USING SPATIOTEMPORAL RELATIONAL RANDOM FORESTS TO IMPROVE OUR UNDERSTANDING OF SEVERE WEATHER PROCESSES

AMY MCGOVERN[1], DAVID JOHN GAGNE II[2], NATHANIEL TROUTMAN[1], RODGER A. BROWN[3], JEFFREY BASARA[4], AND JOHN K. WILLIAMS[5]

Abstract. Major severe weather events can cause a significant loss of life and property. We seek to revolutionize our understanding of and our ability to predict such events through the mining of severe weather data. Because weather is inherently a spatiotemporal phenomenon, mining such data requires a model capable of representing and reasoning about complex spatiotemporal dynamics, including temporally and spatially varying attributes and relationships. We introduce an augmented version of the Spatiotemporal Relational Random Forest, which is a Random Forest that learns with spatiotemporally varying relational data. Our algorithm maintains the strength and performance of Random Forests but extends their applicability, including the estimation of variable importance, to complex spatiotemporal relational domains. We apply the augmented Spatiotemporal Relational Random Forest to three severe weather data sets. These are: predicting atmospheric turbulence across the continental United States, examining the formation of tornadoes near strong frontal boundaries, and understanding the spatial evolution of drought across the southern plains of the United States. The results on such a wide variety of real-world domains demonstrate the extensive applicability of the Spatiotemporal Relational Random Forest. Our long-term goal is to significantly improve the ability to predict and warn about severe weather events. We expect that the tools and techniques we develop will be applicable to a wide range of complex spatiotemporal phenomena.

Keywords: spatiotemporal data mining, statistical relational learning, severe weather, random forests

1. Introduction

The majority of real-world data, such as the weather data studied here, varies as a function of both space and time. For example, a thunderstorm evolves over time and may eventually produce a tornado through the spatiotemporal interaction of components of the storm. In this paper, we introduce and validate a greatly augmented version of the Spatiotemporal Relational Random Forest (SRRF) algorithm for use with severe weather data. The SRRF is a Random Forest (RF) [6] approach that directly reasons with spatiotemporal relational data and is a major contribution to the research in spatiotemporal relational models. Due to the increased complexity introduced by spatiotemporally varying data, most data mining algorithms ignore one or both of these aspects (e.g. temporal only relational models such as [13, 20, 34]) and our recent work is the only work that we know of that addresses both spatiotemporal and relational data [23, 38, 3].

Our work is motivated by and validated in three real-world earth science domains. The first is predicting thunderstorm-induced turbulence as experienced by aircraft, focusing on the continental United States. Such turbulence is inherently spatiotemporal, with thunderstorms causing increased turbulence on a short time scale in the local region around a storm and also on a longer time scale across a greater spatial extent. With this domain, our goal is to enhance the current operational products that provide turbulence prediction to aviation interests by improving the spatiotemporal

[1]School of Computer Science, University of Oklahoma, amcgovern@ou.edu, ntroutman@ou.edu
[2]School of Meteorology, University of Oklahoma, djgagne@ou.edu
[3]NOAA/National Severe Storms Laboratory, Rodger.Brown@noaa.gov
[4]Oklahoma Climatological Survey, jbasara@ou.edu
[5]Research Applications Laboratory, National Center for Atmospheric Research, jkwillia@ucar.edu.

reasoning of the models. Prior work demonstrated that RFs were a promising approach in the turbulence domain [44]. We are currently performing case studies of the SRRF and investigating the possibility of integrating the trained SRRFs into an operational turbulence guidance product.

Spatiotemporal data mining using the SRRFs can aid the development of effective turbulence predictions by uncovering and exploiting relationships between storm features and environmental characteristics that confirm and go beyond mechanisms that are currently understood by atmospheric scientists. In doing so, it has the potential to not only create practical predictive systems, but also to improve scientific understanding of turbulence.

The second domain is that of understanding and predicting tornadoes. The results presented in this paper are a piece of a larger overall project focusing on revolutionizing our understanding of tornadoes. In this paper, we look at the interaction of tornadic and nontornadic supercell thunderstorms with their environment as they moved across the state of Oklahoma over a 10 year period. Prior tornado research has found that 70% of strong tornadoes in 1995 were located within 30 km of a front [22]. The goal of this part of the project is to use SRRFs to perform a climatological study of tornadic supercell thunderstorms and how the variations in the surface environment affect them.

The National Oceanic and Atmospheric Administration's National Weather Service has a goal of developing Warn-on-Forecast capabilities by 2020, instead of the current warn on detection approach [36]. The Warn-on-Forecast concept hopes to increase the lead time of severe weather and tornado warnings by accurately predicting the time and location of severe storms using numerical models. Our data mining approach promises to identify those within-storm features that discriminate between storms that will produce tornadoes and those that will not. It can be directly used within the numerical modeling of storms and given to the weather forecasters who issue the warnings.

In the third domain, we study the progression of droughts across the Southern Great Plains for a 134 year period. Drought is a spatiotemporal phenomenon that operates on a very different time scale than tornadoes or turbulence. While those appear and disappear relatively quickly, drought takes months to years to progress. The goal with this work is to improve the prediction of drought through an improved understanding of how drought moves in each local region.

RFs [6] are a simple and powerful algorithm with a strong track record (e.g., [33, 26, 14, 4, 43]). RFs learn an ensemble of C4.5 [31] trees, each of which is trained on a separate bootstrap resampled dataset and using a different subset of the attributes. The power of the approach comes from the differences in the trees, which enable the forest to capture more expressive concepts than with a single decision tree. Since the trees are each trained on a different subset of the data, they can focus on different aspects of the overall classification problem. In addition to their predictive capabilities, one of the reasons that RFs are so popular is their ability to analyze the variables for their overall importance at predicting the concept.

We introduced a preliminary version of the SRRFs in [38] and preliminary results in these domains in [24]. This paper represents a significant extension as outlined below.

(1) The SRRF algorithm has been extended to address variable importance of spatiotemporal relational data. Since we are working directly with the domain scientists, the human interpretability of the models is critical. A single tree can be examined easily but an entire forest is more difficult to analyze, making the variable importance aspect crucial. This is a major extension from [38] and was reported in [24].

(2) We have further enhanced our ability to analyze the trees by enabling the domain scientists to examine the valid ranges of specific attribute values in each tree in the forest. For example, instead of a distinction that asks "Is there an updraft with maximum speed at least 15.3 m s^{-1}?", we provide the range of the attribute where performance is the same. This is an extension of our work from [24].

(3) Our underlying Spatiotemporal Relational Probability Tree (SRPT, [23]) algorithm has been considerably enhanced to improve the spatiotemporal distinctions. This gives us the ability to represent temporally and spatially varying fields within objects, which significantly augments our ability to mine and understand severe weather. The spatiotemporal fields were

in progress for the [24] paper and all of the results reported here use these fields. This is a significant extension of the work reported in [24].

(4) We thoroughly explore the parameter space of the algorithm on all of our domains. Due to space limitations, these results were not reported in [24] but they are included in their entirety here. We have also expanded the data sets used in the parameter space experiments.

(5) We have significantly extended the application to multiple real-world severe weather domains, which is a significant extension of the work reported in [38]. Although the domains remain the same as those presented in [24], all of the results and analyses are new to this paper.

2. Growing SRRFs

Growing a SRRF is very similar to the approach used to grow an RF [6] with a few critical changes required by the nature of the spatiotemporal relational data. Algorithms 1, 2, and 3 describe the learning process in detail. Before discussing these, we describe how we represent the spatiotemporal relational data for efficient model learning.

The data are represented as spatiotemporal attributed relational graphs, as we first presented in [23]. This representation is an extension of the attributed graph approach [29, 28, 19] to handle spatiotemporally varying data. All *objects*, such as aircraft, tornadoes, or hail cores, are represented by vertices in the graph. *Relationships* between the objects are represented using edges. With the severe weather data, the majority of the relationships are spatial. Both objects and relationships can have *attributes* associated with them and these attributes can vary both spatially and temporally. In the case of a spatially or spatiotemporally varying attribute, the data are represented as either a scalar or a vector field, depending on the nature of the data. The ability to represent spatial fields of objects is a key addition to the SRRF results presented here. This field can be two or three dimensional for space and can also vary as a function of time. In addition to attributes varying over space and time, the existence of objects and relationships can also vary as a function of time. If an object or a relationship is *dynamic*, it has a starting and an ending time associated with it.

To illustrate the data representation, Figure 1 shows the schema for the turbulence data. All objects and relations are required to be assigned a type. In this case, the attributes on the rain, hail, convection, and vertically integrated liquid objects are all temporal or 2-dimensional spatiotemporal scalar fields. The attributes on the aircraft object are all static as they are measured at a single moment in time. Note that the schema shows the types of objects and relationships possible but any specific graph can vary in the number of such objects present. For example, all graphs in the turbulence data will have an aircraft object but they may have any number (including 0) of rain, hail, and convective regions as defined by the weather nearby the aircraft.

Algorithm 1: Grow-SRPT

Input: s = Number of distinctions to sample, D = training data, m = Maximum depth of tree, d = current tree depth, p p-value used to stop tree growth
Output: An SRPT
if $d \leq m$ **then**
  tree $\leftarrow$ **Find-Best-Split**(D,s,p)
  if *tree* $\neq \emptyset$ **then**
    for *all possible values v in split* **do**
      tree.addChild(**Grow-SRPT**(D where split = v))
    end
    Return tree
  end
end
Return leaf node

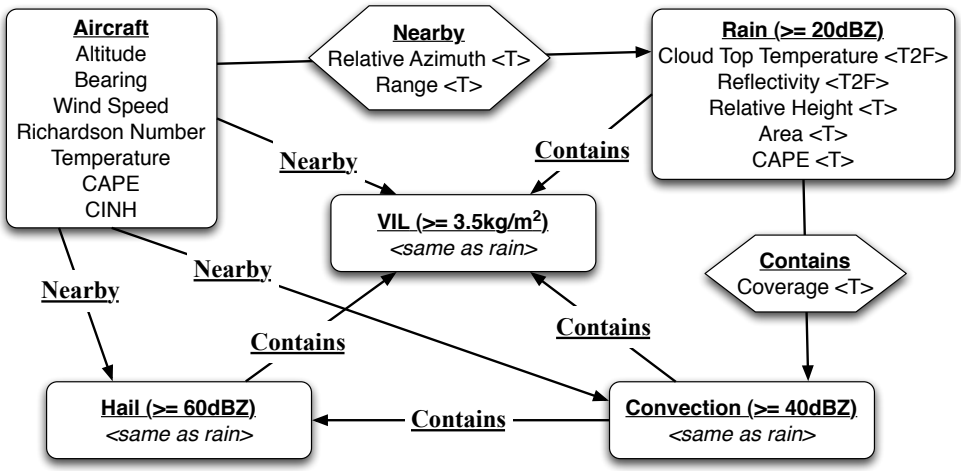


FIGURE 1. Schema for aircraft turbulence data set. Object and relationship types are underlined and bolded. Temporal attributes are denoted with a T and fielded attributes with a F (with 2F specifying 2-dimensional fields). CAPE is the Convective Available Potential Energy, which is a measure of how much thermodynamic energy a storm can use, CINH is the Convective Inhibition, which is a measure of the energy inhibiting storm formation, and VIL is the Vertically Integrated Liquid, which is a radar derived measure of the amount of precipitation contained in a vertical column of air.

Algorithm 2: Find-Best-Split

Input: s = Number of samples, D = training data, p p-value used to stop tree growth
Output: A split if one exists that satisfies the criteria or $\emptyset$ otherwise
best $\leftarrow \emptyset$
for *i = 1 to s* **do**
  split $\leftarrow$ generate random split
  eval $\leftarrow$ evaluate quality of split (using chi-squared)
  if *eval* $< p$ ***and*** *eval < best evaluation so far* **then**
    best $\leftarrow$ split
  end
end
Return best

A SRRF is composed of individual Spatiotemporal Relational Probability Trees (SRPTs) [23], which are probability estimation trees similar to Relational Probability Trees [29] but with the ability to split the data based on spatiotemporal attributes of both objects and relations. Since our initial introduction of SRPTs, their capabilities have significantly expanded. The most significant change is their ability to represent and reason about attribute fields within objects. We summarize the types of questions that the SRPTs can use to make distinctions about the data.

