Proceedings of the 3rd International Conference on Applied Physics and Mathematical Modeling
DOI: 10.54254/2753-8818/2025.DL28337

Analyzing Tornado Genesis Through Multivariate Statistical Modeling of Meteorological Data

Jerry Zhang

Beckman High School, Irvine, USA
jzcw08@gmail.com

Abstract. Tornado forecasting is challenging because the atmosphere involves complex, non-linear interactions among multiple parameters. This study examines tornado records from 2005 to 2023 alongside reanalysis data to identify key parameters of tornado formation and intensity. The analysis focuses on Convective Available Potential Energy (CAPE), Surface Temperature (T), Storm Relative Helicity (SRH), Vertical Wind Shear (VWS), and a derived Dew Point Depression (T–Td) variable. Results demonstrate that tornado events typically occur in environments with higher CAPE, SRH, and VWS, combined with evaluated T and T-Td that indicate greater low-level moisture. Stronger tornadoes were most closely associated with elevated CAPE and SRH. Bayesian logistic regression confirmed that SRH and CAPE were the strongest parameters, with T, VWS, and T-Td providing smaller but consistent contributions. Logistic Regression, Random Forest, and Support Vector Machine models were tested, with Random Forest achieving the best balance of recall and precision due to its ability to capture non-linear relationships. These findings suggest that multivariate, non-linear models offer a more robust framework for enhancing tornado prediction.

Keywords: Tornado forecasting, severe weather prediction, atmospheric parameters, machine learning in meteorology, severe storm environments

1. Introduction

Tornadoes are among the most violent and destructive storms in the atmosphere. They can cause catastrophic damage in just minutes. The U.S. experiences the highest frequency of tornadoes globally, with around 900 tornadoes on average per year. This results in 80 deaths, 1500 injuries, and around 1 billion dollars in damages [1]. Although weather forecasting and local weather monitoring have significantly improved severe weather detection, predicting whether conditions for tornado formation are favorable remains a major challenge in meteorology.

Tornadoes are typically formed from a specific type of thunderstorm called a supercell. A supercell consists of a large, long-lived thunderstorm with a rotating updraft called a mesocyclone. The mesocyclone is the foundation for tornado formation [2]. Research has identified several atmospheric factors connected to tornado formation, such as Convective Available Potential Energy (CAPE), Storm-Relative Helicity (SRH), and Vertical Wind Shear (VWS). However, their ability to predict tornadoes is often limited when used in different mesoscale settings. A better understanding of how these factors interact with the atmosphere could improve forecast lead times, leading to better public safety outcomes.

This study is guided by three research questions. RQ1: What are the key atmospheric conditions needed for tornado formation? RQ2: Do these factors correlate with the intensity of tornadoes? RQ3: How can these factors be used to build an accurate forecasting model for tornado occurrence?

To address these questions, the specific objectives of this work are to identify the atmospheric parameters most frequently associated with tornado formation, collect historical meteorological reanalysis data and tornado event records for analysis, observe the statistical relationships between these parameters and tornado formation, examine their relationships with tornado intensity as measured by the Enhanced Fujita (EF) scale [3], and develop a predictive modeling approach that incorporates multiple atmospheric parameters to enhance tornado forecasting accuracy.

The remainder of this paper is organized as follows. Section 2 looks at the literature on how tornadoes form, covering theoretical frameworks and previous statistical modeling efforts. It also points out the main knowledge gaps that this study seeks to address. Section 3 describes the methods used for data collection, preparation, and multivariate analysis. Section 4 shows the results, including the identified relationships between parameters and how well the models performed. Section 5 concludes the main findings, discussing their impact on forecasting and providing suggestions for future research.

2. Literature review

Several atmospheric parameters have been identified in favor of tornado formation. Convective Available Potential Energy (CAPE), which is the buoyant energy available to accelerate air parcels upward, influences updraft strength [4,5]. Elevated CAPE values have been statistically linked to increased tornado occurrence, although many tornadic storms have been observed in environments with moderate to low CAPE when other favorable conditions exist [6]. Storm-Relative Helicity (SRH) measures the potential for streamwise vorticity into a storm's updraft and is a key parameter of mesocyclone strength and persistence [7]. High SRH values, particularly in the 0–1 mi and 0–2 mi

layers, are strongly correlated with significant tornadoes [8]. Vertical Wind Shear (VWS), especially in the lowest few kilometers of the atmosphere, plays a critical role in updraft rotation and storm organization [9].

