Assessing Machine Learning Probabilistic Forecast Utility for Severe Weather Forecasting

Ian D. Shank[1] and Aaron J. Hill[2]

[1]National Weather Center Research Experiences for Undergraduates Program
Norman, Oklahoma
[2]School of Meteorology, The University of Oklahoma
Norman, Oklahoma

ABSTRACT

This work associated probabilistic values made by an Artificial Intelligence (AI) weather prediction system with historic high-impact severe weather events so that operational forecasters can use AI to help predict high-impact severe weather events in the continental United States (CONUS). This study examined how medium-range (day4-8) machine learning probabilistic forecasts is compared to the Storm Prediction Center (SPC) Day 1 Convective Outlook and observed severe weather reports. Five years of data from 2020-2025 were used to compare probabilities from the GEFS-MLP at different lead times within the medium-range to forecasts made by the SPC Day 1 Convective Outlook. This approach showed which probabilities in the medium-range displayed a better ability in predicting a high-impact severe weather event in CONUS. Higher probabilities in shorter lead times tend to correlate with higher SPC categorical forecasts for high-impact severe weather events. Probabilities in the 5-day lead time of the GEFS-MLP showed the best ability to predict high-impact severe weather events, specifically the 60% probabilistic threshold. Longer lead times, such as day 8 and day 7, had better ability at lower probabilistic thresholds, while shorter lead times, like day 4 and day 5, had better ability at higher probabilistic thresholds. Recognizing which lead times and which probabilities have better correspondence in predicting severe weather can enhance operational forecast products at longer lead times.

1. Introduction

One of the most challenging weather phenomena to predict precisely is severe weather, which includes hail above one inch in diameter, wind gusts 58 miles per hour or greater, or the presence of a tornado (Hill et al. 2020). With the advancements of Artificial Intelligence (AI) and Machine Learning (ML), forecasters can use AI predictions to increase forecast skill. Meteorologists and atmospheric scientists have been utilizing AI/ML to enhance the accuracy and reliability of their forecasts, thereby increasing trust in operational forecasts among the public.

ML, along with traditional numerical prediction models, are used to help operational meteorologists create their forecasts. Some forms of ML in meteorology can use probabilities to predict severe weather. The ML system that is used in this research is the Global Ensemble Forecasting System Machine Learning Probabilities (GEFS-MLP). The GEFS-MLP looks at patterns within historical meteorological data to predict severe weather hazards. The GEFS-MLP creates probabilistic severe weather forecasts for the continental United States for both short-range lead times (day 1-3) and medium-range lead times (day 4-8) (Hill et al. 2020).

Previous research has been done to give more background on how Random Forests (RF) are used to train ML and AI weather prediction models (Hua et al. 2025, Loken et al. 2020). RF trained models, which include the GEFS-MLP that will be used within this research, can often compete with operational severe weather forecasts, including the Storm Prediction Center (SPC) Convective Outlook (Loken et al. 2020).

The SPC Convective Outlook displays categorical forecasts for short-range severe weather outlooks attached to individual risk probabilities made by operational forecasters to convey the overall severe risk in an area. The SPC also creates probabilistic forecasts for the overall severe threat for medium-range forecasts.

[1] *Corresponding author address:* Ian Shank, University of North Carolina at Charlotte, 120 David L. Boren Blvd., Norman, OK 73072, ian.shank71@gmail.com.

Forecasters, including meteorologists at the SPC, already use AI weather prediction models to help create and refine their forecasts, but the goal of this research was to show how forecasters can better utilize the probabilistic forecasts. The GEFS-MLP was made to mimic the SPC Convective Outlook and probabilistically predict the overall severe risk in an area, and the SPC is able to utilize those forecasts as a tool to help adjust their forecasts.

Research already completed on the GEFS-MLP and the SPC shows that the GEFS-MLP has better skill than the SPC Convective Outlook at medium-range lead times, suggesting that ML can successfully analyze patterns from previous events to predict severe occurrences (Hill et al. 2023). Other research has demonstrated how AI models can assist operational forecasters in their process of issuing hazardous weather outlooks (Hill et al. 2020). AI/ML has significantly advanced the prediction of severe weather, including the probabilistic forecasts of the GEFS-MLP. The trust of operational forecasters, along with the utility of the AI weather prediction forecasts, is just as important as the accuracy of the model (McGovern et al. 2023).

