be regionally dependent, with the RF skill pickup being much greater for forecasts made over the Great Plains (GP) region than those made over the Southeast region. Also of note, the

![Graph of AUC for Number of Candidates Used for Splits](image)

![Graph of Max. ETS for Number of Candidates Used for Splits](image)}
RF was somewhat more skillful at predicting daytime MCS-I (AUC, SEDS = 0.77, 0.21) than nighttime MCS-I (AUC, SEDS = 0.75, 0.17), indicating the differing nature/predictability of surface-based and nocturnal (more often elevated) MCS-I.

Further, more detailed manual evaluation of RF performance using slightly less stringent verification (i.e., allowing for some displacement error) revealed that using a RF-based likelihood threshold of 0.1 detects all but 1 of the 550 observed MCS-I events to within 50 km. That is a hit rate of 99.8%! However, this impressive statistic and the RF’s ability to achieve higher hit rates come at the expense of a tendency toward FAs. For example, using the ROC diagram in Fig. 9, it is seen that a hit rate of over 70% can be achieved using RF, but at the expense of the FA rate exceeding 30%.

Reasons for RF’s tendency to FAs are described below using a couple of representative case studies (Figs. 10 and 11). In these figures, forecasts obtained using RF, HRRR reflectivity, and extrapolated

---

**Fig. 8.** (top) Average AUC and (bottom) maximum ETS as a function of the event percentage in the training set. These were tested with a 200-tree RF using two candidate predictors for splitting at each node. The AUC peaks at 0.878 with an event percentage of 40%, and the maximum ETS peaks at 0.407 when the event percentage is 30%.
reflectivity are compared with observed ongoing MCS events (black contours) and instantaneous MCS-I events (magenta contours). The contours given in these figures provide a 125-km extension surrounding each observed MCS and MCS-I event, indicating the size of the region considered positive in the training set. The case shown in Fig. 10 provides an example of simultaneous MCS-I events that occurred around 1800 UTC and spanned regions with vastly different MCS formation mechanisms. The timing of the MCS-I event observed over the southeastern United States on this day was fairly typical (e.g., Geerts 1998), but the MCS-I event occurring over the high plains was unusually early compared to climatology (Carbone and Tuttle 2008), as it was triggered in an area of moderate instability (CAPE $\sim 1500 \text{ J kg}^{-1}$) through the interaction between a stationary front and an old outflow boundary. The MCS-I events observed over the Florida panhandle and Mississippi were well forecasted by both the RF (with likelihoods greater than 0.7) and the HRRR composite reflectivity forecast (areas of 35 dBZ exceeding 100 km in length). Both the HRRR and the RF forecasts provide a much weaker indication of MCS-I in northwestern Kansas with elevated RF likelihood values that peak around 0.4 and the HRRR-forecasted reflectivity being too low to be considered an MCS.

A multistorm MCS-I event over the high plains is shown in Fig. 11. In this case, a long line of convection formed between 2000 and 2100 UTC along a cold front/dryline in an area with no previously existing radar echoes, as evidenced by the lack of extrapolated reflectivity (Fig. 11d). The RF forecast had likelihood values of between 0.05 and 0.20 in the approximate location of the observed MCS-I and with the correct orientation. However, these values were much lower than those routinely obtained for RF-based forecasts of MCS-I in the southeastern United States. The HRRR predicted a broken line of convective cells with the correct orientation but the storm cells were too far apart (i.e., the distance between grid points with 35 dBZ—alogous to VIL of 3.5 kg m$^{-2}$—was greater than 10 km) for this area of convection predicted by the HRRR to be considered an MCS (Fig. 11d). The RF was able to determine that storms larger than those indicated in the HRRR reflectivity forecast were possible in 2 h. In addition, MCS-I likelihoods obtained from the RF forecast issued 1 h later (Fig. 11c) increased dramatically (to over 0.5), “catching up” to the MCS-I event 1 h before it actually occurred (i.e., providing an indication of

![Fig. 9.](image-url)  
Fig. 9. (top) ROC curves for 2-h random forest predictions (red), 4-h forecasts of composite reflectivity from HRRR (black), MCS-I climatology (cyan), and 2-h extrapolations of VIL (green). Data were obtained for the period 12 June–5 August 2013 and were matched for availability. Skillful forecasts lie above the dotted 1:1 line. (bottom) As in the top panel, but zoomed in on the dashed area.

