EVALUATION OF A PROBABILISTIC CONVECTIVE NOWCAST FOR COSPA

D. Ahijevych[1], J. Williams, S. Dettling, H. Cai, and M. Steiner

National Center for Atmospheric Research (NCAR)[2], Boulder, CO

1 INTRODUCTION

This paper describes the evaluation of a probabilistic convective nowcast system developed as part of the Federal Aviation Administration's (FAA) Collaborative Storm Prediction for Aviation (CoSPA; Wolfson et al. 2008, Pinto et al. 2010) effort. Probabilistic convective forecasts are still in their infancy—the research community spends significant efforts toward development and calibration of probabilistic forecasts, while also trying to establish what kind of probabilistic information should be provided. We offer a viable alternative to a spatial smoothing of the extrapolation field or compositing time-lagged model ensembles (two of the methods that have been attempted). Our experimental forecast system is based on random forests (RFs), a data mining technique that uses sets of decision trees trained to nowcast areas of convective weather based on a given set of predictor fields. Individual trees in the forest "vote" on the prediction, and the number of votes is mapped to the likelihood, or probability, that a storm intensity threshold will be exceeded. These probabilities have been evaluated using dichotomous and probabilistic scoring methods and shown to compare favorably to other benchmark forecasts at one and two hour lead times.

One of the probabilistic forecasts used as benchmark was the Localized Aviation MOS Program (LAMP) thunderstorm guidance (Ghirardelli 2005). This product was used both as an independent forecast and as a component of the RF.

The RF was relatively skillful at predicting NWS Video Integrator and Processor (VIP) intensity level exceedance when tested on data from the summer 2009. Based on a sliding probability threshold, the RF's maximum threat score or critical success index (CSI), true skill score (TSS), receiver operating characteristic (ROC) curves, and other dichotomous evaluation scores all exceeded the equivalent scores from LAMP, simple extrapolation, and CoSPA forecasts. The Brier Skill Score also showed that the RF produced skillful probabilistic forecasts of VIP Level exceedance.

One reason the RF surpasses simple extrapolation is that the RF prediction is not completely dependent on the most recent radar field and motion vectors. It incorporates environmental stability fields and satellite trends, and thereby has the potential to predict new areas of convective initiation away from current storms. Examples will be shown in two case studies.

2 DATASET AND METHODOLOGY

2.1 *RANDOM FOREST (RF) TECHNIQUE*

The RF technique used in the present study (Breiman 2001) is a powerful, non-linear statistical analysis or machine learning method that has previously proven useful for the problem of diagnosing regions of atmospheric turbulence that may be hazardous to aviation (Williams et al. 2007; Cotter et al. 2007; Williams et al. 2008). Essentially, RFs are ensembles of weak, weakly-correlated decision trees that "vote" on the correct classification of a given input. The use of an ensemble of such trees minimizes the risk of overfitting the training set, a significant and well-known problem with individual decision trees. In constructing each tree of an RF, one begins with a "training set" containing many instances of predictor variables along with an associated "truth" value (e.g., 0 or 1 depending on whether or not convective initiation did subsequently

[1] NCAR, Boulder, CO, 80307-3000, ahijevyc@ucar.edu
[2] NCAR is sponsored by the National Science Foundation.

occur at the given pixel in the next hour). A "bagged" training sample is selected by drawing a random subset of *n* instances from the *n*-member training set, with replacement after each draw. This means that, on average, each tree is trained on roughly 2/3 of the dataset, including duplicates. Then, at each node of the tree, a subset of only *m* randomly-selected feature variables are chosen as candidates for splitting, contrasting with the usual practice of choosing the best split from all the feature variables. A typical choice for *m* is the square root of the number of predictor fields. Because not all feature variables are used to train each tree, those not used for training (the so-called "out-of-bag" samples) may be used to evaluate the performance of that tree. This allows the RF training process to estimate the importance of each variable based on the degradation in classification performance when the variable's values are randomly permuted among the training instances. Using this technique, the feature variables may be ranked in order of their importance to the RF's performance, providing a helpful tool for performing selection of a minimal skillful set of predictors.

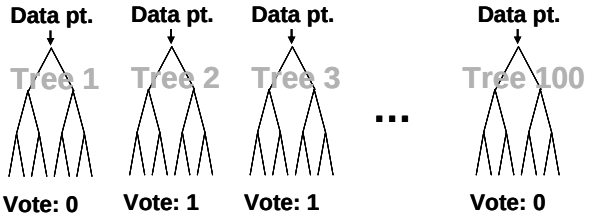


Fig. 1. Conceptual diagram of a random forest, an ensemble of weak, weakly-correlated decision trees that "vote" on the classification of each data point.

Once an RF has been trained, the trees function as an "ensemble of experts" to make predictions. For example, Figure 1 shows a conceptual diagram of a RF with 100 trees. When a new data point (or "feature vector": a set of predictor field values at the point for which the forecast is being made) is presented, each tree will perform a classification. These classification "votes" are then compiled, and can be used to derive a probability for each possible class. For example, if 40 trees vote "0" (no initiation) and 60 vote "1" (initiation), the 60% classification confidence for initiation may be scaled into a probability, as described in a later section.

2.2 *DATASET*

Skill scores are presented from forests that were trained and tested on data from 2009. The training and evaluation period covers 1 July – 19 August 2009.

The spatial domain for the RF training and test sets was over the eastern half of the conterminous U.S. as shown in Figure 2. All fields, both predictors and forecasts, were mapped to the same 0.04° (approximately 4-km) latitude-longitude grid.

