

Forecasting Convective Initiation at a 2-hour Lead Time through the use of Machine Learning Models Trained on RAP Environmental Data

Investigators: Amelia Urquhart¹, Ricky Cavaliero¹

Mentors: Eric Loken²

¹ University of Oklahoma School of Meteorology

² National Severe Storms Laboratory / Cooperative Institute for Severe and High-Impact Weather Research

Project repository on GitHub: https://github.com/a-urq/ml_ci_py.

Abstract:

Current convection-allowing models often struggle to precisely forecast convective initiation. Convective initiation (CI) modeling requires the consideration of both synoptic and mesoscale forcing mechanisms. While full-physics numerical weather prediction models are often quite skillful with representing synoptic features and the forcing associated with them, mesoscale features are much more difficult to resolve. This leads to severe storm events driven by mesoscale forcing to be much more difficult to predict than those driven primarily by synoptic factors. To better resolve the nuances involved in forecasting CI, random forest modeling was used. The model was trained using predictor values drawn from Rapid Refresh (RAP) 13 km model data as hosted on Amazon Web Services and target values derived in an unconventional manner from the ProbSevere version 3 dataset. Initial uncalibrated outputs contained large areas of severe overforecasts, which required a novel “spatial calibration” method to address. While the final version of the ML-CI model still generally overforecasts, the issue was far less severe.

Studies of individual cases show that ML-CI has a basic grasp of the conditions that support and suppress convective initiation, but not the ability to predict the specific location of CI. Additionally, higher elevation regions were generally worse in their tendency for overforecasting. While the model continues to suffer from issues with false positives as well as some misses, it shows some potential for application to operational forecasting. In its current state, the model is best used to determine where CI is possible as opposed to where it is not. Future work will add a few additional predictors to the model to provide information on the state of initiating boundaries and the current stage of the initiating boundary within the CI processes, as well as correcting surface fields with observed weather station data.

1. Introduction

Anticipating convective initiation (CI) has remained a challenge for forecasters, making it difficult to issue accurate public watches and warnings for possible severe weather impacts, such as hail, wind, and tornadoes. Current physics-based models aim to pinpoint the timing and placement of CI, but discrepancies persist. While current convection-allowing models (CAMs) predict convective events with sufficient accuracy overall, they struggle to model discrete CI events (Kain et al. 2013). For example, in the development of the June 20, 2025 Enderlin, North Dakota tornado, the HRRR failed to anticipate the formation of the tornado's parent supercell at a 1-2 hour lead time (Figure A1). Since physics-based CAMs are directly reliant on environmental parameters, any discrepancy between modeled conditions and reality is liable to produce errors in the forecast.

CI is a process that is dependent on multiple environmental parameters, and many depend on the magnitude of synoptic forcing. In weak forcing situations, small mechanisms such as thermal variations and convergence bands provide subtle ascent for initiation, whereas in strong forcing situations, synoptic features provide dramatic ascent (Tempest et al. 2024). Both isolated supercells and linear squalls can be produced regardless of forcing magnitude, but some variations in characteristics exist. Frontal quasi-linear convective systems and fast-moving supercells are more commonly found when forcing is strong, and slow-moving supercells and summertime mesoscale convective systems are usually the product of weak mesoscale forcing removed from synoptic features (Bech et al. 2014, Stensrud 1994) . Overall, strongly forced initiation is easier to predict than weakly forced initiation using NWP models (Tempest et al. 2024, Wilson 2006). Out of all influencing factors, which include lift, dilution, inhibition, and buoyancy, the magnitude of lift has the most significant impact (Lock 2014). To accurately forecast convective initiation, resolving lifting mechanisms can provide the clearest indicator of whether convection will occur.

While physics-based modeling has been used in the past to predict CI, newer models have been developed that utilize additional tactics, such as statistical and machine learning analysis to improve prediction accuracy by more efficiently identifying key relationships. Over the past decade, machine learning (ML) techniques have continued to improve predictive modeling. Machine learning's capability for fast and efficient classification can play an essential role in the classification of convective storms (Jergensen et al. 2019). In the third release of the

ProbSevere model, ML techniques such as random forests and gradient-boosted decision trees were used to improve severe storm threat probability forecasts (Cintineo et al. 2024). These methods can also be used to identify convective initiation from a set of environmental predictors (Viellette et al. 2013, Lee et al. 2021). Past attempts to predict and classify convection using ML techniques have focused on specific types of convection. For instance, when the ProbSevere v2 model was applied to pulse thunderstorm forecasting, performance was limited due to the model's difficulties in resolving the specific characteristics of a pulse thunderstorm environment (Gard et al. 2022).

In CI modeling, the statistical analysis of the environmental predictor set and its overall significance are important considerations. Some predictors may have a degree of proportionality or cross correlation. For instance, mass flux convergence and moisture flux convergence are both proportional to horizontal surface wind convergence, with horizontal mass convergence likely having a clearer link to convective development than moisture flux convergence (Banacos 2005). Another example is the low level jet, common in the South Great Plains, which is linked to horizontal moisture convergence, which in turn supports elevated saturation levels (Zhang et al. 2019). To account for cross correlated variables, permutation importance testing is a sufficient method to utilize. Permutation importance is advantageous because it does not require any assumptions about the distribution of forecast differences, and in the event of cross correlated predictors, multipass permutation can be effectively carried out (Anderson 2001).

This project seeks to explore the potential of ML techniques within CI forecasting and its possible use as an alternative or supplement to traditional forecasting methods. Section 2 of this paper discusses the methods used for data acquisition, the assembly of the predictor and target matrices, and the testing and calibration of the random forest model. Section 3 includes four case studies to demonstrate and qualitatively assess the model's performance in forecasting past events. Section 4 outlines the model verification steps and results as well as the variable significances indicated by the random forest model. Lastly, Section 5 outlines the overall interpretation of the model's performance and possible avenues for improvement.

2. Data and Methods:

For the construction of a training dataset, and analogously a validation and testing dataset, two matrices must be constructed: the predictor matrix and the target matrix. Parameters

known as predictors must be chosen, and target output values must be matched with every set of predictors. The predictors used to train the ML-CI model were drawn from historical Rapid Refresh (RAP) 13-km model data, using the first forecast hour of each model run. No set of gridded model data available for real-time operations is a perfect representation of the state of the atmosphere, and there exist benefits and drawbacks in using either the zero-hour or one-hour RAP output. While the zero-hour forecast is a slightly more accurate representation of the real atmosphere, the one-hour forecast was chosen for construction of the predictor matrix because it most closely mirrored the data available to operational forecasters, especially those on the Storm Prediction Center's Mesoscale Analysis webpage.

Eleven predictors were chosen to represent the factors most relevant in CI. Mixed-layer Convective Available Potential Energy (MLCAPE) was chosen to represent the energy available to developing thunderstorms. Most importantly, the absence of MLCAPE marks areas where deep moist convection is unsustainable, but the inclusion of this parameter also allows the model to account for effects from particularly high amounts of available energy. Mixed-layer Convective Inhibition (MLCIN) represents the energy barrier that suppresses the initiation of deep moist convection. Higher absolute values of MLCIN, or equivalently more negative values, indicate that more forcing is required for an updraft to push through the capping layer. Relative humidity in the layer between the lifted condensation level and the level of free convection (LCL-LFC RH) reflects the effect of dry air entrainment on shallow, narrow updrafts attempting to push through the capping layer. This parameter is not explicitly provided in RAP data files, and although the necessary data was available, calculating it for the tens of millions of records in the training dataset was prohibitively computationally expensive. In the interest of feasibility, considering the short time available to complete this project, 850 mb and 700 mb relative humidity were chosen to roughly represent the LCL-LFC RH.

