NOAA ProbSsevere v2.0 – ProbHail, ProbWind, and ProbTor

John L. Cintineo[†]
Cooperative Institute of Meteorological Satellite Studies, University of Wisconsin—Madison

Michael J. Pavolonis
NOAA/NESDIS/Center for Satellite Applications and Research/Advanced Satellite Products Team

Justin M. Sieglaff
Cooperative Institute of Meteorological Satellite Studies, University of Wisconsin—Madison

Lee Cronce
Cooperative Institute of Meteorological Satellite Studies, University of Wisconsin—Madison

Jason Brunner
Cooperative Institute of Meteorological Satellite Studies, University of Wisconsin—Madison

[†] Corresponding author email address: john.cintineo@ssec.wisc.edu

Early Online Release: This preliminary version has been accepted for publication in *Weather and Forecasting*, may be fully cited, and has been assigned DOI 10.1175/WAF-D-19-0242.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

ABSTRACT

Severe convective storms are hazardous to both life and property and thus their accurate and timely prediction is imperative. In response to this critical need to help fulfill the mission of the National Oceanic and Atmospheric Administration (NOAA), NOAA and the Cooperative Institute for Meteorological Satellite Studies (CIMSS) at the University of Wisconsin (UW) have developed NOAA ProbSevere – an operational short-term forecasting subsystem within the Multi-Radar Multi-Sensor (MRMS) system, providing storm-based probabilistic guidance to severe convective hazards. ProbSevere extracts and integrates pertinent data from a variety of meteorological sources via multi-platform multi-scale storm identification and tracking in order to compute severe hazard probabilities in a statistical framework, using naïve Bayesian classifiers.

Version 1 of ProbSevere (PSv1) employed one model—the "probability of any severe hazard" trained on the U.S. National Weather Service (NWS) criteria. Version 2 of ProbSevere (PSv2) implements four models, three naïve Bayesian classifiers trained to specific hazards: 1) severe hail, 2) severe straight-line wind gusts, 3) tornadoes; and a combined model for any of the aforementioned hazards, which takes the maximum probability of the three classifiers. This paper overviews the ProbSevere system and details the construction and selection of predictors for the models. An evaluation of the four models demonstrated that v2 is more skillful than v1 for each severe hazard with higher critical success index scores and that the optimal probability threshold varies by region of the U.S. The discussion highlights PSv2 in NOAA's Hazardous Weather Testbed (HWT) and current and future research for convective nowcasting.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

1. Introduction

The U.S. National Weather Service (NWS) issues critical severe weather warnings for the public to take mitigating action from hazards such as large hail, strong straight-line winds, and tornadoes. The volume of meteorological data available to forecasters has exploded in recent years with the advent of datasets such as high-resolution numerical weather prediction (NWP) models (e.g., High Resolution Ensemble Forecast System (HREF); Roberts et al. 2019), next generation Geostationary Observational Environmental Satellites (e.g., GOES-16, GOES-17; Schmit et al. 2015), space-borne lightning mappers (e.g., Geostationary Lightning Mapper [GLM]; Rudlosky et al. 2019, Goodman et al. 2013), terrestrial lightning arrays (e.g., Earth Networks Inc. [ENI] Total Lightning Network [ENTLN], Vaisala Global Lightning Dataset GLD360), and Multi-Radar Multi-Sensor Doppler weather radar products (MRMS; Smith et al. 2016). These developments present new capabilities and challenges for severe storm forecasting and warning operations. On one hand, better observing capabilities (spatially, temporally, and increased information-content) should help forecasters better understand thunderstorm processes and severe potential; on the other hand, it is increasingly difficult for forecasters to integrate the pertinent aspects of all of these novel and more frequent observations to capitalize on their advantages and quickly identify threats and issue warnings.

In response to this opportunity and dilemma, the National Oceanic and Atmospheric Administration (NOAA) and Cooperative Institute of Meteorological Satellite Studies (CIMSS) at the University of Wisconsin (UW) developed a system called the NOAA/CIMSS ProbSevere model, building on years of applied and basic research in the fields of satellite, radar, lightning, and NWP meteorology, as well as image science. This model, ProbSevere version 1 (PSv1), fused together a variety of datasets to predict that any given thunderstorm in the contiguous United States

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

(CONUS) would produce any type of severe weather in the near-term (0-60 min). Cintineo et al. (2014; C14 hereafter) and Cintineo et al. (2018; C18 hereafter) describe the methodology and performance of PSv1. PSv1 has been used by the NWS experimentally since at least 2016, with many forecasters using it to increase confidence in their warnings and lead time to severe weather hazards (C18).

Through experiments at the Hazardous Weather Testbed (HWT) and the NOAA Operations Proving Ground (OPG) conducted between 2013 and 2016 (see C18), forecasters expressed a desire for different statistical models for each NWS-defined severe weather hazard, as an enhancement to the "any severe" criterion of PSv1. The three NWS-defined types of severe hazards are: 1) large hail (diameter $\geq$ 1 in), 2) straight-line convective winds (gust $\geq$ 58 mph), and 3) tornadoes. Thus, ProbSevere version 2 (PSv2) was subsequently developed with three models specific to the aforementioned hazards (ProbHail, ProbWind, and ProbTor), as well as a maximum hazard probability model (probSevere), which takes the maximum probability from ProbHail, ProbWind, and ProbTor. PSv2 was evaluated at the HWT in 2017-2019 and is expected to become operational within MRMS in 2020. This paper describes the PSv2 methodology, reports on verification studies, and discusses the potential for further enhancements.

2. ProbSevere system overview

The ProbSevere system has several aspects: 1) observation processing, 2) storm identification and tracking, and 3) application of machine learning.

a) Observation processing

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

The ProbSevere system directly utilizes data from four sources: 1) geostationary satellites, 2) NEXRAD, 3) ground-based lightning, and 4) NWP. The processing of each data stream is run in parallel, enabling more efficient updates to ProbSevere output products. Each data source is described in further detail below.

i) GOES-East Advanced Baseline Imager (ABI)

ProbSevere currently utilizes a single source of satellite data. The GOES-16 (currently also known as GOES-East) satellite, which is located at 75^{o} W, is used because it covers the entire CONUS at 5-min resolution with little degradation in satellite pixel area except for the far western U.S. (i.e., California, Oregon, Washington), where the average pixel area ranges from 16-24 km^2. As resources allow, GOES-17 (currently also designated as GOES-West) can be easily incorporated in ProbSevere, but for the time being, the increase in processing expense outweighs expected gains in using GOES-West ABI CONUS scans. GOES-East CONUS observations, taken by the Advanced Baseline Imager (ABI), occur every 5 min. As in PSv1 (C14), the GOES observations are processed using radiative transfer software to create an infrared "window" based top-of-troposphere emissivity (ε_{tot}) using ABI band 14 (11.2 μm). This field is the emissivity a cloud would have if it were at the tropopause (Pavolonis 2010a) and helps account for seasonal and latitudinal variations in storm top height. The ε_{tot} is subsequently remapped to a cylindrical equidistant projection that covers the CONUS region at 0.02^{o} x 0.02^{o} resolution (approximately 2 km x 2 km in the midlatitudes). The time rate of change of the maximum ε_{tot} within satellite-identified storms, referred to as the "normalized satellite growth rate", is one of the predictors in ProbHail and ProbWind. The $\Delta\varepsilon_{tot}$ is analogous to decreases in the minimum observed 11.2-μm brightness temperature in the cloud object and has been shown to help discern between severe and non-severe convection during the evolution of cumulus to cumulonimbus (Cintineo et al. 2013).

