Temporally Weighting Machine Learning Models for High-Impact Severe Hail Prediction

Amanda Burke
School of Meteorology
University of Oklahoma
Norman, OK 73072
aburke1@ou.edu

Amy McGovern
School of Computer Science
University of Oklahoma
Norman, OK 73072
amcgovern@ou.edu

David John Gagne II
National Center for Atmospheric Research
Boulder, CO
dgagne@ucar.edu

Nathan Snook
Center for Analysis and Prediction of Storms
University of Oklahoma
Norman, OK 73072
nsnook@ou.edu

Abstract

We explore a new method to improve machine-learning (ML) based severe hail predictions. A temporal weighting scheme allows the random forest models to increase importance of relevant feature data while maintaining general information about the problem domain from other feature data. We show that the weighting scheme improves forecast skill and forecaster rust. With a flexible design, this method can produce localized forecasts under multiple different scenarios without increasing computational expense.

1 Introduction

Damage from hail events between 2015-17 were three times more expensive than all other convective hazards combined (Gallo et al., 2019). Accurate forecasts are needed to help mitigate the property and subsequent economic losses associated with hailstorms. To aide forecasters in creating daily severe hail forecasts, Gagne et al. (2017) and Burke et al. (2020) trained random forest (RF, Breiman, 2001) models on numerical weather prediction output to produce probabilistic next-day severe hail guidance. Although skillful at predicting hail over the contiguous United States, neither machine learning (ML) method explicitly represents the spatio-temporal development of severe thunderstorms, a pivotal part of severe weather prediction in local environments (Smith et al., 2012; Allen et al., 2020).

While forecast skill is important, ML models that adhere to forecaster's physical understanding of severe weather formation is essential for providing trustworthy guidance. Other methods explicitly constrain ML models under physical assumptions (e.g., Beucler et al., 2019a,b; Boukabara et al., 2019; Karpatne et al., 2017) to increase meteorological understanding, but this can be computationally expensive and time consuming to tune. Rather than changing the ML models, selecting physically relevant training data increases forecast interpretability without increasing model complexity or computational load on model developers. However, arbitrarily choosing relevant data adds unnecessary bias and may not translate well to different forecast scenarios. Instead, we use an autocorrelation (Huitema and Laraway, 2006) method to define a flexible function that can vary with new data and changes in model persistence. In this paper we demonstrate that including statistically chosen weights improves an already skillful object-based RF framework for predicting severe hail. We compare the weighted model to other hail prediction methods.

AI for Earth Sciences Workshop at NeurIPS 2020.

2 Data Pre-processing

Data from the High-Resolution Ensemble Forecast System version 2 (HREFv2, e.g., Jirak et al., 2018; Burke et al., 2020; Loken et al., 2017), a numerical weather prediction ensemble used by Storm Prediction Center (SPC) forecasters, are input to the ML models as predictors. The observational dataset that relates the input features to severe hail formation is a radar derived product, the Maximum Estimated Size of Hail (MESH, Witt et al., 1998; Zhang et al., 2011; Smith et al., 2016). The HREFv2 and MESH training data encompass 1 April to 31 July 2017 and 1 May to 31 August 2018, the calibration data spans 1 May to 31 August 2019, and finally the testing period comprises data from 1 May to 31 August 2020. The calibration method applied to the RF model output can be found in Burke et al. (2020). As adding weights does not impact the data pre-processing method both the unweighted and weighted ML models are trained, calibrated, and tested over the same data. Both hourly datasets are initially gridded, however we extract storm objects to account for the relative rarity of severe hail across the contiguous United States. Appendix A provides details on the storm-object identification and storm tracking methods.

3 Adding Temporal Weights to Random Forests

Three RF models and an isotonic regression model together form the ML framework for severe hail prediction. First, storm track values and their hail-producing class label are input to a RF classification model, one model for each ensemble member. If a storm track is predicted to produce hail the data are input to two RF regression models, one that predicts the scale parameter of a gamma distribution and the other the shape parameter. Gridded hail tracks are produced by converting the storm track values to percentiles, based on training set values, and extracting the associated percentiles from the predicted hail size gamma distribution.

Grid points with hail sizes exceeding 25 mm (severe hail threshold) are retained from each member and averaged across the ensemble to create neighborhood maximum ensemble probabilities of hail. Finally, the unweighted hail probabilities are calibrated using an isotonic regression model, trained on the calibration dataset, to adjust the non-zero probabilities for more reliable forecasts.