The non-temporal splits are:

- Exists: Does an object or relation of a particular type exist?
- Attribute: Does an object or a relation with attribute a have a [MAX, MIN, AVG, ANY] value $\geq$ than a particular value v?
- Count Conjugate: Are there at least n yes answers to distinction d? Distinction d can be any distinction other than Count Conjugate.

- Structural Conjugate: Is the answer to distinction d related to an object of type t through a relation of type r? Distinction d can be any distinction other than Structural Conjugate.

The temporal splits are:

- Temporal Exists: Does an object or a relation of a particular type exist for time period t?
- Temporal Ordering: Do the matching items from basic distinction a occur in a temporal relationship with the matching items from basic distinction b? The seven types of temporal ordering are: *before*, *meets*, *overlaps*, *equals*, *starts*, *finishes*, and *during* [2].
- Temporal Partial Derivative: Is the partial derivative with respect to time on attribute a on an object or relation of type $t \geq v$?

The spatial and spatiotemporal splits are:

- Spatial Partial Derivative: Is the partial derivative with respect to space of attribute a on object or relation of type $t \geq v$?
- Spatial Curl: Is the curl of fielded attribute $a \geq v$?
- Spatial Gradient: Is the magnitude of the gradient of fielded attribute $a \geq v$?
- Shape: Is the primary 3D shape of a fielded object a cube, sphere, cylinder, or cone? This question also works for 2D objects and uses the corresponding 2D shapes.
- Shape Change: Has the shape of an object changed from one of the primary shapes over to a new shape over the course of t steps?

Algorithm 1 describes the procedure for growing an individual tree. This procedure follows the standard greedy decision tree algorithms with the exception of the sampling of the distinctions. Because there is a very large number of possible instantiations for the split templates listed above, we sample the specific distinctions using a user specified sampling rate. For each sample, a split template is selected randomly and the pieces of the template are filled in using randomly chosen examples in the training data. This process is described in Algorithm 2. The split with the highest chi-squared value is chosen so long as its p-value satisfies the user specified p-value threshold. This threshold can be used to control tree growth, with higher values enabling the growth of deeper trees and lower values enabling potentially higher quality splits but less complicated trees.

Algorithm 3: Growing SRRFs

Input: s = Number of distinctions to sample, n = number of trees in the forest, D = training data
Output: An SRRF
for *i = 1 to n* **do**
    [in-bag-data, out-of-bag-data] ← **Bootstrap-Resample**(D)
    T_i ← **Grow-SRPT**(in-bag-data, s)
end
Return all trees $T_{1...n}$

Algorithm 3 shows the overall learning approach for growing a SRRF. The SRRFs preserve as much of the RF training approach as possible. The training data for each tree in the forest is still created using a bootstrap resampling of the original training data. The difference in the learning methods arises from the nature of the spatiotemporal relational data and the SRPTs versus C4.5 trees. In the RF algorithm, each node of each tree in the forest is trained on a random subset of the available attributes. Since the individual trees are standard C4.5 decision trees, this limits the number of possible splits each tree can make. Because each tree is also trained on a different bootstrap resampled set of the original data, the trees are sufficiently different from one another to make a powerful ensemble. Because there are a very large number of possible splits that the SRPTs can choose from, a SRPT finds the best split through sampling, as described above. Like the original RF trees, SRPTs are still built using the best split identified at each level. With fewer samples, these splits may not be the overall best for a single tree, but they will be sufficiently different across the sets of trees that the power of the ensemble approach will be preserved. However, if the number

of samples is too small, the number of trees needed in the ensemble to obtain good results may be prohibitively large. We examine these hypotheses empirically in the experimental results.

Once a SRRF is learned, it can be used for classification on new data by having each tree in the forest vote on the class label. The standard approach is to have each tree's vote be the class label with the highest probability (e.g. if $p(turbulence = yes) > p(turbulence = no)$, then that tree votes for turbulence). This is the approach that we follow in the results reported here. However, we have also investigated the use of a user-supplied threshold for a tree voting yes. In this case, instead of taking the class with the maximum probability, the tree can be said to vote for the rare class if the probability of that class is greater than the user supplied threshold. We have seen success with this approach in the past and will investigate its use in future work.

For a particular attribute a, RFs measure variable importance by querying each tree in the forest for its vote on the out-of-bag data. Then, the attribute values for attribute a are permuted within the out-of-bag instances, and the tree is re-queried for its vote on the permuted out-of-bag data. The average difference between the votes on the unpermuted data for the correct class and the votes for the correct class on the permuted data is the raw variable importance score. This score can then be computed for each SRRF and tested using a z-test. We have directly converted this approach to the SRRFs and can measure variable importance on any attribute of an object or relation. Spatially and temporally varying attributes are treated as a single entity and permuted across the objects/relations but their spatial and/or temporal ordering is preserved. We examine the variable importance in each of our data sets.

In this paper, we also present viable ranges for questions based on attributes on both objects and relations. For an attribute a that appears in an attribute-based split, the range is identified by testing for a drop in performance in the training set. Given a specific value in the split v, the possible range examined is $v - 10$ to $v + 10$ in increments of 0.1. This range can be expanded in future work but our current results demonstrate that this is a good starting point.

3. Parameter Exploration

While our primary goal is to enable the domain scientists to gain a better understanding for their prediction problems, this information is not likely to be used unless the SRRF is skillfully predicting in the domain. To better understand the effects of each of the parameters on the performance of the SRRF, we performed a combinatorial experiment across all parameters on a synthetic dataset. The primary parameters that affect the performance of the SRRF are the number of distinctions sampled at each node of SRPT training (this is analogous to the number of randomly-selected attributes examined for splitting in a C4.5 tree in constructing a RF), the maximum depth the tree is allowed to reach, the number of trees in the forest, the p-value used to control tree growth (using the chi-squared statistical test), and the types of distinctions the tree can use. Sampling several values from each of these parameters as shown below yields a total of 60 parameter sets, each of which is run 30 times for statistical testing.

- Number of samples (NS): [10, 100, 500, 1000].
- Maximum depth of the tree (MD): [1, 3, 5].
- Number of trees in the forest (NT): [1, 10, 50, 100].
- We fixed the p-value to 0.01

The goal of this section is to understand the impact of the parameters on performance. For non-binary problems, we measured performance using the Gerrity Skill Score (GSS) [15]. We chose GSS because it was designed for measuring performance on imbalanced multi-class classification problems. GSS is Gandin and Murphy's equitable skill score with a scoring matrix derived to satisfy the constraints of symmetry and equitability as stated by Gandin and Murphy. To calculate GSS, let S be an equitable scoring matrix whose calculation is given below and E be a matrix (contingency table) such that $e(i, j)$ is the relative frequency of instances with the true class i being classified as class j. From S and E, GSS = trace($S^{\mathsf{T}}E$). To calculate S, following [16], let $P(r)$ be the relative

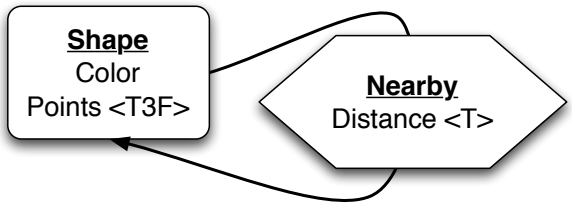


FIGURE 2. Schema for Points dataset: Attributes tagged <T> are temporal, <T3F> are temporal 3D fields, and untagged attributes are discrete.

frequency of class r. Using $P(r)$ define the following:

$$D(n) = \frac{1 - \sum_{r=1}^{n} P(r)}{\sum_{r=1}^{n} P(r)} \quad \text{, and}$$

$$R(n) = \frac{1}{D(n)} \quad .$$

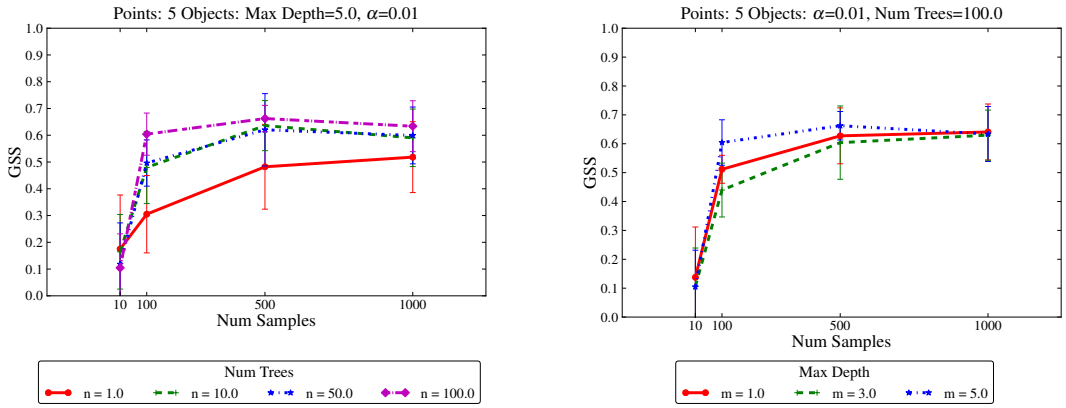
Let K be the number of classes and $\kappa = \frac{1}{K-1}$.The elements of S are calculated as:

$$\begin{aligned}
s_{n,n} &= \kappa \left[\sum_{r=1}^{n-1} R(r) + \sum_{r=n}^{K-1} D(r) \right] && n = (1, 2, \cdots, K) \\
s_{m,n} &= \kappa \left[\sum_{r=1}^{m-1} R(r) + \sum_{r=m}^{n-1} (-1) + \sum_{r=n}^{K-1} D(r) \right] && 1 \leq m < K, m < n \leq K \\
s_{n,m} &= s_{m,n} && 2 \leq n \leq K, 1 \leq m < n
\end{aligned}$$

GSS varies from -1 to 1 [16]. A value of -1 indicates "intentionally" classifying incorrectly; a value of 0 is equivalent to the performance of a random classifier; and a value of 1 indicates perfect classification.

For the binary classification tasks, we also measure Area under the Receiver Operator Characteristic Curve (AUC). AUC has a long, respected history in evaluating machine learning algorithms [5, 12]. AUC measures the robustness of the classifier across the various probability thresholds [30]. AUC varies from 0 to 1, with 1 indicating a perfect classifier and 0.5 indicating a random classifier.

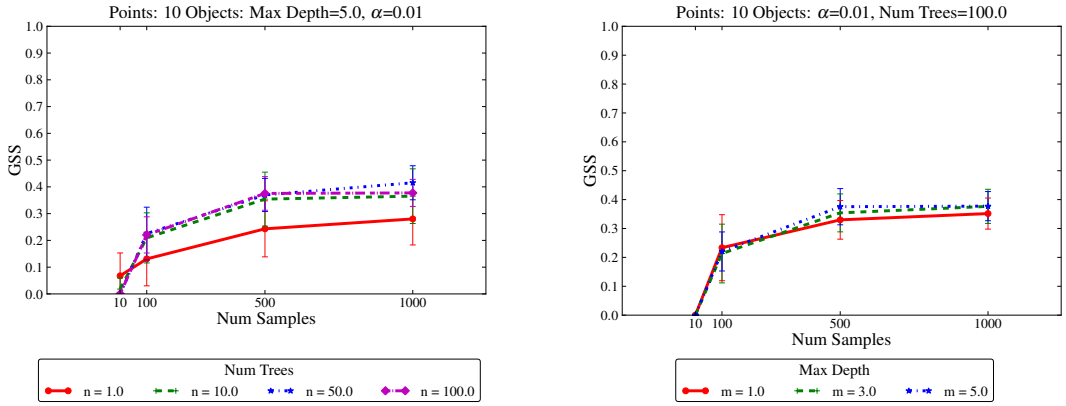
The **Points** data set is a synthetic spatiotemporal relational data set that we designed to highlight the ability of the SRRF to handle spatiotemporally varying fielded attributes. The schema for Points is shown in Figure 2. We generate each graph randomly according to the rules of the classes that we created for this data. Each graph has between 3 and a parameterized maximum number of objects. We varied the maximum number of objects between 5, 10 and 20, with the larger number representing more difficult problems (each graph contains more noise). Each object in the graph has two attributes associated with it: a discrete color and a 3D-field of integers from $\{0, 1\}$. The 3D-field attribute field is used to represent a point cloud with a 1 indicating a point at that location and 0 indicating empty space. This cloud is used to represent the shape of the object. There are three classes in the Points dataset: *change*, *grow*, and *flip*. *Change* has a box turn into a sphere and a cone turn into a cylinder. *Grow* has three blue spheres that grow to have a volume greater than ≈ 167 cubic units (radius of 40 units). *Flip* has a cone that flips 180 degrees along its axis. To generate the data, we randomly created a uniform distribution of class labels. Using the class labels, we filled in the required number of objects with the correct attributes to meet the definition of each



(a) GSS by forest size

(b) GSS by maximum tree depth

FIGURE 3. GSS on the synthetic points dataset with a maximum of 5 objects per graph as a) a function of the number of trees and samples and b) a function of the maximum depth of the trees and samples.



