# UNDERSTANDING SEVERE WEATHER PROCESSES THROUGH SPATIOTEMPORAL RELATIONAL RANDOM FORESTS

AMY MCGOVERN[1], TIMOTHY SUPINIE[2], DAVID JOHN GAGNE II[2], NATHANIEL TROUTMAN[1], MATTHEW COLLIER[3], RODGER A. BROWN[4], JEFFREY BASARA[5], AND JOHN K. WILLIAMS[6]

Abstract. Major severe weather events can cause a significant loss of life and property. We seek to revolutionize our understanding of and 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 translation 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.

## 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) [4] 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 [7, 12, 23]) and our recent work is the the only work that we know of that addresses both spatiotemporal and relational data [15, 26, 2].

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 reasoning of the models. Prior work demonstrated that RFs were a promising approach in the turbulence domain [29]. This summer, we are performing case studies of the SRRF and and investigating the possibility of integrating the trained SRRFs into an operational turbulence guidance product

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

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 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 tornadoes and frontal boundaries 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 [14]. The goal of this part of the project is to use SRRFs and objective front analysis to perform a climatological study of tornadic supercell thunderstorms and how the relative positions of fronts 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 [25]. 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 [4] are a simple and powerful algorithm with a strong track record (e.g., [22, 17, 8, 3, 28]). RFs learn an ensemble of C4.5 [20] 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 [26]. This paper represents a significant extension of that work. The contributions of this paper are: 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. 2) Our underlying Spatiotemporal Relational Probability Tree (SRPT, [15]) 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. 3) We have thoroughly explored the parameter space of the algorithm on all of our domains. 4) We have significantly extended the application to multiple real-world severe weather domains in preparation for extensive field testing occurring in the summer and fall of 2010.

## 2. Growing SRRFs

Growing a SRRF is very similar to the approach used to grow a RF [4] 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 learning.

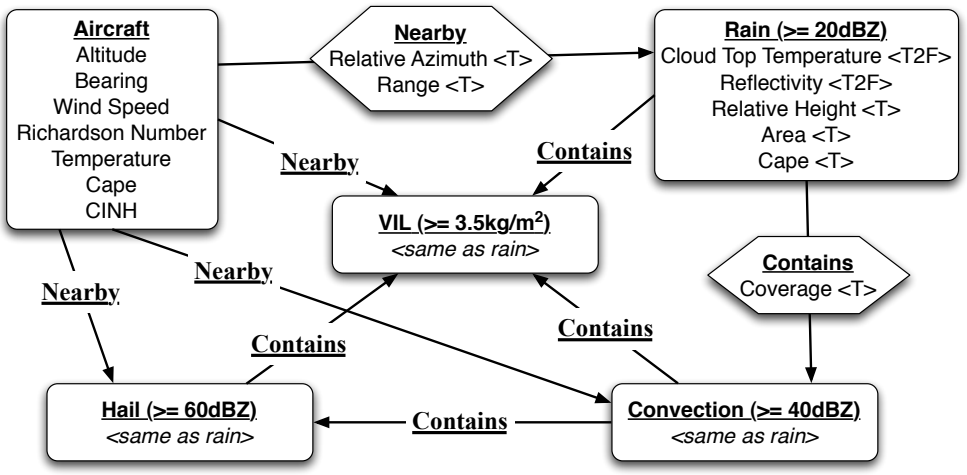



FIGURE 1. Schema for aircraft turbulence data set. Object and relationship types are underlined and bolded. Temporal attributes are denoted with a T and fielded attributes with a F (with 2F specifying 2-dimensional fields).

The data are represented as spatiotemporal attributed relational graphs [15]. This representation is an extension of the attributed graph approach [19, 18, 11] to handle spatiotemporally varying data. All *objects*, such as people, places, or events, 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. 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 typed. In this case, the attributes on the rain, hail, convection, and vertically integrated liquid objects are all 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.

An SRRF is composed of individual Spatiotemporal Relational Probability Trees (SRPTs) [15], which are probability estimation trees similar to Relational Probability Trees [19] 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 $v$?