Differing from the original ML framework, the RF inputs are weighted to produce a localized prediction. As neither the RF hyperparmeters nor the input data are changed, this method does not substantially increase developer load. Weights can be based on timing information, spatial information, a combination of the two, or any other priority type. Only temporal weighting will be presented for this study. While this approach is theoretically simple, determining what data should be prioritized can be a challenge.

We apply an exponentially decreasing function (equation 1) to the input storm tracks, where data within the desired time frame receive full weight while data outside the period receive substantially decreasing weight with time.

$$\text{Weights} = 5e^{\alpha(\text{time difference})}, \alpha = \frac{ln\frac{1}{5}}{\text{Days when Auto-correlation Function below 0}}$$

The number of lags needed for the autocorrelation value of the storm-object tracking variable to drop below 0 indicates at how many days outside a time period the input data receives a weight of 1, or are less likely to be chosen by the RFs as important during training. The weights are multiplied by 5 to heavily prioritize relevant data, which also means the alpha/threshold value must go to 0 when the weights equal the natural log of one-fifth.

Other than this weighting factor, the ML method from Burke et al. (2020) remains the same and will be compared to the weighted training models. Visually, Figure 1 demonstrates how adjusting the alpha parameter creates functions allowing for more data persistence (higher alpha) versus less persistent data (lower alpha). For an alpha of -0.1 data are irrelevant outside 15 days, while an alpha of -0.32 reduces relevance to 5 days. Further explanation of how the weights are implemented in a decision tree and a visual representation of the weighting method applied to the training data is provided in Appendix B.

4 Results

The temporally weighted and unweighted probability forecasts are compared at different time scales to determine the impacts of incorporating weights. Additional hail prediction methods are verified against the ML forecasts, including the SPC day 1 outlook at 1200 UTC and 2-5 km updraft helicity values >75 m^2s^2, which have been related to severe hail prediction (e.g., Sobash et al., 2016; Gagne et al., 2017). A single subjective case study provides a visual analysis of the hail forecasts while objective verification over the testing dataset (1 May to 31 August 2020) describes forecast performance over the spring/summer seasons.

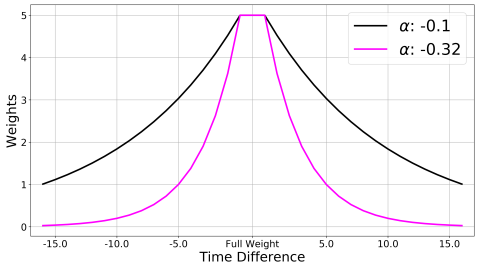

Figure 1: Weighting function at different alpha values. The time difference is in days days from a given period, such as days outside a month

4.1 Subjective Verification

Very large hail impacted parts of the southern plains and Mid-Mississippi Valley on 4 May, 2020. Concentrated regions of observed hail are reported in Kansas, Oklahoma, Arkansas, and Missouri (Fig. 2). Strong boundary-layer heating, dry-line and surface warm front convergence, and intense upper-level shear caused severe convection to initiate in southeastern OK and extend downstream of a 500 hPa speed maximum. The resulting supercells from the initiated convection led to large hail formation in the aforementioned areas, as well as wind damage and a couple of tornadoes.

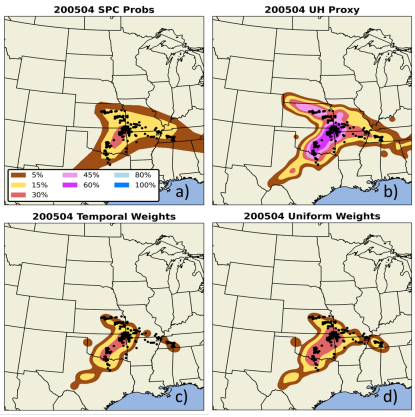

Figure 2: Severe hail probability forecasts on 4 May, 2020 from the (a) SPC day 1 outlook at 1200 UTC, (b) updraft helicity proxy, (c) temporally weighted RF model, and (d) unweighted RF model. Black dots are observed severe hail reports.

The day 1 SPC outlook at 1200 UTC includes severe hail probabilities up to 30 % to account for the likelihood of supercell development and very large hail (Fig. 2a). Similar to the SPC outlook, the updraft helicity proxy (Fig. 2b) also predicted a swath of higher probabilities from northern Texas up to central Missouri, but extended probabilities zonally from western Kansas to eastern Arkansas. The updraft helicity proxy also increased the probability magnitudes for this event, with values exceeding 60 % where the greatest concentration of hail reports are located.