It is already known that GEFS-MLP skill drops within the medium-range; however, there is little research on whether there are still signals within the probabilities in the medium-range that operational forecasters can utilize within their forecasts (Hill et al. 2023). More signals in probabilistic values based on the historical data can create more certainty about how to use the AI forecasts and may continue to add value to those medium-range forecasts. Throughout this research, there will be a focus on probabilities and lead times from AI-based guidance products that would provide additional confidence for operational forecasters that a high-impact severe weather event will occur. Adding value to these probabilistic forecasts from the GEFS-MLP can create a more positive perspective on the utility of AI/ML in operational forecasting. Understanding the distribution of AI probabilistic forecasts per severe weather event can help show forecasters what probabilities and lead times have value towards predicting severe weather. This research seeks to understand how operational forecasters can utilize the medium-range AI forecasts to help increase lead time for a high-impact severe weather event.

2. Data and Methods

This research utilizes three sets of data- the GEFS-MLP forecasts, the SPC Convective Outlooks, and the SPC storm reports. The SPC Storm Reports are used to assess the utility of GEFS-MLP forecasts and SPC Convective Outlooks. The SPC Storm reports were gathered via the SPC Product and Report Archives. These archives allow people to obtain storm reports for a specific date or range of dates. SPC Convective Outlooks were also obtained to look at the SPC Day 1 forecast and compare it to medium-range GEFS-MLP probabilistic forecasts.

The GEFS-MLP forecasts mimic what the SPC Convective Outlook is trying to convey, which is to predict the occurrence of severe weather. The GEFS-MLP uses probabilistic values to predict the overall risk of severe weather and individual hazards in the short and medium range. The biggest difference between the GEFS-MLP forecasts and the SPC outlooks is that the SPC attaches categorical names (marginal, slight, enhanced, moderate, and high) to their short-range convective outlooks (days 1-3) to help relay the severe risk to the public more efficiently.

Another difference between the GEFS-MLP and the SPC Convective Outlook is the SPC only provides outlooks for the short-range (Day 1-3 lead times). In this research, the focus is on the medium-range (Day 4 through Day 8) forecasts of the GEFS-MLP and the day 1 forecasts of the SPC Convective Outlook to look at the signal that a medium-range GEFS-MLP forecast has in predicting a high-impact event. When comparing GEFS-MLP forecasts, SPC forecasts, and storm reports, it is important to know how the Day 1 convective outlook categorical forecasts from the SPC correspond to the probabilistic forecasts from the GEFS-MLP.

To evaluate forecasts, storm report data is gathered for every date between October 2020 and April 2025 from 1200 UTC until 1159 UTC the next day. The start times of the storm report data are matched to the 1200 UTC issued SPC Convective Outlooks. All GEFS-MLP forecasts are valid at 1200 UTC and span a 24-h period, consistent with the SPC outlook and gathered storm reports.

Storm reports and severe weather forecasts from both the GEFS-MLP and the SPC are within the continental United States (CONUS). Severe storm reports were collected from October 1, 2020, through April 30, 2025, to compare with the GEFS-MLP. The GEFS-MLP medium-range forecasts and the SPC Day 1 Convective Outlooks were analyzed

from October 2, 2020, through April 30, 2025, as the GEFS-MLP archives began on October 2, 2020.

The overall approach to this research was to analyze the ML probabilities created from the GEFS-MLP and to see which of the ML probabilities had better accuracy signaling a high-impact severe weather event than the SPC. To do this, the forecasts were visually inspected from recent high-impact events to understand how poor forecast skill was not indicative of potential impacts. Secondly, forecasts were analyzed through the ~5 years of data gathered to understand the distribution of categorical SPC forecasts with respect to the different probabilistic values made by the GEFS-MLP at each lead time day between day 4 and day 8. This shows the distribution of SPC categorical forecasts in 5 years for a specific ML probabilistic value.