**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 ← **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

**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 ← $\emptyset$
**for** *i = 1 to s* **do**
  split ← generate random split
  eval ← evaluate quality of split (using chi-squared)
  **if** *eval* $<$ *p* ***and*** *eval* $<$ *best evaluation so far* **then**
    best ← split
  **end**
**end**
Return best

- 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* [1].
- Temporal Partial Derivative: Is the partial derivative with respect to time on attribute $a$ on 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 describe the procedure for growing an individual tree. This procedure follows the standard greedy decision tree algorithms with the exception of the sampling of the splits. Because there is a very large number of possible instantiations for the split templates listed above, we sample

the specific splits 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 was trained on a different subset of the available attributes. Since the individual trees were standard C4.5 decision trees, this limited the number of possible splits each tree could make. Because each tree was also trained on a different bootstrap resampled set of the original data, the trees were 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, an 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.

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 each 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. 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.

## 3. Parameter Exploration

In order to study the effects of the parameters on the SRRF algorithm, we performed a combinatorial experiment on two datasets. The primary parameters that affect the performance of the SRRF are the number of possible splits each SRPT can examine at each level of tree growth (this is analogous to the number of attributes in a C4.5 tree), 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.

- Number of samples: [10, 100, 500, 1000, 5000].
- Maximum depth of the tree: [1, 3, 5].
- Number of trees in the forest: [1, 10, 50, 100].

- We fixed the p-value to 0.01
- Distinctions: [all, non-temporal only]

This yields 120 parameter sets in total, each of which is run 30 times for statistical testing. Due to space limitations, these results are presented online at `http://idea.cs.ou.edu/cidu2010/`.

## 4. Convectively-induced turbulence

Convectively-induced turbulence (CIT) – 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) [24]. 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 [27]. 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 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 these experiments were collected from United Airlines Boeing 757 aircraft in the summer of 2007. Convection is most prevalent 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 using 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, 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 turbulence (1514 negatives and 541 positives).

The data available for this study comes from a combination of the 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]). This is 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. Each of the objects represents a meteorological concept

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

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

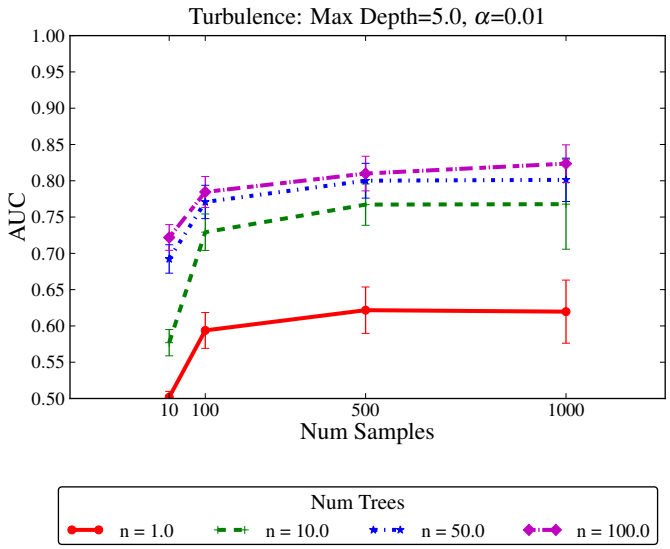



FIGURE 2. AUC for the Turbulence data as a function of the number of splits sampled at each node of tree growth size for 10-, 50-, and 100-tree SRRFs and a single SRPT. The maximum tree depth was fixed at 5 and the chi-squared threshold at 0.01.

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). 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 2 shows the Area Under the Receiver Operating Characteristic Curve (AUC) for the SRRF turbulence predictions on an independent test set as a function of the number of distinctions sampled in the forest. AUC is a standard measure of performance of a probabilistic classifier. An AUC of 1 indicates perfect performance and an AUC of 0.5 indicates random performance. The maximum tree depth for this graph was fixed at 5. As the sample size increases, the performance of the forest increases and then asymptotes. 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. Additionally, increasing the number of trees in the SRRF increases the performance. This behavior is also expected as ensembles with more members are expected to better capture the underlying relationships. 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 trees become more similar. RF performance has also been shown to asymptote as the diversity of the trees in the forests is reduced [4].