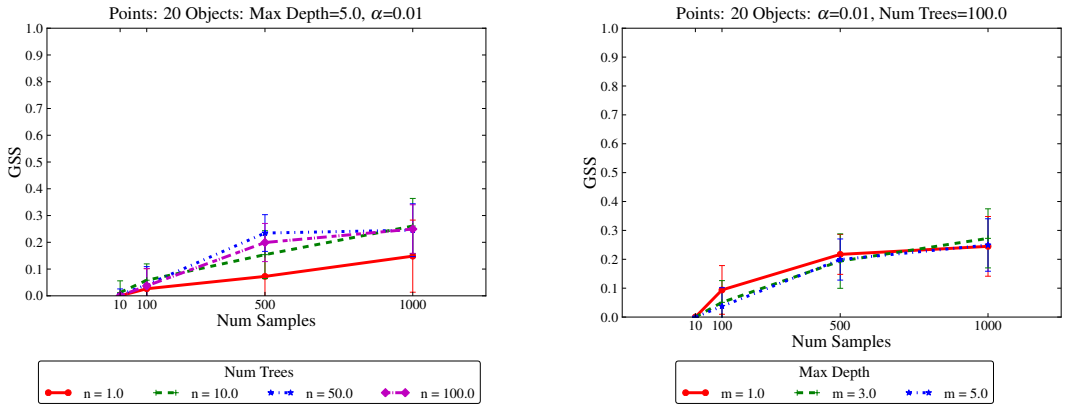
(a) GSS by forest size

(b) GSS by maximum tree depth

FIGURE 4. GSS on the synthetic points dataset with a maximum of 10 objects per graph as a) a function of the number of trees and samples and b) a function of the maximum depth of the trees and samples.

class label. Additional randomly generated objects were added to the graph along with relations that randomly related objects within the graph. These extra objects and relations served to add noise.

Figures 3, 4, and 5 show the performance of the SRRF on the points domain for the case of 5 objects, 10 objects, and 20 objects per graph respectively. Panel a of each figure shows the performance of the SRRF as a function of the number of trees in the forest and the number of samples at each level of tree growth. For panel a, the maximum depth of the tree is fixed to 5. Panel b of each figure shows the performance as a function of the maximum depth and the number of samples. In these graphs, we fixed the number of trees at 100.



(a) GSS by forest size

(b) GSS by maximum tree depth

FIGURE 5. GSS on the synthetic points dataset with a maximum of 20 objects per graph as a) a function of the number of trees and samples and b) a function of the maximum depth of the trees and samples.

The set of graphs in Figures 3, 4, and 5 are instructive about the performance of the SRRF in several ways. First, it is clear that the SRRF is achieving a high level of performance, starting with the easier case of a maximum of 5 objects per graph (Figure 3a). As the number of objects in the graph increases, the task becomes harder because the extra objects add noise. The maximum performance drops as a function of higher noise levels for all sizes of forests and for all levels of sampling. With 5 objects per graph, the maximum GSS is slightly above 0.6. With 10, that drops to approximately 0.4 and with 20 objects, the maximum GSS is about 0.25. While 0.25 still indicates that the SRRF is able to learn a skillful predictor on this domain, it is clearly better able to predict in the case with less noise.

Panel a of Figures 3, 4, and 5 also demonstrate that the individual SRPT is always less skillful than the forest, across all levels of sampling. This is to be expected as ensembling methods have proven quite powerful in a wide variety of domains. In the case with the least noise (Figure 3a), there is a clear differentiation of skill up to forests of size 100 for all levels of sampling. However, in the harder domains the number of trees in the forest provides less of an effect than the number of samples.

Random forests [6] also demonstrated that performance leveled off as the diversity of the trees in the forest decreased. To maintain diversity, Random Forests limited the number of attributes allowed to be chosen at each level of tree growth while the SRRF limits the number of samples. As the number of samples increases, we expect the skill of the forest to increase and then asymptote as the forest diversity decreases. This is also seen in each of the three cases.

Panel b of Figures 3, 4, and 5 provides an alternate view into the same results, with the number of trees fixed at 100 but varying the maximum depth. Somewhat surprisingly, the maximum depth is less important in this domain than the other parameters, most likely because the domain can be solved in shallow trees. With increasing depth, the trees are more likely to overfit, which can decrease performance. Note that the full set of graphs for all parameter variations are available at `http://idea.cs.ou.edu/cidu2010/sadm_extended/`.

Table 1 shows the results of running an Analysis of Variance (ANOVA) across all of the parameters for each of the three variations of points data. In the case with a maximum of 5 objects per graph, the maximum depth parameter is not statistically significant but all of the other parameters and interactions are. The general trend across all the parameters is that both the number of samples

Experiment	MD	NS	NT	MD×NS	MD×NT	NS×NT	MD×NS×NT
5 objects	0.17	**2.2e-16**	**2.2e-16**	**0.00256**	**0.0315**	**2.2e-16**	**0.00147**
10 objects	**3.1e-07**	**2.2e-16**	**2.2e-16**	**0.0053**	0.625	**2.2e-16**	0.149
20 objects	0.136	**2.2e-16**	**2.2e-16**	**0.00502**	**0.0132**	**2.2e-16**	**0.000148**

TABLE 1. ANOVA results for all parameters on all three of the variations on points: with 5 objects maximum, 10 objects, and 20 objects. Statistically significant effects are shown in bold.

and the number of trees have a statistically significant effect on GSS. The cross product of those two as well as the products with the other parameters are also significant, indicating interaction effects across the parameters. This is not surprising since each of the parameters individually affects the overall quality of the tree.

4. Convectively-induced turbulence

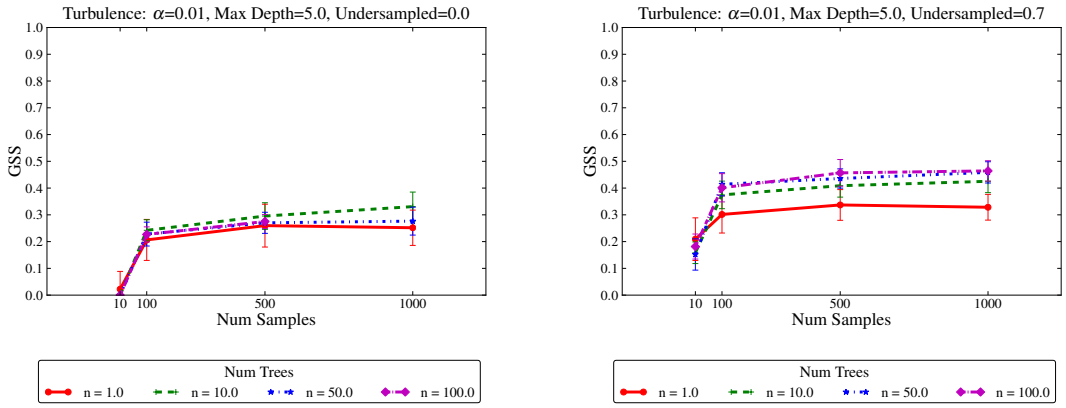
Convectively-induced turbulence (CIT), that is, atmospheric turbulence in and around thunderstorms, is a major hazard for aviation that commonly causes delays, route changes and bumpy rides for passengers, particularly in the summer. Turbulence encounters can cause structural damage to aircraft, serious injuries or fatalities, and frightening experiences for travelers. Better information about likely locations of turbulence is needed for airline dispatchers, air traffic managers and pilots to accurately assess when ground delays are truly necessary, plan efficient routes, and avoid or mitigate turbulence encounters. For these reasons, enhanced prediction of CIT is one of the stated goals of the FAA's current effort to modernize the national air transportation system, called NextGen.

An existing system for forecasting turbulence over the US is called Graphical Turbulence Guidance (GTG) [35]. GTG was developed by the FAA's Aviation Weather Research Program, and currently runs operationally at NOAA's Aviation Weather Center[1]. The GTG algorithm is based on a combination of turbulence "diagnostic" quantities derived from an operational numerical weather prediction (NWP) model's 3-D forecast grids. For example, the Richardson number measures the ratio of atmospheric stability to wind shear; low values of this quantity suggest the transition from laminar to turbulent flow [41]. Unfortunately, operational NWP models run on a grid that is too coarse to resolve thunderstorms, and thus are unable to fully capture CIT generation mechanisms even if they are quite accurate. Therefore, the best hope for CIT prediction is to couple model-derived information about the storm environment and diagnostics of turbulence with timely observations from satellite or radar that characterize the location, shape, and intensity of a storm.

The advent of an automated turbulence reporting system on board some commercial aircraft [10, 11] makes it possible to associate objective atmospheric turbulence measurements with features from NWP models and observations. The system uses rapid measurements of the vertical acceleration of the aircraft to deduce the atmospheric winds, and then performs a statistical analysis of the wind fluctuations to determine the turbulence intensity, which is measured in terms of eddy dissipation rate (EDR) over 1-minute flight segments. The data used in the SRRF experiments below were collected from United Airlines Boeing 757 aircraft in the summer of 2007. Convection is most common in the summer and studying this time period helps to generate a dataset in which convection is the most prevalent source of turbulence.

One difficulty in training intelligent algorithms to predict turbulence is that the data contain an overwhelming number of cases with null or light turbulence reported. Turbulence is a rare phenomenon to begin with, and the data were collected from aircraft whose pilots were doing their best to avoid turbulence so as to maximize passenger comfort and safety. As a result, light-to-moderate or greater (LMOG) turbulence occurs in less than 1% of the data points and an algorithm can achieve 99% accuracy by simply predicting "no turbulence" everywhere. To counteract this,

[1]See http://aviationweather.gov/adds/turbulence/



(a) No undersampling

(b) 0.7 undersampling

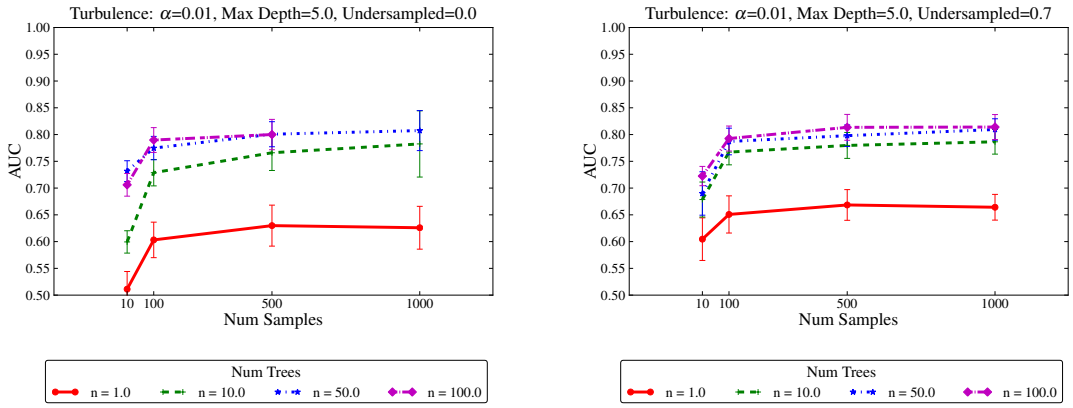
FIGURE 6. Mean GSS over 30 runs for the turbulence prediction problem as a function of sample size and the number of trees in the forest. Error bars indicate 1 standard deviation. Panel a shows the results for no undersampling and panel b shows the results for 0.7 undersampling.

we resampled the data, retaining only 3% of the null or light turbulence cases. The final data set contains 2055 cases, approximately 26% of which are LMOG (positive) turbulence cases and 74% of which are null or light turbulence (negative); the counts are 1514 negatives and 541 positives. After the initial resampling of the full data set, the data were split into training and testing sets. The null or light turbulence cases were additionally undersampled by 0, 70, 80, and 90% to test the effects of changing the ratio of positive to negative instances. However, all evaluations were performed on the original testing set without undersampling.