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

ii) MRMS

Several MRMS products are used in the ProbSevere models. The MRMS products are processed and quality-controlled according to Smith et al. (2016) and include the maximum expected size of hail (MESH; Witt et al. 1998), vertically integrated liquid density (VIL Density), merged composite reflectivity, 0-2 km AGL azimuthal shear (LLAzShear), and 3-6 km AGL azimuthal shear (MLAzShear). The composite reflectivity, VIL Density, and MESH arrive natively in a 0.01^o x 0.01^o cylindrical equidistant projection and are simply cropped to the ProbSevere domain, whereas LLAzShear and MLAzShear arrive in a 0.005^o x 0.005^o cylindrical equidistant projection and are remapped to the 0.01^o x 0.01^o projection and cropped to align with the reflectivity-based fields.

iii) Earth Networks Total Lightning Network

Analogous to PSv1 (C18), ProbSevere ingests observations of total lightning, from the ground based ENTLN, every minute. Each 1-minute ENTLN file contains a list of flashes recorded in the previous minute along with several attributes such as location, height, amplitude, polarity, and the type of flash (intra/inter cloud [IC] or cloud-to-ground [CG]). These binary files are ingested and processed with a Warning Decision Support System – Integrated Information (WDSS-II; Lakshmanan et al. 2007) algorithm (w2ltg), which grids the flashes into a lightning density field with 2-min temporal resolution and 0.01^o x 0.01^o spatial resolution on the same domain as the processed satellite and MRMS gridded products. This total lightning flash density field (LtgFD) has units of flashes min^{-1} km^{-2}.

iv) Rapid Refresh Model (RAP) NWP Output

Rapid Refresh Model (RAP) output from each hourly forecast cycle is obtained from the National Centers for Environmental Prediction. Analysis data and forecast data (1-hour, 2-hour,

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

and 3-hour forecasts) are utilized. See Table 1 for a complete list of RAP based parameters that were considered for inclusion into PSv2. The fields used in predictors in PSv2 are: MUCAPE, MLCAPE (0-90mb AGL), MLCIN (0-90mb AGL), 0-1 km AGL storm-relative helicity (SRH01), 1-3 km AGL mean wind (MW 1-3km), effective bulk shear (EBShear), lowest wet bulb 0°C height (wet bulb zero), CAPE in the -10°C to -30°C layer (generally referred to as "hailCAPE"), and precipitable water (PWAT). These NWP fields are remapped to a 0.04° x 0.04° cylindrical equidistant grid on the ProbSevere CONUS domain. ProbSevere uses a temporal compositing and spatial smoothing technique over 5 forecast times (hours), described in C14, which helps mitigate timing and placement errors in the NWP data. Since RAP data have a latency of about 1 hour, the forecast hours used for this operation are the current analysis (t_0), the previous hour analysis (t_{-1}), and the 1-, 2-, and 3-h forecasts (t_{F1}, t_{F2}, and t_{F3}, respectively). Thus, it is "centered" on t_{F1}, which is approximately the current time when the t_0 data become available. For most fields, the temporal compositing is a maximum operation of the 5 hours, but for wet bulb zero and MLCIN, it is a minimum operation. The MLCIN and SRH01 use a 3-hour temporal compositing (still centered on t_{F1}). The spatial smoothing uses a 5 × 5 grid point (~67.5 km × 67.5 km) Gaussian filter, with a smoothing radius equal to three standard deviations. The RAP data are essentially used as a near-storm environment information. Although mature storms often modify their own environments locally, which may not be represented by the RAP, bulk NWP parameters have been shown to provide reliable guidance on the character of the convective environment in general.

b) Storm identification and tracking

C14 and C18 summarize the tracking procedures in ProbSevere, while Sieglaff et al. (2013) details the satellite tracking methodology. Each model in PSv2 uses identical radar and satellite

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

objects, identified and tracked using the WDSS-II algorithm w2segmotionll (Lakshmanan et al. 2003). Some of the w2segmotionll configuration options have been modified for radar object tracking, so we will summarize those modifications here. The WDSS-II algorithm uses an enhanced watershed algorithm to create radar objects. In ProbSevere, the algorithm searches for local maxima of 40 dBZ ≤ composite reflectivity ≤ 57 dBZ. It should be noted that reflectivity < 40 dBZ is not included in storm objects whereas reflectivity > 57 dBZ is considered to be 57 dBZ. Reflectivity maxima are searched at every 1-dBZ threshold, with the algorithm spatially growing objects in increments of 5 dBZ until a size, or "saliency", of at least 40 pixels is reached (approximately 40 km^2). For example, if a local maximum of 47 dBZ is identified, the algorithm will search for pixels spatially connected to the maximum pixel greater than or equal to 42 dBZ. If this yields an object of at least 40 pixels, the object will stop growing. A second, larger spatial scale is also produced by the enhanced watershed algorithm at a saliency of 200 pixels, using the same object growing criteria as above. The scale_0 (40-pixel saliency) objects are grown to the scale_1 footprints (200-pixel saliency) if the "parent" scale_1 objects only contain one "child" scale_0 object. The scale_0 objects without a scale_1 parent ("orphans") or scale_0 objects with the same scale_1 parent ("siblings") are not modified when merging radar objects. The purpose of this post-processing step of spatially growing certain small objects is to better capture observations related to processes that may be outside the core of a storm (e.g., total lightning flashes, tornadoes). The full w2segmotionll configuration options can be found in Appendix A.

c) Predictor extraction and probability computation

From within the bounds of merged satellite and radar objects, attributes are extracted from the remapped satellite, MRMS, lightning, and NWP fields. The $\Delta\varepsilon_{tot}$ is computed for satellite objects

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

when possible and shared with overlapping radar objects after a parallax correction (a constant cloud height of 9 km is assumed to perform the correction). MRMS, lightning, and NWP attributes are extracted from the radar object footprint (e.g., the spatial maximum, median, or other percentile values). Model predictors are then computed from the extracted observations and the probabilities are computed using a naïve Bayesian classifier (Zhang 2006; Domingos and Pazzani, 1997). ProbHail, ProbWind, and ProbTor are each binary classifiers. Their classes are 'yes' (C_{yes}) or 'no' (C_{no}) for whether the given hazard will occur for a given storm within the next 60 min. Using Bayes' theorem, the probability of a storm producing a targeted hazard, given a set of observed predictors **F**, is defined by

$$P(C_{\text{yes}}|\mathbf{F}) = \frac{P(C_{\text{yes}})P(\mathbf{F}|C_{\text{yes}})}{P(\mathbf{F})}. \quad (1)$$

$P(C_{\text{yes}})$ is the sample frequency of the hazard occurring (the *a priori*). Naturally, $P(C_{\text{no}}|\mathbf{F}) = 1-P(C_{\text{yes}}|\mathbf{F})$. The "naïve" assumption of predictor independence allows for simplification of Bayes' theorem by reduction of dimensionality. Thus, Eqn. (1) can be rewritten as

$$P(C_{\text{yes}}|\mathbf{F}) = \frac{P(C_{\text{yes}})\prod_{i=1}^{N} P(F_i|C_{\text{yes}})}{P(\mathbf{F})}, \quad (2)$$

with F_i denoting the value of the i^{th} predictor, and N the number of predictors. The denominator can be rewritten as

$$P(\mathbf{F}) = P(C_{\text{yes}})\prod_{i=1}^{N} P(F_i|C_{\text{yes}}) + P(C_{\text{no}})\prod_{i=1}^{N} P(F_i|C_{\text{no}}). \quad (3)$$

The Π is the product operator, multiplying the probability of the i^{th} predictor conditional on the storm being a member of C_{yes}. Thus, only the *a priori* and conditional probability distribution for each predictor is needed to compute the final probability conditional on the observed predictor set, **F**. Appendix B shows the evaluation of one of the naïve Bayesian models using example data values. See Kossin and Sitkowski (2009) for details on dimensionality reduction of Bayes'

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

theorem. The assumption of predictor independence is considered a "strong" assumption since it diverges from the reality that many meteorological observations of thunderstorms are indeed correlated. In practice, the naïve Bayesian classifier works well even while violating this assumption. However, the performances of the models do degrade when too many highly correlated predictors are used, as will be elaborated in the next section.

The final probSevere model (maximum hazard probability) of PSv2 simply takes the maximum value of ProbHail, ProbWind, and ProbTor. This was found to have the best skill measured to reports of any hazard type, as opposed to more complex methods that attempt to take into account the dependent nature of hazards (e.g., a joint 3D lookup table of the three naïve Bayesian classifiers).

3. Predictor selection methodology

The classifiers were trained on 167 days of MRMS, ENTLN, and RAP data from 2015 and 2016, encompassing the months of January through November. For the GOES-16 predictors, temporal maximum values of the satellite trends within a 2.5-hr time window were used, as "lifetime" maximum growth rates have been shown to inform the risk of storm severity in the future, when severe weather is manifested (e.g., Cintineo et al. 2013). Lifetime maximum or minimum values were not used for the MRMS, ENTLN, or RAP predictors, but rather the instantaneous extracted values. One advantage of the naïve Bayesian classifier is that training data need not all come from the same samples, or storms. Thus, the training days for GOES-16 were drawn from 2017 since that is when data became available.