The two ML forecasts are similar to the SPC day 1 outlook probability magnitudes and coverage compared to the updraft helicity proxy. Yet even without the added weights, the ML forecasts focus more on the regions of greatest observed hail than the SPC outlook. Adding temporal weights (Fig. 2c) reduces the extent of 30 % probabilities as well as non-zero probabilities to produce a forecast with smaller false alarm area at the expense of missing a few reports the unweighted ML forecast catches (Fig. 2d). Even with the missed reports, the weighted forecast captures the concentration of observed hail across central Oklahoma towards Missouri. Overall, the ML predictions capture the regions of highest hail-likelihood with greater precision compared to the updraft helicity proxy and SPC outlook. Forecasters from the 2020 Hazardous Weather Testbed Spring Forecasting Experiment commented that the 4 May weighted forecast "performed particularly well" (Adam Clark, Personal Communication May 2020) indicating greater trust in the output.

4.2 Objective Verification

A reliability and performance diagram objectively compare the different severe hail predictions for the 2020 spring and summer seasons (Fig. 3). The reliability diagram is accompanied by the Brier Skill Score (Brier, 1950), which ranges from -∞ to 1 where above 0 values are skilful, and a frequency diagram in the top left. The SPC outlook produces the highest Brier Skill Score (0.09) (Fig. 3a) although underforecasts compared to the observed local storm reports. The weighted ML forecast is

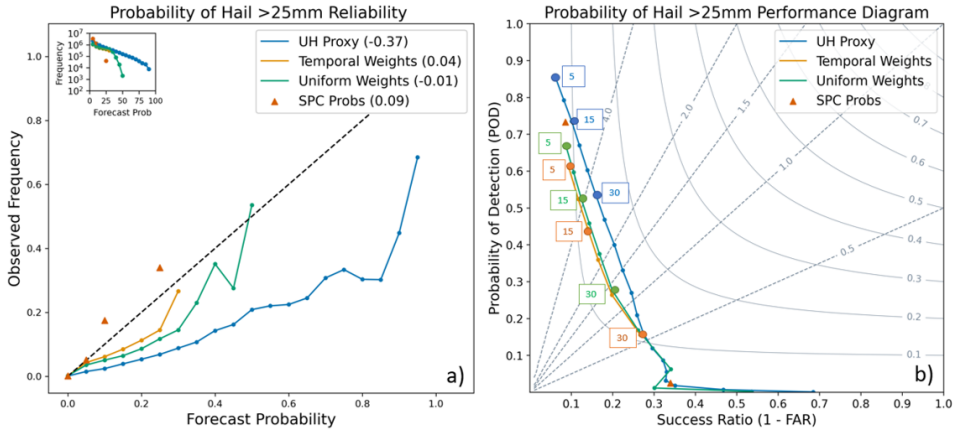

Figure 3: (a) Reliability (forecast frequency in upper left, Brier Skill Score in legend) and (b) performance diagrams verifying severe hail predictions from the SPC day 1 severe outlook at 1200 UTC, updraft helicity proxy, temporally weighted RF model, and unweighted RF model.

second most skillful forecast, with a Brier Skill Score value of 0.04 and more reliable predictions than the unweighted forecasts and updraft helicity proxy, both of which overforecast with lower Brier Skill Score values. With the performance diagram, although the updraft helicity proxy appears to be most skillful overall (Fig. 3b), the individual probabilities show that the ML forecasts have greater critical success index values (CSI, Schaefer, 1990), equivalent to the Intersection over Union score for object segmentation. In addition, the ML forecasts contain lower false alarms (higher success ratio) compared to the updraft helicity proxy at each probability threshold, with the weighted forecast outputting the lowest false alarms and highest CSI values overall.

Combining the subjective and objective results indicates that the ML forecasts are more skillful than the updraft helicity proxy, and show promise in focusing on areas of greater hail threat compared to the SPC outlook. Both the reliability and performance diagrams indicate the weighted ML forecast reduces false alarms compared to the unweighted and updraft helicity proxy predictions at the cost of higher probabilities of detection, similar to Brooks (01 Jun. 2004). Reducing false alarm threat areas, as seen with the weighted ML forecast evaluation, can be important to the *value* (Murphy, 01 Jun. 1993) of a forecast when making decisions (Simmons and Sutter, 2009; Brooks and Correia, 2018). In general, adding temporal weights increases forecast skill in the regions that matter to forecasters.