The other seven predictors were not directly found within RAP GRIB2 files, but rather derived from the data found therein. Upslope flow represents the possibility of subtle orographic forcing created by topographic features within the Great Plains such as the Caprock Escarpment in Texas, Black Mesa in Oklahoma, and the Palmer Divide in eastern Colorado. Moisture flux convergence and mass flux convergence both represent forcing related to horizontal convergence in the atmospheric boundary layer. 500 mb absolute vorticity advection serves as a proxy for the differential vorticity advection term of the quasi-geostrophic omega equation, relying on the

assumption that low-level vorticity advection is approximately zero. The final three predictors are all geopotential height changes. The 1-hour 500 mb and 1-hour 700 mb height changes allow the model to consider the progression of shortwaves, whether subtle or robust, and the 12-hour 500 mb height change represents the general magnitude of synoptic forcing within the local environment, primarily influenced by longwaves. All three rely on the result from quasi-geostrophic theory that within a train of evenly-spaced waves, regions of height falls are colocated with regions of rising motion and vice versa, as well as the assumption that real atmospheric waves are close enough to the ideal case for this colocation of rising motion and height falls to remain approximately true. Formulas for the four RAP-derived non-height change parameters are given in Table 1.

Parameter	Formula
Upslope Flow	$\vec{V}_h \cdot \nabla z_{sfc}$
Moisture Flux Convergence	$-\vec{V}_h \cdot \nabla q - q \nabla \cdot \vec{V}_h$
Mass Flux Convergence	$-\vec{V}_h \cdot \nabla \left(\frac{p}{R_d T_v} \right) - \left(\frac{p}{R_d T_v} \right) \nabla \cdot \vec{V}_h$
Absolute Vorticity Advection	$-\vec{V}_h \cdot \nabla \eta = -\vec{V}_h \cdot \nabla (\zeta + f)$

Table 1: Formulas used for the computation of four RAP-derived parameters. In the interest of clarity, Table A1 has been included in the appendix to explicitly describe what each variable above represents.

The values for the target matrix, representing whether CI occurred, were determined by utilizing the ProbSevere v3 dataset in a somewhat unconventional manner. While all ProbSevere v3 objects contain probability values for the presence of severe weather threats, they were not leveraged. All ProbSevere objects carry a unique identification number that remains the same for the object attached to any given storm throughout its lifecycle, and the presence of every object within the dataset is temporally contiguous. For every hour included in the predictor matrix, a spatial grid matching RAP's gridded data was created and an integer was assigned to every point

within the grid. If a ProbSevere object was newly created within a 40 kilometer radius of a grid point and within the next 2 hours after the time associated with the grid, the point was given a value of 1. If not, it was given a value of 0. In early testing iterations, a third value representing the presence of existing convection was also included, but these values were ultimately removed and replaced with the value representing no new convection in the final version of the target matrix.

At the end of this process, predictor and target matrices representing every hour of the day between March 1 and September 1, 2022 through 2025, had been created. These matrices also included 2021, when RAP data was only sparsely available. These matrices were split into three datasets: a training, validation, and testing set. The training dataset consisted of all data from 2021 to 2023. The validation dataset, used exclusively for calibration, consisted of all 2024 data. The testing dataset, used to ensure that the model was not overfitting, consisted of all 2025 data. The initial version of the training dataset used in the ML-CI model consisted of only 4% positive CI target values and 96% negative values. The resulting random forest returned exclusively near-zero probability values and achieved a ROC-AUC score of 0.96 by stubbornly forecasting no CI. To address this, a new version of the predictor and target matrices was constructed where a 50%-50% balance would be strictly maintained between CI and no-CI target values. To ensure that all days within the dataset were equally represented, the number of CI values within each hour was found and the available no-CI values were randomly sampled to produce a new set of no-CI values equal in size to the CI set. This process was repeated for every hour of every day within the training, validation, and testing datasets to produce balanced training and testing data.

When trained on the balanced dataset, ML-CI's output was much more reasonable, though it tended strongly toward overprediction. At times, the model forecast extreme, unrealistically large areas of high probabilities, covering thousands of square miles. While somewhat interpretable in a heat map plot, it was still egregiously high and would not be readable at all in a contour plot. The standard calibration technique of an isotonic regression, hereafter referred to as "pointwise calibration" when applied in isolation, only made minimal improvements to reliability. Upon the observation that storms were more likely to form near local maxima, or "peaks," in the uncalibrated spatial probability distribution, a new random forest model was trained on a predictor matrix composed of uncalibrated probabilities computed

from the validation dataset as well as the difference of the uncalibrated probabilities from the grid points around it. This was computed by applying a Gaussian filter to the probability distribution with a sigma of 81.27 kilometers, then subtracting the smoothed distribution from its unsmoothed counterpart. The target matrix used is identical to the matrix used for the first random forest. After training the second model, it became clear that the partially calibrated model was reluctant to return any values above 20%. To remedy this, an isotonic regression was applied, and the probability distribution finally seemed qualitatively reasonable. This unorthodox calibration method is hereafter referred to as “spatial calibration.” The top row of Figure 1 shows the uncalibrated probability distribution and the difference of probability from local average, the two predictors in the second random forest. The bottom left plot shows the raw output of the second random forest model, and the bottom right plot shows the same output calibrated with the final isotonic regression step applied.