Preliminary local storm reports (LSRs) from NOAA's Storm Prediction Center (SPC) rough log (unfiltered) for each day were matched up to ProbSevere IDs in time and space in the same

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

manner as C18 (i.e., finding the spatially closest centroid of a ProbSevere storm object to the report location within a +/- 2 min window). The entire history of a storm is labeled as "severe" (or hail-producing, wind-producing, tornado-producing) if one or more of the given report types are associated with the storm at any time. Thus, each set of predictors for each time step of the severe-labeled storm (or non-severe-labeled storm) are added to the appropriate training dataset.

A heuristic approach was taken in determining what predictors to investigate for this training dataset, taking into account a literature review of severe storm forecasting, data availability, and computation time. Table 1 summarizes the parameters explored. For the MRMS-based parameters, several percentile values were considered as predictors for each field (100% [max], 98%, 95%, 90%, 75%, and 50% [median]). These percentiles are computed from the extracted pixels from each ProbSevere object at each valid time. From the ENI-based predictors, the total lightning flash rate (LtgFR) is a sum of LtgFD from within a storm (rounded to the nearest flash), and the *d/dt*(LtgFR) and lightning jump algorithm anomaly (LJA; Shultz et al. 2010) use operators over time on the LtgFR, or series of LtgFR. The maximum value of LtgFD within a storm was also considered. For the RAP-based predictors, the median values within a storm object extracted from the smoothed fields (see section 2aiv) were considered.

As stated previously, the naïve Bayesian classifier performance degrades if too many correlated predictors are used. This usually results in a model that is sharper (i.e., more very high and very low forecasted probabilities) and less reliable (i.e., poor calibration). In light of this limitation, a few rules of thumb were used when constructing models to test. For each model, attempts were made to incorporate: 1) no more than one reflectivity-based MRMS predictor; 2) no more than one lightning-based predictor; 3) no more than one instability-based NWP predictor; and 4) pair together correlated fields into a two-dimensional (2D) predictor when possible, which

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

helps reduce the negative impact of the correlation on the naïve Bayesian classifier. Both one-dimensional (1D) and 2D predictor distributions were smoothed with kernel density estimation (KDE) using a normal kernel function and optimally chosen bandwidths, following the method of Mielniczuk (1997), whereby the chosen bandwidths operate such that the integrated squared error is minimized. The output of the KDE operation is a 1D (2D) conditional probability vector (matrix).

Models were thus constructed in an ad hoc manner, using the most favorable predictors from the training dataset, which were determined by examining the maximum ratio between $P(F_i \mid C_{yes})$ and $P(F_i \mid C_{no})$ and the difference in means of each class. The models were then systematically evaluated on independent data from 2016 and 2017 (over 200 days), by iterating on previous model designs and tests.

4. ProbHail

The probability of severe hail classifier (ProbHail) was trained using two classes: 1) storms that produced severe hail (diameter $\geq$ 1 in) and 2) storms without any severe reports. The latter class excluded severe wind and tornado producing storms in order to help mitigate potential cross-contamination between classes which could occur due to reporting artifacts (e.g., only the most severe hazard gets reported oftentimes [Morgan and Summers 1982]). ProbHail uses the four predictors summarized in Table 2. These include the 1) max MESH / wet bulb zero (Figure 1, row A), 2) LtgFR / EBShear (Figure 1, row B), 3) hailCAPE / PWAT (Figure 1, row C), and 4) $\Delta\varepsilon_{tot}$ (Figure 2). The *a priori* value is 0.03, which is approximately the number of severe hail-producing storms divided by the total number non-severe thunderstorms in the training dataset. Although the most important field in ProbHail is the max MESH, the wet bulb zero increases probabilities when

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

lower than 3000 m AGL and decreases probabilities when greater than 4000 m AGL, which is consistent with physical expectations. The ProbHail model tends to under forecast low-topped storms that generate severe hail because the MESH is generally smaller ($\leq$ 0.5 in) compared to taller storms (the wet bulb zero predictor does not sufficiently compensate for the reduced MESH). Furthermore, low-topped storms also tend to exhibit low LtgFR, which may indicate a shortcoming in the training dataset whereby more storms are needed to populate this area of phase space. Left splits of supercells may also have under-represented MESH, reducing ProbHail probabilities.

5. ProbWind

The probability of severe wind classifier (ProbWind) was trained using two classes: 1) storms that produced severe convective wind gusts (measured or damage-inferred) and 2) storms without any severe reports. Similar to ProbHail, the latter class excluded severe hail and tornado producing storms in order to limit cross-contamination between classes.

There are several mechanisms for severe wind generation including: perturbation pressure forces, condensate loading, dry air entrainment into downdrafts, and evaporative cooling [Wakimoto 2001]. Thus, two classifiers were created for ProbWind—one for "cellular" windstorms and one for "linear" windstorms. While this may seem like a gross simplification, it aligns well when considering scales of motion in the atmosphere. The cellular model roughly encompasses storms on the meso-gamma scale (2 – 20 km), while the linear model encompasses storms from the meso-beta (20 – 200 km) and lower end of the meso-alpha scale (200 – 500 km). The cellular model is appropriate for wet and dry microbursts generated by cellular convection, while the linear model is appropriate for squall lines, bowing segments, quasi-linear convective systems (QLCS) and other mesoscale convective systems (MCS).

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

For wind events in the training data period (2015), regions of severe wind producing storms in the U.S. were manually determined by looking at the SPC's rough log of severe LSRs (NOAA 2016a), the SPC's archived mesoanalysis grids (NOAA 2016b), and archived NEXRAD reflectivity imagery from NOAA (NOAA 2016c). This was performed in order to better train models for the particular wind type (cellular or linear). The severe reports narrowed regions of interest for each day, while the reflectivity and environmental fields (e.g., EBShear, MUCAPE, 0-3 km lapse rate) helped make a final determination of wind type. Latitude-longitude boxes were drawn around regions of the country for each training day, demarking either cellular or linear wind type. Regions that only contained a single wind report and regions where the distinction between cellular and linear convection was unclear were excluded. In general, storms that were relatively small ($\leq$ 20 km in diameter), and circular were considered cellular, while storms that were relatively large or elongated (> 20 km in one dimension) and had EBShear $\geq$ 20 kts were considered linear. The cellular type could have very high EBShear (e.g., supercell environments) or very low EBShear (e.g., wet microbursts in "pulse" storms). The labeled wind reports were then used as the "yes" class for the cellular and linear naïve Bayesian classifiers. In the future, utilization of a storm-typing algorithm may further improve the classification of wind reports.

a) Cellular wind naïve Bayesian classifier

Numerous tests were conducted focusing on severe wind gusts from cellular storms. The most skillful model that performed best on the wide range of cellular storms (e.g., supercells, pulse storms) was the model from C18 (PSv1). The naïve Bayesian classifier optimized for cellular convection has four predictors, which are listed in Table 3. As in C14, the *a priori* is a function of MUCAPE and EBShear (Figure 3), with a climatological frequency built in.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

The other predictors for this naïve Bayesian classifier are the spatial maximum MESH within an object (Figure 4), the LtgFR / EBShear (Figure 1, row B), and the $\Delta\varepsilon_{tot}$ (Figure 2). This naïve Bayesian classifier has been shown to perform well in a number of environments (see C18), but lightning deficient storms and dry microbursts remain a challenge. Dry microbursts are particularly challenging, as severe downbursts can occur in low-reflectivity storms that fail to meet the minimum object identification requirements.

b) Linear wind naïve Bayesian classifier

The linear wind classifier utilizes the four predictors shown in Table 3. The four predictors are 1) the *a priori*, which is a function of MLCAPE and MW 1-3km (Figure 5), 2) maximum VIL Density (Figure 6), 3) 98th percentile LLAzShear / MW 1-3km (Figure 7, row A), and 4) 98th percentile MLAzShear / LtgFR (Figure 7, row B). The 98th percentile AzShear fields were used in lieu of the maximum fields due to the noisy nature of the AzShears and maximum operation.

