Weather Forecasting UNET

Jesus Meza
Stanford University
450 Jane Stanford Way
jemeza@stanford.edu

Noah Anderson
Stanford University
450 Jane Stanford Way
noaha@cs.stanford.edu

Tenzin Tsultrim
Stanford University
450 Jane Stanford Way
ttsult@stanford.edu

Abstract

Effective prediction of tornado occurrences is paramount for proactive emergency planning and response strategies. In this study, we introduce a novel approach by leveraging UNET architecture to forecast tornadoes based on meteorological data. Our model harnesses the power of UNET to analyze spatiotemporal patterns in weather data, thereby generating probabilistic forecasts of tornado occurrence. Key meteorological variables, including Convective Available Potential Energy (CAPE), Convective Inhibition (CIN), Mean Sea Level Pressure (MSLP), Storm Relative Helicity (SRH), Air Temperature, and Geopotential Height serve as input features for our model. We evaluate the model's predictions against retrospective tornado survey data, demonstrating its efficacy in producing accurate tornado forecasts. Our innovative approach offers promising prospects for enhancing disaster preparedness.

1. Introduction

Tornadoes are sudden and destructive natural disasters and pose a substantial risk to people and property in the central and Southeast United State. Accurately predicting when and where there is a change for tornadoes is a difficult task that requires expertise. Discovering new methods to generate tornado forecasts could help with emergency planning and preparedness. An AI model has the potential to analyze large amounts of meteorological data, and process it more quickly than traditional methods before generating a prediction. These predictions could be used to help inform meteorologists tasked with making official forecasts. In this project we intend to generate tornado forecasts from meteorological data. Providing accurate tornado forecasts could significantly help with disaster preparedness and response particularly in rural areas

Our inputs and outputs can be described as follows. For a single input–output pair, our input X is a 3x256x256 matrix, with 3 channels–CAPE, CIN, and Geopotential height. These 256x256 matrices originate from the NOAA Reanalysis data which is described in greater detail in the dataset section. These 256x256 grids are centered on the CONUS or the contiguous United States. Our outputs are "perfect hindcast" probability grids from Gensini et. al.

In terms of results, our model shows promise in some areas, and has shortcomings in others. It almost universally outperforms the baseline models trained on the same data, and also in some cases can be shown to perform better than the prediction made by the Storm Prediction Center for that same day, however, our current model displays some issues–it often struggles to predict significant tornadoes (see experiment section 1), likely due to the sparsity of this data. Further, while it performs relatively well in predicting the locations of threats, it struggles with the magnitude–see experiment 2, and future work.

Our outputs are a 256x256 grid of probabilities representing the chances of a tornado within 20 miles of a given point based on retrospective tornado survey results. The Storm Prediction Center issues daily tornado forecasts defined in the same way, and thanks to the work by Gensini et al[11], we have y labels for every day between 1979 to 2022 describing what would have been the perfect SPC tornado forecast based on their own criterion: give the probability of a tornado within 25 miles of any given point. This criterion also includes a "sigtor" (or significant tornado threshold, often displayed as black hatching on an SPC forecast), which represents the same probabilities, but specifically for tornadoes that are considered "significant" or EF2-EF5 on the Enhanced Fujita scale.

1.1. Related Work

1.1.1 Direct Relation to our Method

The study conducted by [2] employs reanalysis data (various localized parameters) and Satellite data to predict significant hail. It underscores the credibility of utilizing reanalysis data for machine learning. However, the study's reliance on a Multilayer Perceptron (MLP) for point-wise classification poses limitations, particularly in capturing the intricate spatial and temporal dynamics inherent in weather

forecasting. Further, while MLPs may offer efficacy in hail prediction tasks, their inadequacy in comprehensively addressing the complex atmospheric conditions associated with tornado occurrences warrants further exploration into more sophisticated modeling approaches.

[3] applies neural networks to forecast convection up to 2 hours ahead. The predictors consist of a time series of brightness-temperature grids derived from seven infrared bands on the Himawari-8 satellite, with the output being a grid of convection probabilities at the specified lead time. This work serves as a compelling proof of concept, demonstrating the feasibility of employing U-Nets on multi-modal spatial weather data. Specifically, the study showcases the capability of U-Nets to analyze time-series infrared satellite photos of a given region and predict weather conditions forward in time. Building upon this concept, we propose a similar approach, albeit with a distinct focus. While Lagerquist et al. [3] predict convection probabilities 120 minutes into the future, our research endeavors to predict tornado probabilities approximately 12 hours ahead. This extension presents an opportunity to leverage U-Nets for longer-term forecasting of severe weather events, offering valuable insights into potential forecasting capabilities and challenges associated with extended lead times.