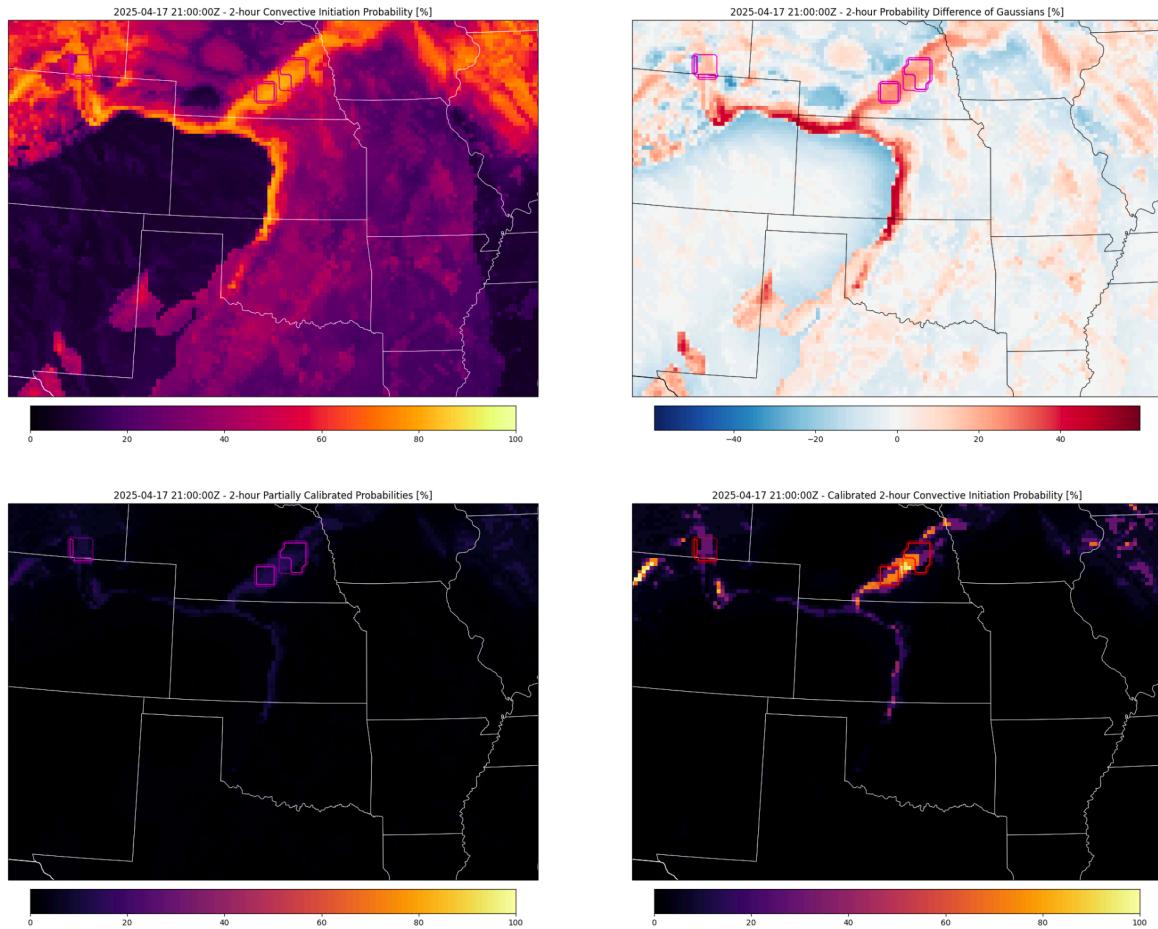


Figure 1: Visual aids for a more intuitive understanding of the spatial calibration process.

3. Results:

Once the calibration steps were completed, the model output was visualized as a set of probability contours overlying MRMS seamless composite reflectivity data and a basemap of United States political subdivision boundaries. The model's skill was qualitatively evaluated by applying it to four past events. The model was trained without the use of latitude and longitude as predictors, so two out-of-domain events were chosen to investigate the possibility that the model performed well in parts of the Great Plains not included in the training domain. The Cole, OK and Morton, TX events were within the training domain while the Enderlin, ND and Mobridge, SD events were outside of the domain.

On April 19, 2023, a significant tornado outbreak occurred in Oklahoma, producing numerous notable tornadoes, including the EF3 Cole, Oklahoma tornado. This model was applied to this case, and the output is shown in Figure 3.

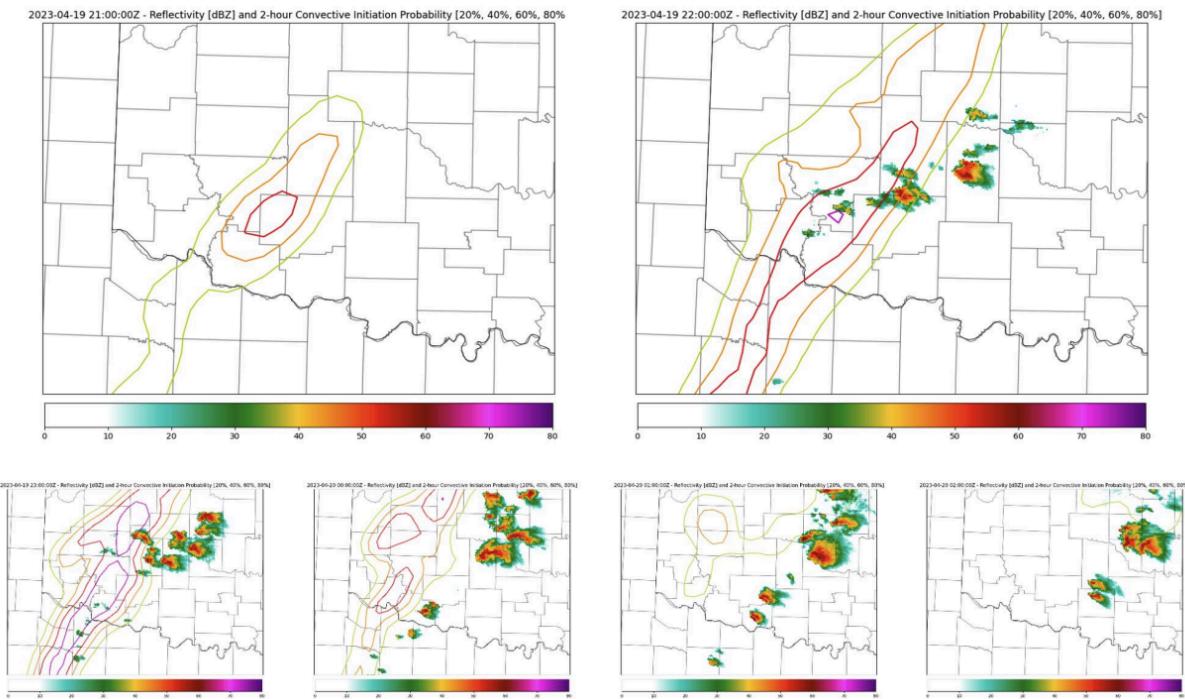


Figure 2: ML-CI outputs for April 19, 2023 in central and southwest Oklahoma, the event that produced the Cole, Oklahoma EF3 tornado.

In the top left plot, a large zone of high CI probabilities of up to 60% was centered over southwestern Oklahoma. One hour later, storms initiated directly within the Comanche County

bullseye and had propagated east-northeast by a few miles. Overall, this model reasonably predicted the general region where most storms initiated. However, there is a notable issue with a large zone of high CI probabilities that persists around the same region, propagating slowly westward before disappearing. This is likely to be a major contributing factor in the model's tendency to overforecast.

On June 5, 2025, a tornado outbreak occurred across the Texas Panhandle and eastern New Mexico, producing significant tornadoes, including the EF2 tornado in Morton, Texas. The sequence of convective initiation is shown in Figure 3.