The data available for this study comes from a combination of the measurements collected from the United aircraft, archived weather observations for the same time period, and archived real-time NWP model data (Rapid Update Cycle[2]). These data are transformed to a spatiotemporal relational representation using the schema shown in Figure 1. The in-situ aircraft data and the interpolated NWP model data were used to make the aircraft objects, and the gridded model and observation data were used to make the other objects. We only retained observations from the aircraft when it was above 15,000 feet flight level. This was done for two reasons. First, passengers are usually belted in during takeoff and landing, so avoidance of turbulence at low altitudes is not as important as at upper altitudes. Second, the in-situ data may be contaminated by aircraft loads and maneuvers during the lower-altitude phases of flight, so the EDR reports may be less accurate there. Each of the objects other than the aircraft represents a meteorological concept or distinct region. We applied thresholds to the radar reflectivity data to obtain connected areas greater than 20 dBZ ("rain" objects), 40 dBZ ("convection" objects) and 60 dBZ ("hail" objects); note that these descriptions are suggestive rather than meteorologically precise. We then extracted connected regions within 40 nautical miles of the aircraft, and co-located them with infrared satellite and NWP model data. The same method was used for radar-derived vertically integrated liquid (VIL), with a threshold of 3.5 kg m^{-2}. The aircraft objects are static and the observations are available only at the same time that the turbulence was measured. The other objects are tracked for 30 minutes, in 5 minute increments.

Figure 6 shows the GSS on an independent test set of the turbulence data as a function of the number of distinctions sampled and the number of trees in the forest. Panel a shows the results for no additional undersampling and panel b shows the results for 0.7 undersampling. The undersampling

[2]http://ruc.noaa.gov/



(a) No undersampling

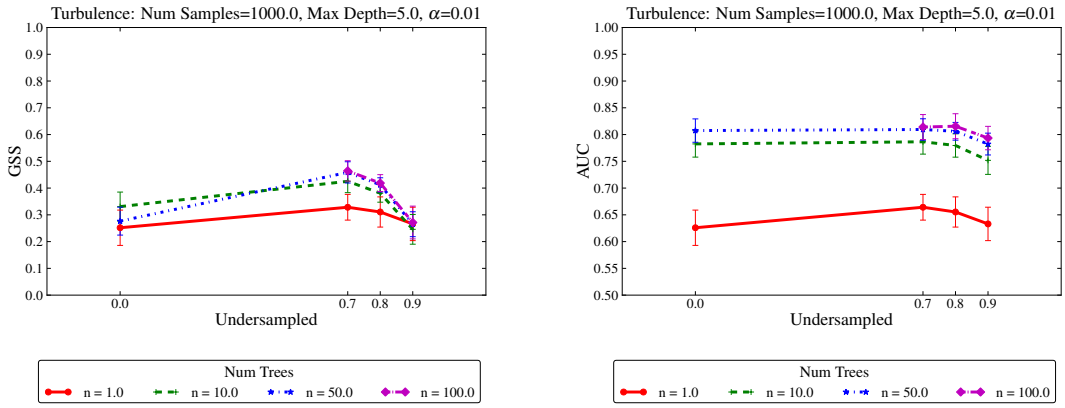
(b) 0.7 undersampling

FIGURE 7. Mean AUC for the turbulence problem as a function of sample size and the number of trees in the forest. Error bars indicate 1 standard deviation. Panel a shows the results for no undersampling and panel b shows the results for 0.7 undersampling.

clearly improves the skill scores as it goes from 0 to 0.7. Another improvement that can be seen is the difference between a single tree forest (an SRPT) and a multi-tree forest. In all cases, the multi-tree forest outperforms the single-tree forest. In addition, the performance for all sizes of forests improves initially as the number of samples is increased. As the number of samples increases from 10 to 100, the quality of the trees improves significantly. However, the performance asymptotes fairly quickly after 100 samples. This is expected, as increasing the number of samples increases the probability that the tree will ask a question that splits the data well, but eventually also reduces the diversity of the forest and increases the risk of overfitting. The asymptotic behavior of the performance occurs because if the sample size is large enough, the trees have probably examined all the best distinctions.

In addition to examining the skill of the forests using GSS, we also examined the robustness using AUC. Figure 7 shows the AUC for the same two cases where we examined GSS. Here the difference between the varying forests' sizes is more clearly seen. As the size of the forest increases, the algorithm performs more robustly across a wide variety of parameters. This behavior is also expected, since ensembles with more members are expected to better capture the underlying relationships between the data. Increasing the number of trees in the SRRF also appears to yield an asymptotic performance gain. This is likely occurring for two reasons. The first is that bootstrap sampling becomes more uniform with the larger number of ensemble members, so the effectiveness of the ensemble is reduced. The second is that, as the number of samples increases, the ensemble becomes more uniform, that is, it becomes more difficult to add a tree that provides new information. RF performance has also been shown to asymptote as the diversity of the trees in the forests is reduced [6]. Also, consistent with the GSS results, as the number of distinctions sampled at each node increases, the performance of the forest increases and then asymptotes.

Because it is clear that the amount of undersampling affects the results, we also examined the GSS and the AUC as a function of the amount of undersampling and the size of the forest. Figure 8 shows these results with panel a showing GSS and panel b showing AUC. There is a jump in skill as measured by GSS with the sampling but the gain quickly drops off as the undersampling becomes too large. The forests remain relatively robust for different amounts of undersampling, but their performance drops at the highest levels of undersampling as the number of positive cases begin to dominate the negatives. Also notable is the very significant difference in performance between the single SRPT and the multiple-tree forests: the ensembles dramatically outperform the single tree



(a) GSS (b) AUC

FIGURE 8. (a) GSS and (b) AUC for the turbulence data as a function of the amount of undersampling. Error bars indicate 1 standard deviation.

TABLE 2. Top 10 statistically significant important attributes ($\alpha = 0.05$) in the turbulence data for a forest with 100 trees, 500 samples at each node, max tree depth of 5. This is computed over 30 runs.

Attribute	Mean Variable Importance	Standard Deviation
Rain.Aircraft Relative Height	0.796	0.157
Aircraft.CAPE	0.596	0.132
Aircraft.Altitude	0.500	0.104
Aircraft.Temperature	0.411	0.147
VIL.Aircraft Relative Height	0.297	0.110
Aircraft.CIN	0.280	0.139
Convection.Cloud Top Temperature	0.275	0.086
VIL.Cloud Top Temperature	0.259	0.111
Convection.Aircraft Relative Height	0.220	0.088
VIL.Area	0.190	0.115
Aircraft.Richardson Number	0.156	0.125
Rain.Cloud Top Temperature	0.145	0.126

model at all levels of undersampling. This underscores the value of the SRRFs ensemble learning method.

Table 2 shows the importance values for the top 12 attributes in the turbulence data. Attributes are appended to the object they are associated with (e.g. Aircraft.CAPE means the CAPE attribute of objects of type Aircraft). Two of the top twelve attributes directly involve the flight altitude of the aircraft: Aircraft.Altitude and Aircraft.Temperature, which is closely related to altitude since higher altitudes have colder temperatures. The relationship of altitude to turbulence frequency and severity has been established in separate studies based on climatologies of pilot reports of turbulence [45]. In particular, a higher incidence of turbulence is observed around the tropopause, which over the US is often between 30,000 and 42,000, near the altitudes at which many aircraft cruise. However, an examination of the turbulence data used in the present study, summarized in Figure 9(c), indicates a higher likelihood of LMOG turbulence encounters between 15,000 and 25,000 ft. This distribution is not consistent with published turbulence climatologies and may be due to contamination of the

EDR data by aircraft vertical acceleration and maneuvers in the climb and descent phases of flight. Future studies using these data will focus on turbulence prediction above 25,000 ft. This outcome illustrates how SRRF importance results can inspire a closer critical examination of the data and refine the definition of the problem being studied.

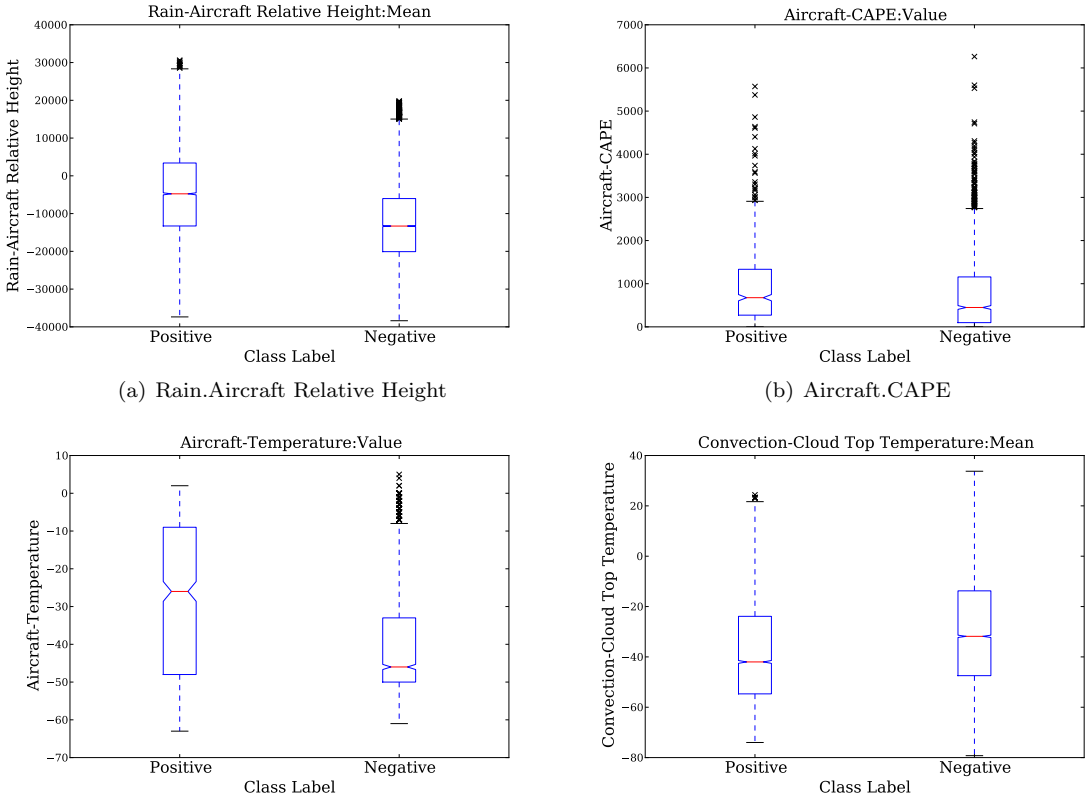
Three additional attributes in the top twelve are related to the aircraft's altitude relative to storm tops: Rain.Aircraft Relative Height, VIL.Aircraft Relative Height, and Convection.Aircraft Relative Height. These represent the difference between the radar echo top within the weather object and the aircrafts flight altitude. As shown in Figure 9(a), LMOG turbulence encounters are more likely to be associated with larger relative altitudes, which indicate aircraft flight levels nearer or below the radar echo tops. This result is consistent with previous research and with the FAAs thunderstorm avoidance guidelines [3], which proscribe flights above active thunderstorms. However, the fact that the highest score is for the rain object attribute comes as a bit of a surprise, since not all rain objects are associated with mature convection. This result suggests that weaker rain processes or the early or late stages of convection may also have important connections with turbulence. The Cloud Top Temperature attributes of Convection, VIL, and Rain objects also fall within the top twelve most important. As shown in Figure 9(d), lower values of the satellite-measured cloud-top temperature are associated with a higher likelihood of turbulence, since colder storm tops indicate deeper, more active storms. The importance of VIL Area is likely due to its representation of the size of convectively active storms in the vicinity of the aircraft. Larger storms are often more intense and longer-lived, providing a greater perturbation of the environment.