Kuchera and Parker (2006) found that the wind in the highest positively buoyant level in the surface inflow layer discriminated well between non-severe convection and severe wind producing convection. The MW 1-3km field in ProbWind is computed over 1-3 km AGL, which likely contains the top of the inflow layer of many storms. Thus, the MW 1-3km is likely correlated with the surface inflow wind field from Kuchera and Parker (2006). The MRMS AzShear products were found to be robust indicators of short-term severe wind gust potential. However, when the azimuthal gradient is not present in the NEXRAD velocity field, ProbWind may under forecast severe wind. Although the LLAzShear, MLAzShear, and MW 1-3km help improve severe wind prediction relative to PSv1, the forecast skill for lightning deficient linear systems is limited.

c) Final ProbWind

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

To create a final probability value for ProbWind, different thresholds of EBShear and MW 1-3km were evaluated in order to discern when the cellular or linear wind models should be applied. However, no set of values that discriminated well enough for linear and cellular storms was found. This could be because severe wind gusts occur on a continuum of MW 1-3km and EBShear phase space. It is also possible that using spatial metrics such as storm size and aspect ratio (which was done qualitatively to gather training datasets) may better discriminate between linear and cellular storms and thus help determine when to compute the appropriate classifier. Given the challenge of automated discrimination of between wind type, a 2D lookup table was created using the joint distributions of computed cellular and linear models for thunderstorms (Figure 8). This was computed in the same fashion as the *a priori* predictor in the cellular wind model, except it is conditional on the naïve Bayesian classifier output from the cellular and linear models instead of physical quantities. Figure 8 is the final lookup table to compute ProbWind. The joint lookup table of the linear and cellular models produced the best skill when considering all severe wind events in the validation data from 2016 and 2017 (not shown). The increased performance is possibly due to the fact that some storms have both 'cellular' and 'linear' features.

6. ProbTor

The probability of tornado classifier (ProbTor) was trained using two classes: 1) storms that produced a confirmed tornado (EF0+) and 2) storms associated with reports of severe hail and/or severe wind reports, but no tornado reports. Unlike ProbHail and ProbWind, the null class for this model contains severe, but non-tornadic storms (hereafter, 'non-tornadic storms'). This was done in order to better simulate what forecasters must discern when issuing severe weather warnings—

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

367 forecasters often ask the question, "why do some *severe* storms produce tornadoes and others do
368 not".

369 ProbTor utilizes the six predictors summarized in Table 4. These include the 1) MLCAPE /
370 MLCIN (Figure 9), 2) max LLAzShear (Figure 10), 3) 98th percentile LLAzShear / SRH01 (Figure
371 11, row A), 4) 98th percentile MLAzShear / LtgFD (Figure 11, row B), and 5) EBShear / MW 1-
372 3km (Figure 11, row C). The *a priori* of 0.01 is approximately the number of tornadic storms
373 divided by the total number thunderstorms in the training dataset.

374 When initially developed, the ProbTor false alarm ratio was large, particularly in regions of
375 high MLCIN and/or low MLCAPE. It is generally understood that tornadoes occur less frequently
376 when surface-based CAPE is absent or is located above a deep layer of CIN (e.g., Davies 2003).
377 The MLCAPE / MLCIN predictor was constructed in a unique way, compared to the other
378 predictors in PSv2. Cumulative distribution functions (CDFs) were created for the MLCIN and
379 MLCAPE fields for the tornadic class of storms. The *shapes* of the CDFs were shifted to lower
380 ranges of MLCAPE (closer to zero) and MLCIN (more negative) where frequent false alarms were
381 occurring. These shifted CDFs act as physically based "fuzzy" functions (see Figure 9), and agree
382 well with previous sounding-derived research (e.g., Thompson et al. 2003, Rasmussen and
383 Blandchard 1998). The shifts 1) minimize impacts due to uncertainty in NWP-modeled MLCIN
384 and MLCAPE, 2) maintain physical consistency with previous research (i.e., low MLCIN and
385 sufficient MLCAPE are necessary but not sufficient conditions for tornadogenesis), and 3)
386 mitigates impact of relatively small tornado sample size. The minimum factor between the MLCIN
387 and MLCAPE functions is multiplied by the *a priori*. This factor may vary between the
388 heuristically determined floor of 0.1 and 1.0. Thus, the final ProbTor *a priori* can vary between
389 0.001 and 0.01, meaning large MLCIN or small MLCAPE can only reduce the *a priori*. Using the

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

conditional probability distributions for these two fields did not show good discrimination between tornadic and non-tornadic storms but implementing these fuzzy functions reduced the false alarm ratio markedly, improving the overall skill of ProbTor. The MLCIN and MLCAPE CDFs produced from the tornadic class agreed well with previous research relating frequency of tornado to those fields (Brotzge et al. 2013, Davies 2003, Thompson et al. 2003, Rasmussen and Blanchard 1998).

The NEXRAD based MRMS velocity fields, which capture storm rotation, are essential to ProbTor (particularly the LLAzShear) despite increased noise relative to the reflectivity-based fields (e.g., Miller et al. 2013). The increased noise is due to a number of effects, including radar return echoes from non-meteorological targets due to anomalous propagation (e.g., ducting in the atmosphere), wind farms, automobiles, birds, and insects, as well as interference from the sun and microwave frequency towers. From a meteorological perspective, strong turbulence within storms and strong flow that is not aligned with storm motion (e.g., ahead of a squall line) may create erroneous regions of high AzShear despite no organized rotation within the storm. Furthermore, the lowest elevation tilt of 0.5º of NEXRAD radars often overshoots low-level rotation within storms with the current CONUS NEXRAD coverage, missing potential tornadic threats. Nonetheless, the MRMS AzShear is a skillful predictor for classifying between tornadic and non-tornadic storms.

The environmental EBShear and SRH01 have been shown to help discriminate between tornadic and non-tornadic storms (Thompson et al. 2007). Although the MW 1-3 km was evaluated mainly for the sake of being a potential predictor in ProbWind, it also stood out as an excellent NWP-based predictor for tornadoes. The MW 1-3km captures strong low-level jets, which increase the storm-relative inflow and supercell potential (Markowski and Richardson 2010) and is a key factor in the formation of strong tornadoes (Broyles et al. 2018). Another possibility is that MW

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

1-3km is a proxy for strong mid-level storm-relative flow that is important for strengthening the mesocyclone. Regardless, it is a useful field in ProbTor and is coupled with the EBShear, with which it shares only a weak correlation (Pearson correlation = 0.17 for tornadic storms).

7. Validation

a) Method

After the initial training and validation of PSv2 using SPC storm reports from 2015 and 2016/2017, respectively, a final evaluation of PSv2 models was performed and validated with 2018 data from *Storm Data*, the National Centers for Environmental Information (NCEI) publication that aggregates and quality-controls official storm reports from the NWS field offices for many phenomena, including severe hail, severe convective wind gusts, and tornadoes (NOAA 2019). ProbHail was validated with severe hail reports, ProbWind was validated with severe wind reports (indicated by damage or measured gusts), ProbTor was validated with tornado reports, and probSevere was validated with any severe reports. These reports were associated with ProbSevere objects in space and time as explained in section 3 and C18. Using a history of each object, the probability of detection (POD), false alarm ratio (FAR), and critical success index (CSI) for the different models were computed. From the contingency table (see Table 5), we can define the metrics thusly:

$$POD = \frac{A_e}{(A_e+B)} \quad , \tag{4}$$

$$FAR = \frac{C}{(A_w+C)} \quad , \tag{5}$$

$$CSI = ((POD)^{-1} + (1 - FAR)^{-1} - 1)^{-1} \quad , \tag{6}$$

where A_e is the number of warned events (e.g., hail, wind, or tornado reports), A_w is the number of verified warnings, B is the number of unwarned events (misses), and C is the number of unverified

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

436 warnings (false alarms). Although ProbSevere is not generating warnings, mapping the ProbSevere
437 probabilistic output to a yes/no "warning" facilitates the comparison to NWS severe thunderstorm
438 and tornado warnings.

439 In order to align the ProbSevere validation analysis with NWS warning and verification
440 practices, initial "warning" times for ProbSevere objects were artificially created upon attainment
441 of a probability threshold. For hail (ProbHail) and wind (ProbWind) the ProbSevere "warnings"
442 were taken to be valid for 45 min. For ProbTor verification, each "warning" was assigned an
443 expiration time of 30 min. The 45 and 30-minute criteria represent the midpoints of the range in
444 warning lifespan given in the NWS Weather Forecast Office Severe Weather Products
445 Specification document (NWS 2018). After a given ProbSevere "warning" expires, for a given
446 probability threshold, the "warning" can then be reissued if the probability threshold is
447 subsequently met after the initial "issuance" time. Thus, a single storm can generate multiple
448 warnings.

449 *Storm Data* reports and ProbSevere data were obtained from January through December
450 2018 for a total of 227 days (see Table 6 for a list of dates). The ProbSevere data were saved from
451 near real-time processing at UW-CIMSS during 2018. Each date represents the "convective day"
452 defined as starting at 12 UTC of the given date and ending at 11:59 UTC of the following day.
453 This validation dataset resulted in nearly 10,800 severe thunderstorms (3,150 hailstorms, 8,250
454 windstorms, and 840 tornadic storms) and 25,000 severe hail, wind, or tornado reports.