Not only was the distribution of the SPC categorical forecasts evaluated, but so were the severe storm reports. Using graphical analysis and statistical evaluation of verified storm reports from CONUS, the severe storm reports helped demonstrate which probabilistic values from the GEFS-MLP were verified in terms of high-impact severe weather events, meaning severe weather events that caused significant loss of life, property damage, or economic loss. This analysis, along with the analysis done with the SPC Day 1 Convective Outlook, illustrated which probabilistic value and which lead time had a good signal of a high-impact severe weather event occurring within CONUS.

3. Results

When comparing GEFS-MLP medium-range probabilistic forecasts with verified storm reports from the SPC, more high-impact severe weather events occur during days 4 and 5 60% risk for severe weather (Figure 1). There are more severe storm reports on average for days 7 and 8 at lower probabilities, but the same signal exists at days 4 and 5 for larger probabilities. This is what is expected as the probabilities of severe weather are trained to increase as the risk of severe weather increases with time. All lead time days have low storm report averages between 100-300 reports at low probability values, but continue to increase until the 30% probability threshold. The average number of storm reports at the day-8 lead-time decreases as probabilities get larger, while days 6 and 7 see a drastic decrease in average storm reports at the 60% probability threshold, as there are fewer days with those probabilities at longer lead times. Days 4 and 5 drastically increase at 60% as there is a higher number of days that have higher probabilities of severe weather at shorter lead times.

After looking at the probabilities from the GEFS-MLP and how they correlated to severe storm reports, it was crucial to analyze how high-impact events occurred within SPC Convective Outlooks and how those forecasts correlated with the GEFS-MLP forecasts. During this comparison of the GEFS-MLP forecasts and the SPC Day 1 Convective Outlook, one specific day was analyzed to show the spatial probabilistic differences between the two forecasts (Figure 2). The GEFS-MLP did a good job predicting the severity of the event, based on the severity that the SPC predicted at Day 1, as well as the general location at the 5-day lead time, which continues to indicate that the GEFS-MLP has good skill for medium-range forecasts. The biggest difference between the GEFS-MLP forecast and the day 1 SPC Convective Outlook for this date is the orientation of the moderate risk from the SPC being more southeast from the areas of highest probabilities from the GEFS-MLP, as well as the northwestward extension of the marginal risk. Other specific case studies that couldn't be included in the paper show similar results, such as April 27, 2024, April 2, 2025, and May 26, 2024, which show that GEFS-MLP forecasts predict severe weather in medium-range forecasts in a spatially and probabilistically accurate manner.

The same results can be aggregated when all dates between October 2, 2020 and April 30, 2025 were compared. In Figure 3, the maximum probability for day-4 forecasts increases, and the percentage of each bin that has a higher categorical forecast from the SPC Day 1 Convective Outlook increases. This means that the percentage of enhanced, moderate, and high risks increase as probabilities increase. There is a positive correspondence between the GEFS-MLP forecasts and the SPC Day 1 Convective Outlook - as maximum probabilities increase at day 4, so do the day 1 categorical forecasts from the SPC. At the highest probability bin (>60%), there are only enhanced, moderate, and high risks issued by the SPC at Day 1.

The highest variation in categorical forecasts is the 45-60% bin, which has all SPC categorical forecasts from marginal to high risks, indicating that