Three model-derived quantities fill out the list of the top 12 attributes: Aircraft.CAPE, Aircraft.CIN and Aircraft.Richardson Number. CAPE, or convective available potential energy, is related to turbulence intensity because higher CAPE environments tend to produce more vigorous storms with more intense updrafts, downdrafts, anvils and gravity wave generation (see Figure 9(b)). CIN, or convective inhibition, plays the opposite role. Finally, the Richardson number is the ratio of model-resolved stability to shear; low values indicate the possibility of a breakdown from laminar flow to turbulent eddies. Therefore, Richardson number may help diagnose non-convective turbulence from sources such as the jet stream. It may also help characterize the upper troposphere and lower stratospheres' susceptibility to turbulence as they are modified by gravity waves and anvil outflows produced by thunderstorms [39]. These results suggest the value of combining model output and observational data.

The discussion above has emphasized the plausibility of the SRRF importance results by describing conditional distributions of the individual attributes. However, it should be noted that the SRRF is also discovering and utilizing complex, nonlinear relationships between the various attributes. Thus, an attribute that on its own shows no correlation with turbulence could still show high importance, since in combination with other attributes it could have significant value. Indeed, the conditional distributions for CAPE in Figure 9(b) suggest less independent discrimination capability than some other attributes that are ranked lower. On the other hand, the attributes which show up as most important may sometimes be highly correlated, so that not all are really needed in the final SRRF model. In this case, the importance results may be used to guide a domain scientist in determining which of several attributes containing similar information have the least value and should be considered for elimination from the model.

In order to more concretely illustrate the SRPTs that comprise the SRRF ensemble, Figure 10 shows the yes branch of the tree with the highest GSS out of the 1000 sample single tree forests. The no branch was not included due to space limitations. For each question regarding an attribute, the question's split value was varied to measure the sensitivity of the skill to changes in the split value as described above. The larger ranges are less performance sensitive and smaller ranges are more performance sensitive. Some nodes are sensitive in one direction but not the other. The node double circled in Fig. 10 regarding Max VIL Area is an example of this case. Larger VIL area is

[3]FAA Advisory Circular 00-24, available at www.airweb.faa.gov/Regulatory_and_Guidance_Library/rgAdvisoryCircular.nsf/, and FAA Aeronautical Information Manual section 7-1-30, available from www.faa.gov/atpubs/AIM/



(a) Rain.Aircraft Relative Height

(b) Aircraft.CAPE

(c) Aircraft.Temperature

(d) Convection.Cloud Top Temperature

FIGURE 9. Schematic plots [42] depicting the distributions of selected attributes in the Turbulence domain. The Positive label is linked to moderate or greater turbulence, and the Negative label is for light or no turbulence. The red line is the median, and the upper and lower edges of the blue box are the upper and lower quartiles of the distribution. The indentations around the median line indicate the range for 95% statistical significance.

correlated with larger updrafts since more precipitation is being held aloft. Given that only 5 cases fell below the given split value, and they were all negative, decreasing the split value below 8.24 likely caused a significant number of false positives. The split value range from 8.04 to 18.9 is due to either most of the VIL Areas being above 18.9 or most of those that were in that range being negative cases.

5. Storm environment and surface boundary effects on tornadogenesis

The thermodynamic environment surrounding an evolving supercell thunderstorm can vary greatly on both small spatial and temporal scales. General variations in thermodynamic variables such as temperature, pressure, and relative humidity within the storm region could be used as the basis of a model to distinguish tornadic and nontornadic supercells and gain a better understanding of the mechanisms leading to tornadogenesis realizable from operational mesoscale observing networks. Another important component of the storm environment is the surface boundary. When different air masses meet, such as along a warm front or a cold front, boundary regions exist. Given that air mixes continuously, the transition zone along the boundary is not instantaneous and includes

FIGURE 10. Left most subtree of the most highly ranked SRPT from the turbulence domain for 1000 samples. The tree shows the ranges on the attribute questions. Due to space limitations, we focused on the left-most subtree. The node that is double circled is discussed in the text. P(T) denotes the probability of a LMOG turbulence event.

regions of strong temperature and moisture gradients. In addition to fronts, boundaries also occur along drylines or due to outflow from thunderstorms. While boundaries are commonly associated with the generation of storms through the lifting of warm, moist air over cool, dry air, their overall impact on the generation of tornadoes is not well understood. Markowski et al. [22] describe how boundaries can yield a zone of enhanced horizontal rotation. A supercell thunderstorm with a strong updraft moving through the zone can vertically tilt and stretch the enhanced horizontal rotation which assists with the process of producing a tornado. That study analyzed strong tornadic supercell thunderstorms over a one-year period and found that 70% occurred near frontal boundaries. However, due to the limited sample size and time period, further study was needed to quantify the relationship between boundaries and tornadoes over longer periods.

The data for our study was obtained from a climatology of 926 supercells within Oklahoma from 1994-2003 by Hocker and Basara [17]. To obtain the frontal passage information, surface observations from the Oklahoma Mesonet [25] were analyzed to a grid through bilinear interpolation. The gridded field within a 50 km radius of the storm track was kept as a fielded object. Surface frontal boundaries associated with each supercell were analyzed from the grid using objective front analysis techniques [32, 18]. Each group of supercells and frontal boundaries was labeled based on whether or not the supercell produced a tornado. The front and supercell data were related using the schema shown in Figure 11, where Nearby relationships indicated storms and fronts less than 40 km apart and

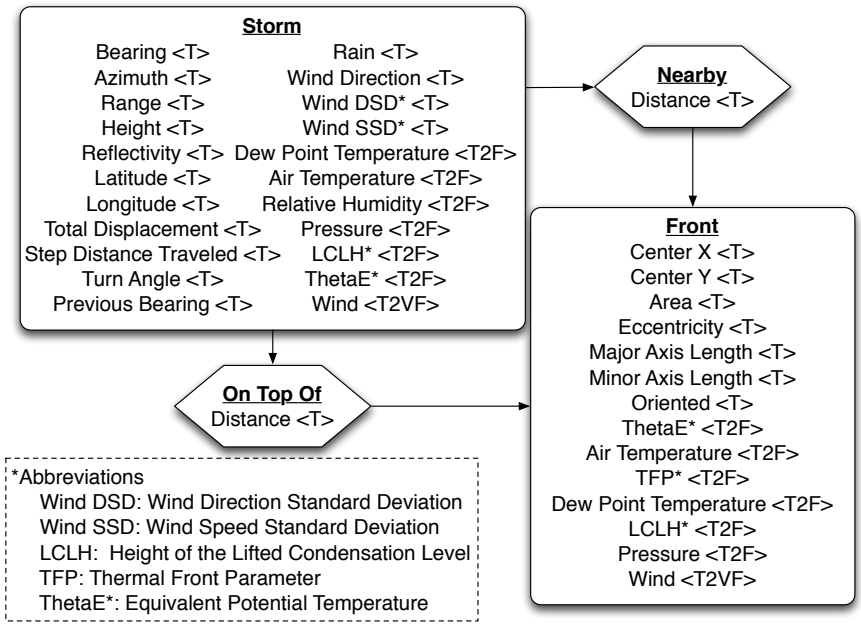


FIGURE 11. Schema for tornadogenesis data. Temporal data is denoted with a <T> and 2-dimensional fielded data with a <T2F>.

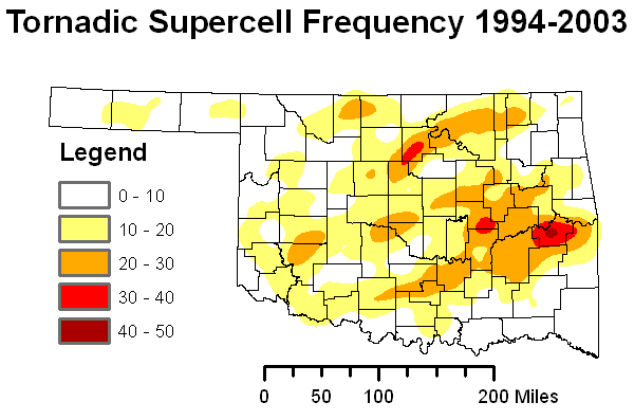


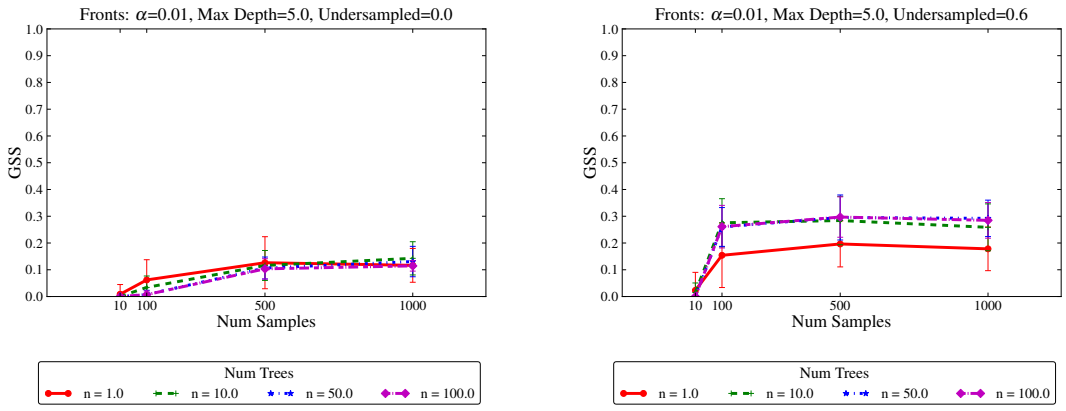
FIGURE 12. Number of tornadic supercells that have passed within 30 km of a point from 1994-2003.

On Top Of relationships indicated a distance of less than 10 km apart, the typical diameter of a supercell thunderstorm. This data included a wide variety of temporal and spatial attribute fields.

Table 3 shows the class distribution of the supercell thunderstorms and Figure 12 shows the spatial distribution of tornadic supercells in Oklahoma. While most supercells in the data were found to be non-tornadic, the tornadic supercells were found to last an hour longer on average than non-tornadic supercells, a significant (p=0.01) difference. Although duration is well correlated with

TABLE 3. The distribution of tornadic and non-tornadic supercell durations.

	Tornadic	**Non-Tornadic**
Count	215	711
Proportion	0.235	0.765
Median Duration (hr)	2.71	1.71
Mean Duration (hr)	2.90	1.96
Std. Dev. Duration (hr)	1.48	1.09
Max. Duration (hr)	9.33	7.06
Min. Duration (hr)	0.32	0.08



(a) No undersampling

(b) 0.6 undersampling

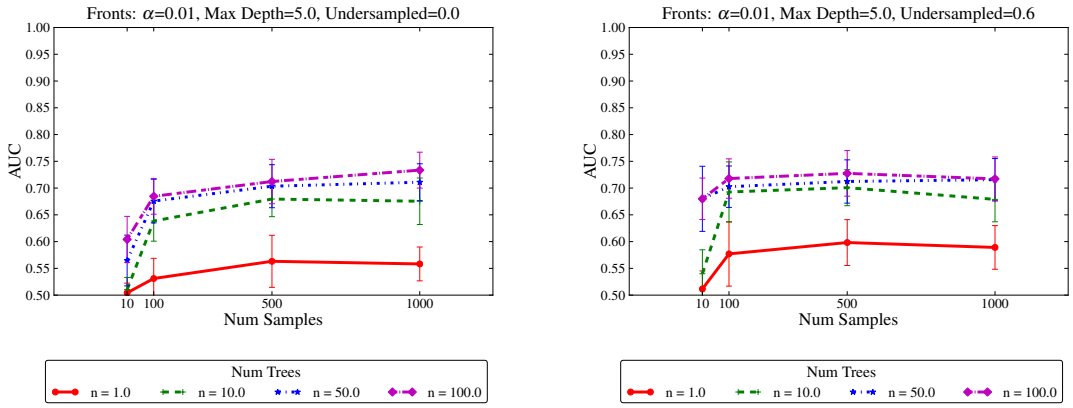
FIGURE 13. GSS for the fronts domain as a function of sample size. Error bars indicate 1 standard deviation. Panel a shows the results for no undersampling and panel b shows the results for 0.6 undersampling.

tornadic supercells, it is not a predictive variable and is not useful while a storm is developing as its final duration is not known until the storm has ended.