455 To mitigate storm object mergers and splits and better link together broken storm tracks,
456 the python library "besttrack" was employed (Harrison 2018; Lakshmanan et al. 2015). This
457 library utilizes the Theil-Sen estimator, which fits a line to storm centroid points (i.e., a storm
458 track) by choosing the median of the slopes of all lines through pairs of points. This automated

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

process helps mitigate but does not completely eliminate broken tracks. Thus, several longevity thresholds were placed on ProbSevere storm objects in order to ignore segments of storms that change object ID frequently: 15, 30, 45, and 60 min. Applying the different longevity thresholds, as well as a lightning activity threshold of 2 flashes per minute, yielded 190,300, 166,800, 110,900, and 81,300 non-severe thunderstorms in the dataset, respectively. Short-lived severe storms were not screened out in order to ensure that all events (i.e., severe reports) were included in scoring.

b) Results

From the performance diagrams in Figure 12, for any given storm longevity threshold, the maximum hazard probability of PSv2 ("probSevere") improves upon PSv1 at every forecast probability bin (between 10% and 90%), by as much as 0.08 CSI. As a function of increasing storm longevity threshold, the CSI increases for each PSv2 model, albeit by a small amount (range is $\leq$ 0.03 CSI), suggesting that the validation is fairly insensitive to the threshold value. ProbHail is the best performing model, followed by probSevere, ProbWind, and ProbTor. While not explicitly shown, PSv2 improves upon the maximum CSI of PSv1 for each hazard type, by 0.05 CSI for ProbHail, 0.04 CSI for ProbWind, and 0.1 CSI for ProbTor at the 45 min storm longevity threshold (although PSv1 only provided the probability of any severe weather).

In Figure 12, the 45 min storm longevity threshold performance diagram is annotated with a black box to highlight the PSv2 comparison with the NWS, because the 45 min constraint yielded the ratio of severe storms to non-severe storms, 8.9%, which was closest to historical climatology, 10% (NWS 2010, FEMA 2007) . Unofficial NWS validation metrics were obtained from the Iowa State Mesonet for all CONUS NWS offices for the dates in Table 6 (Iowa State University 2019). Comparing the 80% probability threshold for probSevere and NWS severe thunderstorm and tornado warnings (orange triangle), probSevere has slightly lower FAR compared to the NWS

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

(0.47 vs. 0.53), but much lower POD (0.48 vs. 0.76), yielding a CSI of 0.34 compared to 0.41 for the NWS. Comparing the 50% probability threshold for ProbTor and NWS tornado warnings only (upside-down red triangle), ProbTor has a higher FAR (0.80 vs. 0.74) and much lower POD (0.33 vs 0.58), yielding a CSI of 0.14 compared to 0.22 for the NWS. While NWS warnings are more skillful, PSv2 can highlight storms early and increase warning lead time and confidence (see discussion in section 8). Another source of uncertainty in comparing PSv2 with NWS is the fact that PSv1 was used experimentally throughout the NWS in 2018. Thus, some portion of the NWS warnings could be influenced by ProbSevere, since PSv1 and PSv2 use similar approaches, thus making the comparison between PSv2 and NWS not completely independent. One last, and perhaps significant, reason for differences between the skill of PSv2 and NWS is the quality of severe reports, particularly wind reports. The quality of estimated wind reports has been found to be suspect, due to population density, time of day, wind gust estimate inflation, and warning-verification biases (e.g., Edwards et al. 2018, Trapp et al. 2006). Training and evaluating a model based on higher quality measured wind reports may yield a better result.

The reliability diagram for probSevere (maximum hazard probability) shows an over forecasting bias above the 40% forecast probability threshold (Figure 13). Using the manual verification analysis techniques in C18 on storms from 2017 demonstrated generally better forecast calibration (yellow line in Figure 13), which is likely due to more accurate storm track associations using the manual method, rather than the automated method of this paper. Nevertheless, efforts are ongoing to improve the reliability of PSv2, including utilization of more advanced machine learning techniques that are less affected by correlated predictors.

In general, the process of matching up reports to ProbSevere object centroids (see section 3) works well most of the time, but it has the drawback of erroneously assigning reports to spatially

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

closer storms if the actual parent storm centroid was further away, as for example, in a squall line, where the centroid may be displaced a considerable distance from the actual severe weather report. This may be one source of error for ProbSevere (v1 and v2) that can create a double penalty effect of increased false alarms and increased misses, hampering perceived performance. Future work may attempt to assign "valid areas" or "polygons" to ProbSevere "warnings" based on the object shape and motion.

An analysis of the annual cycle shows that the CSI of PSv2 peaks in March/April (Figure 14). As the number of storms increase in May/June/July, the CSI for the overall probability of severe decreases to about 0.35. A further decrease in CSI is observed in the August/September/October timeframe. The decrease in CSI is most likely linked to a decrease in environmental shear due to the subtropical jet stream migrating northward. With less synoptic forcing, "pulse" severe storms and multicell thunderstorms are the predominant storm types in much of the summer months, which are notoriously more difficult to forecast than supercells and strong convective lines (e.g., Miller and Mote 2017) and exhibit less distinct characteristics than non-severe storms with respect to radar, lightning, satellite, and NWP features. In November the CSI increases to 0.3, possibly due to increased shear from stronger upper level jets coupled with sufficient thermodynamic parameters. As the number of storms decreases in December/January/February, the CSI drops to about 0.2. The annual cycle in maximum CSI for each of the three severe hazards follows a very similar pattern as the maximum hazard probability, probSevere. The seasonality of CSI for the combined NWS severe thunderstorm and tornado warnings exhibits a very similar behavior to probSevere, and the seasonality of CSI for NWS tornado warnings is similar to that of ProbTor, with September and November being exceptions.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

527 Notably, the maximum CSI for ProbTor occurs in the transition from winter to spring and autumn
528 to winter when strong synoptic forcing and instability are most likely to be collocated.

529 Figure 15 illustrates the relationship between PSv2 performance and NWS county warning
530 area (CWA). For each ProbSevere storm object, the mean centroid position over its lifetime was
531 used to place it within a CWA. For each CWA, storms were then aggregated using the given CWA
532 and all spatially adjacent CWAs. For example, the Milwaukee, WI CWA aggregation would
533 include the CWAs of Milwaukee WI, La Crosse WI, Green Bay WI, Quad Cities IA/IL, and
534 Chicago IL. The spatial aggregation helps ensure that statistical relationships are more robust.
535 Verification metrics for each geographic region were then calculated as previously described. The
536 bottom right image shows the overall probability of severe threshold that maximizes the CSI. The
537 optimal probability threshold is a strong function of geography, where 30-40% thresholds are
538 common in the Northeast, 40-70% thresholds are typical in the Southeast, and thresholds ranging
539 from 60% to 90% are prevalent west of the Appalachian Mountains to the Great Plains. In the
540 mountainous regions of the western U.S., the CSI is maximized when the probability is around
541 80%. Along the west coast, the optimal probability threshold varies significantly due to limited
542 sampling (fewer observed storms). The largest values of CSI are found in the northern/southern
543 Plains and northeast (0.35 – 0.50), although the probability threshold required for maximizing the
544 CSI varies considerably as noted earlier. The Southeast exhibits lower CSI, approximately 0.25 to
545 0.3, owing to a higher FAR. This could be due to a number of reasons, including tall storms that
546 grow quickly from a satellite perspective, lightning-rich storms, erroneously high values in the
547 MESH algorithm, or an inflated number of wind reports based on minor tree damage and warning-
548 verification bias (Edwards et al. 2018). The Intermountain West and West Coast have notably