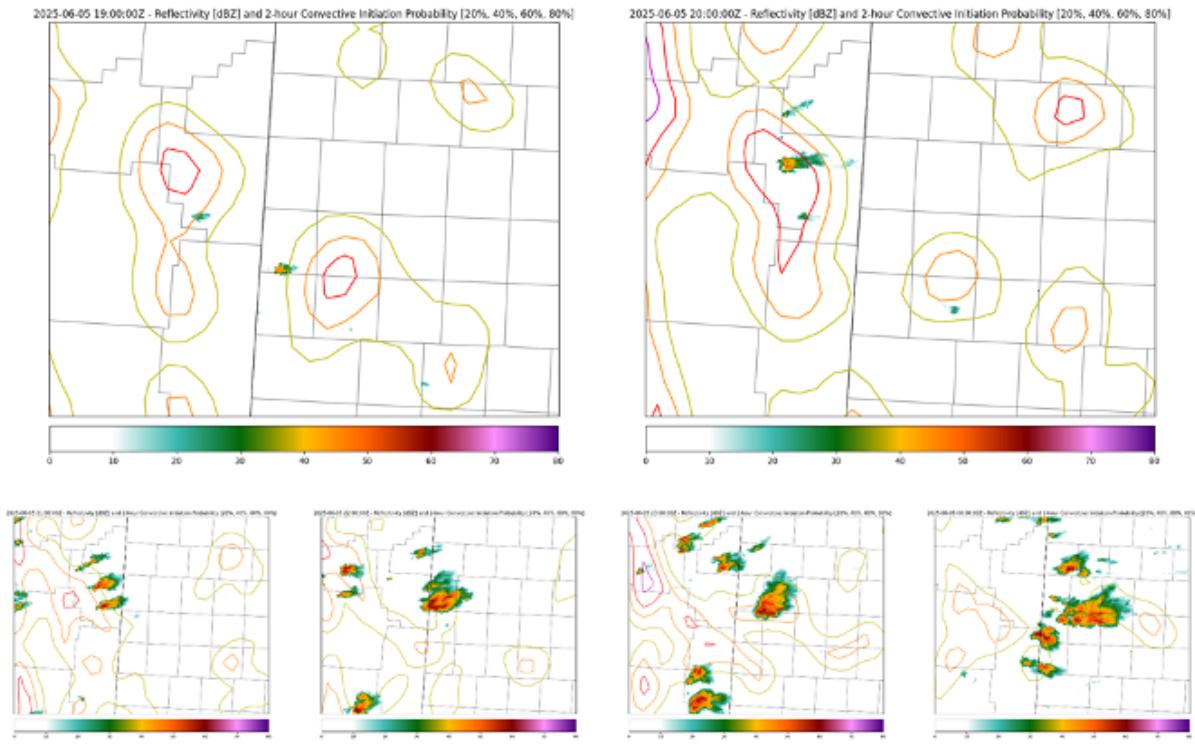


Figure 3: ML-CI outputs for June 5, 2025, the event that produced the Morton, Texas EF2 tornado.

At the start of initiation, there were an abundance of high CI probability contours within the development region. Initial convection, shown in the top two plots, mainly occurred within high CI probability regions. However, many additional high-CI regions did not produce convection. In this region, higher altitudes saw a greater level of overforecasting. Despite this issue, convection was reasonably well resolved by the model, possibly due to the case's location within the training domain.

On June 20, 2025, a tornado outbreak occurred over North Dakota, which eventually spawned the Enderlin EF5 tornado. In this weak forcing environment, existing CAMs struggled to accurately predict storm initiation. This model also experienced forecasting difficulties, as shown in Figure 4.

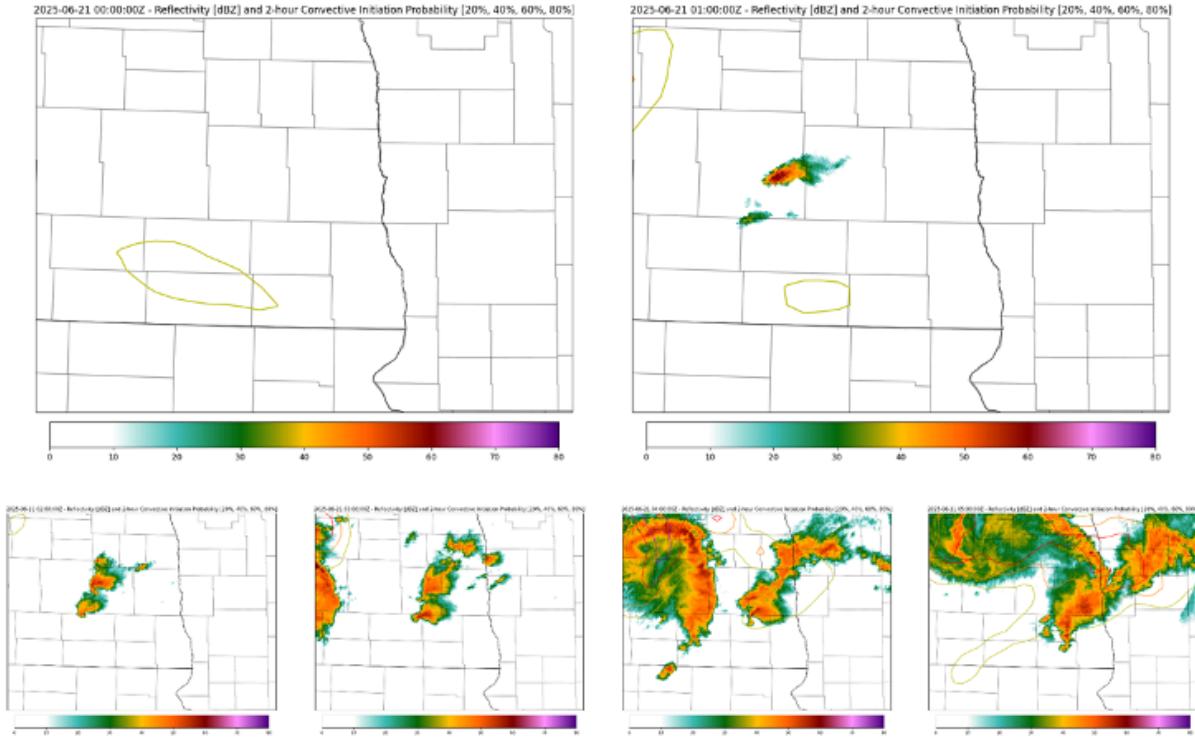


Figure 4: ML-CI outputs for June 20, 2025, the event that produced the Enderlin, North Dakota EF5 tornado.

In this case, the model did a poor job at resolving convective initiation. Few CI regions were mapped, and although the convection shown in the top right plot occurred north of a 20% contour, this does not indicate that the model resolved CI well. Since this event occurred outside of the training domain, the degradation of model performance is not entirely surprising. It is also important to note that full-physics CAMs also failed to resolve the potential for convection in eastern North Dakota in the hours before the event, suggesting that the source of error here may not be an inability of ML-CI to handle out-of-domain events, but rather a set of initial conditions that poorly represented the state of the atmosphere in eastern North Dakota at this time.

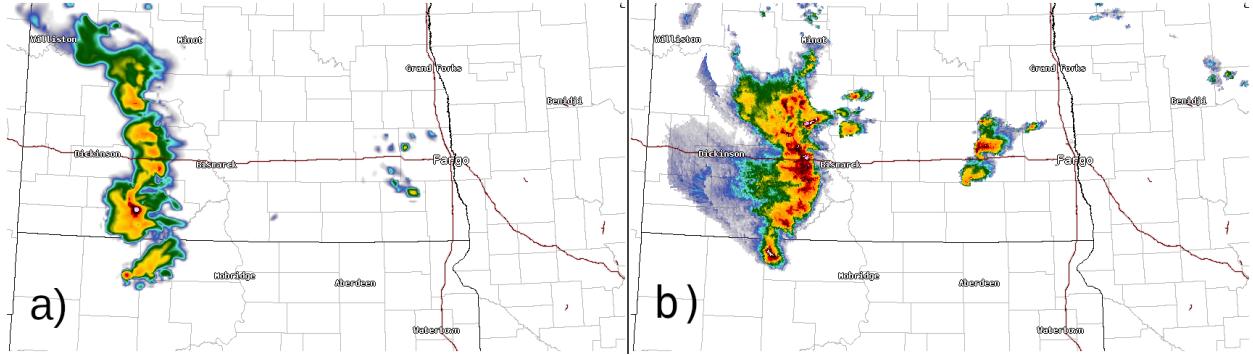


Figure 5: a) HRRR simulated reflectivity and b) MRMS seamless hybrid scan reflectivity valid 04:00Z, 11 minutes before the beginning of the EF5 tornado. Noticeably, no storms were produced by the HRRR in eastern North Dakota despite the fact that rapid vertical development of towering cumulus was ongoing at the time of initialization. See Figure A1 in the appendix.