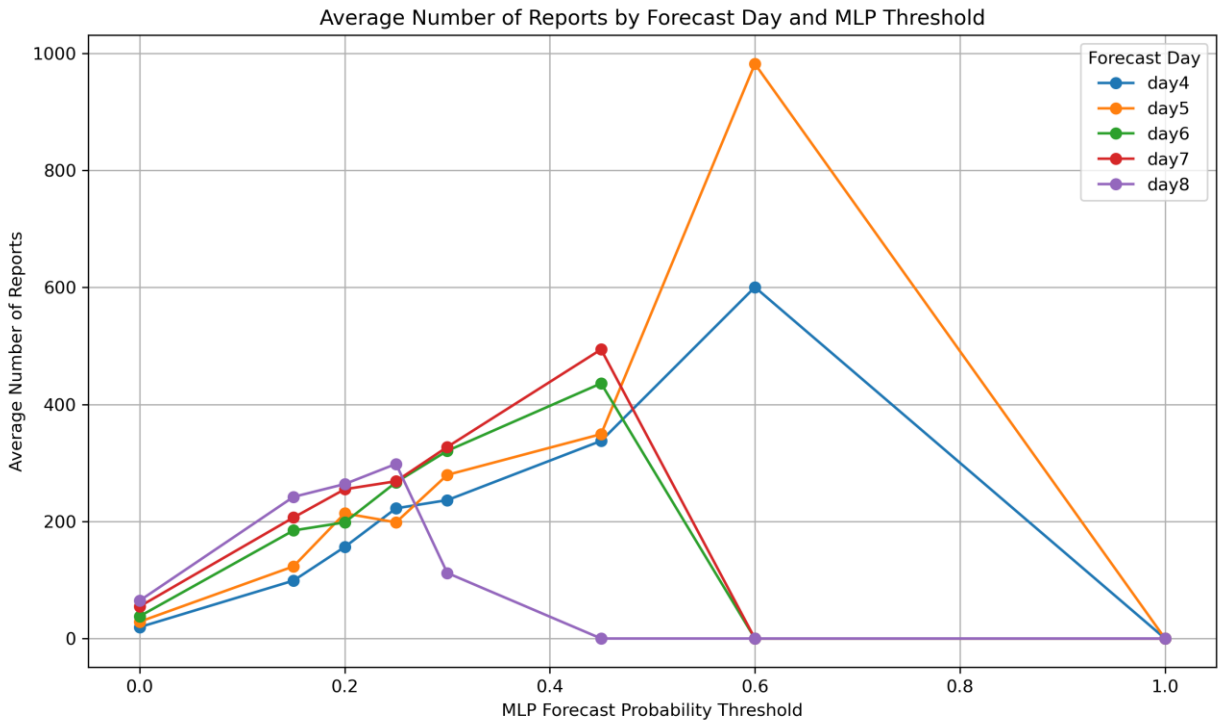

Figure 1. Average number of severe storm reports per probability threshold for each lead time, including days 4, 5, 6, 7, and 8.

this probability bin does not do as well for the prediction of high-impact events as the highest probability bin does.

Correspondence between SPC outlooks and day 5 GEFS-MLP forecasts is similar (Figure 4). The biggest difference between the day 4 and day 5 probabilistic forecasts is the last two probability bins. The 45-60% bin has high risks, but the 60-99% bin has no high risks for the day 5 probabilities, which is different from the day 4 probabilities. Although this can be a perplexing finding, this correlates with the biggest finding in Figure 1, which shows that day 5 60% probabilistic forecasts from the GEFS-MLP have the highest number of severe storm reports on average. This finding indicates that a day 5 60% probabilistic forecast from the GEFS-MLP is a good indicator of a high-impact severe weather event. Even though there are