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

549 lower CSI likely due to differences in storm type (e.g., relatively more dry microbursts), poorer
550 radar coverage, lower overall population density, and smaller sample size.
551
552 8. Discussion & Conclusion
553 ProbSevere version 2 (PSv2) was developed using an ad hoc ingredients-based approach (e.g.,
554 Doswell et al. 1996) in a statistical framework. Unlike PSv1, PSv2 provides probabilistic guidance
555 for specific hazards in accordance with National Oceanic and Atmospheric Administration
556 (NOAA) severe weather criteria (large hail, strong wind gusts, and tornadoes). When compared to
557 storm reports, the critical success index (CSI) of PSv2 is reasonably comparable to the CSI of
558 official NOAA National Weather Service (NWS) forecasts. As expected, human experts at the
559 NOAA NWS determine which storms warrant severe thunderstorm or tornado warnings with
560 greater accuracy than PSv2. PSv2, however, often highlights storms that go on to produce severe
561 weather at an earlier stage of development, thereby giving forecasters insight that allows for more
562 confident and timely warning decisions.
563 PSv2 was formally evaluated in the NOAA Experimental Warning Program (EWP) at the
564 Hazardous Weather Testbed (HWT) from 2017-2019. The model output was evaluated between 4
565 and 6 weeks each spring during the three years, where 68 forecasters answered four survey
566 questions at the end of each day of operations in the HWT, which yielded 232 forecaster individual
567 sets of daily responses. From the forecaster feedback, ~80-90% of forecaster responses said that
568 ProbSevere increases confidence in their warning decision making, while ~60-70% indicated that
569 ProbSevere increases lead time to severe hazards. For ProbTor, ~40-60% and ~20-50% of
570 forecasters indicated increased confidence and lead time, respectively, depending on the year of
571 the experiment (Figure 16). The percentage of 'yes' responses to the question, "does ProbTor

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

572 increase lead time for your tornado warnings" fluctuated year-to-year more so than responses to
573 other questions. This is possibly due to high variability in the number and intensity of tornadoes
574 across the CONUS during the three years of ProbTor being evaluated in the EWP. For instance,
575 while the number of inflation-adjusted tornadoes (documented by the NOAA Storm Prediction
576 Center [NOAA 2019b]) in May 2017 was approximately average, the tornado count was well
577 below average for May 2018, and well above average for May 2019 (the HWT-EWP operated
578 during most weeks in May). In fact, an unusually active late May 2019 led to an extended tornado
579 outbreak across the CONUS (Gensini et al. 2019). The forecasters also had an opportunity for
580 writing open-ended feedback related to PSv2. Of the daily written responses, 16% specifically
581 highlighted upward or downward "trends" in model probabilities in being a useful aspect of PSv2
582 models. Future work should evaluate how time tendencies of PSv2 model probabilities may inform
583 subsequent occurrence or non-occurrence of severe weather.

584 A primary objective of ProbSevere is to provide information in support of determining which
585 developing storms will become severe in the next 60 minutes or so. The NWS does not
586 systematically collect statistics on lead time relative to the first report of severe weather associated
587 with a given storm, so comparisons of lead time between ProbSevere and official NWS warnings
588 relative to the first report of severe weather, for a given storm, requires time-consuming manual
589 analysis as reported in Cintineo et al. (2018). Cintineo et al. (2018) showed that the ProbSevere
590 Version 1 lead time to the first report of severe weather was 34 minutes, compared to 16 minutes
591 for official NWS warnings. PSv2 provides similar additional lead time relative to the first report
592 of severe weather (30 – 34 min, depending on probability threshold). In practice, it is difficult to
593 quantify the impact of ProbSevere on severe weather warning operations, as ProbSevere is just
594 one information source out of several that influence decision making. However, past research and

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

595 recent forecaster feedback suggest that PSv2 can alert forecasters at least 5 – 10 min earlier than
596 they may otherwise have had with only radar interrogation, in many cases. Nevertheless, more
597 research is needed to quantify how PSv2 affects lead time statistics in the NWS.

598 With vastly increasing quantity and quality of meteorological data and improving computing
599 capacities, there are several areas of research that seem fruitful for making rapid progress in the
600 field of severe storm nowcasting. One such area is that of "deep learning", which, for example,
601 can operate convolutional neural networks (CNN) on sequences of meteorological images of cloud
602 fields or storms in radar, satellite, lightning, or NWP output to help automatically identify salient
603 features conducive to severe hazard development or maintenance. The main advantage of CNNs
604 is that they may learn from images without the need to calculate features beforehand (e.g.,
605 mesocyclone detection, above anvil cirrus plume detection). This is an area of active research in
606 severe storm prediction (e.g., McGovern et al. 2019, Gagne et al. 2019), tropical cyclone intensity
607 estimation (Wimmers et al. 2019), and storm-top satellite feature identification (Bedka et al. 2018,
608 Cintineo et al. 2020). CNNs may also be an excellent method to extract value from the spatial
609 information of lightning that the GLM provides.

610 Although the naïve Bayesian method has been effective despite its simplicity, other machine
611 learning techniques are generally more robust for large quantities of correlated data. Random
612 forests or gradient boosted machines, for example, train an ensemble of models (often decision
613 trees) to make classifications or predictions. Random forests have been employed in a variety of
614 fields with much success and offer a more systematic approach to optimization than the naïve
615 Bayes (e.g., Lagerquist et al 2017, McGovern et al. 2014). Tools such as principal component
616 analysis and multipass permutation tests (e.g., McGovern et al. 2019) help to objectively rank the
617 importance of features. Future work will evaluate how ProbSevere can be improved through

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

utilization of a more sophisticated machine learning model and tuned objectively through multipass permutation tests. A more advanced technique may also alleviate the need for multiple conceptual models, such as is currently the case for the ProbWind component of PSv2.

Several observational data sources might also serve to improve the ProbSevere models. For instance, recent work has highlighted how polarimetric variables can enhance hail size prediction (such as the hail differential reflectivity H_{DR}, Murillo and Homeyer 2019). With respect to ProbTor, recent research has suggested that radar-observed radial divergence and azimuthal shear from the upper levels of storms can be excellent indicators of imminent or ongoing tornadic activity (Sandmæl et al. 2019). Such metrics should be prioritized for future investigation and potential inclusion into ProbSevere and MRMS.

Lastly, bridging the nexus between the "nowcast" period and "forecast" period of severe storm prediction is an area of active research (e.g., Rothfusz et al. 2014, Lawson 2018, Karstens et al., 2018). Roughly speaking, for severe storm prediction, this is the 1-6 hour time frame before initial severe hazards are observed. Blending the physically based NWP models (e.g., convection-allowing models) with empirical models like ProbSevere to provide forecasters with consistent and continuous guidance is a challenge. Identifying situations when (or how much) to "trust" the NWP guidance (or ensemble guidance) and when not is a common task for forecasters. Automated methods to determine the inherent predictability of a given scenario will benefit forecasters directly as well as via improved statistical model guidance that could incorporate this information with streaming meteorological observations to provide more accurate guidance, sooner.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

Acknowledgements

The NOAA grant (number NA15NES4320001) for the Cooperative Institute for Meteorological Satellite Studies (CIMSS) supported this work. The authors acknowledge David Harrison of NOAA/SPC for guidance on running the besttrack code and Jordan Gerth for aid in AWIPSII plugin development. The authors are also grateful for Emma Sinclair, Jennifer Lake, and Benjamin Rodenkirch of the University of Wisconsin – Madison for aid in the manual validation of ProbSevere data from 2016 and 2017. The authors also acknowledge Dr. Christopher Karstens and two anonymous reviewers who helped improve the quality of this manuscript. The *GOES-16* data used in this study can be freely obtained from NOAA's Comprehensive Large Array Data Stewardship System (CLASS; online at https://www.class.ncdc.noaa.gov). The Rapid Refresh NWP data can be freely obtained from NOAA's National Centers for Environmental Information (NCEI; online at https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/rapid-refresh-rap). The MRMS data in this study can be obtained from UW-CIMSS upon request of the corresponding author. The ENI total lightning data can be obtained from UW-CIMSS upon request of the corresponding author, pending confirmation of NOAA partnership and expressed written consent from Earth Networks Inc. The views, opinions, and findings contained in this paper are those of the authors and should not be construed as an official National Oceanic and Atmospheric Administration or U.S. government position, policy, or decision.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

Appendix A

w2segmotionll is a WDSS-II executable used for the identification and tracking of radar and satellite objects in ProbSevere. Here are the configuration options for each use of the executable:

1. Radar identification and tracking
    - trackedProductName (-T): MergedReflectivityQCComposite
    - "min max incr maxdepth" (-d): "40 57 5 -1"
    - prunerSizeParameters (-p): 40,200,300,0:0,0,0
    - smoothing filters (-k): percent:75:4:0:4
    - clusterIDMatchingMethod (-m): MULTISTAGE:2:10:0

2. Satellite identification and tracking
    - trackedProductName (-T): emiss11_tot
    - "min max incr maxdepth" (-d): "40 80 9 -1"
    - prunerSizeParameters (-p): 15,60,120,240:0:0,0,0,0
    - smoothing filters (-k): percent:50:1:0.33:1,percent:100:1:0.33:1,percent:50:1:0.33:1,percent:30:1:0.33:1,threshold:40:80
    - clusterIDMatchingMethod (-m): MULTISTAGE:2:15:0

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

Appendix B

This is an example evaluation of the ProbHail naïve Bayesian classifier using example data and the lookup tables found in Figures 1 and 2.