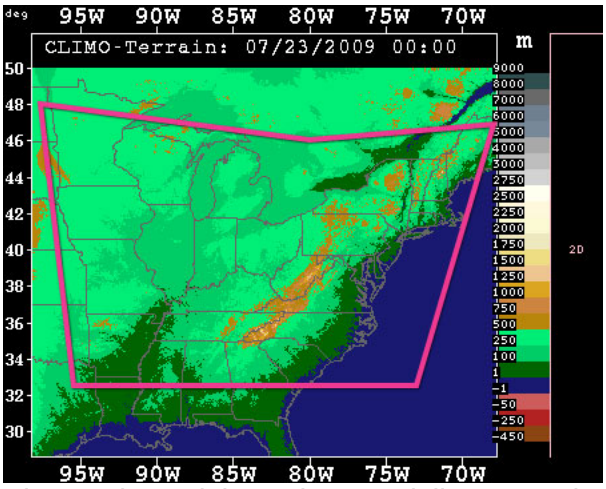


Fig. 2. The pink polygon delineates the domain used for this study. Terrain height [m] is in the background.

2.3 *TRUTH FIELD*

The truth fields, or the events that we try to predict, are VIP level exceedance. VIP levels are categorized intervals of vertically integrated liquid (VIL)[1] that range from VIP 1 to VIP 6. VIP level 1 is associated with very light precipitation and VIP level 3 is associated with moderate to heavy rain.

[1] The underlying field for VIP level in this study was VIL, which was derived from radar reflectivity as in Robinson et al. (2002).

Table 1 shows the range of VIL and composite reflectivity that corresponds to each VIP level. VIP level 3+ exceedance is generally considered a proxy for hazardous convective weather that should be avoided by airline pilots. The RF methodology has been applied to all 6 VIP thresholds, but we will focus on results for VIP 1 and VIP 3 for simplicity.

Table 1. Categories of VIP level with their equivalent intervals of VIL and radar reflectivity.

VIP level	VIL interval	Composite reflectivity interval
0	< 0.14 kg m^{-2}	< 18 dBZ
1	0.14-0.76 kg m^{-2}	18-30 dBZ
2	0.76-3.5 kg m^{-2}	30-38 dBZ
3	3.5-6.9 kg m^{-2}	38-44 dBZ
4	6.9-12 kg m^{-2}	44-50 dBZ
5	12-32 kg m^{-2}	50-57 dBZ
6	>32 kg m^{-2}	>57 dBZ

2.4 *PREDICTORS*

Part of the RF training process involves estimating the importance of the predictors, as described in Section 2.1. To isolate the most helpful predictors and reduce the computational overhead, only the predictors with significant importance were retained. Sets of similar predictors, such as the standard deviation of satellite-measured infrared (IR) radiance temperature within a 10 km radius and the standard deviation of IR radiance temperature within a 20 km radius, were reduced to the most important member. In this fashion, 300+ predictors were gradually winnowed down to a set of 35.

The LAMP thunderstorm product was used both as a predictor within the RFs and an independent probability forecast. LAMP is a statistical forecast system run at NCEP that uses multiple linear regression equations to update the Global Forecast System (GFS) Model Output Statistics (MOS). Along with temperature and precipitation grids, the LAMP produces probabilistic thunderstorm forecasts for two hour windows beginning at 1-3 h. For

Table 2. The RFs were trained with these 36 predictors. Additional descriptions of the individual fields may be found in Williams et al. (2008).

Predictor Fields for the Random Forest
LAMP TSTM 1-3 h forecast
Accumulated precipitation from last 3 h
Max. echo top height within 40 km
MIT/LL environmental stability mask
MIT/LL satellite peaks (max. within 40km)
MIT/LL growth/decay (max. within 40km)
MIT/LL growth/decay (standard dev. within 40km)
MIT/LL “air mass” storm indicator
MIT/LL weather type (22 types)
RUC13 relative humidity (900-700mb average)
RUC13 most unstable CAPE
RUC13 filtered frontal likelihood
RUC13 700-200 mb mean U and V wind (2 fields)
RUC13 low-level wind shear (975-725 mb)
RUC13 deep wind shear (1000-350 mb)
RUC13 mid-level lapse rate (700-350mb temp. change)
NRL cloud classification
NCAR cloud classification
Satellite IR radiance 13.3 μm
Satellite IR radiance 11 μm (min. within 40 km)
Satellite IR radiance 11 μm
Satellite water vapor radiance 6.7μm (min. within 40km)
Satellite IR radiance 3.9 μm
Satellite radiance difference (11 μm -13.3 μm)
Satellite radiance difference (11 μm -6.7 μm)
Satellite visible albedo
Distance to VIP level 1-4 (4 separate fields)
Max. VIP level within 40 km
Standard dev. of VIP level within 40 km
Max. VIP level within 20 km
NSSL radar reflectivity smoothed with 5km filter
Local solar time

LAMP, a thunderstorm is defined as one or more lightning strikes in a 20-km grid box

(Charba and Samplatsky, 2009). While the LAMP thunderstorm forecast is not designed to predict VIP level exceedance, and its forecast time window is not strictly equivalent to the 2-h lead time for the RF, we offer it as a loose benchmark for performance.

Table 2 lists 36 fields that were used to train the RF.