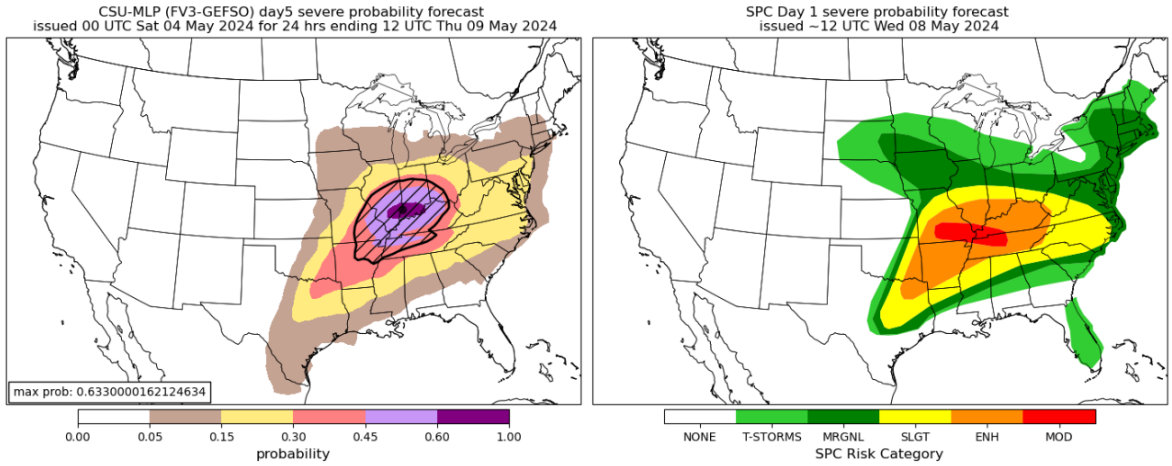

Figure 2. Visual Comparison of the GEFS-MLP at a 5-Day lead time versus the SPC Day 1 Convective Outlook. GEFS-MLP forecast for May 8, 2024 was issued on May 4, 2024 at 1200 UTC, and the SPC Day 1 Convective Outlook was issued at 1200 UTC on May 8, 2024.

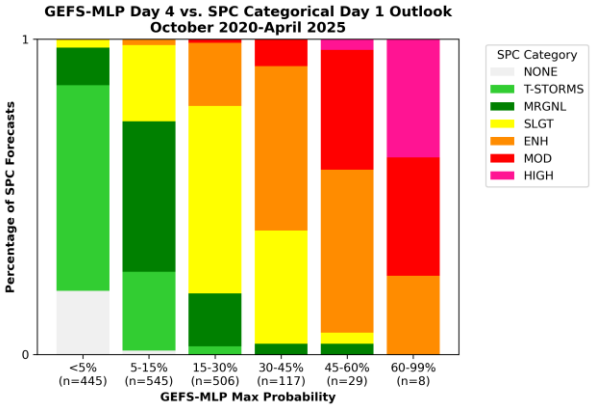

Figure 3. Shows the Normalized Distribution of SPC Categorical Forecasts for each bin for each Range of Probabilistic Values from the day-4 Forecast from the GEFS-MLP.

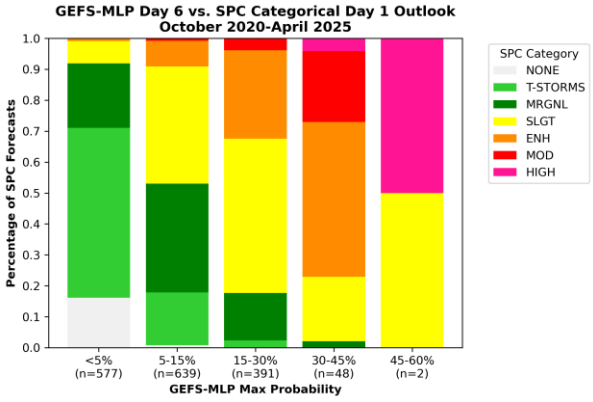

Figure 5. Shows the Normalized Distribution of SPC Categorical Forecasts for each bin for each Range of Probabilistic Values from the day-6 Forecast from the GEFS-MLP.

enhanced risks issued at Day 1 for maximum probabilities of less than 5%, there is a higher percentage in the higher probability bins, indicating that those probabilistic values do a better job at predicting a higher-severity event.

The day 6 probabilities and SPC forecasts, as seen in Figure 5, show the same positive correlation that was seen in days 4 and 5. All lead times, including day 4 through day 6, have more marginal severe risks and general thunderstorm risks in bins <5% and 5-15%. Day 6 has more variability in the last two bins, including all categorical forecasts from the SPC Day 1 Convective Outlook, ranging from marginal to high risk in the bin ranging from 45-60%. The last bin has only slight and high risks, skipping enhanced. This is due to the sample size of days that we have within this probability bin. There are only two days in that bin, and since they had drastically different SPC categorical forecasts, it can be concluded that bin does not predict the severity of a severe weather event well.