The final case is August 28, 2024, where a series of storms initiated along the North Dakota-South Dakota border, eventually producing an EF1 tornado which tracked near Mobridge, South Dakota. In this somewhat more synoptically forced initiation, the model's performance, shown in Figure 6, remains significant despite being outside the training domain.

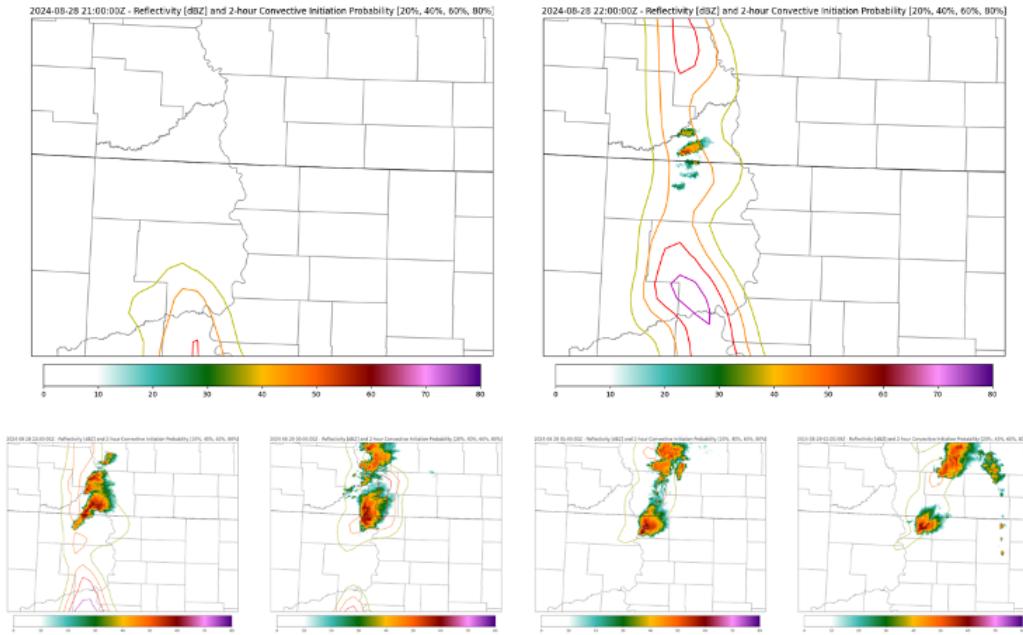


Figure 6: ML-CI outputs for August 28, 2024, the event that produced the Mobridge, South Dakota EF1 tornado.

In the top right plot, a clear boundary of high CI probabilities appears, and within this region, north of a local max in CI probs, initiation occurs. Since this initiation occurred within a region of high CI probs, this shows that the model had some success in resolving this event despite being outside the training domain. This suggests that, in strong, synoptically forced CI cases, the model can still perform well even outside the training domain. However, the highest CI probabilities were located well south, closer to Eagle Butte, and no storms initiated near there. The model overall shows a tendency to correctly identify features capable of providing the mesoscale forcing necessary for CI, but not the exact locations within those features where CI is most likely to occur. Some areas not marked for CI potential also formed storms, but this was much less common.

4. Discussion:

4.1 Model Verification:

As previously discussed, spatial calibration significantly improves the model's performance. This improvement is especially apparent below the 60% probability range, as displayed in Figure 7. While it still generally overforecasts, the severity of overforecasting is seriously improved, and small but noticeable improvements are also seen in the 60% to 90% range. Contrastingly, pointwise calibration showed only very small improvements in the 0% to 50% range, and in fact worsened the tendency to overforecast above 50%. The spatially calibrated model is highly reliable for modeled probabilities below 20%, and in fact slightly underforecasts between 3% and 15%. Despite improvements from spatial calibration, the model's performance is still severely limited by its strong tendency to overforecast, in large part because it assigns high probabilities to boundaries that have already initiated storms and are not primed to do so again. All of the reliability curves for the uncalibrated and pointwise-calibrated models are far below the no-skill line, and a large portion of the spatially calibrated model's reliability curve falls below it as well.

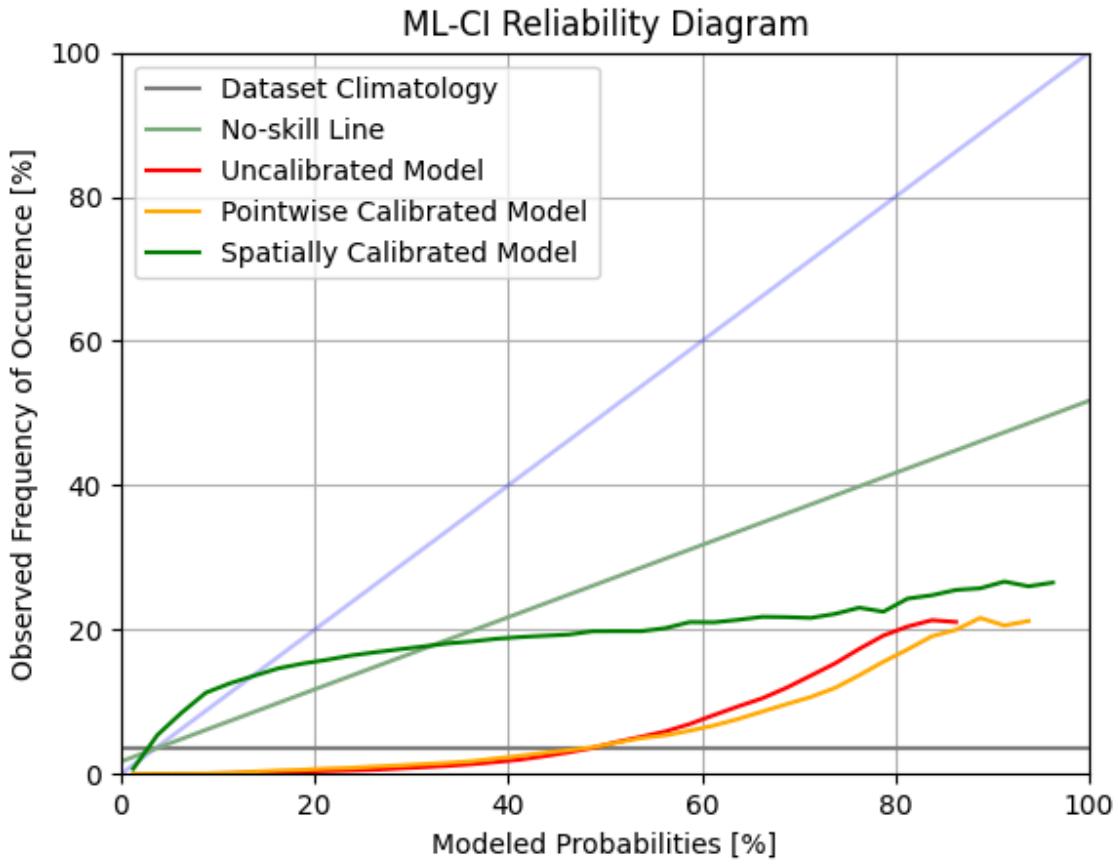


Figure 7: Reliability diagram of the ML-CI model's performance. Uncalibrated, pointwise-calibrated, and spatially calibrated versions are all included here.