Beyond these parameters, composite indices such as the Energy Helicity Index and the Significant Tornado Parameter have been developed to include multiple atmospheric parameters into a single forecast index [10]. These indices are utilized in forecasting severe weather and tornadoes; however, there are some aspects of an index that could be more significant for one kind of tornadic type and less significant for others [11].

Despite these advances, several knowledge gaps remain. Most studies have primarily focused on soundings from the U.S. Great Plains, with fewer analyses examining tornado-prone regions such as Dixie Alley or the Northeast United States. In addition, statistical models often overlook the combined effects of parameter interactions, instead treating each parameter in isolation.

This study addresses these gaps by conducting a multivariate statistical analysis of historical tornado events with their atmospheric environment. By utilizing CAPE, SRH, VWS, T, and T–Td into a combined structure, this work aims to analyze their relationship to both tornado formation and intensity.

3. Methodology

This study uses a multivariate statistical modeling approach to differentiate between tornado and non-tornado environments based on selected atmospheric parameters. It assumes that the atmosphere behaves differently before a tornado compared to the atmosphere where no tornado occurs. Factors such as instability, wind shear, and moisture are expected to show clear differences in pre-tornadic environments.

The following procedure was carried out to address the claim. First, historical event records from 2005 to 2023 were downloaded from the NOAA Storm Prediction Center (SPC) database [12]. This dataset contained approximately 22,000 tornado events with key variables such as event time, location, rating, and path length and width. In parallel, historical atmospheric data for the same period (2005–2023) were obtained from the National Centers for Environmental Prediction (NCEP), specifically from the North American Regional Reanalysis (NARR) dataset [13,14]. This dataset, stored in a gridded format using the Lambert Conformal projection, contains atmospheric measurements at multiple pressure levels and time intervals.

The atmospheric data grid was structured as a 349 × 277 matrix, with an approximate resolution of 0.3° (~20 mi) at the lowest latitudes [15]. A lookup table was created to map each grid point's (y, x) indices to its corresponding latitude and longitude using inverse projection calculations. To refine the study area, only tornadoes occurring east of the Rocky Mountains were considered. Specifically, the filtered dataset included events where the latitude was within [24.54, 49] and the longitude was within [-105.5, -44.86].

Each tornado was then mapped to its nearest atmospheric grid point. This was achieved by identifying all grid points within a ±1.5° latitude and longitude range, calculating the Haversine distance from the tornado's location to each candidate grid point, and selecting the nearest grid point (y, x) as the location for atmospheric conditions. Tornado event timestamps were converted to integer format compatible with atmospheric time indices. For each tornado, atmospheric parameter values from the preceding 12 hours were extracted at the matched (time, y, x) coordinates, allowing for analysis of pre-existing conditions that may have contributed to tornado formation.

To provide a balanced comparison set, a random sample of 25,000 time-space points (time, y, x) not associated with tornadoes was selected. For each of these sampled points, the same set of atmospheric parameters was collected. The tornado and non-tornado data were then combined into a single dataset. Based on the availability of data within the dataset, the following atmospheric parameters were selected for analysis: CAPE, surface dew point, temperature at 2m, vertical wind shear (0–4 mi), and storm relative helicity (0–2 mi).

To address non-normal distributions and enhance model performance, a preprocessing procedure is set up. First, data cleaning was performed by removing all samples with missing values to ensure a complete and reliable dataset. Next, logarithmic transformations were applied to variables observed with strong positive skew, specifically CAPE, SRH, and VWS, using the natural log function with a small constant (log(x+1)) to accommodate zero values while preserving relative differences. In addition, a new feature, T-Td (Dew Point Depression), was introduced. These variables measure the dryness of the lower atmosphere, with smaller spreads indicating moist conditions and larger spreads reflecting drier air [16]. Because low-level moisture is a key ingredient for tornado development, this feature captures atmospheric dryness more directly than using temperature and dew point separately. It also resolved the multicollinearity issue between these two variables, making the model more stable and interpretable. After these transformations, all parameters were standardized using z-score normalization (subtracting the mean and dividing by the standard deviation) to place them on a common scale. This preprocessing approach improved model performance, reduced the influence of skewed distributions, and provided a more meaningful assessment of each parameter's contribution to tornado probability.