[4] demonstrates that UNets can be effectively utilized to produce spatially aware probabilistic predictions for severe weather (in their case, severe hail), however for the purpose of making useful forward-in-time forecasts, their inputs consist of HREF guidance, which necessitates utilizing predictions of atmospheric conditions generated by a numerical model forecasting approximately 13 hours ahead. Therefore, if the numerical models are off, and they frequently are, the Unet's predictions will be off. Our method on the other hand relies on using current atmospheric conditions to make predictions about future tornado probabilities.

Gensini et al. [5, 11], demonstrates the validity of using SPC storm report data, from which the "Perfect Hindcast" dataset [3] is derived, as an output or target set for various classification models using reanalysis as inputs, however they do not explore the effectiveness of UNets for producing forecast-like predictions that emulate those of the storm prediction center. Moreover, like previously discussed related works, Gensini et Al. are not trying to forecast forward in time. In other words, their methods are using data from T=0 to make predictions about the immediate environment.

Finally, in [6], McGuire and Moore demonstrate the validity of CNN-based architectures for prediction of high-end tornado events, and of using input features such as 500mb Geopotential height, which point-wise models like logistic regression and decision trees cannot appropriately utilize due to the importance of spatial features that appear within a given 2D snapshot. They also provide a framework for future work, as they perform feature engineering to enhance the edges and shapes in geopotential height plots, thus improving the model's ability to learn and train on these features.

1.1.2 State-of-the-art Relation

Related work in terms of state-of-the-art: what ML models are currently being used by forecasters.

The paper [7] discusses a state-of-the-art decision-tree-based model, currently utilized by the Storm Prediction Center (SPC) as a key tool in their forecasting suite to facilitate informed predictions. As it is a decision-tree model, with each gridded prediction being independent of the other, it is incapable performing forward-in-time prediction using current (T=0) data, and must also rely on short and medium term numerical forecast models like the NAM (North American Mesoscale Model) , the HRRR (High Resolution Rapid Refresh Model), and the RAP (Rapid Refresh Model) to provide numerical inputs that themselves predict the raw conditions of, say, T= +12hrs.

Another SOA forecasting tool that uses some form of ML methods is the CIPS Analog Guidance model [8]. This model performs a meta-analysis of current atmospheric conditions, to find matching atmospheric conditions from the past (so-called "analogs") to make probabilistic predictions about severe weather likelihood. It is considered a respected forecasting tool used by the Storm Prediction Center. The model is, however limited, of course, by quality of matching analogs, and so its usefulness can vary wildly depending on the nature of the given severe weather threat.

1.1.3 SOA Techniques

Finally, given the relative sparsity of exploration of this problem setup in meteorological literature, it's worth briefly discussing SOA techniques in other fields that better match the modality of our problem.

Namely, [9] shows the validity of using U-Nets for localizing risk. Moreover, [10] provides us with future considerations. Using Multimodal magnetic resonance inputs and ensembled UNets, they were able to make more specialized, multi-faceted predictions about the location and nature of brain tumors. We believe that there is an analogous nature between these considerations and those of probabilistic severe weather forecasting.

1.2. Methods (2 pages)

Three different models were trained on our dataset, the first two were a logistic regression model and decision tree model for the purpose of having baselines, and we note from our related work section that both of these models have been

used for severe weather prediction. The second was a pre-trained U-NET, which we fine-tuned. [13].

The pre-trained U-NET chosen was a model that was used for semantic segmentation on cars trained on the Carvana dataset [15]. We chose this model because it had been pre-trained on a large dataset and it's output channels and input channels matched our desired input and output channels. Our input data was a 3 channel image where the first channel was cin (convective inhibition), the second channel was mean sea level pressure, and the third channel was CAPE (Convective Available Potential Energy). These channels were chosen because . The labels we used were two channel ¡images? tensors?¿ where the first channel and second channel were the probabilities of tornado and significant tornado across the United States. These labels were generated by [11] and as previously stated, represent the probability of a tornado within 25 miles of any given point.