From the paper text, equation 3 can be substituted into equation 2:

$$P(C_{\text{yes}}|\mathbf{F}) = \frac{P(C_{\text{yes}})\prod_{i=1}^{N} P(F_i|C_{\text{yes}})}{P(C_{\text{yes}})\prod_{i=1}^{N} P(F_i|C_{\text{yes}}) + P(C_{\text{no}})\prod_{i=1}^{N} P(F_i|C_{\text{no}})} \qquad (1)$$

$P(C_{\text{yes}}|\mathbf{F})$ is the final probability of severe hail.

$P(C_{\text{yes}})$ is the *a priori*, which is 0.03 for ProbHail.

$P(C_{\text{no}}) = 1 - P(C_{\text{yes}}) = 0.97$

Let *N* be the number of predictors for ProbHail (4),

- i = 1: Max MESH / wet bulb 0ºC height (see Figure 1, row A)
- i = 2: ENI flash rate / effective bulk shear (see Figure 1, row B)
- i = 3: CAPE -10ºC to -30ºC / precipitable water (see Figure 1, row C)
- i = 4: Normalized satellite growth rate from GOES-16 (see Figure 2)

For an example storm, let

- Max MESH = 1.0 in
- Wet bulb 0ºC height = 3000 m
- ENI flash rate = 40 fl /min
- Effective bulk shear = 40 kts
- CAPE -10ºC to -30ºC = 600 J / kg
- Precipitable water = 1.2 in
- Normalized satellite growth rate = 3 % / min

In the numerator and the first term of the denominator:

$$\prod_{i=1}^{4} P(F_i|C_{\text{yes}}) = P(F_1|C_{\text{yes}}) * P(F_2|C_{\text{yes}}) * P(F_3|C_{\text{yes}}) * P(F_4|C_{\text{yes}})$$

- From Figure 1, cell A2, a combination of Max MESH = 1.0 and wet bulb 0ºC height = 3000 yields $P(F_1|C_{\text{yes}}) = 0.002089$
- From Figure 1, cell B2, a combination of ENI flash rate = 40 and effective bulk shear = 40 yields $P(F_1|C_{\text{yes}}) = 0.000806$
- From Figure 1, cell C2, a combination of CAPE -10ºC to -30ºC = 600 and precipitable water = 1.2 yields $P(F_1|C_{\text{yes}}) = 0.003785$
- From Figure 2 (red line), normalized satellite growth rate = 3 yields $P(F_1|C_{\text{yes}}) = 0.010978$

After substitutions,

$$\prod_{i=1}^{4} P(F_i|C_{\text{yes}}) = 0.002089 * 0.000806 * 0.003785 * 0.010978 = 6.996 * 10^{-11}$$

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

In the second term of the denominator:

$$\prod_{i=1}^{4} P(F_i|C_{\text{no}}) = P(F_1|C_{\text{no}}) * P(F_2|C_{\text{no}}) * P(F_3|C_{\text{no}}) * P(F_4|C_{\text{no}})$$

- From Figure 1, cell A1, a combination of Max MESH = 1.0 and wet bulb 0ºC height = 3000 yields $P(F_1|C_{\text{no}}) = 0.000267$
- From Figure 1, cell B1, a combination of ENI flash rate = 40 and effective bulk shear = 40 yields $P(F_1|C_{\text{no}}) = 0.000208$
- From Figure 1, cell C1, a combination of CAPE -10ºC to -30ºC = 600 and precipitable water = 1.2 yields $P(F_1|C_{\text{no}}) = 0.001348$
- From Figure 2 (blue line), normalized satellite growth rate = 3 yields $P(F_1|C_{\text{no}}) = 0.004094$

After substitutions,

$$\prod_{i=1}^{4} P(F_i|C_{\text{no}}) = 0.000267 * 0.000208 * 0.001348 * 0.004094 = 3.065 * 10^{-13}$$

Substituting back into (1),

$$P(C_{\text{yes}}|\mathbf{F}) = \frac{P(C_{\text{yes}}) \prod_{i=1}^{N} P(F_i|C_{\text{yes}})}{P(C_{\text{yes}}) \prod_{i=1}^{N} P(F_i|C_{\text{yes}}) + P(C_{\text{no}}) \prod_{i=1}^{N} P(F_i|C_{\text{no}})}$$

$$P(C_{\text{yes}}|\mathbf{F}) = \frac{0.03 * (6.996 * 10^{-11})}{0.03 * (6.996 * 10^{-11}) + 0.97 * (3.065 * 10^{-13})}$$

$P(C_{\text{yes}}|\mathbf{F})$ = 0.8759

The probability of severe hail for this storm is approximately 88%.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

References

Atkins, N. T., and R. M. Wakimoto, 1991: WET MICROBURST ACTIVITY OVER THE SOUTHEASTERN UNITED-STATES - IMPLICATIONS FOR FORECASTING. *Weather and Forecasting*, **6,** 470-482.

Bedka, K., E. M. Murillo, C. R. Homeyer, B. Scarino, and H. Mersiovsky, 2018: The Above-Anvil Cirrus Plume: An Important Severe Weather Indicator in Visible and Infrared Satelite Imagery. *Weather and Forecasting*, **33,** 1159-1181.

Broyles, C., C. K. Potkin, C. Crosbie, R. M. Rabin, and P. Skinner, 2018: Location and Frequency of Surface Lows and Lower-Tropospheric Jets for U.S. Violent Tornadoes. *29th Conference on Severe Local Storms,* American Meteorological Society.

Cintineo, J. L., M. J. Pavolonis, J. M. Sieglaff, and A. K. Heidinger, 2013: Evolution of Severe and Nonsevere Convection Inferred from GOES-Derived Cloud Properties. *Journal of Applied Meteorology and Climatology*, **52,** 2009-2023.

Cintineo, J. L., M. J. Pavolonis, J. M. Sieglaff, and D. T. Lindsey, 2014: An Empirical Model for Assessing the Severe Weather Potential of Developing Convection. *Weather and Forecasting*, **29,** 639-653.

Cintineo, J. L., M. P. Pavolonis, J. M. Sieglaff, A. Wimmers, J. C. Brunner, and W. Bellon, 2020: A deep learning model for automated detection of mid-latitude convection using geostationary satellite images. *Weather and Forecasting,* **submitted**

Cintineo, J. L., and Coauthors, 2018: The NOAA/CIMSS ProbSevere Model: Incorporation of Total Lightning and Validation. *Weather and Forecasting*, **33,** 331-345.

Davies, J. M., 2004: Estimations of CIN and LFC associated with tornadic and nontornadic supercells. *Weather and Forecasting*, **19,** 714-726.

Doswell, C. A., H. E. Brooks, and R. A. Maddox, 1996: Flash flood forecasting: An ingredients-based methodology. *Weather and Forecasting*, **11,** 560-581.

Edwards, R., J. T. Allen, and G. W. Carbin, 2018: Reliability and Climatological Impacts of Convective Wind Estimations. Journal of Applied Meteorology and Climatology, 57, 1825-1845.

(FEMA), F. E. M. A., 2007: Thunderstorms Fact Sheet. FEMA 557 ed., FEMA, Ed., U.S. Department of Homeland Security.

Gagne, D. J., S. E. Haupt, D. W. Nychka, and G. Thompson, 2019: Interpretable Deep Learning for Spatial Analysis of Severe Hailstorms. Monthly Weather Review, 147, 2827-2845.

Gensini, V. A., D. Gold, J. T. Allen, and B. S. Barret, 2019: Extended US tornado outbreak

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

during late May 2019: a forecast of opportunity. Geophysical Research Letters, 46, 10150 - 10158.

Goodman, S. J., and Coauthors, 2013: The GOES-R Geostationary Lightning Mapper (GLM). *Atmospheric Research*, **125–126,** 34-49.

Harrison, D., 2018: CORRECTING, IMPROVING, AND VERIFYING AUTOMATED GUIDANCE IN A NEW WARNING PARADIGM, School of Meteorology, University of Oklahoma, 97 pp.

Karstens, C. D., and Coauthors, 2018: A Conceptual Model for Generating a Continuous Flow of Information for Severe Convective Events *Weather and Forecasting*.

Kuchera, E. L., and M. D. Parker, 2006: Severe convective wind environments. *Weather and Forecasting*, **21,** 595-612.