### Table 3. Hit rate, AUC, and SEDS obtained for times in which all forecasts were available between 11 Jun and 5 Aug 2013. In column headers, eUS stands for eastern United States, SE stands for Southeast, and GP stands for Great Plains.

<table>
<thead>
<tr>
<th></th>
<th>RF eUS</th>
<th>RF SE</th>
<th>RF GP</th>
<th>HRRR reflectivity eUS</th>
<th>HRRR reflectivity SE</th>
<th>HRRR reflectivity GP</th>
<th>Extrapolated reflectivity eUS</th>
<th>Extrapolated reflectivity SE</th>
<th>Extrapolated reflectivity GP</th>
<th>Observed climatological MCS-I frequency eUS</th>
<th>Observed climatological MCS-I frequency SE</th>
<th>Observed climatological MCS-I frequency GP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hit rate for FA rate $= 0.1$</td>
<td>0.31</td>
<td>0.27</td>
<td>0.28</td>
<td>0.24</td>
<td>0.21</td>
<td>0.23</td>
<td>0.20</td>
<td>0.14</td>
<td>0.19</td>
<td>0.22</td>
<td>0.21</td>
<td>0.12</td>
</tr>
<tr>
<td>AUC</td>
<td>0.76</td>
<td>0.73</td>
<td>0.75</td>
<td>0.59</td>
<td>0.57</td>
<td>0.60</td>
<td>0.55</td>
<td>0.52</td>
<td>0.53</td>
<td>0.61</td>
<td>0.65</td>
<td>0.52</td>
</tr>
<tr>
<td>Max SEDS</td>
<td>0.19</td>
<td>0.17</td>
<td>0.18</td>
<td>0.14</td>
<td>0.13</td>
<td>0.14</td>
<td>0.11</td>
<td>0.06</td>
<td>0.11</td>
<td>0.15</td>
<td>0.13</td>
<td>0.03</td>
</tr>
</tbody>
</table>
increased confidence in the likelihood of an MCS-I event). The increase in RF likelihood values between successive RF forecasts issued before an observed MCS-I event happened a number of times during the evaluation period, indicating that the trend in the RF likelihood forecasts may provide additional information that could be used by forecasters to ascertain whether or not an MCS may initiate soon.

While the RF was able to capture nearly every MCS-I event, this forecast tool requires forecaster intervention because of the high FA rate. It turns out that the RF forecasts have three failure modes as listed in Table 4. The most common cause of these FAs (class 1) results from the inability of the RF to distinguish between an MCS-I event and an ongoing MCS. Detailed analyses reveal that ongoing MCSs are nearly always coincident with RF likelihoods of greater than 0.5. Examples of the RF’s tendency to remain elevated for ongoing MCSs are evident within the black contours (previously existing MCS locations) of Figs. 10 and 11. Oftentimes, the RF likelihood values will remain elevated up to 3 h after the MCS has dissipated, further exacerbating the FA problem. This failure mode can easily be recognized by forecasters, who thus may ignore these elevated RF values in areas of ongoing or recently decayed MCSs.

The second most common cause for FAs (class 2) results from the RF’s tendency to predict MCS-I earlier than observed. An example of this type of FA is shown in Figs. 10a and 10c for the two MCS-I events that occurred in the Southeast. In this case, the RF forecast valid 1 h prior to the observed MCS-I events observed over Mississippi and Florida exhibited likelihood values of up to 0.7 (Fig. 10a), rising to over 0.8 at the observed time of MCS-I (Fig. 10c). While this type of forecast bias leads to FAs, it could also be used constructively by providing forecasters early warning of areas worth further exploration. The third type of FAs (class 3) seen in the RF MCS-I forecasts occurs in areas where convective storms are observed, but do not reach MCS size criteria. An example of class 3 FAs is evident in Fig. 10, where an arc of elevated RF likelihood values extends from the Gulf Stream across eastern Georgia and the western portions of the Carolinas. Convective storms are evident in eastern Georgia, but fail to reach MCS

![Figure 10](image_url)
size criteria. The HRRR model forecast had small-scale storms throughout this region that, when coupled with the other HRRR predictor fields and SBTD, yielded MCS-I likelihood values exceeding 0.6 in this region. In this case, the RF could not discriminate the environmental conditions responsible for upscale growth from those that inhibit upscale growth. Nonetheless, the warning information could be evaluated by a forecaster who, in turn, may decide if the possibility of MCS-I warranted more attention or not.