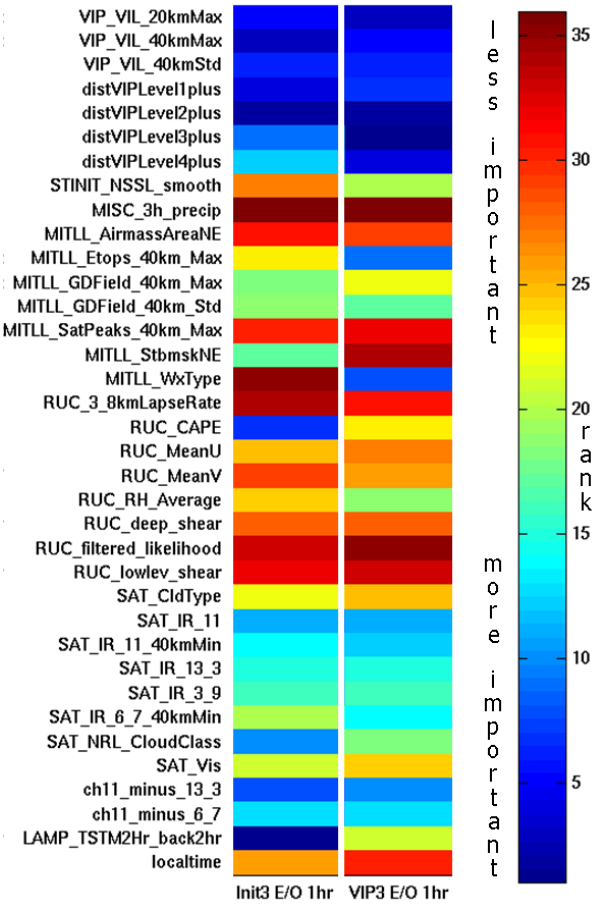


Fig. 3. Colored importance ranks for the RF predictors in the convective initiation regime (first column) and growth/decay regime (second column). Blue denotes a lower rank, and thus higher importance. Red denotes a higher rank and lower importance. Below the first column the label "Init3 E/O 1hr" signifies VIP level 3 initiation; the RF was trained on even Julian days, while odd Julian days were held out for testing; and the lead time is 1 h. The meaning of the other label is the same except "VIP" stands for the growth/decay regime.

2.5 *CONVECTIVE REGIMES*

Although a single RF is capable of predicting VIP level for all regimes, it is useful to train separate RFs in regions of isolated convective initiation and regions that already have existing storms. These two regimes are sufficiently different that training two separate forests is more effective. This notion was supported by the RF predictor importance analysis, which showed very different rankings for predictors in the near-storm environment and in the far-storm environment (Fig. 3). For example, convective available potential energy (RUC_CAPE) is colored blue in the convective initiation (CI) regime and yellow in the growth/decay regime. This indicates a very high importance in the convective initiation regime versus the growth/decay regime. The same relation holds true for the LAMP thunderstorm product (LAMP_TSTM2Hr_back2hr). It is one of the top predictors in the CI regime, but is in the middle of the pack for the growth/decay regime. It makes sense meteorologically for the RF to rely more on broad environmental predictors like CAPE for the CI regime than for the growth/decay regime. The RF in the growth/decay regime can rely on the location of existing convective weather, but the CI forest cannot. On the other hand, the current precipitating weather type (MITLL_WxType) is irrelevant to convective initiation, so its importance is very low (dark red). However it is very important in the growth/decay regime (blue).

A distance threshold of 40 km was used to separate the near-storm from far-storm environments. If a pixel exceeded VIP 3 at a target time, then it was considered convective initiation if its position was at least 40 km away from any other VIP 3+ pixels at the beginning of the time window. To obtain the position of the pixel at the beginning of the time window, the storm was advected backwards in space using MIT/LL storm motion vectors (Chornoboy et al., 1994; Wolfson and Clark, 2006). In the present paper, the near-storm environment is called the growth/decay regime and the far-storm

environment is the convective initiation regime.

2.6 *RESAMPLING AND CALIBRATION*

In order to cross validate and get a sense of the robustness of our statistics, the data were divided into three mutually exclusive subsets. Then the pixels from even Julian days were separated from the odd Julian days. One set was used for training and the other for testing. The opposite was done also—odd Julian days were used for testing and even Julian days were used for training—so that a total of six subsets were available for cross validation. Thus, for each forecast problem, there were six evaluations.

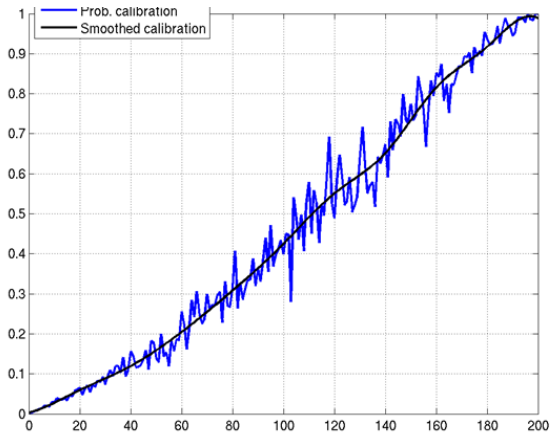


Fig. 4. Calibration curve for the VIP 1+ exceedance at 1 h lead time in the growth/decay regime. The observed frequency of the event is shown as a function of the number of RF votes. Raw counts are shown in blue and the smoothed calibration curve is in black.

Computational resources limited the number of training samples (pixels) to something on the order of 60,000 pixels. This number was a compromise between having a large enough sample to be representative and finishing the training in a reasonable amount of time. Besides the pixel count, additional trees and predictor fields also increased the training time. To speed up the training process and increase the sensitivity of the trained RF, the data were selectively resampled so that the proportion of positive events (VIP level exceedance) in our training set was much higher than climatology. For example, the climatology based on June-August 2007 and 2008 showed an actual observed event frequency of 7% for VIP level 1+ and 0.8% for VIP level 3+. By resampling the data so that the training set had a larger proportion of events than the climatology, the RF was able to train more efficiently on the important events, maintaining better sensitivity. Sensitivity tests revealed that the final performance results were relatively insensitive to this ratio. A range of ratios from 5% event frequency to 50% was tested. The results presented in this paper come from forests trained with a 10% ratio of events to non-events in the training set.