To determine what impact environmental variables have on the distribution of tornadic supercells, we applied the SRRFs to this data. While the supercall data set was also skewed, it was not nearly as dramatic as the turbulence data which was over 99% one class. Even so, we utilized undersampling of the majority class with ratios of 0.0 to 1.0 in increments of 0.1. The full set of results graphs are again found at `http://idea.cs.ou.edu/cidu2010/sadm_extended/`.

As with the previous experiment, we examined the GSS and the AUC as a function of the number of trees in the forest, the number of distinctions sampled at each level, and the amount of undersampling. Also as before, the test set is not undersampled. Figure 13 shows the GSS for no undersampling (panel a) and 0.6 undersampling (panel b). In the case of no undersampling, there is very little skill indicated by either the trees or the forest. When the undersampling is increased to 0.6, the skill level increased significantly and developed a clear separation between the SRPT and the SRRF.

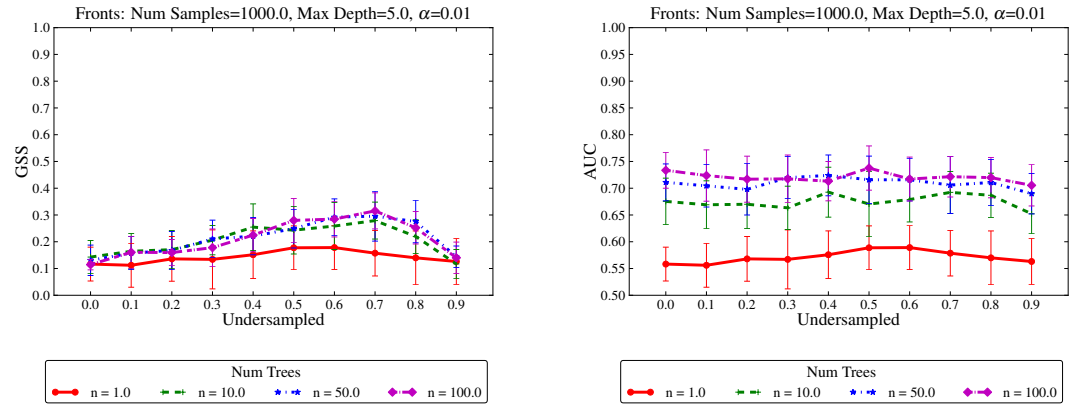
We also examined the performance of the SRRFs for each level of undersampling using AUC. The AUCs as a function of the number of trees and number of samples for no undersampling and undersampling 0.6 are both shown in Figure 14. The AUC indicates that this is a robust classifier and the forests again were able to outperform the single SRPT. Also, as with the turbulence data,



(a) No undersampling

(b) 0.6 undersampling

FIGURE 14. AUC for the fronts domain as a function of sample size. Error bars indicate 1 standard deviation. Panel a shows the results for no undersampling and panel b shows the results for 0.6 undersampling.



(a) GSS

(b) AUC

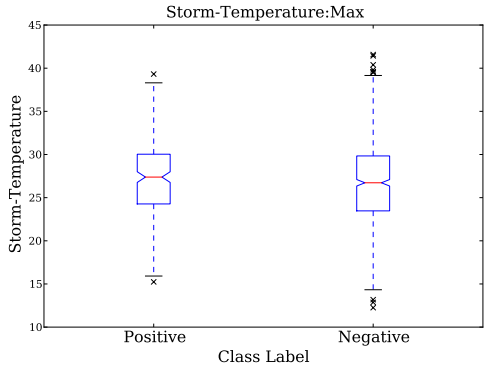
FIGURE 15. (a) GSS and (b) AUC for the fronts domain as a function of the amount of undersampling. Error bars indicate 1 standard deviation from mean value.

the performance asymptotes as a function of the number of trees in the forest and as the number of splits sampled at each level of tree growth increases.

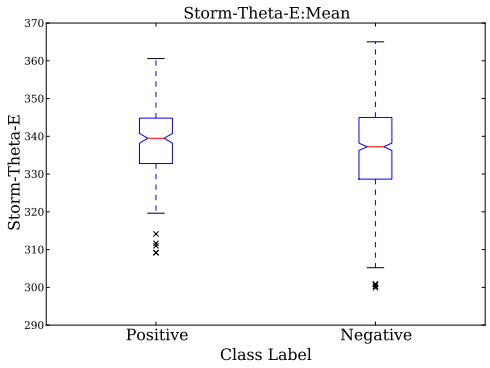
Figure 15 shows the GSS and the AUC as a function of the levels of undersampling. In the case of GSS, there is a noticable increase in skill for the forests at the mid-levels of undersampling. However, there is no change in performance for the SRPT (single tree forest). The differences between the GSS (panel a) and the AUC (panel b) graphs is dramatic. Whereas there is a clear increase in skill as a function of undersampling for the multi-tree forests, the AUC remains relatively constant. This indicates that the forests are robust even if the skill level is sometimes lower than desired. Also, again the forests clearly outperform the single SRPT and there is a visual difference between forests of size 10 and forests of size 50.

TABLE 4. The top 12 most important variables for the fronts domain, averaged over 30 runs of a 100-tree SRRF with a sample size of 1000 and a maximum tree depth of 5 and 70% undersampling of nontornadic cases.

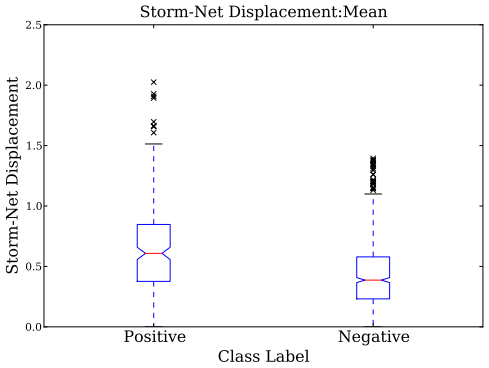
Attribute	Mean Variable Importance	Standard deviation
Storm.AirTemperature	0.283	0.138
Storm.Theta-E	0.198	0.097
Storm.RelativeHumidity	0.137	0.146
Storm.NetDisplacement	0.118	0.095
Storm.LiftedCondensationLevelHeight	0.115	0.111
Storm.DewPoint	0.109	0.123
Storm.Pressure	0.101	0.128
Front.ThermalFrontParameter	0.066	0.103
Front.MinimumAxisLength	0.066	0.103
Front.AirTemperature	0.065	0.101
Front.Area	0.061	0.100
Front.Orientation	0.058	0.100



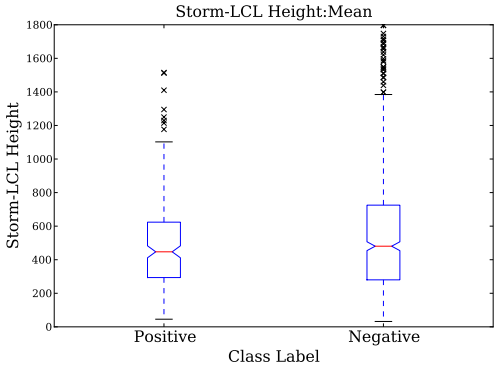
(a) Storm.Temperature



(b) Storm.Theta-E



(c) Storm.Net Displacement



(d) Storm.LCL Height

FIGURE 16. The distributions of selected important front attributes shown as schematic plots. Positive storms are tornadic and negative are nontornadic. The red line is the median, and the upper and lower edges of the blue box are the upper and lower quartiles of the distribution. The indentations around the median line indicate the range for 95% statistical significance.

To understand which variables are the most important in determining whether a supercell is tornadic, we calculated the variable importance for the resampled data, as shown in Table 4. Seven of the top twelve variables were associated with the storm only, indicating that characteristics of the storm environment are generally more influential than conditions along surrounding boundaries. Air temperature, equivalent potential temperature (theta-e), and relative humidity were the most important variables. For a greater understanding of why some of these variables are important, we produced schematic plots [42] that visually compare the medians, upper and lower quartiles, and full range of the distributions for both types of storms (Fig. 16). Although the distributions for the mean and maximum values of temperature, theta-e, and relative humidity were similar for tornadic and nontornadic storms, the medians of those quantities were significantly higher for tornadic storms (Fig. 16). In general, warmer and more moist environments tend to be more unstable thereby enhancing the potential for tornadic storms. Net displacement is significantly higher for tornadic storms (Fig. 16(c)) and is tied to the duration of the storm, which is consistent with the findings of [8] that long duration supercells are more likely to be tornadic. Storm pressure is related to the intensity of the storm. Lifted Condensation Level (LCL) Height estimates the distance from the cloud base to the ground and is proportional to the dew point depression. Bunkers [8] and others have shown that lower LCL heights are associated with weaker downdrafts and cold pools, leading to longer-lasting supercell storms and more favorable environments for tornadoes. In the dataset, the mean LCL heights distribution was more concentrated in the lower values for tornadic supercells (Fig. 16(d)) although the medians were not significantly different. As shown by the selection of important variables, the SRRF confirms trends discussed in the literature.

Figure 17 shows the tree with the highest GSS on the training set in the 1000 sample single-tree forests. Because it is difficult to fit even a single tree on the page, we focused only on the single-tree forests for this figure. Each of the attribute questions show the viable range for that split. For example, the double circled node splits the data by asking "Is the storm's maximum ThetaE ≥ 351.88?" The range for this split $[349.38, 352.08]$, which means that ThetaE can vary by about ± 2 and the performance of the tree will remain the same. In a physical context, the tree is capturing the potential breakpoint for a critical thermodynamic variable associated with intense convective thunderstorms. ThetaE is a conserved thermodynamic quantity that takes into account the combined temperature and moisture properties of the atmosphere. As ThetaE increases, the potential instability of the atmosphere increases and the likelihood of intense convection also increases. For this study, the value of 351.88 yields a node that identifies significantly greater occurrences of tornadic supercells given the observed values of ThetaE which further reinforces the notion that surface-based observations can yield predictive skill in identifying environmental conditions that can differentiate between tornadic an non-tornadic storms.

6. Drought

Drought, loosely defined as insufficient water for normal purposes, has one of the highest costs of any natural event in terms of socioeconomic loss. In the United States alone, drought has cost the economy over $5B annually on average since 1980 and extreme drought events rival hurricanes in their destructive potential [21]. Although drought differs significantly from the previous application domains, the impact demonstrates that there is a need for an improved understanding of drought. One of the interesting differences for SRRFs is that drought acts on a much slower temporal and much wider spatial scale.