In explaining our approach, we want to emphasize that the vast majority of widely use machine learning applications to severe weather prediction are performed in a point-wise, purely ingredient-based manner [7, 8]. Unfortunately, severe weather and especially tornadoes translate across space. For this reason, tornado-predictions using point-wise methods will always be at a disadvantage. To remedy this disadvantage these papers will spend a lot of time and effort attempting to encode spatial relationships within the point-wise data. As an example [7] encodes latitude and longitude as features to try to teach the DT model about climatological trends. This is inherently flawed in our opinion, and we believe the the UNet's CNN-based architecture can more adequately learn and capture the intricacies of severe weather.

1.3. Dataset and Features

Category/Num examples	**Years**
Validation (94,170)	2005, 1992, 1984, 1999, 2014, 2008
Test (94,170)	2017, 2018, 2019, 2020, 2021, 2022
Training (470,850)	Rest

Table 1: Data breakdown from 1980 - 2022

In our project, we use datasets from the NOAA Physical Sciences Laboratory [1]. The resolution of all images is 256x256–this is the result of a resize and crop, to better fit our model and also to be better centered on the CONUS. We chose five features: (1). CAPE (convective available potential energy), mean, surface, per-day. (2). CIN (convective inhibition), ensemble mean, surface, 8x daily (12Z). (3). Geopotential Height, ensemble spread, tropopause, 8x daily (12Z). (4). MSLP (mean sea level pressure), mean, 8x daily (12Z). (5). SRH (storm relative helicity), individual obs, 8x daily (12Z). (6). Air temperature, individual orbs, pressure levels, 8x daily (12Z). We conducted minimal data processing on the dataset, often simply dividing them by their global max to promote numerical stability. We resized the input images to only cover the United States. We also normalized the dataset by scaling the values of the different features to a common range. Further explanation of each of these features is explained below.

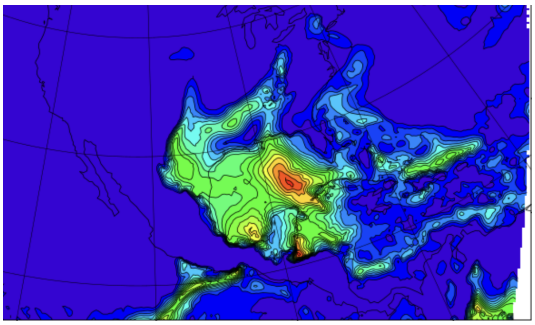

Figure 1: This is an example of an image from the CAPE dataset

CAPE, or convective available potential energy–is measured in Joules per Kg of atmosphere, and also must be anchored to a certain parcel starting height. In our case, we are using surface-based CAPE. CAPE is highly correlated with the atmosphere's ability to maintain storms, but says little to nothing about if a storm can actually form or if the storm will produce tornadoes.

CIN, or convective inhibition is the negative of CAPE, and is measured in negative Joules per Kg of atmosphere. It similarly must be anchored to a particular parcel starting point. CIN can prevent storm formation which can be to the detriment of tornadic activity, but it can also allow for more isolated storms which may promote tornadic activity. Furthermore, some amount of CIN in the atmosphere prior to storm initiation can allow CAPE to build to explosive levels prior to storm formation, resulting in stronger updrafts, and potentially a higher chance for tornadoes.

Geopotential Height of a particular pressure level is an approximation of the altitude at which that pressure level can be found. For example, if the 500mb geopotential height at Palo Alto is 5400 meters, then we know that half of the atmosphere's mass column is above 5400 meters and half is below 5400 meters. Geopotential height plots are useful for understanding the conditions of the Jet-Stream, which is the primary driver for the majority of extreme weather events and tornadoes in the mid-latitudes. [12]

MSLP, or mean sea level pressure, is typically in millibars. MSLP is not directly measured but is rather calculated based on the observed pressure at a given location,

adjusted to sea level. MSLP serves as a key indicator of atmospheric circulation patterns and weather systems. It provides valuable information about the overall pressure distribution across a region, which influences wind patterns and weather conditions.

SRH, or storm relative helicity, is a measure of the potential for cyclonic updraft rotation in supercell thunderstorms. It quantifies the amount of horizontal vorticity that can be tilted and stretched by a thunderstorm's updraft to produce rotation. SRH is typically measured in m^2/s^2. High values of SRH are associated with an increased likelihood of severe weather events, such as tornadoes. The feature provides crucial insights into the dynamics of storm development and helps in forecasting severe weather conditions.