Table 1 gives the importance of the top 10 attributes in the turbulence data. Attributes on objects list the object they are associated with (e.g. VIL.Area means the area attribute of objects of type VIL) and attributes listed with an arrow are on the relations. For example, the contains relationship

TABLE 1. Top 10 statistically significant important attributes ($\alpha = 0.05$) in the turbulence data for a forest with 100 trees, 1000 samples at each node, max tree depth of 5. This is computed over 30 runs.

| Attribute | Mean Variable Importance |
|---|---|
| VIL.Area | 0.198 |
| Aircraft.RichardsonNumber | 0.106 |
| Rain→Contains.Coverage→VIL | 0.085 |
| Rain→Contains.Coverage→Convection | 0.084 |
| Convection→Contains.Coverage→VIL | 0.061 |
| Rain.Area | 0.056 |
| Hail.CloudTopTemperature | 0.054 |
| Aircraft→Nearby.Range→Rain | 0.053 |
| VIL.CloudTopTemperature | 0.052 |
| Aircraft→Nearby.Range→Vil | 0.048 |

between Rain and VIL objects has a Coverage attribute that is the second most important attribute. Most of these attributes characterize the storm environment. The most important attribute, the area on VIL objects, reflects the size of active thunderstorms in the vicinity of the aircraft. Large thunderstorms are often more intense and longer-lived, with greater outflow and environmental disturbance than smaller storms. The Rain object's area may play a similar role, though somewhat less effectively. Cloud top temperatures within Hail and VIL objects provide additional indications of storm severity; cold cloud tops suggest deeper clouds and potentially more powerful updrafts, downdrafts and gravity waves. The range attribute on the relationship between aircraft and both rain and VIL objects indicates the proximity of the plane to precipitating cloud or active convection, and hence is related to the storm's ability to influence it. The coverage attribute on the contains relationship between Rain and VIL and Convection objects denotes the fraction of active convection the larger rain regions, which may help distinguish rain due to convection from less turbulence-prone stratiform or orographic rainfall. Only one of these top 10 attributes is derived from the NWP model analysis: the Richardson number at the plane location indicates both turbulence due to the model-resolved storm and also non-CIT turbulence related to environmental factors such as the jet stream that may also occur in the dataset.

## 5. Surface boundaries and tornadogenesis

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 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. [14] 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 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.

Our data was created from a ten-year analysis of supercell thunderstorms and surface boundaries in the state of Oklahoma. The supercell data came from a climatology of 926 Oklahoma supercells from 1994-2003 by Hocker and Basara [9]. Surface frontal boundaries associated with each supercell

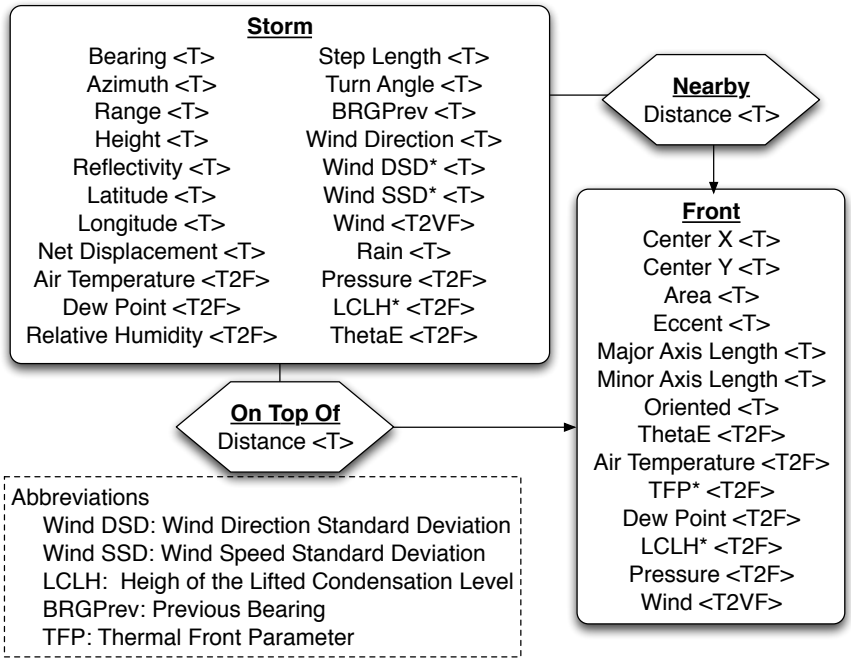



FIGURE 3. Schema for tornadogenesis data. Temporal data is denoted with a <T> and 2-dimensional fielded data with a <T2F>.