GEFS-MLP data at the day-7 lead time shows promising results for lower-range probabilities. As lead-time increases, so does uncertainty, meaning that the results in days 7 and 8 begin to show low-probability results (Figures 6). Most maximum probabilities issued by the GEFS-MLP at day 7 were less than 45%. There was only one day that had a maximum probability greater than 45% for day 7, and it had a moderate risk issued by the SPC at day 1. The next probability bin (30-45% maximum probability) showed promising results with a high percentage of enhanced, moderate, and high risks issued by the SPC at day 1. This

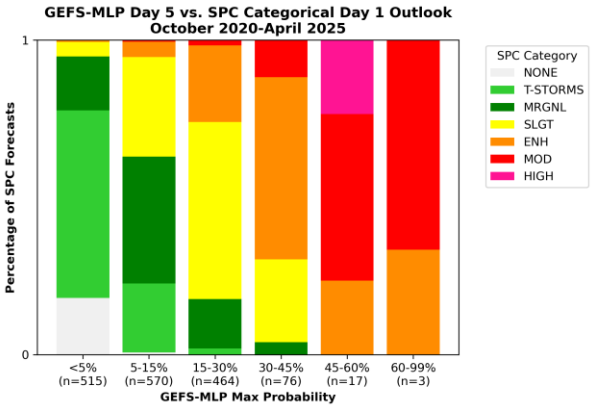

Figure 4. Shows the Normalized Distribution of SPC Categorical Forecasts for each bin for each Range of Probabilistic Values from the day-5 Forecast from the GEFS-MLP.

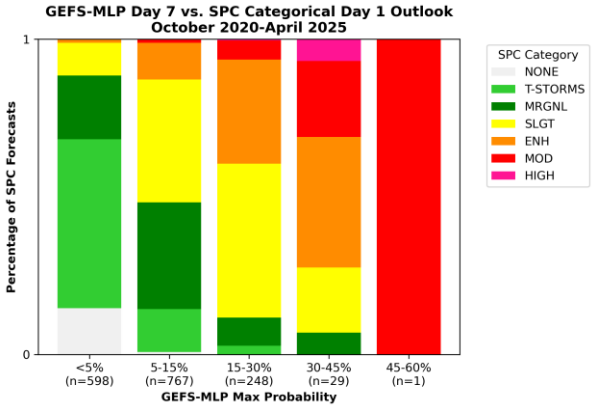

Figure 6. Shows the Normalized Distribution of SPC Categorical Forecasts for each bin for each Range of Probabilistic Values from the day-7 Forecast from the GEFS-MLP.

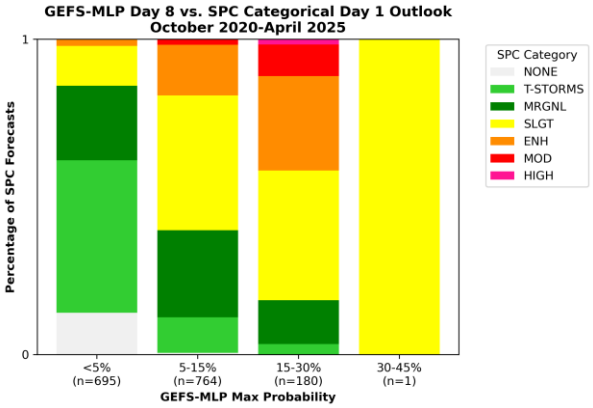

Figure 7. Shows the Normalized Distribution of SPC Categorical Forecasts for each bin for each Range of Probabilistic Values from the day-8 Forecast from the GEFS-MLP.

indicates that the GEFS-MLP probabilistic forecasts at day 7 might signal a high-impact event at medium probabilistic thresholds.

Day 8 forecasts, on the other hand, showed a lot more uncertainty in predicting the severity of a severe weather event, as there was only a singular day that had a maximum probability in the 30-40% probability bin, and that day was issued a slight at day 1 (Figure 7). The rest of the days within the data had a maximum probability of less than 30% issued at day 8; however, there is a lot of variability between the categorical forecasts issued at day 1 by the SPC in the 15-30% bin. The large number of days that have a smaller severe weather probabilistic forecast at day 8 is high due to the large amounts of uncertainty at such a long lead time. This is shown in the results as most of the days are in the lower probability bins, with only one day at a day 8 lead-time that had a maximum probability of over 30%.