Following this, visualizations were created to better understand the relationships between the parameters and tornado formation as well as tornado intensity. Boxplots were generated to compare parameter distributions for tornado versus non-tornado cases and for different tornado intensity levels.

Finally, predictive modeling was conducted using three approaches. A logistic regression model was developed as a baseline to interpret the relationship between atmospheric parameters and tornado formation probability. Additionally, a random forest classifier and an SVM classifier were built to capture non-linear interactions and more complex relationships between variables. All models were evaluated using data from 2005 to 2022 with balanced samples, applying cross-validation techniques and performance metrics such as recall, precision, F1 score, and accuracy. Additionally, data from 2023 with unbalanced samples were used to provide an additional evaluation of model robustness in real-world conditions.

4. Result

Figure 1 shows boxplot comparisons of key atmospheric parameters for tornado versus non-tornado environments. Across all six parameters, tornado cases consistently exhibit higher values in the distributions. CAPE values are elevated in tornado environments, with values above 1000 J/kg being particularly favorable for tornado development. SRH is also higher during tornado events, with values ranging from 250 to 400 m²/s², which appear favorable for tornado development. VWS shows broader variability in tornadic cases yet still trends higher than in non-tornadic cases. T is generally concentrated between 75–85°F for tornado events, while non-tornado cases display a wider spread with many lower outliers. Similarly, Td values are higher in tornado environments, typically ranging from 63°F to 73°F, compared with a broader and lower distribution in non-tornado cases. Finally, the T-Td is higher for tornado events, indicating a more humid boundary layer compared to non-tornadic environments. These patterns confirm that tornado occurrence is strongly associated with higher instability, stronger shear, and greater availability of low-level moisture.

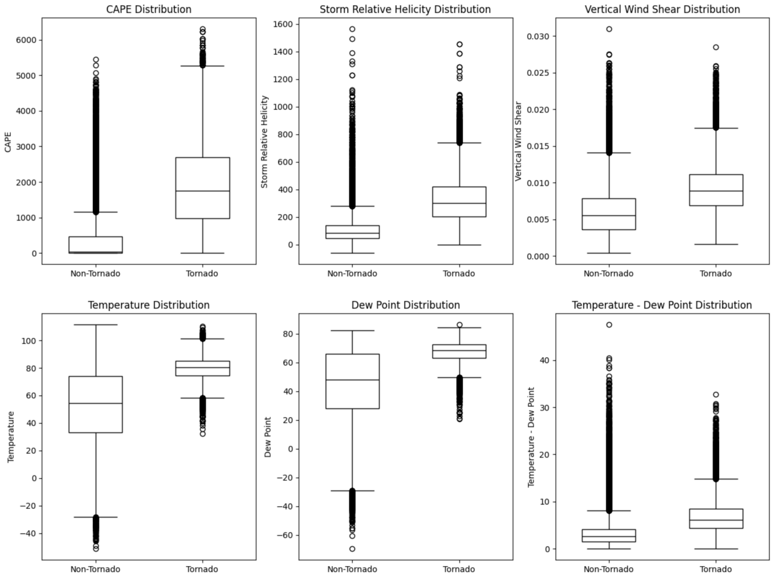

Figure 1. Distributions of atmospheric parameters for tornado and non-tornado events (picture credit: original)

Further analysis of tornado intensity revealed distinct patterns across various atmospheric parameters (Figure 2). CAPE showed a clear upward trend with tornado strength, with stronger tornadoes typically forming in environments with higher instability. Td values remained relatively stable across categories but displayed a modest increase for the most intense tornadoes (EF3–EF5), suggesting that higher low-level moisture can contribute to sustaining violent storms. SRH demonstrated the strongest relationship with tornado intensity, as median SRH values increased substantially with tornado strength, highlighting the importance of enhanced low-level wind shear and storm-relative inflow in producing stronger tornadoes. In contrast, T distributions remained relatively consistent across tornado magnitudes, indicating that a moderate boundary-layer temperature range is sufficient for tornado genesis regardless of intensity. VWS displayed only modest differences, though the highest category (EF5) was associated with slightly elevated shear values compared to weaker tornadoes.