5 Discussion and Future Work

With the addition of a weighting parameter in a previously skillful ML framework, we have demonstrated that the weighted forecasts better highlight areas of high hail threat without compromising forecast performance. When comparing the ML forecasts, weighted and unweighted, against two baseline verification datasets, both ML predictions reduce false alarm area coverage and increasing performance. The weighted ML forecasts were found to be more reliable and with increased Brier Skill Score values during the 2020 spring to summer, with similar skill and bias based on the probability of detection and success ratio of the ML forecasts.

The increased skill with ML forecasts that are weighted shows the feasibility of adding physical information to the ML framework without high computational expenses (See more in Appendix B). Additionally, the weighting parameter is flexible and can be tailored to different scenarios without the need for new data, making it important for future localized ML model deployment. In the future we plan to test different weighting functions and strategies to identify an optimal configuration.

Acknowledgements

This work was primarily supported by the Joint Technology Transfer Initiative (JTTI) Grant NA16OAR4590239 provided by NOAA. This material is based upon work supported by the National

Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977.

References

Allen, J. T., I. M. Giammanco, M. R. Kumjian, H. Jurgen Punge, Q. Zhang, P. Groenemeijer, M. Kunz, and K. Ortega, 2020: Understanding hail in the earth system. *Reviews of Geophysics*, **58 (1)**, e2019RG000 665, doi:10.1029/2019RG000665.

Beucler, T., M. Pritchard, S. Rasp, J. Ott, P. Baldi, and P. Gentine, 2019a: Enforcing analytic constraints in neural-networks emulating physical systems.

Beucler, T., S. Rasp, M. Pritchard, and P. Gentine, 2019b: Achieving conservation of energy in neural network emulators for climate modeling. 1906.06622.

Boukabara, S.-A., V. Krasnopolsky, J. Q. Stewart, E. S. Maddy, N. Shahroudi, and R. N. Hoffman, 2019: Leveraging modern artificial intelligence for remote sensing and nwp: Benefits and challenges. *Bulletin of the American Meteorological Society*, **100 (12)**, ES473–ES491, doi:10.1175/BAMS-D-18-0324.1.

Breiman, L., 2001: Random forests. *Machine Learning*, **45 (1)**, 5–32, doi:10.1023/A:1010933404324.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**, 1–3, doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Brooks, H. E., 01 Jun. 2004: Tornado-warning performance in the past and future: A perspective from signal detection theory. *Bulletin of the American Meteorological Society*, **85 (6)**, 837 – 844, doi:10.1175/BAMS-85-6-837, URL https://journals.ametsoc.org/view/journals/bams/85/6/bams-85-6-837.xml.

Brooks, H. E., and J. Correia, 2018: Long-term performance metrics for national weather service tornado warnings. *Weather and Forecasting*, **33 (6)**, 1501 – 1511, doi:10.1175/WAF-D-18-0120.1, URL https://journals.ametsoc.org/view/journals/wefo/33/6/waf-d-18-0120_1.xml.

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. *Weather and Forecasting*, **35 (1)**, 149–168, doi:10.1175/WAF-D-19-0105.1.

Cintineo, J. L., T. M. Smith, V. Lakshmanan, H. E. Brooks, and K. L. Ortega, 2012: An Objective High-Resolution Hail Climatology of the Contiguous United States. *Wea. Forecasting*, **27**, 1235–1248, doi:10.1175/WAF-D-11-00151.1.

Gagne, D. J., A. McGovern, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017: Storm-Based Probabilistic Hail Forecasting with Machine Learning Applied to Convection-Allowing Ensembles. *Wea. Forecasting*, **32**, 1819–1840, doi:10.1175/WAF-D-17-0010.1.

Gallo, K., P. Schumacher, J. Boustead, and A. Ferguson, 2019: Validation of Satellite Observations of Storm Damage to Cropland with Digital Photographs. *Weather and Forecasting*, 435–446, doi:10.1175/WAF-D-18-0059.1.

Huitema, B., and S. Laraway, 2006: *Encyclopedia of Measurement and Statistics*, chap. Autocorrelation.

Jirak, I. L., A. J. Clark, B. Roberts, B. T. Gallo, and S. J. Weiss, 2018: Exploring the Optimal Configuration of the High Resolution Ensemble Forecast System. *25th Conf. on Numerical Weather Prediction*, Denver, CO, Amer. Meteor. Soc., [Available online at https://ams.confex.com/ams/29WAF25NWP/webprogram/Paper345640.html.].