Air temperature is measured at different pressure levels. It is typically recorded in degrees Celsius (°C) or Kelvin (K). Air temperature affects a wide range of atmospheric processes, including the formation of weather systems, precipitation patterns, and atmospheric stability. Variations in air temperature can influence convection, cloud formation, and the development of weather fronts.

To expand on the previous point further [16] and many others have demonstrated that things such as frontal orientation–e.g. if a warm front or cold front is oriented north-south or east-west or something else–can have significant impacts on tornadic potential, and CNNs seems particularly well suited to capturing this complexity. This is also why one of the key features we trained out models on was MSLP, or Mean Sea Level Pressure, which is on its own sufficient for finding fronts and frontal orientations for any reasonably skilled forecaster.

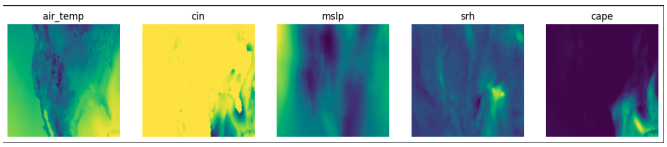

Figure 2: visual of five channels on 2021-03-25

1.4. Experiments

When training our model, we chose to use a learning rate of 1e-4 using the adam optimizer with a mini-batch of 64. We tried out various batch sizes starting from 4 until working up to 64. We didn't try batch sizes past 64 given constraints on computation. A learning rate of 1e-4 with adam worked best and provided smooth training.

When analyzing the performance of our models we decided to use KL-divergence:

$$D_{\mathrm{KL}}(P\|Q) = \sum_{x\in\mathcal{X}} P(x)\log\left(\frac{P(x)}{Q(x)}\right)$$

KL-divergence is a standard method used to compare the difference and similarity between two different probability distributions. The output of our model had two channels the first was a point wise probability distribution of generic tornado probabilities–weak or strong tornadoes included–and the second was that of a significant tornado. We compared both separately to their corresponding label.

We provide a qualitative analysis of our data demonstrating both failure cases and success cases on various different days by showing the true label, the U-Net's prediction, and the Storm Prediction Center's prediction [14].

1.4.1 Experiment 1: Comparison with Baseline Model

For the first experiment a comparison with the baseline models is provided. For experiment 1, we wanted to see if the model was capable of not just predicting tornadoes, but predicting significant tornadoes. This meant that our output dim was 256x256x2.

In terms of results, we note a set of success cases, and a set of failure cases. For a remarkable success, we can see our prediction made by our model 15 has the sig probabilities located remarkably where the sigtor probabilities are in the perfect true label 14. Furthermore, we note that this far outperforms the Storm Prediction Center's prediction from the same day 16.

For failure mode, however, we note that overall, the model generally failed to output significant tornado probabilities at all. We posit that due to the very sparse nature of this data in particular, the model learned to simply avoid outputting these values most of the time. See figures 6, 5, 7.

Finally, we should also note 4, which shows that though our model trained on sig tor data was a bit disappointing, it far outperformed the baseline models. Finally, we should also note 3, which shows our model far outperforms baseline models.

1.4.2 Experiment 2: Significant Tornado Prediction

For our second experiment we thoroughly analyzed the performance of our U-Net in predicting the probability of significant Tornados. We show one failure case and one success case.

One failure case is for March 25th, 2021(Day A), the true label, the U-Net's prediction, and the Storm Prediction Center's (SPC) prediction are shown in 5, 6, and 7 respectively. This covers one of the failure cases in which our model fails to predict the occurrence of a significant tornado. Note that significant tornado predictions are indicated by a hatch, as you can see in 6, the U-Net's prediction does not have a hatch.