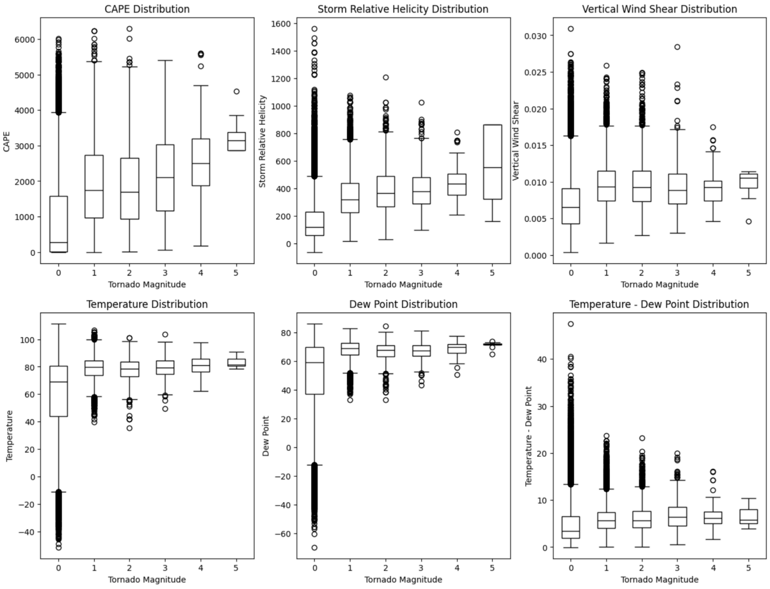

Figure 2. Atmospheric parameter distributions by tornado intensity. (picture credit: original)

Table 1. Ranked importance of atmospheric parameters from bayesian logistic regression

Rank	Parameter	Posterior Mean (β)	95% HDI	Significance
1	Storm Relative Helicity (log)	2.88	[2.81, 2.95]	Very strong
2	CAPE (log)	2.86	[2.73, 2.97]	Very strong
3	Surface Temperature	1.6	[1.43, 1.77]	Strong
4	Vertical Wind Shear (log)	0.88	[0.82, 0.93]	Moderate
5	Dew Point Depression	0.62	[0.57, 0.67]	Moderate

Bayesian logistic regression was also applied to evaluate the relative contributions of the five selected parameters. Posterior mean estimates showed that SRH (β = 2.88) and CAPE (β = 2.86) were the strongest parameters of tornado formation, followed by T (β = 1.60), VWS (β = 0.88), and T-Td (β = 0.62). All parameters had positive coefficients with narrow highest density intervals, indicating consistent positive associations with tornado environments. Table 1 summarizes the ranked importance of these parameters.

Table 2. Performance comparison of LR, RF, and SVM models on test and evaluation datasets

	Test Set (2005 - 2022)				Evaluation Set (2023)			
	Recall	Precision	F1-score	Accuracy	Recall	Precision	F1-score	Accuracy
LR	94.50%	94.00%	94.25%	94.59%	96.04%	13.29%	23.36%	93.21%
RF	94.50%	95.42%	94.96%	95.30%	94.13%	17.51%	29.53%	95.18%
SVM	95.70%	94.81%	95.25%	95.25%	96.69%	15.82%	27.18%	94.44%

Table 2 shows the model performance on the balanced test set from 2005 to 2022, where all three classifiers (Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM)) achieved high accuracy and balanced performance metrics. LR achieved 94.6% accuracy with an F1-score of 94.3%, while RF slightly outperformed it with an F1-score of 95.0% and 95.3% accuracy. The SVM model achieved the best overall performance on the test set, with a recall of 95.7% and an F1-score of 95.3%, indicating a strong ability to correctly identify tornadic environments without sacrificing precision.

When applied to the unbalanced 2023 evaluation set, differences between the models became more obvious. All models maintained high accuracy (93–95%), but precision dropped mainly due to the rarity of tornado events in the unbalanced dataset. LR showed the weakest generalization, with precision falling to 13.3% and an F1-score of only 23.4%. RF improved on this, reaching 17.5% precision and a higher F1-score of 29.5%. SVM achieved intermediate performance with 15.8% precision and a 27.2% F1-score. Notably, recall remained high across all three models (>94%), meaning that most tornado cases were detected, but at the result of many false positives.