The geographical extent of our drought analysis roughly corresponds to the Southern Great Plains of the United States. Because we have previously demonstrated [9] that the Palmer Drought Severity Index (PDSI) exhibits strong spatial and temporal structure in terms of its predictability, we continue to focus on the PDSI data. The PDSI drought data is provided on a 2.5 degree geographic coordinate grid and each coordinate has 134 years of data recorded in one month intervals[4]. Incomplete data

[4] http://iridl.ldeo.columbia.edu/

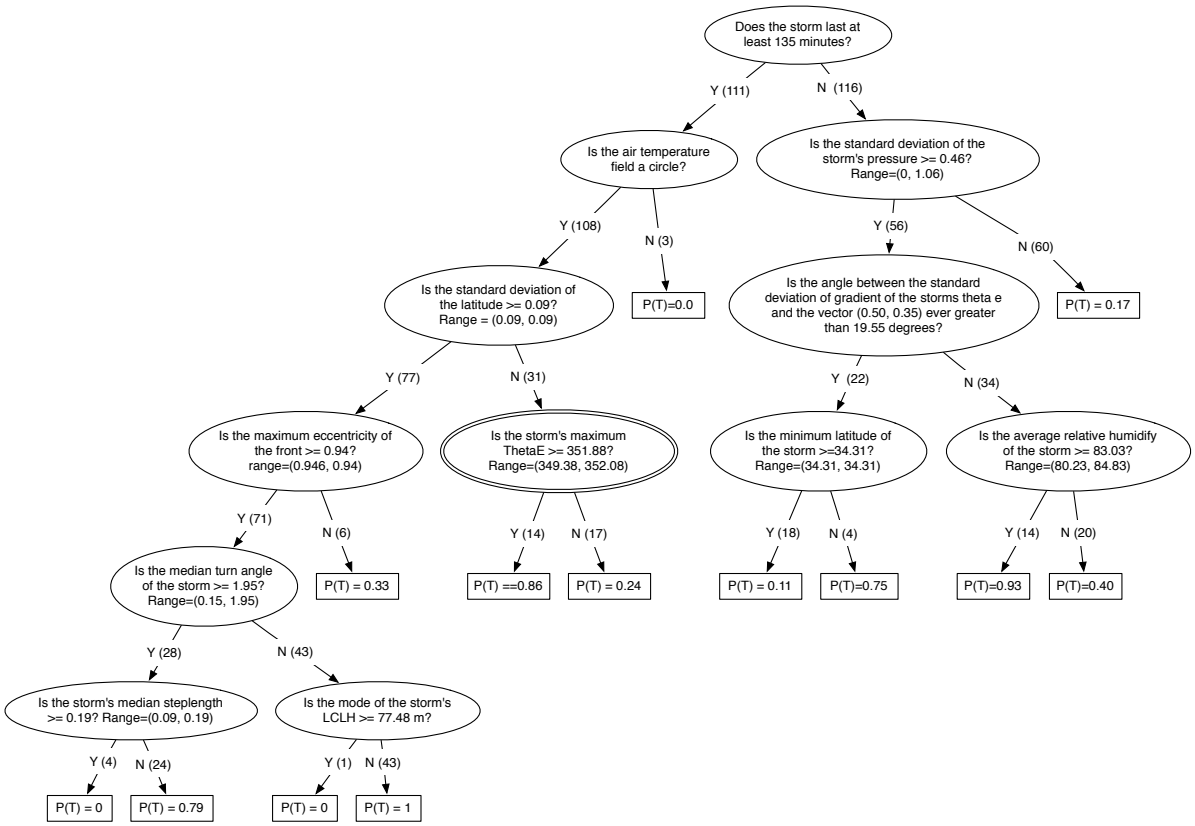


FIGURE 17. Most highly ranked SRPT from the fronts domain for 1000 samples. The tree shows the ranges on the attribute questions. The node that is double circled is discussed in the text. P(T) denotes the probability of a tornadic supercell event.

records due to the presence of bodies of water and the slow early establishment of meteorological records reduce the number of useful grid cells around the edges. The locations of the data points are shown in Figure 18.

Figure 19 shows the schema for the spatiotemporal relational data used to study the PDSI. The inherent gridded nature of the data logically leads to using each grid point as an object and the relations are the spatial relationships between the grid points. We focus on labeling the center point of a 3x3 spatial grid given the PDSI value over the previous 3 months at all neighboring locations. A graph is labeled as positive if the center grid point is in drought in the current month. With 134 years of data, we have approximately 1600 graphs for each location.

For the drought data, we performed several experiments. First, we varied the number of trees in the forest and the number of samples as described for the previous domains. For this experiment, we focused on the location of Tulsa, Oklahoma. The reason for running this experiment on only one location was to find the best set of parameters and then repeat those parameters across the entire data set, focusing on the variable importance analysis.

Figure 20 shows the GSS and AUC as a function of the number of distinctions sampled and the number of trees in the forest. As with the previous domains, performance increases as the number of trees increases although there is a quick asymptote with anything greater than 10 trees. Performance also improves as the number of samples increases to 100 and then asymptotes. This is true for both

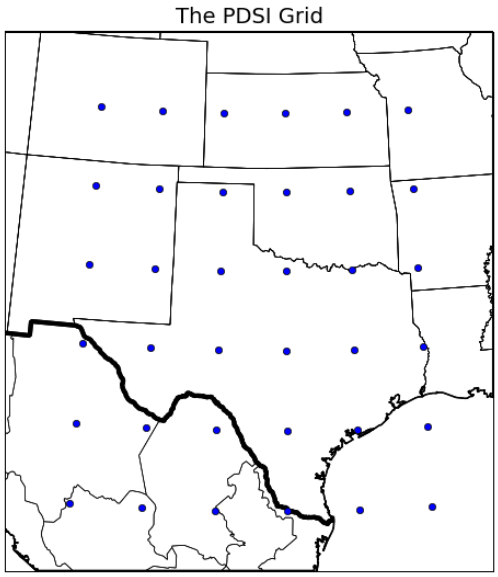


FIGURE 18. Domain of study for drought. Each of the dots shows the location of the PDSI drought data. We focus on the southern plains of the United States for this work.

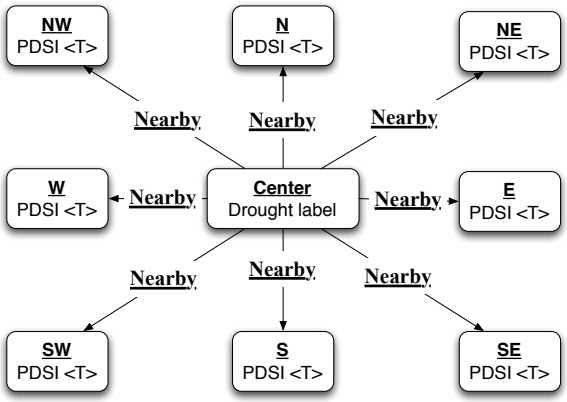
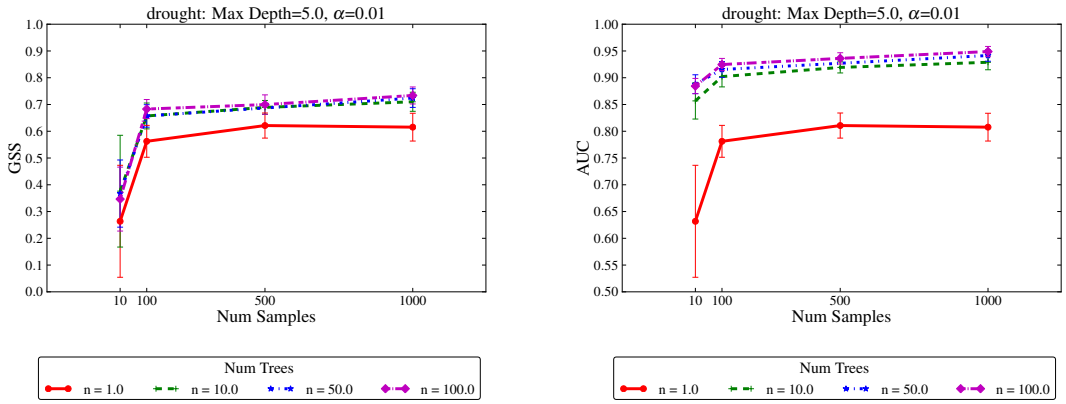


FIGURE 19. Schema for the drought data

performance measures. As expected, a single tree forest is less robust than a multi-tree forest, as can be seen in the AUC graph.

Domain scientists want to be able to use a model such as the SRRF to better understand drought, not just to predict it. We focus on the variable importance for this aspect. For this experiment, we trained a SRRF with 50 trees and 100 samples for all 18 locations that have sufficient data at all neighboring locations. We ran 30 runs of this training with the same parameter set and used variable importance to analyze which direction is most important in predicting drought.

Figure 21 shows the corresponding map of the results obtained using the SRRF. The length of the arrows emanating from each grid point indicates the variable's importance. For example, a



(a) GSS for drought at Tulsa, Oklahoma

(b) AUC for drought at Tulsa, Oklahoma

FIGURE 20. a) GSS and b) AUC for predicting drought in Tulsa, Oklahoma as a function of the number of distinctions sampled and number of trees in the forest.

FIGURE 21. Importance of spatiotemporal information, as a function of direction, in the prediction of future states of drought.

long arrow pointing towards the southeast would indicate that spatiotemporal information to the southeast of the center grid point is more useful in predicting the future occurrence of drought in the center than a direction that exhibited a lower variable importance (shorter arrow).

It is immediately seen that spatiotemporal structure exists in the abilities of the various cardinal and inter-cardinal directions to predict the presence of drought at the center grid points. This highlights the potential ability of the SRRF algorithm to aid in drought response planning and mitigation over short time spans. However, not only does Figure 21 demonstrate the ability to predict, it also begins to hint at geographic structure with regards to how drought responds to its spatiotemporal informational surroundings. This is most clearly seen from the similarity of the rosettes of variable importance surrounding the sites in Eastern and Central Kansas. Their qualitative similarity is suggestive that drought behaves similarly across this geographic region. Other potential regions may be seen in the Western Oklahoma/Northern Texas Panhandle, and in the Southeastern New Mexico/Southern Texas Panhandle rosettes.

Our results are encouraging and warrant further investigation into the strength of the similarity between rosettes, the inclusion of seasonality into the study, and the variations that drought indices different from the PDSI might present. And finally, as nearly all geographic regions exhibit individualized behavior, rather than relying upon Tulsa to calibrate the experimental parameters, each grid cell should be examined for its own set of "best parameters."

7. Conclusions

We have introduced and validated a significantly augmented Spatiotemporal Relational Random Forest, a new Random Forest based algorithm that learns with spatiotemporally varying relational data. We have focused our application of the SRRF algorithm on three real-world severe weather domains: turbulence, tornadoes and drought. In each domain, we demonstrated that the SRRF is a strong predictor and that the variable importance analysis significantly aids human understanding of the results. The contributions of this paper include the enhanced SRRF algorithm, the variable importance analysis for spatiotemporally varying relational data, the enhancements of the underlying SRPT, parameter exploration, and a thorough validation on real-world severe weather data.

The current FAA turbulence prediction algorithm, GTG [35], is based primarily on NWP model data, though efforts are underway to integrate observations to better diagnose convective turbulence [44]. We anticipate that the SRRF will aid in this improvement by uncovering new spatiotemporal relationships with predictive value via the variable importance analyses. Furthermore, to evaluate its potential to become a useful component of the prediction algorithm, we are evaluating gridded predictions made by the SRRF on case studies drawn from selected days. We hope to integrate the SRRF into the next-generation turbulence prediction product.

Our work in the tornado domain is a piece of a larger project focusing on understanding the formation of tornadoes through high resolution simulations as well as the analysis of observational data. Future work on this same 10-year climatological data set includes extending the time period, extending the period before each storm, and expanding the set of environmental variables. All of our work on tornadoes will also be immediately relevant for the Warn-on-Forecast models being developed for the National Weather Service. Our study of a 10 year dataset of tornadoes in Oklahoma is helping to better understand "what" atmospheric variables are critical "when." This provides basic new insights into the overall set of processes related to the occurrence of tornadic supercells. In the future, this will be integrated with the knowledge gained through field studies such as VORTEX 2[5].

Our drought application is also a piece of a larger project studying the predictability of drought in the continental United States using a variety of data mining techniques. The goal of this project is to improve our understanding of how drought moves and thus to improve the predictions of drought, enabling those affected by it to mitigate the impact.