One of the model's best predictions was for May 18th, 2017 (Day B), the true label, the U-Net's prediction, and the Storm Prediction Center's prediction are shown in 8, 9, and 10 respectively. In this example the model performed

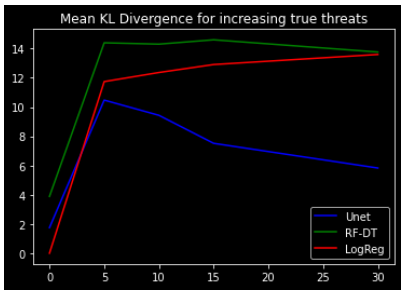

Figure 3: Regular Tornado Probability Results

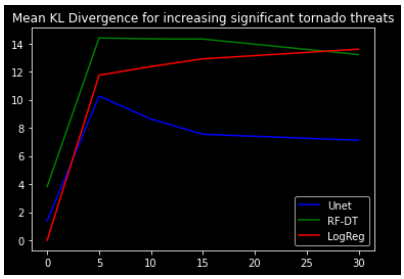

Figure 4: Significant Regular Tornado Probability Results

strongly and significantly outperformed the SPC's prediction.

1.4.3 Experiment 3 Tornado

For our third experiment we analyzed the performance of our model in predicting the probability of tornados. We present one success case and one failure case.

One failure case is for January 10th, 2021(Day C), the true label, the U-Net's prediction, and the Storm Prediction Center's (SPC) prediction are shown in 11 **??**, and 13 respectively. Though our model correctly localizes the tornado, it does not predict proper probabilities.

One of the model's best predictions was for January 11th, 2020, Day D the true label, the U-Net's prediction, and the Storm Prediction Center's prediction are shown in 14, 15, and **??** respectively. In this example the model performed strongly and significantly outperformed the SPC's prediction. However, once again the U-Net does extremely well in localizing where the tornado will occur but does not predict the correct tornado probabilities.

1.5. Conclusion/Future Work

The analysis revealed notable promise in certain instances, where the machine learning models exhibited predictive capabilities comparable to traditional storm prediction methods. This convergence of predictions with established storm prediction techniques underscores the potential of advanced computational methods in accurately forecasting tornado events. The ability of the models to align closely with storm prediction outcomes suggests that they

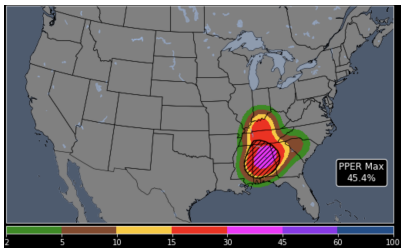

Figure 5: High Sig Tornado Day A True Label

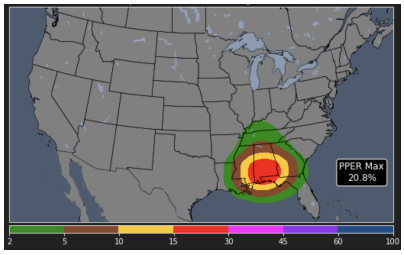

Figure 6: High Sig Tornado Day A U-Net Prediction

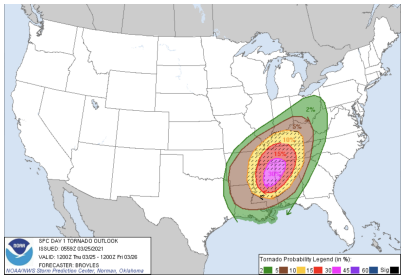

Figure 7: High Sig Tornado Day A SPC Prediction

capture essential meteorological variables and patterns associated with tornado formation and development. Reflecting, the model could have used more features and cleaner features. In terms of performance, The UNET was the highest performing algorithm. When comparing the UNET to the other algorithms, see the Methods section for a description.

In the future, we believe we can expand in multiple ways: (1). Conduct a comprehensive analysis of feature importance to identify the most informative meteorological variables for tornado forecasting. (2). We could also extend the forecasting horizon to predict tornado occurrences beyond the current time frame, possibly exploring multi-step forecasting approaches. (3). Use a loss function better suited for the task, such as weighted binary cross-entropy or focal loss. The new loss function will help ensure that the model learns to prioritize the correct identification of tornado events. (4). Use Sigmoid, the sigmoid activation facilitates the interpretation of model outputs in terms of calibrated possibilites, enabling stakeholders to make informed decisions based on the forecasted tornado probabilities. (5). Normalize the labels, normalizing the labels helps mitigate