After the forests were trained on the resampled data, their vote counts on the independent testing set were compared to actual observed event frequency so that reliable probability forecasts could be obtained. For example, Fig. 4 shows the relation between vote count and observed event frequency for VIP level 1+. This particular curve is for a 200-tree forest trained in the growth and decay regime for forecasts with 1 h lead time. The blue line shows the raw observed frequencies for each vote count and the smooth black curve shows the final fitted calibration curve, which may be used to map vote count to probability. For every 100 forecasts of 70% probability, one could expect the event to occur 70 times. In this way the RF forecast was trained to be statistically reliable. Similar calibration curves were also made for the convective initiation regime, 2 h lead times, and also for VIP level 3+. Unfortunately, there were so few convective initiation events for VIP level 3+ at 2 h lead time that a reasonable calibration curve could not be automatically fitted to the raw data. For this reason, the evaluation of the VIP level 3+ probability forecast was not available; for case studies, the calibration curve was fit manually.

2.7 *VERIFICATION*

This section describes the verification methods used to evaluate the deterministic and probabilistic nowcasts.

The standard contingency table scores such as probability of detection (POD), probability of false detection (POFD, or false alarm rate), probability of false alarm (POFA, or false alarm ratio), and bias are all defined in terms of the 2x2 contingency table (Wilks, 2006). There are four possible outcomes for each forecast: a hit, miss, false alarm, or correct null. If we let a denote hits; b false alarms; c missed events; and d correct negatives, then POD= $a/(a+c)$, POFD= $b/(b+d)$, POFA= $b/(a+b)$, and bias= $(a+b)/(a+c)$. Other scores are defined below. The threat score or critical success index (CSI) is

$$CSI = \frac{a}{a+b+c}$$

and the true skill statistic (TSS) is

$$TSS = \frac{ad-bc}{(a+c)(b+d)}.$$

CSI is the fraction of observed and forecasted events that were correctly forecasted, and TSS measures how well the forecast separates "yes" events from "no" events. Note TSS is equivalent to POD - POFD.

We begin with these scores because (a) they are familiar to most users and (b) they are directly applicable to the deterministic forecasts such as simple extrapolation and CoSPA, thereby facilitating a comparison with the RF and LAMP forecasts. Probabilistic measures will be discussed later. First, in order to convert the probabilistic forecasts (RF and LAMP) to binary yes/no forecasts, we calculated the scores at all probability thresholds. The maximum possible skill scores were used as summary statistics. Fig. 5 illustrates the maximum CSI for one of the 2-h RF forecasts. For low thresholds, the score suffers from many false alarms and for the high thresholds it suffers from a low hit rate. In between, the CSI reaches a maximum of 0.46, and TSS reaches 0.78. These statistics are summarized for all the forecasts in the Results section.

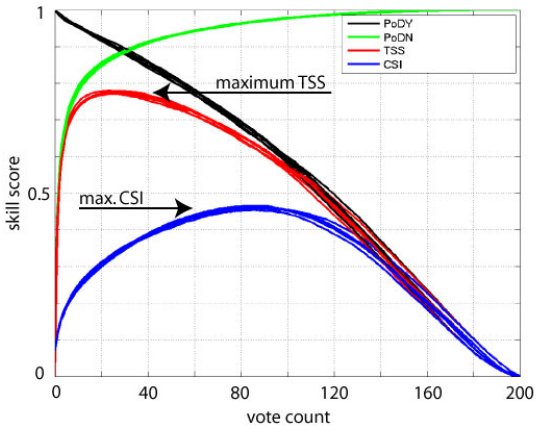


Fig. 5. Illustration of how max. CSI and max TSS were calculated from the probabilistic RF forecasts. This example happens to be for the RF designed for VIP level 1+, 2 h lead time, and the growth/decay regime. There is a curve for each of the six cross-validation subsets.

Probabilistic measures were also used to evaluate the forecasts. Reliability diagrams similar to Fig. 4, but with probability along the *x*-axis, were used to show whether forecasts in a particular probability bin actually coincide with the observed frequency of events. Ideally, the points in the reliability diagram should fall along the diagonal 1:1 line. Numerically, the performance of a probability forecast can be summarized by the Brier score (BS). It is the weighted average of the squared difference between the forecasted probability (f) and the binary observation (o; 0 or 1).

$$BS = \frac{1}{n}\sum_{k=1}^{n}(f_k - o_k)^2$$

The Brier Skill Score (BSS) uses the climatological event frequency as a baseline to assess a forecast. The BSS can be broken into three components consisting of (i) reliability, (ii) ability to resolve different periods with low and high probabilities, and (iii) closeness to a climatological probability of 50% (Wilks, 2006). Events that have a climatological probability close to 50% are

more difficult to forecast than events close to 0% or 100% because a guess based on the most likely outcome is less likely to be accurate by chance.

