Using the XSEDE Supercomputing and Visualization Resources to Improve Tornado Prediction Using Data Mining

Bradley Pirtle
School of Electrical and Computer Engineering
University of Oklahoma
Bradley.L.Pirtle-1@ou.edu

Ross Kimes
School of Meteorology
University of Oklahoma
Ross.Kimes@ou.edu

Amy McGovern
School of Computer Science
University of Oklahoma
amcgovern@ou.edu

Rodger Brown
NOAA/National Severe Storms Laboratory
Rodger.Brown@noaa.gov

ABSTRACT
In this paper we introduce the use of XSEDE resources and mathematical models for the simulation of tornadoes, as well as novel techniques for analyzing the results of these simulations.

Categories and Subject Descriptors
I.6.6 **[SIMULATION AND MODELING]**: Simulation Output Analysis

General Terms
Experimentation, Human Factors

Keywords
Data mining, Simulations, Spatiotemoral, Relational

1. INTRODUCTION
Responsible for over four-hundred deaths in the United States in 2011, tornadoes are a significant source of preventable deaths. The reasons for the high number of fatalities are complex and manifold. Arguably the most significant reason is simply a limited understanding of tornadogenisis, the process by which tornadoes form.

The relative infrequency of tornadoes coupled with fundamental limitations of radar has resulted in a limited amount of real-world data. Because tornadoes form by a complex sequence of spatiotemporal events, and little real-world data has been collected, a complete understanding of tornadogenisis has eluded domain scientists.

2. XSEDE RESOURCES
The absence of real-world data can be addressed by running highly sophisticated and computationally intensive mathematical models on Kraken, a high-performance computer (HPC) provided by XSEDE, to numerically simulate supercell thunderstorms. By using simulated data in lieu of real-world data, datasets of arbitrary resolution and scale can be generated as needed, as can be seen in Figure 1.

Our approach is to generate over 100 high resolution simulations of supercell thunderstorms using the Advanced Regional Prediction System (ARPS) [5,6,7]. ARPS is one of the leading systems for numerically simulating mesoscale storms. To properly resolve the tornadoes, each of our simulations uses a horizontal resolution of 75 m. Each simulation requires 3,000 cores on Kraken and runs for approximately 35 hours, and is spread across multiple jobs.

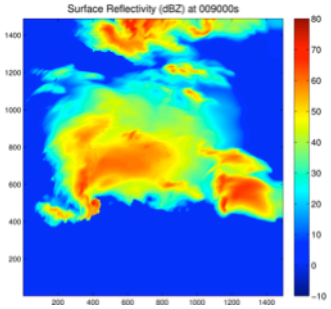

Figure 1: Reflectivity of simulated

In addition to the simulations on Kraken, we use Nautilus, a high performance visualization resource provided by XSEDE. Nautilus provides 4 TB of shared memory, which is crucial for the visualization and data mining of the simulations. We use MATLAB for visualization and data processing, and we use Java for performing data mining.

An additional, crucial XSEDE resource we use for storm simulations and data mining is the High Performance Storage System (HPSS). Each simulation generates over 1 TB of data and HPSS provides a secure place to store the data, as well as an easy way to transfer data from Kraken to Nautilus.

The team at NICS provided advanced user support and was able to significantly improve the efficiency of our workflow on both Kraken and Nautilus. Without this help, our approach would have been untenable.

3. SPATIOTEMPORAL RELATIONAL PROBABILITY TREES AND FORRESTS
Because the dataset will exceed 100 TB, which is far too large for any individual to analyze, we developed Spatiotemporal Relational Probability Trees (SRPTs) and Spatiotemporal Relational Random Forests (SRRFs) [3,4,8]. SRPTs and SRRFs are capable of autonomously analyzing the large dataset and extracting meaningful patterns useful for understanding tornadogenisis.

SRPTs are a spatiotemporal extension of Relational Probability Trees (RPTs) [1]. Relational probability trees were chosen for their strong predictive ability, efficiency, and human-readability. SRPTs differ from RPTs in several respects, however. The most critical difference is SRPTs are capable of creating spatial, temporal, and relational distinctions within each tree. This uniquely gives SRPTs the ability to reason about the spatial, temporal, and relational

Conference'10, Month 1–2, 2010, City, State, Country.

dimensions of a dataset in concert. Because the natural world is inherently spatiotemporal and relational, and many scientific datasets share these properties, this greatly enhances the strength and applicability of SRPTs to real-world datasets.

While SRPTs are powerful predictors, they do suffer from one major weakness: overfitting. Overfitting occurs when an SRPT ceases to discover meaningful, generalizable patterns within a dataset, and instead begins fitting to noise.