5. Discussion

This study examines key atmospheric parameters that influence tornado genesis, investigates their relationship to tornado intensity, and explores their integration into a predictive modeling framework. The results confirm that certain variables, like CAPE, SRH, T, VWS, and T-Td play significant roles in differentiating tornado from non-tornado environments.

Bayesian logistic regression was used to quantify the relative contributions of the five atmospheric parameters to tornado occurrence. Posterior mean estimates (Table 1) indicated that SRH ($\beta = 2.88$) and CAPE ($\beta = 2.86$) were the most influential parameters, highlighting the critical roles of rotational shear and instability in distinguishing tornado from non-tornado environments. T ($\beta = 1.60$) showed a moderate positive effect, suggesting that boundary-layer warmth supports convective initiation, while VWS ($\beta = 0.88$) contributed a smaller but robust influence consistent with its role in storm organization. T-Td ($\beta = 0.62$) was the weakest parameter but still displayed a positive association, reinforcing that lower atmospheric moisture remains an essential ingredient.

All parameters had positive coefficients with narrow 95% highest density intervals (HDIs), providing strong probabilistic evidence that each factor makes a meaningful contribution to tornado genesis. These results confirm that tornado occurrence is best explained by the combined influence of instability, shear, and moisture rather than any single atmospheric parameter in isolation.

The correlation analysis between these parameters and tornado intensity (EF scale) showed that higher CAPE and SRH values were often connected with more intense tornadoes. This supports earlier studies linking strong instability and enhanced storm rotation to severe tornado outbreaks. However, the relationships were not perfectly linear, implying that mesoscale and storm-scale processes, like boundary interactions, storm mode, and atmospheric lift, also contribute to intensity outcomes.

Overall, the model performance results indicate that while all three models perform well on balanced datasets, Random Forest and SVM generalize better to real-world, imbalanced conditions. RF offered the best balance between recall and precision, making it the most effective model for forecasting applications where both detection and reliability are critical. Its advantage is likely to be a result of its ability to capture non-linear interactions and complex dependencies among atmospheric parameters, which simpler models like Logistic Regression cannot represent.

Despite these promising results, limitations remain. The model's performance partially depends on the frequency of input datasets. Using maximum CAPE values, for example, over a 12-hour pre-event window, while practical, may overlook significant short-term changes that could influence storm evolution. The tornado vs. non-tornado data was artificially balanced for modeling purposes, which might not reflect real-world frequencies and could affect model calibration. In addition, the dataset included only a limited set of atmospheric parameters that proved necessary for tornado formation. While variables such as CAPE, SRH, VWS, and T-Td provided valuable insights, many other available parameters contributed little predictive value. The parameters that had an effect on tornado formation were not included in the dataset.

Future work should explore several directions to further improve tornado prediction and understanding of storm environments. One area involves incorporating higher-frequency sampling, such as hourly atmospheric parameter trends, to better capture the evolving details of storm–environment interactions. In addition, adding mesoscale boundary information and radar-derived parameters could enhance the ability to predict tornado intensity, which remains a difficult challenge. Model robustness could also be evaluated through testing on independent years or different geographic regions, ensuring that predictive skill works beyond the training domain. Finally, developing probabilistic forecasting approaches that include uncertainty estimates would provide a more realistic and useful framework for decision-making.

6. Conclusion

This study investigated the atmospheric parameters strongly associated with tornado formation and intensity and evaluated their usefulness in predictive modeling. By analyzing historical tornado records alongside reanalysis data from 2005 to 2023, it is found that higher values of CAPE, SRH, and VWS, along with elevated T and T-Td, were consistently linked to tornado events. Moreover, tornado intensity showed a strong relationship with SRH and CAPE, highlighting the importance of both instability and low-level shear in producing stronger tornadoes.

Bayesian logistic regression further quantified the influence of each parameter. SRH and CAPE emerged as the most dominant parameters, while T, VWS, and T-Td provided secondary but consistent contributions. These probabilistic results reinforce weather forecasting theories by confirming that instability, shear, and moisture together create the most favorable environments for tornado genesis.