[5] http://www.vortex2.org/home/

Our long term goals include integration of the SRRF and/or the knowledge gained from the SRRF into current prediction algorithms. As part of this, we need to demonstrate that our algorithms are skillful in whatever way that skill is measured for each domain and we need to compare to existing algorithms. For tornadoes, skill is most often measured using probability of detection (POD) and false alarm ratio (FAR) but there is not an a-priori level of POD/FAR or AUC/GSS such that an algorithm can be automatically used to issue tornado warnings. Studies such as [7] indicate that these may not even be the best skill measures to use. Currently, tornado warnings are issued by a human forecaster who examines the weather using all of the available data sources. The human may use current tornado warning algorithms such as [37, 27] to inform their decision but the ultimate warning is based on the human's forecasting knowledge. One of NOAA's long term goals is to warn on computer forecasts[6] and our algorithms will integrate well into this paradigm. We also can't compare directly to the algorithms used by the human forecasters as we work on radically different data. Both of the algorithms in use by the weather service that are cited above are designed for radar-based detections. We have a significantly augmented data set (Mesonet) but this data is only available in Oklahoma and not yet deployed across the United States. This means that any comparison would be on very limited data (Oklahoma only) and would be unfair (comparing apples to oranges).

It is also unclear what the critical skill level for turbulence is and what the right comparison algorithm would be. For skill, there are draft performance requirements available at [1] but they include statements like, "The NextGen NAS shall forecast CIT of EDR .375 or greater in Super Density Terminal Airspace to verify greater than or equal to 95% for forecasts from 0 to 45 minutes." It is not clear what "verify greater than or equal to 95%" means. If 95% is the POD, it begs the question of what the FAR limit is. In a NASA Airborne Turbulence Detection Systems project that NCAR participated in in the early 2000s, the goal was to use airborne radar data to achieve a PoD of $> 80\%$ with a false alert rate $< 20\%$ in regions where the radar was able to take measurements with a signal to noise ratio > 15 dB. These were object-based statistics, not pointwise statistics, however, so they may not compare directly to our analysis. In addition, the AUC and GSS scores do not lend themselves directly to analyzing the POD/FAR tradeoffs.

On the question of comparing to existing algorithms for turbulence, Sharman [35] reports an AUC of 0.878 for upper levels and 0.818 for midlevels for GTG. These were sufficient for GTG to be approved by the FAA for operational use. CIT is more challenging than the clear-air turbulence predicted by GTG, so the SRRF AUCs > 0.8 show good promise for operational usefulness. However, this is not an apples-to-apples comparison, since GTG was verified over a different time period and with pilot reports, not the automated in-situ EDR that we are using.

Research Reproducibility: All of the graphs and data from the parameter exploration studies, the full set of graphs for all domains, and the code used for all of the experiments are available at: `http://idea.cs.ou.edu/cidu2010/sadm_extended`

8. Acknowledgments

We would like to thank Matthew Collier for his work in the drought domain and Timothy Supinie for his work in the turbulence domain. This material is based upon work supported by the National Science Foundation under Grant No. NSF/IIS/0746816 and related REU supplements NSF/IIS/0840956 and NSF /IIS/0938138. This research was supported in part by NASA under Grants No. NNS06AA61A and NNX08AL89G. The Oklahoma Mesonet is funded by the taxpayers of Oklahoma through the Oklahoma State Regents for Higher Education and the Oklahoma Department of Public Safety.

References

[1] Nextgen wx performance requirement - forecast. http://www.jpdo.gov/library/Appendix

[6] http://www.nssl.noaa.gov/projects/wof/

[2] J. F. Allen. Time and time again: The many ways to represent time. *International Journal of Intelligent Systems*, 6(4):341–355, 1991.

[3] M. Bodenhamer, S. Bleckley, D. Fennelly, A. H. Fagg, and A. McGovern. Spatio-temporal multi-dimensional relational framework trees. In *Proceedings of the International Workshop on Spatial and Spatiotemporal Data Mining, IEEE Conference on Data Mining*, 2009. electronically published.

[4] A. Bosch, A. Zisserman, and X. Munoz. Image classification using random forests and ferns. In *Proceedings of the International Conference on Computer Vision*, 2007.

[5] A. P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.

[6] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[7] H. E. Brooks. Tornado-warning performance in the past and future. *Bulletin of the American Meteorological Society*, pages 837–843, 2004.

[8] M. J. Bunkers, J. S. Johnson, L. J. Czepyha, J. M. Grzywacz, B. A. Klimowski, and M. R. Hjelmfelt. An observational examination of long-lived supercells. part II: Environmental conditions and forecasting. *Weather and Forecasting*, 21:689–714, 2006.

[9] M. Collier and A. McGovern. Mining spatiotemporal data to map drought transitions. *International Journal of Geographical Information Science*, in preparation.

[10] L. B. Cornman, G. Meymaris, and M. Limber. An update on the faa aviation weather research program's in situ turbulence measurement and reporting system. In *AMS 11th Conf. Aviat. Range Aerospace Meteorol.*, volume 11, page 4.3, 2004.

[11] L. B. Cornman, C. S. Morse, and G. Cunning. Real-time estimation of atmospheric turbulence severity from in-situ aircraft measurements. *Journal of Aircraft*, 32:171–177, 1995.

[12] J. P. Egan. *Signal Detection Theory and ROC Analysis*. Series in Cognition and Perception. Academic Press, New York, 1975.

[13] A. Fern. A simple-transition model for relational sequences. In *Proc. of the Intl. Joint Conference on Artificial Intelligence*, pages 696–701, 2005.

[14] P. O. Fislason, J. A. Benediktsson, and J. Sveinsson. Random forests for land cover classification. *Pattern Recognition Letters*, 27(4):294–300, 2006.

[15] L. S. Gandin and A. H. Murphy. Equitable skill scores for categorical forecasts. *Monthly Weather Review*, 120(2):361–370, 1992.

[16] J. P. Gerrity. A note on Gandin and Murphy's equitable skill score. *Monthy Weather Review*, 120(11):2709–2712, 1992.

[17] J. Hocker and J. Basara. A geographic information systems-based analysis of supercells across Oklahoma from 1994-2003. *J. Appl. Meteor. Climatol.*, 47:1518–1538, 2008.

[18] J. Jenkner, M. Sprenger, I. Schwenk, C. Schwierz, S. Dierer, and D. Leuenberger. Detection and climatology of fronts in a high-resolution model reanalysis over the Alps. *Meteorological Applications*, 2010.

[19] D. Jensen. Proximity knowledge discovery system. kdl.cs.umass.edu/proximity, 2005.

[20] K. Kersting, L. De Raedt, and T. Raiko. Logical hidden Markov models. *Journal of Artificial Intelligence Research (JAIR)*, 25(425-456), 2006.

[21] N. Lott and T. Ross. Tracking and evaluating U.S. billion dollar weather disasters. In *Preprints of the 86th Annual Meeting of the American Meteorological Society*, Atlanta, GA, 2006.

[22] P. Markowski, E. Rasmussen, and J. Straka. The occurrence of tornadoes in supercells interacting with boundaries during vortex-95. *Wea. Forecasting*, 13:852–859, September 1998.

[23] A. McGovern, N. Hiers, M. Collier, D. J. Gagne II, and R. A. Brown. Spatiotemporal relational probability trees. In *Proceedings of the 2008 IEEE International Conference on Data Mining*, pages 935–940, Pisa, Italy, 2008.

[24] A. McGovern, T. Supinie, D. J. Gagne II, N. Troutman, M. Collier, R. A. Brown, J. Basara, and J. Williams. Understanding severe weather processes through spatiotemporal relational random forests. In *Proceedings of the 2010 NASA Conference on Intelligent Data Understanding*, pages 213–227, 2010.

[25] R. A. McPherson, C. A. Fiebrich, K. C. Crawford, R. L. Elliott, J. R. Kilby, D. L. Grimsley, J. E. Martinez, J. B. Basara, B. G. Illston, D. A. Morris, K. A. Kloesel, S. J. Stadler, A. D. Melvin, A. J. Sutherland, H. Shrivastava, J. D. Carlson, J. M. Wolfinbarger, J. P. Bostic, and D. B. Demko. Statewide monitoring of the mesoscale environment: A technical update on the Oklahoma Mesonet. *J. of Atmos. and Oceanic Technology*, 24:301–321, 2007.

[26] N. Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999, 2006.
[27] E. D. Mitchell, S. V. Vasiloff, G. J. Stumpf, A. Witt, M. D. Eilts, J. Johnson, and K. W. Thomas. The National Severe Storms Laboratory tornado detection algorithm. *Weather and Forecasting*, 13(2):352–366, 1998.
[28] J. Neville and D. Jensen. Dependency networks for relational data. In *Proceedings of the Fourth IEEE International Conference on Data Mining*, pages 170–177, 2004.
[29] J. Neville, D. Jensen, L. Friedland, and M. Hay. Learning relational probability trees. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 625–630, 2003.
[30] F. Provost and T. Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42(3):203–231, 2001.
[31] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.
[32] R. Renard and L. Clarke. Experiments in numerical objective frontal analysis. *Mon. Wea. Rev.*, 93:547–556, 1965.
[33] M. R. Segal. Machine learning benchmarks and random forest regression. Technical report, Center for Bioinformatics and Molecular Biostatistics, April 14 2004.
[34] U. Sharan and J. Neville. Temporal-relational classifiers for prediction in evolving domains. In *Proceedings of the IEEE International Conference on Data Mining*, 2008.
[35] R. Sharman, C. Tebaldi, G. Wiener, and J. Wolff. An integrated approach to mid- and upper-level turbulence forecasting. *Weather and Forecasting*, 21:268–287, 2006.
[36] D. J. Stensrud, M. Xue, L. J. Wicker, K. E. Kelleher, M. P. Foster, J. T. Schaefer, R. S. Schneider, S. G. Benjamin, S. S. Weygandt, J. T. Ferree, and J. P. Tuell. Convective-scale warn on forecast system: A vision for 2020. *Bulletin of the American Meteorological Society*, 90:1487–1499, 2009.
[37] G. J. Stumpf, A. Witt, E. D. Mitchell, P. L. Spencer, J. Johnson, M. D. Eilts, K. W. Thomas, and D. W. Burgess. The National Severe Storms Laboratory mesocyclone detection algorithm for the WSR-88D. *Weather and Forecasting*, 13(2):304–326, 1998.
[38] T. Supinie, A. McGovern, J. Williams, and J. Abernethy. Spatiotemporal relational random forests. In *Proceedings of the IEEE International Conference on Data Mining (ICDM) workshop on Spatiotemporal Data Mining*, page electronically published, 2009.
[39] S. B. Trier and R. Sharman. Convection-permitting simulations of the environment supporting widespread turbulence within the upper-level outflow of a mesoscale convective system. *Mon. Wea. Rev.*, 137:1972–1990, 2009.
[40] J. Trueblood, T. Sliwinski, D. J. Gagne II, A. McGovern, J. K. Williams, and J. Abernethy. Spatiotemporal relational random forest (srrf) prediction of convectively-induced turbulence: a severe encounter case study. In *Presented at the Ninth Conference on Artificial Intelligence and its Applications to the Environmental Sciences*, 2011.
[41] J. M. Wallace and P. V. Hobbs. *Atmospheric Science: An Introductory Survey*. Elsevier, New York, second edition, 2006.
[42] D. S. Wilks. *Statistical Methods in the Atmospheric Sciences*. Academic Press, 2nd edition, 2006.
[43] J. Williams, D. Ahijevych, S. Dettling, and M. Steiner. Combining observations and model data for short-term storm forecasting. *W. Feltz and J. Murray, Eds., Remote Sensing Applications for Aviation Weather Hazard Detection and Decision Support. Proceedings of SPIE*, 7088:paper 708805, August 2008.
[44] J. K. Williams, R. Sharman, J. Craig, and G. Blackburn. Remote detection and diagnosis of thunderstorm turbulence. In *Proceedings of SPIE*, volume 7088. Remote Sensing Applications for Aviation Weather Hazard Detection and Decision Support, 2008.
[45] J. Wolff and R. Sharman. Climatology of upper-level turbulence over the continental united states. *J. Appl. Meteor. Climatol.*, 47:2198–2214, 2008.