However, other evaluations of model performance exist. Figure 8 shows the relative operating characteristic (ROC) curves and performance diagrams for all three versions of the model. The uncalibrated and pointwise-calibrated curves are identical in both plots, as they are one-to-one mappings of the same underlying probability distribution. Thus, the binarization of pointwise-calibrated model outputs is exactly identical to a binarization of uncalibrated model outputs using another probability threshold determined by the shape of the isotonic regression curve. Even though there are three versions of the model being plotted below, there are effectively only two curves to analyze per diagram.

ROC curves can be thought of as showing the model's ability to resolve between two outcomes, without making any considerations for bias. It evaluates the distribution's shape more than the individual modeled probabilities at each grid point. All three models show quite good

performance in the ROC curve plot, and all stay well above the no-skill line here, distinct from the reliability diagram's no-skill line. This indicates that ML-CI may still have some utility for operational forecasters despite its serious flaws. However, the performance diagram serves to temper expectations. While the probability of detection remains relatively high at low probability thresholds, the success ratio, defined as the percentage of positive model outputs that verify, remains very low and only inches upward as probability thresholds increase and detection probabilities decrease.

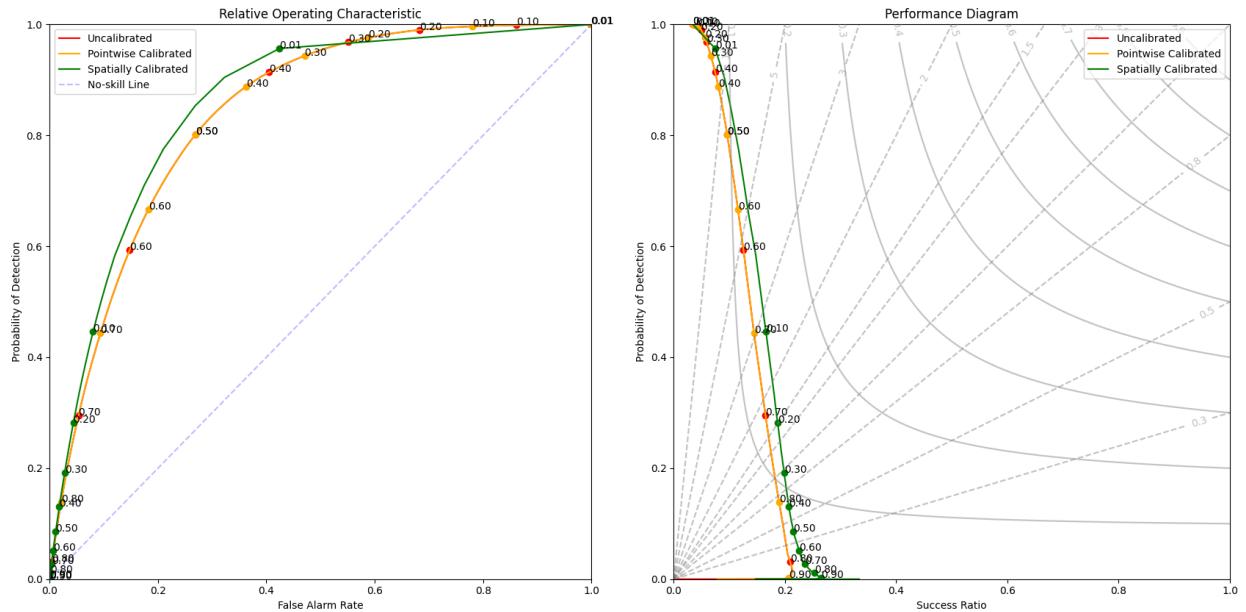


Figure 8: ROC curves and performance diagrams for the ML-CI model. Uncalibrated, pointwise-calibrated, and spatially calibrated versions are all included here.

The performance diagram, despite its fairly harsh verdict, can also be used to tune probability thresholds based on individual users' priorities. The straight dashed lines are contours of model bias, and the diagonal line with a slope of 1 represents a forecast that neither overforecasts nor underforecasts at large. If a perfectly balanced forecast at large is desired, a probability threshold of 0.29 may be used. This probability threshold has an associated probability of detection of 19% and a success ratio of 19%. Contrastingly, if the best overall performance is desired, the roughly hyperbolic solid gray lines representing the critical success index (CSI), alternatively known as the threat index, may be used. CSI is defined as the number of true positive model outputs divided by the sum of true positives, false positives (false alarms),

and false negatives (misses). If maximizing CSI is appropriate, a probability threshold of 0.15 may be used. The probability of detection using this threshold is 37% and the success ratio is 18%.

4.2 Variable Significance:

In creating the CI model, the contributions of each environmental parameter were analyzed to understand the significance of certain environmental variables towards convective initiation. To start, the mean decrease in impurity (MDI) was calculated for the fitted random forest model. This parameter analyzed the effect of each variable in determining whether a CI code of 1 or 0 would occur. The plot of this parameter is shown in Figure 9.

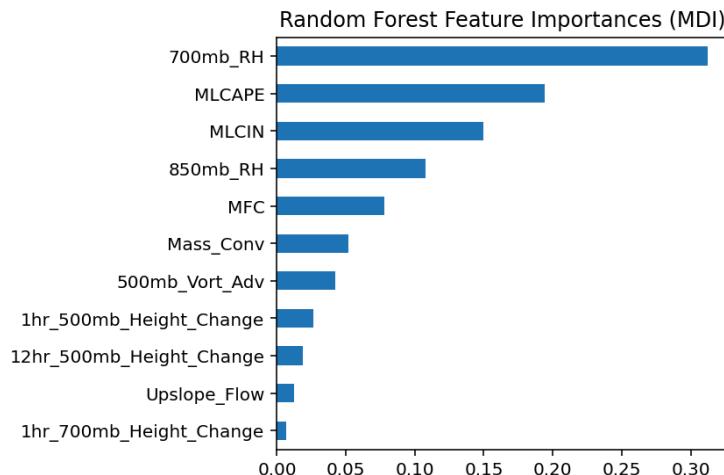


Figure 9: Plot of Random Forest MDI feature importances for each environmental variable used.

Overall, 700 mb RH and MLCAPE had the highest overall feature importance while upslope flow and 1-hour 700 mb height change had the least importance. However, to assess the specific importance of the environmental parameters in predicting new convective initiation, backwards multipass permutation significance testing was conducted. Multipass permutation testing, rather than single-pass permutation testing, was necessitated by the existence of some cross-correlations between environmental variables, as shown in Figure 10.

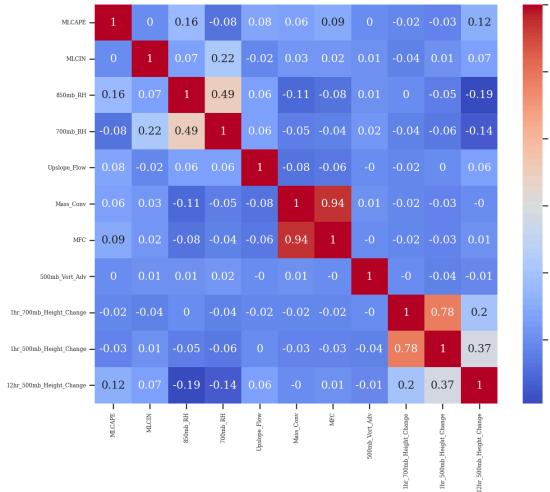


Figure 10: Correlation matrix for the predictor set.