Predictive modeling results further demonstrated that machine learning approaches, particularly Random Forest and SVM, outperformed Logistic Regression when tested on unbalanced real-world data. Although accuracy remained high across all models, precision dropped substantially in the evaluation set, underlining the challenge of forecasting rare events like tornadoes. The Random Forest model offered the best balance between recall and precision, largely because of its ability to capture non-linear interactions among atmospheric parameters. This suggests that ensemble approaches capable of modeling complex patterns may be especially valuable for forecasting.

Overall, the findings highlight the importance of combining multiple atmospheric parameters within a multivariate framework rather than treating them in isolation. While limitations remain, such as reliance on a limited set of parameters and the use of artificially balanced training data, this study demonstrates that statistical approaches can meaningfully improve the understanding of tornado genesis and enhance predictive power. Future work incorporating higher-frequency data, mesoscale boundary information, and probabilistic models has the potential to further strengthen tornado forecasting and, ultimately, improve public safety outcomes.

References

[1] National Centers for Environmental Information. (2022). Storm Events Database. National Oceanic and Atmospheric Administration.

[2] Fischer, J., Dahl, J. M. L., Coffer, B. E., Houser, J. L., Markowski, P. M., Parker, M. D., Weiss, C. C., & Schueth, A. (2024). Supercell Tornadogenesis: Recent Progress in Our State of Understanding. Bulletin of the American Meteorological Society, 105(7), E1084–E1097. https: //doi.org/10.1175/BAMS-D-23-0031.1.

[3] National Weather Service. (n.d.). The Enhanced Fujita Scale (EF Scale). Retrieved August 29, 2025, from https: //www.weather.gov/oun/efscale

[4] Moncrieff, M. W. and Miller, M. (1976). The dynamics and simulation of tropical cumulonimbus and squall lines. Quarterly Journal of the Royal Meteorological Society, 102, 373–394.

[5] Johns, R. H., & Doswell III, C. A. (1992). Severe Local Storms Forecasting. Weather and Forecasting, 7(4), 588-612.

[6] Rasmussen, E. N., & Blanchard, D. O. (1998). A Baseline Climatology of Sounding-Derived Supercell and Tornado Forecast Parameters. Weather and Forecasting, 13(4), 1148-1158.

[7] Davies-Jones, R. P., D. W. Burgess, and M. Foster, 1990: Test of helicity as a tornado forecast parameter. Preprints, 16th Conference on Severe Local Storms, Kananaskis Park, AB, Canada, Amer. Meteor. Soc., 588–592.

[8] Thompson, R. L., Edwards, R., & Hart, J. A. (2003). Close Proximity Soundings within Supercell Environments Obtained from the Rapid Update Cycle. Weather and Forecasting, 18(6), 1243-1253.

[9] Weisman, M. L., & Klemp, J. B. (1984). The structure and classification of numerically simulated convective storms in directionally varying wind shears. Monthly Weather Review, 112(12), 2419-2438.

[10] Hart, J. A., & Korotky, W. (1991). The SHARP workstation v1.50 users' guide. National Weather Service.

[11] Doswell, C. A. III, & Schultz, D. M. (2006). On the use of indices and parameters in forecasting severe storms. Electronic Journal of Severe Storms Meteorology, 1(3), 1–22.

[12] NOAA/NWS Storm Prediction Center. (n.d.). Severe Weather Database Files. Retrieved August 28, 2025, available at: https: //www.spc.noaa.gov/wcm/index.html#data

[13] NCEP North American Regional Reanalysis (NARR), NOAA Physical Sciences Laboratory, available at: https: //psl.noaa.gov/data/gridded/data.narr.html

[14] Explanation of SPC Severe Weather Parameters, NOAA/NWS Storm Prediction Center, available at: https: //www.spc.noaa.gov/exper/mesoanalysis/help/begin.html

[15] Lambert Conformal Format, NOAA Physical Sciences Laboratory, available at: https: //psl.noaa.gov/data/narr/format.html

[16] Wallace, J. M., and Hobbs, P. V. (2006). Atmospheric Science: An Introductory Survey. 2nd ed., Academic Press, Amsterdam, 483 pp.