The analyses discussed above clearly indicate that the RF-based MCS-I forecasts add value over the HRRR model forecasts in the short term. It does this by effectively combining available data (both observations and model data) to predict the likelihood of an MCS-I event. Key features of the RF technique include its ability to remove biases in each predictor field and to form complex nonlinear relationships between the predictors as part of the training process, thereby condensing a great deal of data into a single probabilistic forecast that can

TABLE 4. Classification of FA in RF forecasts of MCS-I.

<table>
<thead>
<tr>
<th>Class</th>
<th>Description</th>
<th>Comment</th>
<th>Region of most common occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>RF probabilities are always elevated in areas of ongoing MCSs</td>
<td>Use RF along with current radar observations to disregard these areas</td>
<td>Entire domain</td>
</tr>
<tr>
<td>2</td>
<td>RF probabilities are often elevated in forecasts valid up to 2 h prior to the MCS-I event</td>
<td>RF provides an early indication of regions where MCS-I is possible in the next 2–4 h</td>
<td>Southeastern United States</td>
</tr>
<tr>
<td>3</td>
<td>Elevated RF probabilities can occur in areas where convective storms form but fail to reach MCS size criteria</td>
<td>Reflects the difficulty of predicting upscale growth of storm cells into clusters large enough to be considered MCSs in weakly forced environments</td>
<td>Southeastern United States</td>
</tr>
</tbody>
</table>
be used as guidance for the short-term prediction of discrete high-impact events like the initiation of MCSs. Predicting the exact timing and location of an MCS-I event is an extremely challenging problem due to the discrete nature of the predictand and the complex, nonlinear, and somewhat chaotic processes responsible for the development of MCSs. Thus, the success of the RF on this problem suggests that it should be broadly applicable.

5. Summary and conclusions

An RF data mining technique was used to objectively rank a set of predictor fields and evaluate their potential to predict MCS-I. Predicting the initiation of MCSs is an extremely challenging forecast problem owing to its discrete nature (i.e., occurring at a specific instant in time) and infrequency (occurring in about 0.3% of the sample points across the eastern two-thirds of the United States during the period June–August in 2011). As a proof of concept, the RF was trained to predict MCS-I using a set of 2D fields that are available from the HRRR model. An iterative method for selecting which variables have the most predictive skill was described. It was found that precipitable water was the most useful predictor of MCS-I, with local solar time and surface pressure (i.e., terrain height) ranked highly as well. Interestingly, soil temperature also ranked very highly, while 2-m moisture variables were found to be less useful. In addition, it was found that CAPE was a good predictor of MCS-I while CIN was not. It is not clear why the model-derived CIN provided little in the way of predictive skill in nowcasting MCS-I. One possible explanation is that surface-based CIN calculations underestimate the true large-scale stability of the atmosphere, especially during daytime hours when a superadiabatic layer often exists at the surface.

Adding extrapolated radar reflectivity to the set of predictors significantly increased the RF’s skill, while adding SBTD did not. It is believed that the radar reflectivity helped to capture the more slowly evolving MCS-I events that occur in the southeastern United States. It was somewhat surprising that the SBTD did not improve the skill of the RF forecasts, as a recent study by Mecikalski et al. (2015) indicated it had value in nowcasting convective initiation (albeit at much smaller scales and shorter lead times than explored in this study). It is also possible that the CIWS motion vectors were not as well suited for extrapolating SBTD as they were for radar reflectivity. Further studies are needed to assess the utility of the full range of satellite measurements in the forecasting of MCS-I, but such work is beyond the scope of this paper.