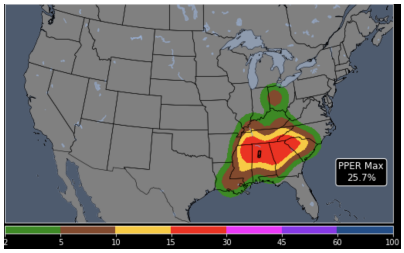

Figure 8: High Sig Tornado Day B True Label

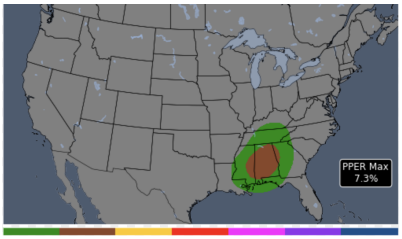

Figure 9: High Sig Tornado Day B U-Net Prediction

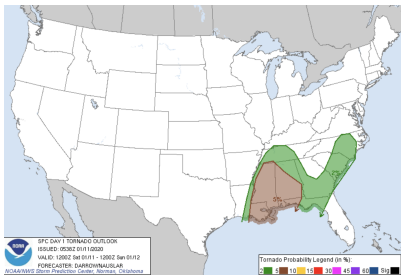

Figure 10: High Sig Tornado Day B SPC Prediction

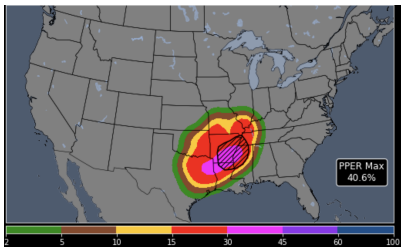

Figure 11: High Tornado Day C True Label

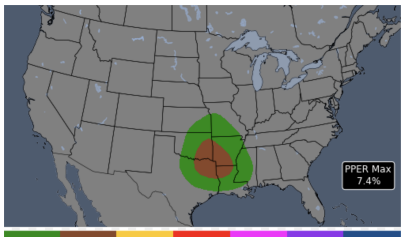

Figure 12: High Tornado Day C U-Net Prediction

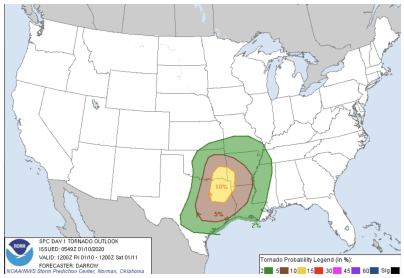

Figure 13: High Tornado Day C SPC Prediction

potential biases or inconsistencies in the original tornado occurrence data, such as variations in reporting practices or data collection methodologies.

1.6. Contributions and Acknowledgements

Jesus Meza implemented and debugged the U-Net architecture, delivering the theoretical analyses presented in the Methods section. He also developed the training function and conducted numerous experiments on the Modal Platform, significantly enriching the Experiments/Results/Discussion (ERD) section. Additionally, Jesus maintained the GitHub repository, ensuring smooth development throughout the project.

Noah Anderson's expertise in tornado tracking was integral to the group's understanding of the various factors impacting predictions and addressing related questions. He performed the baseline tasks and gathered the dataset used for all experiments in this paper. Noah also delivered an erudite and exhaustive analysis within the Methods and ERD sections.

Tenzin Tsultrim wrote scripts to gather images for the dataset, developed the test function for the Modal Platform, and assisted in running various experiments. He covered the Dataset and Features section and provided support to the team as needed.

[13], is the original repo, we used for the UNET.

We would like to acknowledge our mentor Samir Agarwala, the CS231N course staff for support and guidance, as well as Modal for computing resources.

References

[1] National Oceanic and Atmospheric Administration. "NOAA Physical Sciences Laboratory." NOAA, https://www.psl.noaa.gov/.

[2] Scarino, B., K. Itterly, K. Bedka, C. R. Homeyer, J. Allen, S. Bang, and D. Cecil, 2023: Deriving Severe Hail Likelihood from Satellite Observations and Model Reanalysis Parameters Using a Deep Neural Network. Artif. Intell. Earth Syst., 2, 220042, https://doi.org/10.1175/AIES-D-22-0042.1.