TABLE 2. 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 |

were analyzed from Oklahoma Mesonet surface observations [16] using objective front analysis techniques [21, 10]. 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 3, where Nearby relationships indicated storms and fronts less than 40 km apart and 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 attributes.

Table 2 shows the class distribution of the supercell thunderstorms and Figure 4 shows the spatial distribution of tornadic supercells in Oklahoma. Most supercells in the data were found to be non-tornadic. 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 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. As with the previous experiment, we examined the AUC as a function of the number of trees in the forest and the number of distinctions sampled at each level.

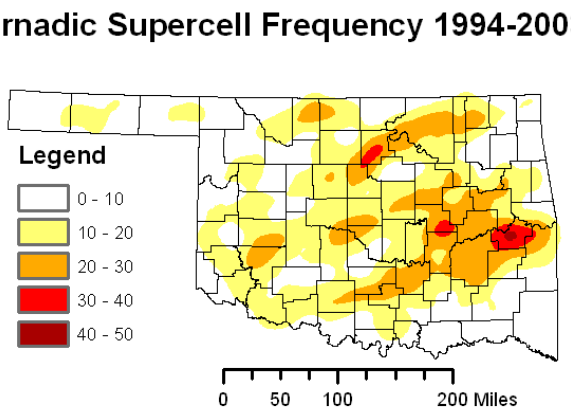



FIGURE 4. Number of tornadic supercells that have passed within 30 km of a point from 1994-2003.

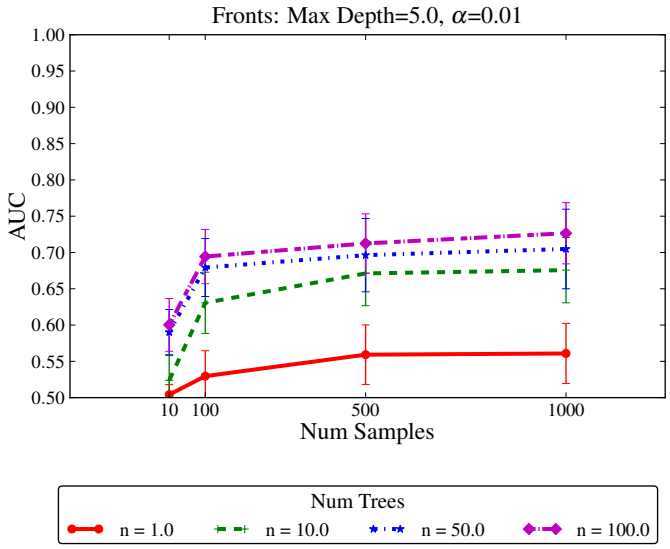



FIGURE 5. AUC for the Fronts and Tornado Data as a function of sample size for 10- and 50-tree SRRFs and a single SRPT. Error bars indicated 95% confidence intervals.

These results are shown in Figure 5. The AUC indicates that this is a robust classifier and the forests are again able to outperform the single SRPT. Also, as with the turbulence data, 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.

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 3. Seven of the top ten variables were associated with the storm only, indicating that characteristics of the

TABLE 3. The top 10 most important variables for the front and tornado data, averaged over 30 runs of a 100-tree SRRF with a sample size of 1000 and a maximum tree depth of 5.