The sensitivity of RF forecast skill to several tuning parameters was explored. Results were most sensitive to the fraction of events to nonevents in the training set. Our best results came when 30% of the training set consisted of MCS-I events, which is 100 times the climatological frequency of 0.3%. The RF skill increased with more trees, but there was definitely a point of diminishing return. A forest size of 200 trees was found to have AUC and ETS values that were roughly 99% of those obtained for a 500-tree forest. The best number of candidate variables to split on was 2 or 3, depending on the verification metric. It should be noted that the optimal number of candidate split variables will depend on the number and type of predictor fields used in the training set.

Case studies were used to demonstrate the strengths and weaknesses of the RF in the prediction of MCS-I. The probabilistic RF forecasts captured nearly all of the 654 MCS-I events observed during the 6-week evaluation period timed to the closest hour and to within 50 km. In many cases, the RF was able to detect MCS-I events that were not explicitly predicted by the deterministic HRRR forecast used as input to the RF. The RF is able to do this by accounting for biases in the model and by developing nonlinear relationships between the HRRR-based predictor fields and the two observational inputs. While the RF was able to detect a large percentage of the MCS-I events observed during the evaluation period, it also produced a larger than optimal FA rate. The largest class of FA (termed class 1) was when RF forecasted likelihoods remaining high in the vicinity of existing MCSs. While this high FA rate contributed to overprediction of MCS-I (i.e., a high bias), these forecasts could be automatically masked out of an operational product with existing MCSs. Class 2 FAs happened when elevated RF forecast likelihoods occurred prior to the observed MCS-I event by 1–2 h. However, this feature could be considered a strength of the RF MCS-I forecasts by providing advanced notice for the potential of an MCS-I event.

The basic process of training and optimizing the RF was discussed here; however, a number of additional pre- and postprocessing steps could be employed to further enhance performance. Both terrain and time of day ranked highly in terms of importance as predictors in the training set. Thus, one area of future research would be to assess the value of developing separate training sets by region and time of day. It was also found in comparing Figs. 5 and 9 that the skill of the RF increases notably when using more recently obtained training data. The ROC curves obtained in Fig. 5 were obtained using training data from the even days of JJA 2011 to make forecasts on the odd days of JJA 2011,
while the ROC curves shown in Fig. 9 were obtained using RF’s trained using 2011 data and used to make forecasts employing 2013 data. In fact, an ideal approach for operational use would be to retrain the RF each day using the latest available datasets. To do this, one would have to determine the ideal length for the period of recent data used to train the RF. There are trade-offs one must consider in doing this: one would like the training period to be long enough to capture the full range of conditions that lead to the event occurring, while at the same time, one would like the training period to be short enough for the RF to respond to changes in the skill of the predictor fields (e.g., as a result of changing NWP models or evolving weather regimes). This might be accomplished by sampling a training set more heavily from instances that occurred near the current Julian date, and more from the current convective season than from previous years.

Finally, if desired, the RF forecast likelihood fields can be calibrated by relating the forecast categories to observed frequencies; however, because of the high biases described above, the calibration process would necessarily reduce the dynamic range of likelihood values.

The findings presented herein, along with the positive results of Mecikalski et al. (2015) of the prediction of small-scale storm initiation, and Williams et al. (2008c) and Williams (2014) in the diagnosis of turbulence, demonstrate the potential benefit of using RF techniques for difficult nowcasting problems that require analysis and interpretation of large amounts of data in a short amount of time to predict discrete high-impact events. As such, the RF represents a class of data mining techniques that can be used to digest the ever-increasing wealth of observational and model data into a single probabilistic product that alerts forecasters to the possibility of an impending high-impact event that warrants further attention.

Acknowledgments. This research is in response to requirements and funding by the Federal Aviation Administration (FAA). Partial support also came from the National Science Foundation. The views expressed are those of the authors and do not necessarily represent the official policy or position of the FAA or NSF. We thank Drs. Stan Benjamin, Curtis Alexander, and Steven Weygandt of NOAA/GSD for providing the HRRR data and Dr. Haig Iskenderian of MIT/LL for providing access to the CIWS VIL data used to generate the MCS-I truth dataset and the CIWS motion vectors used to extrapolate the satellite and radar reflectivity to the forecast valid time. The authors also thank Dr. Stan Trier (NCAR/MMM) and three anonymous reviewers for insightful reviews and constructive comments that helped improve the manuscript.

REFERENCES