Single-pass permutation testing requires all variables to be independent, and multipass permutation testing can bypass this issue. This test is carried out by first ranging over the predictor matrix and permuting each value to identify which, when removed, is associated with the most significant drop in accuracy. Then, this variable is permanently permuted and the cycle continues until only one parameter is left which is the least significant one. In Figure 11, backwards multipass testing was conducted, which is carried out the same way as described earlier but backwards, starting with the least important variable and concluding with the most important one.

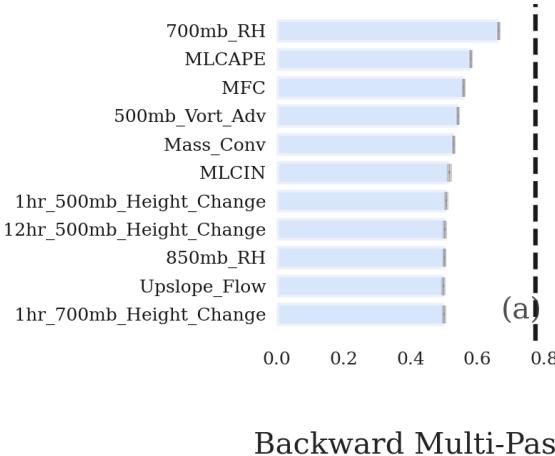


Figure 11: Plot of the Backwards multipass permutation test evaluated with the “Area under the Curve” Function and conducted with the predictor set.

Notably, the top two and bottom two variables in Figure 11 are the same in Figure 9. This demonstrates the model's ability to differentiate between CI and non-CI events and confirms which predictors are the most influential in producing ML-CI's forecast. However, since these tests were conducted on the uncalibrated random forest model, there may be some discrepancies between these analyses and the final model, and any existing discrepancies would be very difficult to address with current methods.

5. Conclusion and Future Work:

Most broadly, the ML-CI model is best for identifying the locations of mesoscale boundaries and other features that could initiate a convective thunderstorm. It is far less skillful at predicting precise locations of CI, and cannot be reliably used to determine the specific parts of a boundary that will successfully initiate convection or which isolated topographical features and confluence bands will fire storms as opposed to which ones will not. This may possibly be caused by the limitations of RAP data as an accurate representation of the atmosphere. However, the ROC curve shows that the model has a fairly strong ability to resolve the probability of convective initiation over the spatial domain, even if the individual values contained within the distribution are not at all accurate.

The inclusion of 850 mb and 700 mb relative humidity predictors also present a possible source of major error. The 700 mb relative humidity is meant to exclusively represent the relative humidity in the lower free troposphere. This assumption generally holds true in lower elevations such as Central Oklahoma where storm environments are dominated by maritime tropical airmasses and atmospheric boundary layers are confined within the lowest kilometer or two of the atmosphere. However, the higher elevation western areas are characterized by both lower surface pressures and deeper boundary layers, caused in part by the increased influence of continental tropical airmasses. The top of the boundary layer in the summer can easily rise above the 700 mb level, and the predictor intended to represent the humidity within the capping layer can instead be replaced by a value representing the humidity within the boundary layer. This problem is exacerbated by the high importance of the 700 mb relative humidity predictor, as shown by the variable significance analysis performed in section 4.2. Future iterations of this model will replace 850 mb and 700 mb with an explicitly computed mean relative humidity

between the lifted condensation level and the level of free convection, with the intention that lower free tropospheric relative humidity and thus the magnitude of dry air entrainment will be more accurately represented to the ML-CI model.

Future iterations of ML-CI will also include observed data to supplement model data. Corrections to surface fields derived from surface observations, in a manner similar to the Storm Prediction Center's Mesoscale Analysis, will be performed to improve the accuracy of the atmospheric data used to train the model. Geostationary satellite data may also be included to represent information relating to the presence of already-existing clouds, including agitated cumulus fields and towering cumulus clouds. This may allow ML-CI to identify regions where CI is already in progress. A few other predictors may also be considered for inclusion, such as equivalent potential temperature advection. If included, this indicates that the airmass may tend to destabilize over the next few hours. The model, in its current state, is not ready for operations outside of use as an experimental forecasting aid, but it also shows the potential for improvement with additional time to continue construction and refinement.

6. Acknowledgements:

This project would not have been possible without the assistance and mentorship of Dr. Eric Loken. Dr. Loken helped this team access archived ProbSevere v3 data that was not publicly available and assisted in the production of ProbSevere-derived CI target matrices. The datasets curated and hosted online by Amazon Web Services (RAP 13 km and MRMS), National Severe Storms Laboratory (MRMS), and Storm Chase Archive were essential in the qualitative evaluation process of the ML-CI model.

7. References:

- Anderson, M. J., 2001: Permutation tests for univariate or multivariate analysis of variance and regression. *Canadian Journal of Fisheries and Aquatic Sciences*, **58**, 626–639. [doi:10.1139/cjfas-58-3-626](https://doi.org/10.1139/cjfas-58-3-626).
- Banacos, P. C. and D. M. Schultz, 2005: The use of moisture flux convergence in forecasting convective initiation: Historical and operational perspectives. *Wea. Forecasting*, **20**, 351–366. [doi:10.1175/waf858.1](https://doi.org/10.1175/waf858.1).
- Bech, J., J. Arús, S. Castán, N. Pineda, T. Rigo, J. Montanyà, and O. van der Velde, 2014: A study of the 21 March 2012 tornadic quasi linear convective system in Catalonia. *Atmospheric Research*, **158–159**, 192–209. [doi:10.1016/j.atmosres.2014.08.009](https://doi.org/10.1016/j.atmosres.2014.08.009).
- Cintineo, J. L., M. J. Pavolonis, and J. M Sieglaff, 2024: ProbSevere Version 3: Improved exploitation of data fusion and machine learning for nowcasting severe weather. *Wea. Forecasting*, **39**, 1937–1958. [doi:10.1175/waf-d-24-0076.1](https://doi.org/10.1175/waf-d-24-0076.1).
- Gard, T. L., H. E. Fuelberg, and J. L. Cintineo, 2022: The utility of ProbSevere v2.0 for predicting pulse severe thunderstorms. *Wea. Forecasting*, **37**, 1601–1613. [doi:10.1175/waf-d-22-0003.1](https://doi.org/10.1175/waf-d-22-0003.1).
- Jergensen, G. E., A. McGovern, R. Lagerquist, and T. Smith, 2019: Classifying convective storms using machine learning. *Wea. Forecasting*, **35**, 537–559. [doi:10.1175/waf-d-19-0170.1](https://doi.org/10.1175/waf-d-19-0170.1).
- Kain, J. S. and Coauthors, 2013: A feasibility study for probabilistic convection initiation forecasts based on explicit numerical guidance. *Bull. Amer. Meteor. Soc.*, **94**, 1213–1225. [doi:10.1175/bams-d-11-00264.1](https://doi.org/10.1175/bams-d-11-00264.1).
- Lee, Y., C. D. Kummerow, and I. Ebert-Uphoff, 2021: Applying machine learning methods to detect convection using geostationary operational environmental satellite-16 (GOES-16) advanced baseline imager (ABI) data. *Atmospheric Measurement Techniques*, **14**, 2699–2716. [doi:10.5194/amt-14-2699-2021](https://doi.org/10.5194/amt-14-2699-2021).
- Lock, N. A., and A. L. Houston, 2014: Empirical examination of the factors regulating thunderstorm initiation. *Mon. Wea. Rev.*, **142**, 240–258. [doi:10.1175/mwr-d-13-00082.1](https://doi.org/10.1175/mwr-d-13-00082.1).
- Stensrud, D. J., and J. M. Fritsch, 1994: Mesoscale convective systems in weakly forced large-scale environments. Part III: Numerical simulations and implications for