4. Discussion

It can be seen between Figures 1 and 4 that the day 5 GEFS-MLP forecasts are very important in forecasting high-impact severe weather events. There are more high and moderate-risk days in the 45-60% bin in Figure 4 compared to the 60-99% bin, which shows that days with probabilistic forecasts between 45-60% can predict the severity of an event better, which aligns with the positive skill seen in the model at these lead times (Hill et al. 2023). There is more value added to these forecasts as forecasters are able to see a day 5 60% probability of severe weather from the GEFS-MLP and determine that a high-impact severe event is more likely based on past events.

Based on Figure 1, days 7 and 8 have the ability to indicate a higher impact severe weather event at lower probability thresholds compared to days 4 and 5. This result means that forecasts for days 7 and 8 are more valuable at lower probabilistic thresholds than days 4 and 5 because they could be signaling a higher impact severe weather event at longer lead times. At days 4 and 5, probabilistic thresholds gain more value as they increase after the 45% threshold, as the average number of severe weather reports increases. This is in part due to the increase in certainty as the lead-time shortens, but understanding severe impacts from ML probabilistic forecasts can create longer lead-times for severe weather events across CONUS.

The increase in days with higher probabilities as lead times decrease can be seen in Figures 3 and 4 as the days 4 and 5 GEFS-MLP probabilistic forecasts show that more intense SPC categorical forecasts from the Day 1 Convective Outlook occur in the 45-60% and 60-99% bins. Operational forecasters can use this to increase lead time for high-impact severe weather forecasts by looking at how the probabilities evolve during the medium-range. These forecasts are important, especially day 5 forecasts, in predicting high-impact severe weather events, as the average number of severe storm reports increases as probabilities increase during these lead times.

Day 6 shows the least impressive data out of all the lead times from the GEFS-MLP data as the day 7 data always has a higher average number of severe storm reports in accordance to the probabilistic threshold (Figure 1). Day 5 even surpasses the average number of severe storm reports compared to day 6 GEFS-MLP data in the 20% threshold (Figure 1). This could mean that day 6 has the least amount of value compared to the other medium-range forecasts from the GEFS-MLP data.

Day 7 shows interesting results as it looks like there is a higher percentage of moderate risks in the highest probability bin (45-60%), but there is only one day that had a maximum probability at day 7 that fit into that bin (Figure 6). Looking at the second largest probability bin at day 7, there is a slightly higher percentage of enhanced, moderate, and high risks. This could indicate that day 7 does slightly better than day 6; however, day 8 does not show the same results.

Day 8 shows no probability bin that illustrates that day 8 does well predicting a high-impact event

as there is no signal that is able to be seen. Day 8 is similar to the results at day 6 where there is little indication that a high-impact event will occur at these lead times. There is no probability bin that has a high percentage of enhanced, moderate, or high risks for days 6 and 8. An operational forecaster can use these results to know which probabilities at different lead times are more likely to signal a high-impact severe weather event.

There are important characteristics within the data that need to be mentioned to look at different biases or forms of disadvantages that this data presents. Firstly, population bias is going to influence the storm report data as densely populated areas are going to have more confirmed severe storm reports than rural areas. Secondly, even though this research uses four and a half years of data, that is not a lot of data to really investigate the distribution of SPC categorical forecasts in relation to GEFS-MLP probabilistic forecasts. A larger dataset would be more beneficial in the future to help see how ML forecasts can aid operational severe weather forecasting.

5. Summary and Conclusion

This study looked at how medium-range (day 4-8) ML probabilistic forecasts from the GEFS-MLP compared to short-range (day 1) operational forecasts. This study showed what forecasts to pay attention to when predicting a high-impact severe weather event. The GEFS-MLP forecasts for each lead-time day were compared through their verification of the average number of storm reports to show which forecasts can predict high-impact severe weather events. These results were then compared through map analysis of the medium-range forecasts to the short-range forecasts of the SPC Convective Outlook. The SPC Day 1 Convective Outlook for every day between October 2020 and April 2025 was compared graphically to the highest probabilistic threshold from the GEFS-MLP for that day to show the distribution of the highest SPC categorical forecast for each probabilistic threshold.