| Attribute | Mean Variable Importance |
|---|---|
| Storm.AirTempature | 0.162 |
| Storm.ThetaE | 0.130 |
| Storm.NetDisplacement | 0.120 |
| Storm→Nearby.RelativeAzimuth →Front | 0.117 |
| Storm.DewPoint | 0.112 |
| Storm.Bearing | 0.109 |
| Front.ThetaE | 0.088 |
| Front.ThermalFrontParameter | 0.085 |
| Storm.Pressure | 0.085 |
| Storm.LiftedCondensationLevelHeight | 0.079 |

storm environment are generally more influential than conditions along surrounding boundaries. Air temperature, equivalent potential temperature (theta-e), and dewpoint were all among the most important variables, which is potentially indicative that storms have different tornadic probabilities given different moisture and heating conditions. Net displacement is tied to the duration of the storm, which is consistent with the findings of [5] that long duration supercells are more likely to be tornadic. The angle between the storm and the front and the bearing of the storm considered highly important but not the distance to the front, so how the storm moves relative to local boundaries is more indicative of tornadic potential than how far away a boundary is. A storm's motion can affect how long it remains in a favorable environment and from there affect the tornadic potential. 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 directly related to the dew point depression. Bunkers [5] 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. As shown by the selection of important variables, the SRRF confirms trends discussed in the literature for the studied domain.

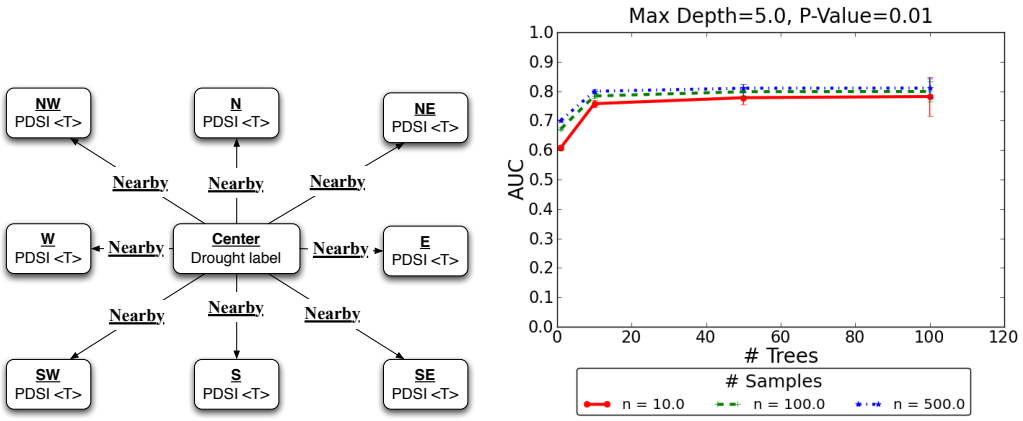
## 6. DROUGHT

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 [13]. 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 [6] 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[3]. Incomplete data 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.

Figure 6a 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

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




(a) Schema for drought data

(b) AUC on drought data

FIGURE 6. a) Schema for the drought data. b) AUC as a function of the number of distinctions sampled and the number of trees in the forest at Tulsa, Oklahoma.

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 all of 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 6b shows the AUC as a function of the number of distinctions samples and the number of trees in the forest. As with the previous domains, performance increases as the number of trees increases and asymptotes around 50 to 100 trees. Performance also improves as a function of the number of samples while asymptoting around 500 samples.

Domain scientists want to be able to use such a model 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 7 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 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 7 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.

FIGURE 7. Importance of spatiotemporal information, as a function of direction, in the prediction of future states of drought.

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 [24], is based primarily on NWP model data, though efforts are underway to integrate observations to better diagnose convective turbulence [29]. 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. Based on the results of this study, we hope to integrate the SRRF into the current prediction product in the Fall of 2010.

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[4].

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.

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

## 8. Acknowledgments

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

[2] 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.

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

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

[5] 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.

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

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

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

[9] 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.

[10] 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.

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

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

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] N. Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999, 2006.

[18] 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.

[19] 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.

[20] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[21] R. Renard and L. Clarke. Experiments in numerical objective frontal analysis. *Mon. Wea. Rev.*, 93:547–556, 1965.

[22] M. R. Segal. Machine learning benchmarks and random forest regression. Technical report, Center for Bioinformatics and Molecular Biostatistics, April 14 2004.

[23] U. Sharan and J. Neville. Temporal-relational classifiers for prediction in evolving domains. In *Proceedings of the IEEE International Conference on Data Mining*, 2008.

[24] 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.

[25] 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.

[26] 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.

[27] J. M. Wallace and P. V. Hobbs. *Atmospheric Science: An Introductory Survey*. Elsevier, New York, second edition, 2006.

[28] 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.

[29] 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.