$$BSS = 1 - \frac{BS}{BS_{ref}}$$

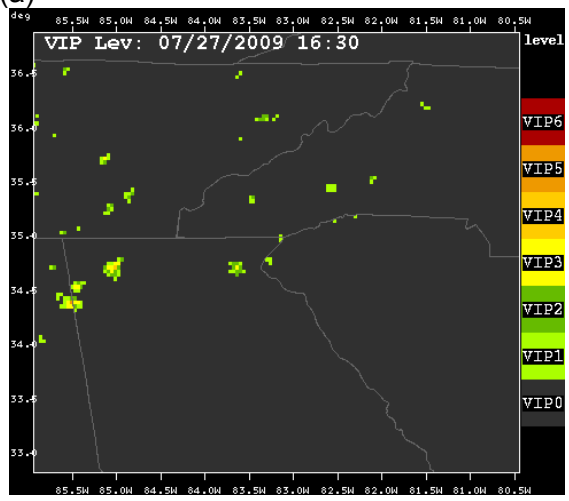
where BS_{ref} is the Brier Score using the same climatological event frequency for each probability forecast. BSS ranges from minus infinity to 1, with negative values associated with forecasts that are poorer than climatology, positive values with forecasts superior to climatology, and 1 with a perfect probability forecast.

The final metric we show is the area under the receiver operating characteristic, or ROC, curve. The ROC curve shows how the detection rate increases as a function of the false alarm rate (POFD). It illustrates the tradeoff between hits and false alarms. The area under the ROC curve, or AUC, ranges from 0 to 1, with 0.5 indicating a forecast with no skill and 1 indicating a forecast with perfect discrimination. Higher AUC is better.

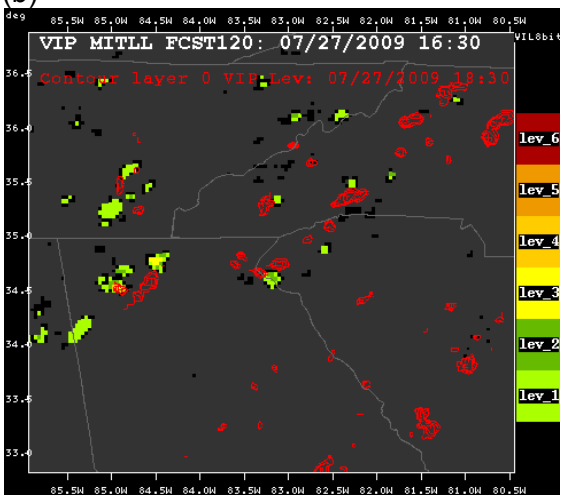
3 RESULTS

Before the overall statistics are presented, two case studies are offered for the CoSPA and RF forecasts.

(a)



(b)



(c)

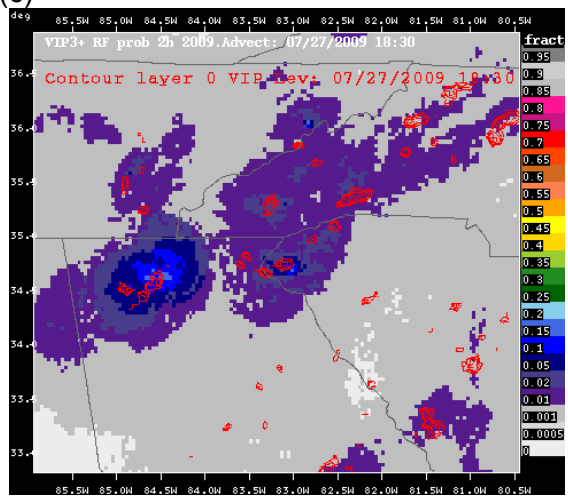


Fig. 6. Three panel plot over the southeast U.S. showing (a) VIP level observations at 1630 UTC, 27 July 2009, the issue time of the forecasts, (b) deterministic 2-h forecast from CoSPA and (c) 2-h forecast from the RF (probability of VIP 3+ exceedance). Overlaid in red are contours of VIP level at the valid time, 1830 UTC. The contours span VIP levels 3 through 6 in increments of 1. Many of the VIP level 3+ storms observed at 1830 UTC developed in the preceding two hours, and were not predicted by CoSPA.