To address this issue, Spatiotemporal Relational Random Forests (SRRFs) were developed. Much like a traditional random forest, an SRRF is an ensemble of SRPTs. By growing hundreds of individual SRPTs and intelligently combining each tree's prediction, an SRRF can discover far more complex patterns within the data, while largely mitigating issues of overfitting [8].

4. RESULTS

A preliminary dataset of 253 simulations, each with a horizontal resolution of 500-meters, was generated and analyzed using ten SRRFs. Each SRRF consisted of 96 individual SRPTs, whose distinctions were selected using the chi-squared statistic with a fixed p-value of 0.01, and a maximum depth of ten [2]. An example of an individual tree can be seen below in Figure 2.

Figure 2: Truncated example of an SRPT

The relative predictive strength of each SRRF was evaluated using the Gerrity Skill Score (GSS) metric. GSS varies between -1 and 1, where a value of -1 indicates a perfectly incorrect classifier, a value of 0 indicates a random classifier, and a value of 1 indicates a perfect classifier. The average GSS score across all ten runs of the SRRF was 0.64, which indicates the algorithm is discovering salient patterns within the dataset, which are useful for classifying storms.

5. DISCUSSION AND FUTURE WORK

While the preliminary results for the 500-meter resolution dataset are encouraging, further progress with regards to understanding tornado formation will more likely result from the analysis of simulations with a much greater resolution. Current and future-work consist of continuing to generate a dataset of approximately 100 simulations, each with a 75-meter horizontal resolution, allowing for far more complex patterns to be captured within the dataset. We have spent the past four years creating this dataset, and we expect to complete it within the following year.

Additionally, new spatiotemporal distinctions are being added to the SPRTs, allowing for the analysis of more expressive patterns. Finally, variable importance is being implemented with greater efficiency, which will directly facilitate the discovery and understanding of conditions critical for tornado formation.

By training SRRFs on the simulated tornado dataset, we hope to discover patterns and conditions necessary for the formation of tornadoes. Discovering and understanding these patterns may improve traditional tornado forecasting dramatically. By doing so, we may ultimately reduce the number of preventable fatalities caused by tornadoes.

6. REFERENCES

1. J. Neville, D. Jensen, L. Friedland, and M. Hay. Learning Relational Probability Trees. In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 625–630, 2003.

2. Kimes, Ross, 2012: Spatiotemporal Relational Data Mining of High Resolution Supercell Simulations. Capstone, University of Oklahoma, School of Meteorology.

3. McGovern, Amy and Gagne II, David John and Troutman, Nathaniel and Brown, Rodger A. and Basara, Jeffrey and Williams, John. (2011) Using Spatiotemporal Relational Random Forests to Improve our Understanding of Severe Weather Processes. Statistical Analysis and Data Mining, special issue on the best of the 2010 NASA Conference on Intelligent Data Understanding. Vol 4, Issue 4, pages 407-429.

4. McGovern, Amy; Supinie, Timothy; Gagne II, David John; Troutman, Nathaniel; Collier, Matthew; Brown, Rodger A.; Basara, Jeffrey; Williams, John. (2010) Understanding Severe Weather Processes through Spatiotemporal Relational Random Forests. Proceedings of the NASA Conference on Intelligent Data Understanding: CIDU 2010.

5. Ming Xue and Donghai Wang and Jidong Gao and Keith Brewster and Kelvin K. Droegemeier. "The Advanced Regional Prediction System (ARPS) - storm-scale numerical weather prediction and data assimilation."Meteorology and Atmospheric Physics 82 (2003): 161-193.

6. Ming Xue, Kevin Droegemeier, and V. Wong. "The Advanced Regional Prediction System (ARPS) - A Multiscale Nonhydrostatic Atmospheric Simulation and Prediction Model. Part 1: Model Dynamics and Verfication."Meteorology and Atmospheric Physics 75 (2000): 161-193.

7. Ming Xue, Kelvin K. Droegemeier, V. Wong and A. Shapiro and Keith Brewster and Fred Carr and D. Weber and Y. Liu and D. Wang. "The Advanced Regional Prediction System (ARPS) - A multiscale nonhydrostatic atmospheric simulation and prediction tool. Part 2: Model physics and applications. "Meteorology and Atmospheric Physics 76 (2001): 143-165.

8. Supinie, Timothy and McGovern, Amy and Williams, John and Abernethy, Jennifer. Spatiotemporal Relational Random Forests. Proceedings of the 2009 IEEE International Conference on Data Mining (ICDM) workshop on Spatiotemporal Data Mining, electronically published.