- operational forecasting. *Mon. Wea. Rev.*, **122**, 2084–2104.
[doi:10.1175/1520-0493\(1994\)122<2084:MCSIWF>2.0.CO;2](https://doi.org/10.1175/1520-0493(1994)122<2084:MCSIWF>2.0.CO;2).
- Tempest, K. I., G. C. Craig, M. Puh, and C. Keil, 2024: Convergence of ensemble forecast distributions in weak and strong forcing convective weather regimes. *Quarterly Journal of the Royal Meteorological Society*, **150**, 3220–3237. [doi:10.1002/qj.4684](https://doi.org/10.1002/qj.4684).
- Veillette, M. S., H. Iskenderian, P. M. Lamey, and L. J. Bickmeier, 2013: Convective initiation forecasts through the use of machine learning methods. *11th conf. on Artificial and Computational Intelligence and its Applications to the Environmental Sciences*, <https://www.ll.mit.edu/r-d/publications/convective-initiation-forecasts-through-use-machine-learning-methods>.
- Wilson, J. W., and R. D. Roberts, 2006: Summary of convective storm initiation and evolution during IHOP: Observational and Modeling Perspective. *Mon. Wea. Rev.*, **134**, 23–47. [doi:10.1175/mwr3069.1](https://doi.org/10.1175/mwr3069.1).
- Zhang, M., Z. Meng, Y. Huang, and D. Wang, 2019: The mechanism and predictability of an elevated convection initiation event in a weak-lifting environment in central-eastern China. *Mon. Wea. Rev.*, **147**, 1823–1841. [doi:10.1175/mwr-d-18-0400.1](https://doi.org/10.1175/mwr-d-18-0400.1).

8. Generative AI Statement:

Generative AI was used sparingly in this research project, exclusively within the code debugging process. Generative AI was not used to write any of this paper's contents nor to produce any figures included in this paper.

9. Appendix:

For the sake of clarity in Table 1, a list of all variables and their physical meanings has been provided in Table A1.

Variable	Description
\vec{V}_h	Horizontal component of the wind. Used here primarily to represent the 10-meter above ground level (AGL) wind, but also used for the 500 mb wind in the case of 500 mb absolute vorticity advection.
z_{sfc}	Elevation of surface topography above mean sea level.
q	Specific humidity.
p	Air pressure, surface-level in all uses here.
R_d	The dry air gas constant; 287.04 J kg^{-1} .
$.T_v$	Virtual temperature.
η	Absolute vorticity.
ζ	Relative vorticity.
f	Earth vorticity, or alternatively the Coriolis parameter.

Table A1: Description of variables used in the formulas of Table 1.

It is not evident anywhere within Figure 5 that towering cumulus was present in eastern North Dakota on June 21, 2025 at 00:00 UTC. Therefore, to substantiate this claim, a radar and satellite plot from the Storm Chase Archive has been included as Figure A1. The towering cumulus cloud can be seen as an area of slightly lower visual reflectance tilting away from the line of sight toward the north in south-central Stutsman County. Further satellite data available on the Storm Chase Archive website shows that this towering cumulus cloud later became the parent supercell of the Enderlin, North Dakota EF5 tornado. The radar and satellite loops only cover a six-hour time period centered on the time of the most significant event within a case, so this satellite image was pulled from the Spiritwood, North Dakota EF3 case page.

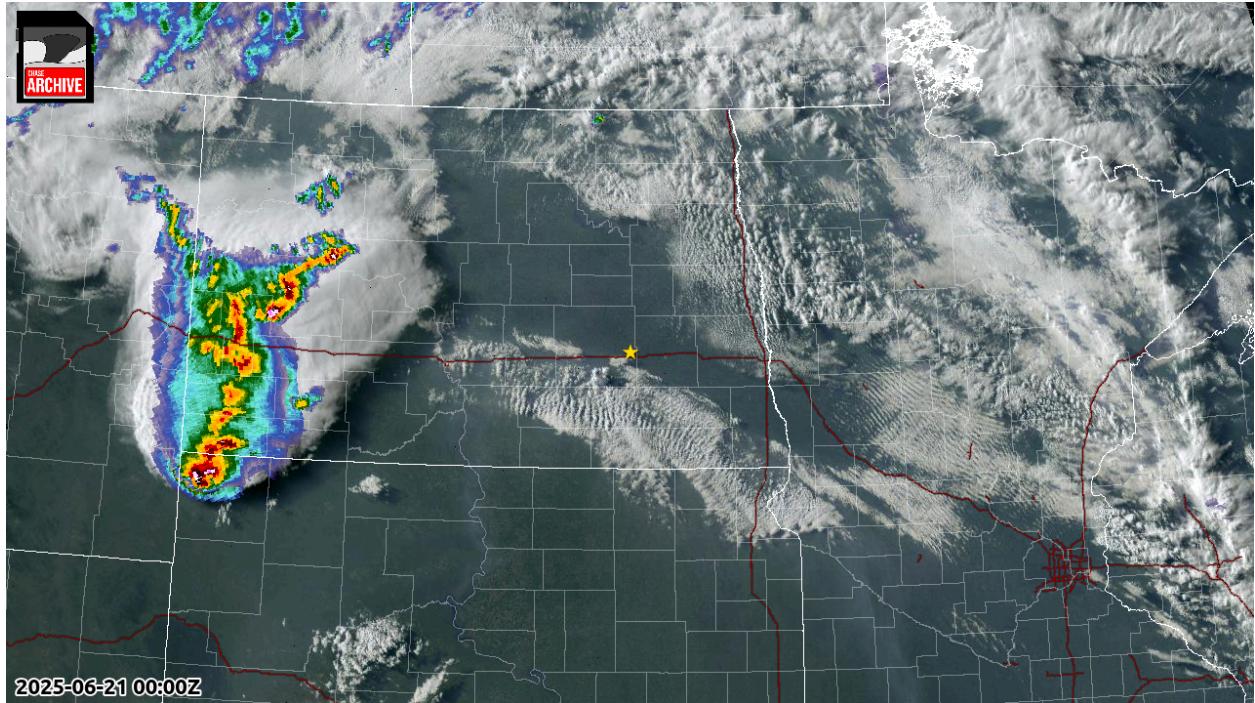


Figure A1: True-color satellite and MRMS hybrid scan reflectivity valid at 2025-06-21 00:00 UTC. The yellow star marks Spiritwood, North Dakota, 48 miles west-northwest of Enderlin and the location of an EF3 tornado in the same outbreak. The updraft that would later become the parent storm of the Enderlin EF5 was already visibly developing and near the point of convective initiation. Plot courtesy of the Storm Chase Archive.

The [Enderlin, North Dakota EF5](#) and [Spiritwood, North Dakota EF3](#) case pages, which contain all satellite data required to validate this analysis, can both be viewed by either following the included hyperlinks or on the Archive's website at <https://app.chasearchive.com>.