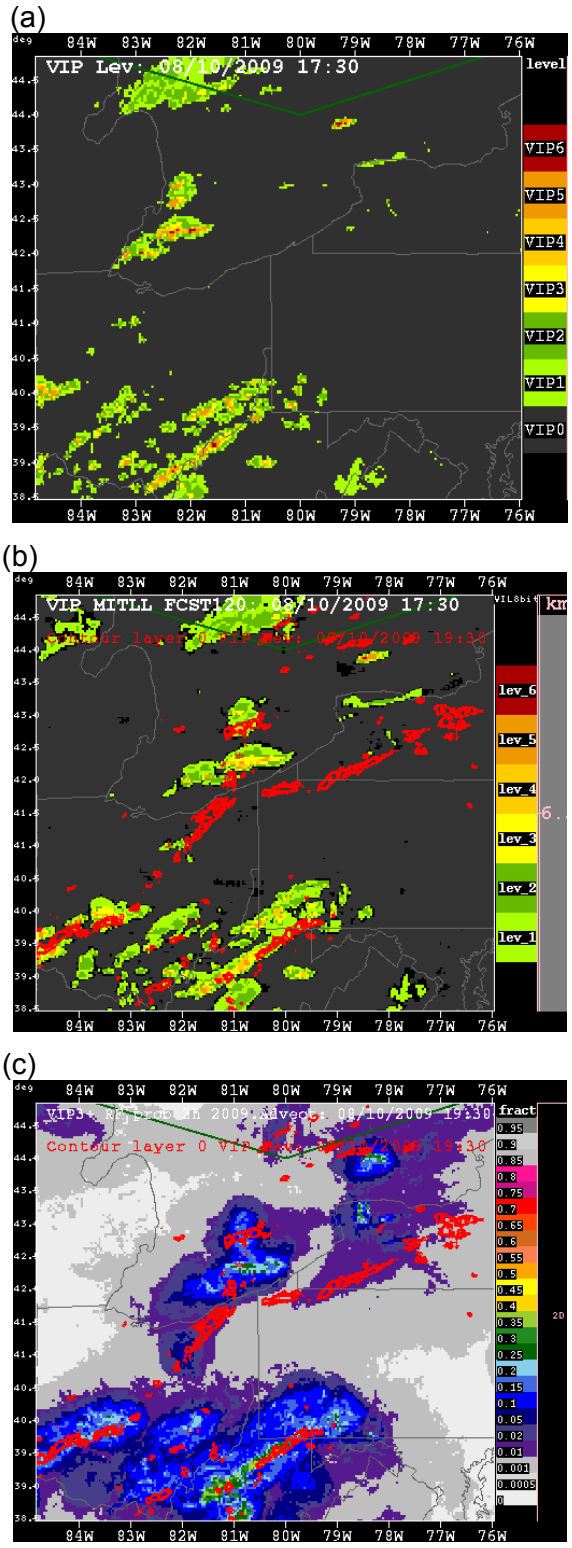


Fig. 7. Case study over Lake Erie. The first panel is the observed VIP level at 1730 UTC on 10 August 2009. Panels (b) and (c) show the 2-h forecasts from CoSPA and the RF, respectively, issued at the same time. The red contour overlay represents VIP level contours 3 through 6 at the valid time of the forecasts (1930 UTC).

3.1 *CASE STUDIES*

The first case study is during the early afternoon of 27 July 2009. The 2 h time window covers 1630-1830 UTC. Fig. 6 shows the VIP level observations at 1630 UTC (panel a). The 2-h forecast from CoSPA is shown in panel (b). It is overlaid with the observed regions of VIP level 3+ (red contours) at the valid time of the forecast (1830 UTC). The only pixels of VIP level 3+ that were correctly predicted were associated with the small storm in extreme north central Georgia. Based on CoSPA extrapolation, almost none of the new convective elements were predicted. There were several dozen storms that initiated from Tennessee to South Carolina, resulting in many missed events. On the other hand, the RF probability forecast for the same time (panel c of Fig. 6) shows elevated probability of VIP level 3+ in northeastern GA, western NC, western SC, and southeast SC where new storms were observed at the forecast valid time. Some of the storms were completely missed in central GA and SC, but many of them are consistent with localized peaks in the RF probability field.

The second case study is 1730-1930 UTC on 10 August 2009. The three panel figures follow the same fields as the first case study, this time over Lake Erie and surrounding states. The first panel shows the VIP level at 1730 UTC and the next panel shows the 2-h CoSPA forecast issued at the same time. The red contours in panels (b) and (c) show the actual areas of VIP level 3+ at 1930 UTC. There was a lot of convective initiation in northeast OH, western NY and PA. The CoSPA forecast had difficulty predicting that new convection, because the

storms did not exist at 1730 UTC (panel a); the RF forecast did somewhat better (panel c). The probabilities are particularly high near storms that already existed at the issue time, which is expected, but the RF forecast also shows isolated areas of higher probabilities (>1%) across northeast OH, western NY and PA, where new storms actually developed. There is certainly room for improvement in the RF forecast and more work must be done to sharpen the probability field. However, to the authors' knowledge, no other product offers a similarly reliable VIP level probability nowcast at this resolution.

Table 3. Scores from the growth/decay regime and VIP level 1+. BSS is not used for the deterministic forecasts, hence not applicable (NA) is used.

Forecast type	MaxCSI	MaxTSS	AUC	BSS
1h simple extrapolation	0.52 ±0.02	0.73 ±0.01	0.90 ±0.00	NA
1h CoSPA	0.53 ±0.01	0.74 ±0.01	0.88 ±0.01	NA
1h RF	**0.57 ±0.01**	**0.84 ±0.01**	**0.97 ±0.00**	**0.57 ±0.01**
2h simple extrapolation	0.39 ±0.01	0.61 ±0.02	0.84 ±0.01	NA
LAMP 1-3h (2hr)	0.17 ±0.03	0.24 ±0.07	0.62 ±0.05	-0.07 ±0.01
2h RF	**0.46 ±0.01**	**0.78 ±0.01**	**0.95 ±0.00**	**0.44 ±0.01**

The case studies illustrate how the RF probability forecast may provide uncertainty information during periods of rapid storm development when the CoSPA forecast needs it the most. The deterministic CoSPA forecast is based almost exclusively on extrapolation during the first 2 h, so new storms are invariably missed. The random forecast can produce a probability forecast that highlights potential areas of storm growth and initiation.

The following statistics demonstrate that the relatively good performance of the RF is not limited to one or two case studies, but is reflected in the overall verification scores as well.

Table 4. Scores from the convective initiation regime and VIP level 1+. There were too few CI events to calculate reliable BSS. The simple extrapolation scores were not available at the time of this publication.