[3] Lagerquist, R., J. Q. Stewart, I. Ebert-Uphoff, and C. Kumler, 2021: Using Deep Learning to Nowcast

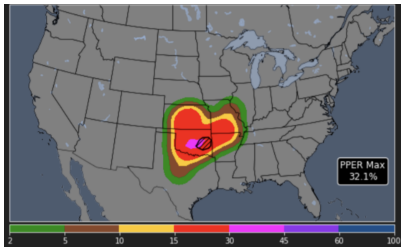

Figure 14: Moderate Tornado Day D True Label

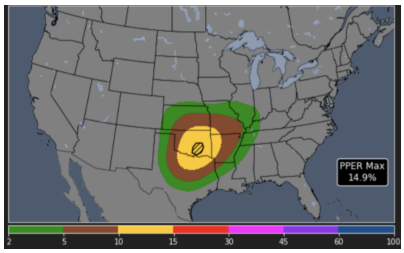

Figure 15: Moderate Tornado Day D U-Net Prediction

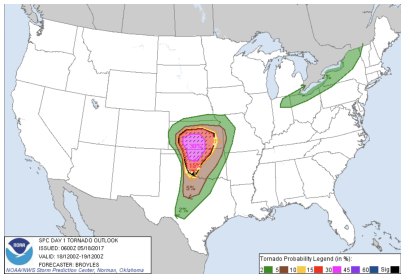

Figure 16: High Sig Tornado Day D SPC Prediction

the Spatial Coverage of Convection from Himawari-8 Satellite Data. Mon. Wea. Rev., 149, 3897–3921, https://doi.org/10.1175/MWR-D-21-0096.1.

[4] Burke, A., N. Snook, D. J. Gagne II, S. McCorkle, and A. McGovern, 2020: Calibration of Machine Learning–Based Probabilistic Hail Predictions for Operational Forecasting. Wea. Forecasting, 35, 149–168, https://doi.org/10.1175/WAF-D-19-0105.1.

[5] Gensini, V. A., C. Converse, W. S. Ashley, and M. Taszarek, 2021: Machine Learning Classification of Significant Tornadoes and Hail in the United States Using ERA5 Proximity Soundings. Wea. Forecasting, 36, 2143–2160, https://doi.org/10.1175/WAF-D-21-0056.1

[6] McGuire, M. P., & Moore, T. W. (2022). Prediction of tornado days in the United states with deep convolutional neural networks. Computers & Geosciences, 159, 104990.https://doi.org/10.1016/j.cageo.2021.104990

[7] Hill, A. J., G. R. Herman, and R. S. Schumacher, 2020: Forecasting Severe Weather with Random Forests. Mon. Wea. Rev., 148, 2135–2161, https://doi.org/10.1175/MWR-D-19-0344.1.

[8] Atmospheric Numerical Analysis Group. ANALOG: Atmospheric Numerical Analysis Group - Analog Ensemble Forecasting. `https://www.eas.slu.edu/CIPS/ANALOG/analog.php`

[9] Al Nasim, Md Abdullah, et al. "Brain tumor segmentation using enhanced u-net model with empirical analysis." 2022 25th International Conference on Computer and Information Technology (ICCIT). IEEE, 2022.

[10] Zhang Y, Zhong P, Jie D, Wu J, Zeng S, Chu J, Liu Y, Wu EX, Tang X. Brain Tumor Segmentation From Multi-Modal MR Images via Ensembling UNets. Front Radiol. 2021 Oct 21;1:704888. doi: 10.3389/fradi.2021.704888. PMID: 37492172; PMCID: PMC10365098.

[11] Gensini, V.A., Haberlie, A.M., & Marsh, P.T. (2020). Practically Perfect Hindcasts of Severe Convective Storms. Bulletin of the American Meteorological Society, 101(8), E1259-E1278. doi:10.1175/BAMS-D-19-0321.1

[12] Grams, Jeremy & Thompson, Richard & Snively, Darren & Prentice, Jayson & Hodges, Gina & Reames, Larissa. (2012). A Climatology and Comparison of Parameters for Significant Tornado Events in the United States. Weather and Forecasting. 27. 106-123. 10.1175/WAF-D-11-00008.1.

[13] https://github.com/milesial/Pytorch-UNet?tab=readme-ov-file

[14] NOAA/NWS Storm Prediction Center. *NOAA/NWS Storm Prediction Center*. Retrieved from `www.spc.noaa.gov/`.

[15] Brian Shaler, DanGill, Maggie, Mark McDonald, Patricia, Will Cukierski. (2017). Carvana Image Masking Challenge. Kaggle. https://kaggle.com/competitions/carvana-image-masking-challenge

[16] Clark MR, Parker DJ. Synoptic-scale and mesoscale controls for tornadogenesis on cold fronts: A generalised measure of tornado risk and identification of synoptic types. Q J R Meteorol Soc. 2020; 146: 4195–4225. https://doi.org/10.1002/qj.3898