Karpatne, A., W. Watkins, J. Read, and V. Kumar, 2017: Physics-guided neural networks (pgnn): An application in lake temperature modeling. 1710.11431.

Loken, E. D., A. J. Clark, M. Xue, and F. Kong, 2017: Comparison of Next-Day Probabilistic Severe Weather Forecasts from Coarse- and Fine-Resolution CAMs and a Convection-Allowing Ensemble. *Wea. Forecasting*, **32**, 1403–1421, doi:10.1175/WAF-D-16-0200.1.

McGovern, A., R. Lagerquist, D. J. Gagne, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019: Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, **100 (11)**, 2175 – 2199, doi:10.1175/BAMS-D-18-0195.1, URL https://journals.ametsoc.org/view/journals/bams/100/11/bams-d-18-0195.1.xml.

Murphy, A. H., 01 Jun. 1993: What is a good forecast? an essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, **8 (2)**, 281 – 293, doi:10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2, URL https://journals.ametsoc.org/view/journals/wefo/8/2/1520-0434_1993_008_0281_wiagfa_2_0_co_2.xml.

Schaefer, J. T., 1990: The Critical Success Index as an Indicator of Warning Skill. *Weather and Forecasting*, **5**, 570–575, doi:10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.CO;2.

Simmons, K. M., and D. Sutter, 2009: False alarms, tornado warnings, and tornado casualties. *Weather, Climate, and Society*, **1 (1)**, 38 – 53, doi:10.1175/2009WCAS1005.1, URL https://journals.ametsoc.org/view/journals/wcas/1/1/2009wcas1005_1.xml.

Smith, B. T., R. L. Thompson, J. S. Grams, C. Broyles, and H. E. Brooks, 2012: Convective modes for significant severe thunderstorms in the contiguous united states. part i: Storm classification and climatology. *Weather and Forecasting*, **27 (5)**, 1114–1135, doi:10.1175/WAF-D-11-00115.1.

Smith, T. M., and Coauthors, 2016: Multi-radar multi-sensor (mrms) severe weather and aviation products: Initial operating capabilities. *Bulletin of the American Meteorological Society*, **97**, 1617–1630, doi:10.1175/BAMS-D-14-00173.1.

Sobash, R. A., C. S. Schwartz, G. S. Romine, K. R. Fossell, and M. L. Weisman, 2016: Severe Weather Prediction Using Storm Surrogates from an Ensemble Forecasting System. *Weather and Forecasting*, **31**, 255–271, doi:10.1175/WAF-D-15-0138.1.

Witt, A., M. D. Eilts, G. J. Stumpf, J. T. Johnson, E. D. W. Mitchell, and K. W. Thomas, 1998: An Enhanced Hail Detection Algorithm for the WSR-88d. *Wea. Forecasting*, **13**, 286–303, doi:10.1175/1520-0434(1998)013<0286:AEHDAF>2.0.CO;2.

Zhang, J., and Coauthors, 2011: National Mosaic and Multi-Sensor QPE (NMQ) System: Description, Results, and Future Plans. *Bull. Amer. Meteor.*, **92**, 1321–1338, doi:10.1175/2011BAMS-D-11-00047.1.

A APPENDIX: DATA PRE-PROCESSING DETAILS

Storm objects are created where gridded data values exceed a user-defined threshold at a single time step, we use the maximum hourly upward vertical velocity (MAXUVV) ≥ 10 ms^{-1}. The exceeding points are grown out using an enhanced watershed algorithm to create storm objects (Gagne et al., 2017). Objects that are close in time and space are connected to create storm tracks. Variables other than MAXUVV are extracted from the grid points within the storm tracks, including geopotential height, U wind, and V wind, MAXUVV, temperature, etc. at various levels throughout the atmosphere (Burke et al., 2020). To reduce the gridded storm track data to singular points at each time step, the predictor values are approximated using the mean, maximum, minimum, skewness, standard deviation, 10th percentile, 50th percentile, and 90th percentile.

The predictor (HREFv2 data) storm tracks in close relative distance to the observation tracks (MESH ≥ 12 mm) are matched with a binary classification. The tracks classified as hail-producing are additionally labeled using the shape and scale values of a gamma distribution that approximates the hail sizes within a matched MESH track. These storm track labels (binary classification, shape and scale parameters) are available hourly across the entire contiguous United States. These three parameters are associated with each storm track example and become the labels for a random forest classification model and two random forest regression models that predict the shape and scale parameter. Gagne et al. (2017) describes the storm-object identification process in more detail with additional changes in Burke et al. (2020).