Forecast type	MaxCSI	MaxTSS	AUC	BSS
1h simple extrapolation	NA	NA	NA	NA
1h CoSPA	0.006 ±0.001	0.02 ±0.01	0.51 ±0.01	NA
1h RF	**0.029 ±0.009**	**0.56 ±0.06**	**0.85 ±0.03**	NA
2h simple extrapolation	NA	NA	NA	NA
LAMP 1-3h (2hr)	0.016 ±0.002	0.20 ±0.02	0.60 ±0.02	NA
2h RF	**0.029 ±0.006**	**0.56 ±0.05**	**0.85 ±0.03**	NA

3.2 *SUMMARY STATISTICS*

This section quantifies the performance of the 2009 RFs alongside other forecast benchmarks such as simple extrapolation, LAMP, and the CoSPA forecast. The results are stratified by regime so that the skill of the forecasts in the convective initiation regime can be isolated from the growth/decay regime, which would otherwise dominate the statistics. In Table 3, we show forecast metrics from the growth and decay regime and an intensity threshold of VIP level 1+. In Table 4, we show results for the convective initiation regime and VIP level 1+. Tables 5 and 6 show the same respective skill scores, but for VIP level 3+. The best score for each lead time (1 and 2 h) is shown in bold typeface. As described in Section 2.6, the data were divided into six mutually exclusive cross validation subsets and the mean score for the six subsets is listed in the tables. The plus/minus value following each number is a

rough indication of uncertainty from 3 times the standard deviation of the six scores from the six cross validation sets.

Table 5. Scores from the growth/decay regime and VIP level 3+.

Forecast type	MaxCSI	MaxTSS	AUC	BSS
1h simple extrapolation	0.21 ±0.01	0.73 ±0.02	0.90 ±0.01	NA
1h CoSPA	0.23 ±0.01	0.74 ±0.02	0.90 ±0.01	NA
1h RF	**0.26 ±0.01**	**0.87 ±0.01**	**0.98 ±0.00**	**0.22 ±0.04**
2h simple extrapolation	0.12 ±0.01	0.56 ±0.03	0.82 ±0.01	NA
LAMP 1-3h (2hr)	0.09 ±0.00	0.48 ±0.06	0.74 ±0.06	-0.01 ±0.00
2h RF	**0.19 ±0.01**	**0.81 ±0.01**	**0.96 ±0.00**	**0.13 ±0.03**

Table 6. Scores from the convective initiation regime and VIP level 3+. The BSS is not available for the convective initiation regime at VIP level 3+ because the event was too rare for the automated calibration to give good results.

Forecast type	MaxCSI	MaxTSS	AUC	BSS
1h simple extrapolation	0.008 ±0.002	0.23 ±0.02	0.65 ±0.01	NA
1h CoSPA	0.010 ±0.003	0.21 ±0.04	0.61 ±0.02	NA
1h RF	**0.022** ±0.008	**0.75** ±0.02	**0.94** ±0.01	NA
2h simple extrapolation	0.008 ±0.001	0.23 ±0.02	0.65 ±0.01	NA
LAMP 1-3h (2hr)	0.008 ±0.001	0.39 ±0.10	0.69 ±0.07	NA
2h RF	**0.022** ±0.005	**0.75** ±0.02	**0.94** ±0.00	NA

4 DISCUSSION

The RF forecasts had the best CSI and TSS for both convective intensity thresholds and both lead times. In the growth and decay regime (Table 3), the CSI was 0.57 for the 1 h forecast and 0.46 for the 2 h forecast. This beats simple extrapolation (0.52) and CoSPA (0.53) at 1 h and simple extrapolation at 2 h (0.39). Due to a data archive issue at the time of publication, the 2 h CoSPA scores weren't available. These scores apply to the growth/decay regime for VIP level 1+ (Table 3). While the skills were much lower for convective initiation than for growth and decay, the relative ranking of the forecasts was consistent in Table 4. For the VIP level 3+ threshold, all scores were much lower than the VIP level 1+ scores, but the ranks were similar. A typical CSI drop was from 0.57 to 0.26 due to the greater difficulty of predicting convection at this higher threshold.

CoSPA and simple extrapolation are very similar, both in terms of the way they are constructed and their skill scores. CoSPA differs only in that it has a growth and decay component. Based on the performance results, this trending component did add some skill. CoSPA's CSI and TSS are modestly, but consistently, better than simple extrapolation (Tables 3 - 6).

The RF forecasts were the only probabilistic ones available at 1 h lead time, so they have the only BSS results. The other forecasts performance statistics are NA, or not applicable.

In terms of CSI and TSS, LAMP skillfully predicted VIP level exceedence in all regimes. But its scores were significantly lower than the other forecasts. This is because LAMP is not designed to provide a VIP level exceedence forecast for a given lead time; it is a thunderstorm guidance product valid over a time window. The probabilistic nature of the LAMP forecast, however, permitted calculation of a BSS. The BSS indicated negative skill (-0.01), which means less skill than a constant climatological

probability forecast. Admittedly, this is not a fair test for LAMP, but it does serve as a "strawman" benchmark for RF probabilistic nowcasts. LAMP still had useful information in it. Note the blue color, signifying high importance in the convective initiation regime (Fig. 3).

The RF analyses presented here were done offline and thus don't reflect a real time environment. Since the RFs were trained on archived data, they were able to use the model analyses, as opposed to the model forecasts. There is always a latency associated with the latest model run, so the best way to estimate the current conditions in real time is to use the best available forecast. It is unclear how much of an effect this had, but it might have helped the RF results. Future analyses should be performed using artificially-delayed model data to replicate the latency that is always present in real-time systems.

The RF forecasts of VIP level exceedance are heavily weighted toward low probability values and the areas are fairly broad except for the vicinity of previously existing convection. In some cases, the RF results look like a smoothed version of the extrapolation. However, through selected case studies and statistics it has been shown that the RF prediction is more skillful than a simple spatial smoothing of the deterministic forecast. Part of the lack of sharpness of the RF predictions is due to the lack of small-scale information on convective triggers. This deficiency may be remedied by including new fields, such as the SATellite Convection Analysis and Tracking system (SATCAST; Mecikalski and Bedka 2006; Iskenderian et al. 2010), and improved motion vectors in the training process.