Days 4 and 5 showed the most value when it came to forecasting high-impact severe weather events (Figure 1). They both showed a higher average number of severe weather reports within the probabilistic threshold of 60% (Figure 1). Figure 4 shows how the 45-60% maximum probability threshold has the most intense categorical forecasts from the SPC Day 1 Convective Outlook for Day 5. Forecasters can use this information to help increase lead time for a high-impact severe weather event in CONUS. Day 6 has the least amount of value when it comes to predicting a high-impact severe weather event, although it can predict severe weather at its appropriate lead time (Figure 5).

Days 7 and 8 had similar results as they had most days with a maximum probability of severe weather less than 30% (Figures 6 and 7). Day 7 showed interesting results as there was a higher percentage of enhanced, moderate, and high risks in the higher probability bins, but day 8 did not show the same results. Day 7 gives a good indication of a high-impact severe weather event based on SPC categorical forecasts, even with the small samples in the highest probability bins. Day 8 has a lot more variability in all probability bins, indicating that there is a lot of uncertainty in the prediction of a high-impact event for a lead time of 8 days.

Forecasters, such as the SPC, are able to look at the value of the day 4 and 5 forecasts to determine whether to increase lead times for what they think would be a high-impact severe weather event. Further research is needed to examine exactly how the storm reports correlate within the medium-range probabilistic thresholds for each lead time. For example, examining the average number of storm reports for each category in each bin would be more concise in determining the importance of each bin for each lead time day.

This study gives a new perspective to operational forecasters so that they may be able to look at AI/ML probabilistic forecasts, like the GEFS-MLP. This study gave value to specific probabilistic thresholds within the GEFS-MLP medium-range forecasts so that operational forecasters can increase lead time for high-impact severe weather events. When operational forecasters are able to use AI/ML in an efficient manner to help effectively predict severe weather, lives and property can be saved.

6. Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. AGS-2050267. Thank you to the University of Oklahoma and the Research Experience for Undergraduates at the National Weather Center. Gratitude is expressed towards Evan Sudler and the CHAOS group for their contributions to this work.

7. References

Cains, M. G., C. D. Wirz, J. L. Demuth, A. Bostrom, A. McGovern, D. J. Gagne, R. Sobash, and D. Madlambayan, 2024: Exploring NWS forecasters' assessment of AI guidance trustworthiness. Wea. Forecasting, 39, 1219–1240, https://doi.org/10.1175/WAF-D-23-0180.1.

Hill, A. J., R. S. Schumacher, and I. L. Jirak, 2023: A new paradigm for medium-range severe weather forecasts: Probabilistic random forest–based predictions. Wea. Forecasting, 38, 251–272, https://doi.org/10.1175/WAF-D-22-0143.1.

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.

Hua, Z., R. A. Sobash, D. J. Gagne II, Y. Sha, and A. Anderson-Frey, 2025: Improving Medium Range Severe Weather Prediction Through Transformer Post-Processing of AI Weather Forecasts. 1-9, https://arxiv.org/abs/2505.11750.

Loken, E. D., A. J. Clark, and C. D. Karstens, 2020: Generating probabilistic next-day severe weather forecasts from convection-allowing ensembles using random forests. Wea. Forecasting, 35, 1605–1631, https://doi.org/10.1175/WAFD-19-0258.1.

McGovern, A., R. J. Chase, M. Flora, D. J. Gagne II, R. Lagerquist, C. K. Potvin, N. Snook, and E. Loken, 2023: A Review of Machine Learning for Convective Weather. Artif. Intell. Earth Syst., 2, 1-24, https://doi.org/10.1175/AIES-D-22-0077.1.

Sobash, R. A., G. S. Romine, and C. S. Schwartz, 2020: A comparison of neural-network and surrogate-severe probabilistic convective hazard guidance derived from a convection allowing model. Wea. Forecasting, 35, 1981–2000, https://doi.org/10.1175/WAF-D-20-0036.1.

Wirz, C. D., J. L. Demuth, M. G. Cains, M. White, J. Radford, and A. Bostrom, 2024: National Weather Service (NWS) Forecasters' Perceptions of AI/ML and Its Use in Operational Forecasting, Bulletin American Meteorological Society, E2194-E2215, https://doi.org/10.1175/BAMS-D-24-0044.1.