B APPENDIX: WEIGHTING EXAMPLES

Weights are implemented within the splitting process of each decision tree in the RF through the "minimum impurity decrease" parameter. A node is split if the improvement in impurity equation is below or equal to this parameter which is usually set to 0.

$$\text{Improvement in Impurity} = \frac{N_{parent}}{N_{total}} * (\text{impurity}_{parent} - \frac{N_{right}}{N_{parent}} * \text{impurity}_{right} - \frac{N_{left}}{N_{parent}} * \text{impurity}_{left})$$

The values for $N_{total}, N_{parent}, N_{left}$, and N_{right} are counts of each example. However, when adding weights the different N fields become a weighted sum, placing more importance on the higher weighted samples. For example, let the impurity values be 0.21, 0.5, and 0.10 for the parent, right, and left nodes. If the N field values are $N_{total} : 100, N_{parent} : 50, N_{left} : 37$, and $N_{right} : 13$, and no weights are applied the improvement in impurity is:

$$\text{Improvement in Impurity} = \frac{50}{100}(0.21 - \frac{13}{50} * 0.5 - \frac{37}{50} * 0.10) = 0.0003$$

If weights are included such that the N fields are higher because more relevant datapoints are included, $N_{total} : 250, N_{parent} : 100, N_{left} : 62$, and $N_{right} : 38$, the improvement in impurity is:

$$\text{Improvement in Impurity} = \frac{100}{250}(0.21 - \frac{38}{100} * 0.5 - \frac{62}{100} * 0.10) = -0.0168$$

This value is at or below the minimum impurity decrease of 0, indicating that adding weights greater than 1 increases the likelihood a split will occur if relevant data are present. The weights, however, do not impact the actual number of examples or the impurity score.

We can further visualize how the weights apply to storm objects within the RF model for May, June, July, August. In Figure 4 the larger the dot, the higher the weight. The regions of highest weight for each month compare well to the monthly MESH climatology findings in Cintineo et al. (2012), meaning that the weighting scheme is realistically adding spatio-temporal thunderstorm development to the RFs without explicitly adding physical constraints that add to developer workloads. Further, adding greater physical understanding of the data the RFs are more likely to chose for decision splits increases the transparency of the ML process which is critical in trustworthy ML model deployment within the meteorology community (McGovern et al., 2019).

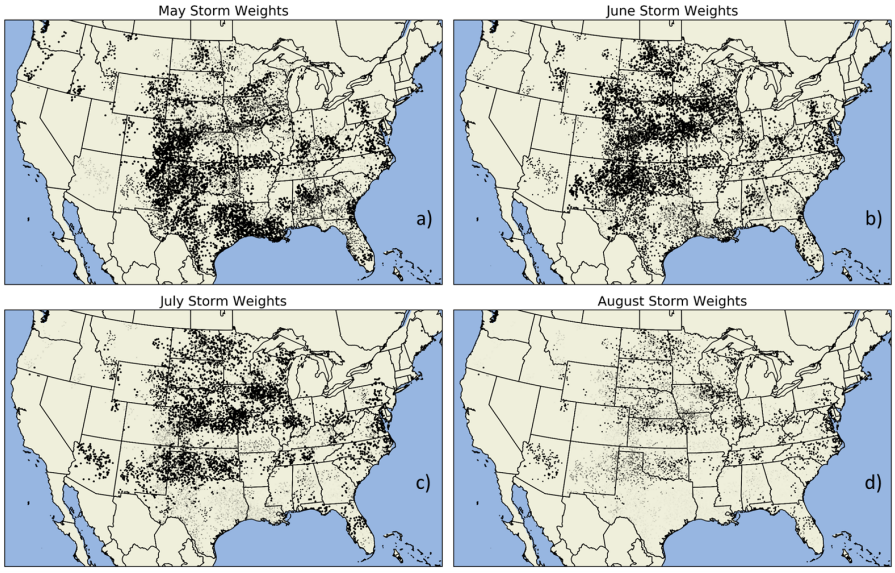

Figure 4: Weights applied to storm objects from (a) May, (b) June, (c) July, and (d) August. The size of the dot represents the weight it would receive in a random forest.