As a reminder, reliable probabilistic nowcasts are still in their infancy, and users may have to expect significant uncertainty in them for some time to come. While there are certainly many avenues that can be explored and possibly lead to sharper RF forecasts, a properly calibrated probabilistic forecast should be expected to look quite different than a deterministic one. After all, there is a lot of inherent uncertainty in the deterministic forecasts, even though it may not be evident; such forecasts rarely get the convective details exactly right.

5 SUMMARY

In the summer of 2009, probabilistic nowcasts of VIP level 1 and VIP level 3 exceedance were developed for 1 and 2-h lead times utilizing the random forest (RF) technique. The RF nowcasts were evaluated by varying a probability threshold and scoring the resulting deterministic predictions alongside deterministic forecasts. They were also scored as probabilistic forecasts, enabling comparisons of reliability and resolution with the LAMP thunderstorm guidance. These evaluations demonstrate that the RF nowcasts perform comparatively well. RF probabilistic nowcasts have the potential to add value to CoSPA by providing an independent analysis of the input data along with valuable quantitative uncertainty estimates that could be essential to users.

ACKNOWLEDGEMENTS

This research has been in response to requirements and funding by the Federal Aviation Administration (FAA). The views expressed are those of the authors and do not necessarily represent the official policy or position of the FAA. Data and other support were provided by MIT/LL and NOAA/GSD. Thanks to Haig Iskenderian (MIT/LL) for help with references, and many others who contributed directly and indirectly to this work.

REFERENCES

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

Charba, J.P., and F.G. Samplatsky, 2009: Operational 2-H Thunderstorm Guidance Forecasts to 24 Hours on a 20-km Grid. *23rd Conference on Weather Analysis and Forecasting/19th Conference on Numerical Weather Prediction*, Amer. Meteor. Soc., 15B.5.

Cotter, A., J. K. Williams, R. K. Goodrich and J. Craig, 2007: A Random Forest Turbulence Prediction Algorithm. *5th Conference on Artificial Intelligence Applications to Environmental Science*, Amer. Meteor. Soc., 1.3.

Chornoboy, E. S., A. M. Matlin, J. P. Morgan, 1994: Automated Storm Tracking for Terminal Air Traffic Control, *Lincoln Laboratory Journal,* **7**(2), 427-448. http://www.ll.mit.edu/mission/aviation/publications/publication-files/journal-articles/Chornoboy_1994_JA-7198.pdf

Ghirardelli, Judy E., 2005: An Overview of the Redeveloped Localized Aviation MOS Program (LAMP) for Short-Range Forecasting. Preprints, *21st Conference on Weather Analysis and Forecasting*. Washington, D.C., Amer. Meteor. Soc., 13B.5.

Iskenderian, H., J. R. Mecikalski, K. M. Bedka, C. Ivaldi, J. Sieglaff, W. Feltz, M. M. Wolfson, and W. M. MacKenzie, 2010: Satellite data applications for nowcasting of convective initiation. *14th Conference on Aviation, Range, and Aerospace Meteor*., Atlanta, GA, Amer. Met Soc., 5.2.

Hsu, W.-R., and A. H. Murphy, 1986: The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *Int J. Forecasting*, **2**, 285-293.

Mecikalski, J. R. and K. M. Bedka, 2006: Forecasting convective initiation by monitoring the evolution of moving cumulus in daytime GOES imagery. *Mon. Wea. Rev.*, **134**, 49-78.

Pinto, J., J. Williams, M. Steiner, D. Albo, S. Dettling, W. Dupree, D. Morse, H. Iskenderian, T. Xiaofeng, M. Wolfson, C. Reiche, S. Weygandt, S. Benjamin, and C. Alexander, 2010: Advances in the Collaborative Storm Prediction for Aviation (CoSPA). Extended Abstracts, *14th Conference on Aviation, Range, and Aerospace Meteorology*, Atlanta, GA, Amer. Meteor. Soc., J11.2.

Robinson, M., J. E. Evans, and B. A. Crowe, 2002: En Route Weather Depiction Benefits of the NEXRAD Vertically Integrated Liquid Water Product Utilized by the Corridor Integrated Weather System, *10th Conf. on Aviation, Range, and Aerospace Meteorology,* Portland, OR, Amer. Meteor. Soc., 5.2.

Wilks, D.S., 2006: Statistical Methods in the Atmospheric Sciences. 2nd ed. Academic Press, 627 pp.

Williams, J. K., J. Craig, A. Cotter, and J. K. Wolff, 2007: A hybrid machine learning and fuzzy logic approach to CIT diagnostic development. *Fifth Conference on Artificial Intelligence Applications to Environmental Science*, San Antonio, TX, Amer. Meteor. Soc., 1.2.

Williams, J. K., D. A. Ahijevych, S. Dettling, and M. Steiner, 2008: Combining observations and model data for short-term storm forecasting. *SPIE Conference*, 7088, doi: 10.1117/12.795737, 12 pp.

Wolfson, M. M., D. A. Clark, 2006: Advanced Aviation Weather Forecasts, *Lincoln Laboratory Journal*, **16**(1), 31-58.

Wolfson, M. M., R. M. Rasmussen and S. G. Wolfson, M. M., R. M. Rasmussen and S. G. Benjamin, 2008: Consolidated Storm Prediction for Aviation (CoSPA). *13th Conference on Aviation, Range, and Aerospace Meteorology*, Amer. Meteor. Soc., J6.5.