Lagerquist, R., A. McGovern, and T. Smith, 2017: Machine Learning for Real-Time Prediction of Damaging Straight-Line Convective Wind. Weather and Forecasting, 32, 2175-2193.

Lagerquist, R., and D. J. Gagne II, 2019. Interpretation of deep learning for predicting thunderstorm rotation: Python tutorial. GitHub, https://github.com/djgagne/ams-ml-python-course/blob/master/module_4/ML_Short_Course_Module_4_Interpretation.ipynb

Lakshmanan, V., R. Rabin, and V. DeBrunner, 2003: Multiscale storm identification and forecast. *Atmospheric Research*, **67-8,** 367-380.

Lakshmanan, V., K. Hondl, and R. Rabin, 2009: An Efficient, General-Purpose Technique for Identifying Storm Cells in Geospatial Images. *Journal of Atmospheric and Oceanic Technology*, **26,** 523-537.

Lakshmanan, V., B. Herzog, and D. Kingfield, 2015: A Method for Extracting Postevent Storm Tracks. *Journal of Applied Meteorology and Climatology*, **54,** 451-462.

Lakshmanan, V., T. Smith, G. Stumpf, and K. Hondl, 2007: The Warning Decision Support System-Integrated Information. *Weather and Forecasting*, **22**.

Lawson, J. R., J. S. Kain, N. Yussouf, D. C. Dowell, D. M. Wheatley, K. H. Knopfmeier, and T. A. Jones, 2018: Advancing from Convection-Allowing NWP to Warn-on-Forecast: Evidence of Progress. *Weather and Forecasting*, **33,** 599-607.

Markowski, P., and Y. Richardson, 2010: *Mesoscale Meteorology in Midlatitudes.* Wiley-Blackwell, 407 pp.

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

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

Implications of Machine Learning. Bulletin of the American Meteorological Society, 100, 2175-2199.

McGovern, A., D. J. Gagne, J. K. Williams, R. A. Brown, and J. B. Basara, 2014: Enhancing understanding and improving prediction of severe weather through spatiotemporal relational learning. Machine Learning, 95, 27-50.

Miller, M. L., V. Lakshmanan, and T. M. Smith, 2013: An Automated Method for Depicting Mesocyclone Paths and Intensities. *Weather and Forecasting*, **28,** 570-585.

Miller, P. W., and T. L. Mote, 2017: A Climatology of Weakly Forced and Pulse Thunderstorms in the Southeast United States. *Journal of Applied Meteorology and Climatology,* **56,** 3017-3033.

Morgan, G. M., Jr., and P. W. Summers, 1982: Hailfall and hailstorm characteristics. Thunderstorms: A Social, Scientific and Technological Documentary. *Thunderstorm Morphology and Dynamics*, E. Kessler, Ed., U.S. Government Printing Office, 363-408.

Murillo, E. M., and C. R. Homeyer, 2019: Severe Hail Fall and Hailstorm Detection Using Remote Sensing Observations. Journal of Applied Meteorology and Climatology, 58, 947-970.

NOAA, cited 2016a: Storm Prediction Center. [Available online at http://www.spc.noaa.gov/climo/online/.]

——, cited 2016b: Storm Prediction Center. [Available online at https://www.spc.noaa.gov/exper/ma_archive/.]

——, cited 2016c: National Centers for Environmental Information. [Available online at https://gis.ncdc.noaa.gov/maps/ncei/radar.]

——, cited 2019a: Storm Events Database. [Available online at https://www.ncdc.noaa.gov/stormevents/.]

——, cited 2019b: Storm Prediction Center WCM page. [Available online at https://www.spc.noaa.gov/wcm/.]

NWS, 2010: Thunderstorms, Tornadoes, Lightning. Nature's Most Violent Storms., N. O. a. A. Administration, Ed., U.S. Department of Commerce.

NWS: WFO severe weather products specification. National Weather Service Instruction 10-511. [Available online at http://www.nws.noaa.gov/directives/sym/pd01005011curr.pdf.]

Pavolonis, M. J., 2010: Advances in Extracting Cloud Composition Information from Spaceborne Infrared Radiances-A Robust Alternative to Brightness Temperatures. Part I: Theory. *Journal of Applied Meteorology and Climatology*, **49,** 1992-2012.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

Rasmussen, E. N., and D. O. Blanchard, 1998: A baseline climatology of sounding-derived supercell and tornado forecast parameters. *Weather and Forecasting*, **13,**1148-1164.

Roberts, B., I. L. Jirak, A. J. Clark, S. J. Weiss, and J. S. Kain, 2019: PostProcessing and Visualization Techniques for Convection-Allowing Ensembles. Bulletin of the American Meteorological Society, 100, 1245-1258.

Rothfusz, L. P., P. T. Schlatter, E. Jacks, and T. M. Smith, 2014: A Future Warning Concept: Forecasting a Continuum of Environmental Threats (FACETs). *94th American Meteorological Society Annual Meeting*.

Rudlosky, S. D., S. J. Goodman, K. S. Virts, and E. C. Bruning, 2019: Initial Geostationary Lightning Mapper Observations. *Geophysical Research Letters*, **46,** 1097-1104.

Sandmæl, T. N., C. R. Homeyer, K. M. Bedka, J. M. Apke, J. R. Mecikalski, and K. Khlopenkov, 2019: Evaluating the Ability of Remote Sensing Observations to Identify Significantly Severe and Potentially Tornadic Storms. Journal of Applied Meteorology and Climatology, 58, 2569-2590.

Schmit, T. J., and Coauthors, 2015: Rapid Refresh Information of Significant Events: Preparing Users for the Next Generation of Geostationary Operational Satellites. *Bulletin of the American Meteorological Society*, **96,** 561-576.

Schultz, C. J., W. A. Petersen, and L. D. Carey, 2011: Lightning and Severe Weather: A Comparison between Total and Cloud-to-Ground Lightning Trends. *Weather and Forecasting*, **26,** 744-755.

Sieglaff, J. M., D. C. Hartung, W. F. Feltz, L. M. Cronce, and V. Lakshmanan, 2013: A Satellite-Based Convective Cloud Object Tracking and Multipurpose Data Fusion Tool with Application to Developing Convection. *Journal of Atmospheric and Oceanic Technology*, **30,** 510-525.

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.

Thompson, R. L., C. M. Mead, and R. Edwards, 2007: Effective Storm-Relative Helicity and Bulk Shear in Supercell Thunderstorm Environments. *Weather and Forecasting*, **22,** 102-115.
Thompson, R. L., R. Edwards, J. A. Hart, K. L. Elmore, and P. Markowski, 2003: Close proximity soundings within supercell environments obtained from the rapid update cycle. *Weather and Forecasting*, **18,** 1243-1261.

Trapp, R. J., D. M. Wheatley, N. T. Atkins, R. W. Przybylinski, and R. Wolf, 2006: Buyer beware: Some words of caution on the use of severe wind reports in postevent assessment and research. Weather and Forecasting, 21, 408-415.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

University, I. S.: Iowa Environmental Mesonet Cow (NWS Storm-Based Warning Verification). [Available online at https://mesonet.agron.iastate.edu/cow/.]

Wakimoto, R. M., 2001: Convectively Driven High Wind Events. *Severe Convective Storms*, C. A. Doswell III, Ed., American Meteorological Society, 255-298.

Wimmers, A., C. Velden, and J. H. Cossuth, 2019: Using Deep Learning to Estimate Tropical Cyclone Intensity from Satellite Passive Microwave Imagery. Monthly Weather Review, 147, 2261-2282.

Witt, A., M. D. Eilts, G. J. Stumpf, E. D. Mitchell, J. T. Johnson, and K. W. Thomas, 1998: Evaluating the performance of WSR-88D severe storm detection algorithms. *Weather and Forecasting*, **13**.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

939

940 *Table 1: Tested parameters in the ProbSevere models and their data sources. Bolded parameters were selected to be*
941 *incorporated into predictors of the ProbSevere models.*

Data source	Parameters
Radar – MRMS	• 30 dBZ echo top height • 50 dBZ echo top height • Height of the 50 dBZ above 0°C • Height of the 50 dBZ above -20°C • Height of the 60 dBZ above 0°C • Height of the 60 dBZ above -20°C • **0-2 km AGL azimuthal shear (LLAzShear)** • **3-6 km AGL azimuthal shear (MLAzShear)** • MergedReflectivityQCComposite • **Maximum expected size of hail (MESH)** • Probability of severe hail (POSH) • Reflectivity at -10°C isotherm • Reflectivity at -20°C isotherm • Vertically integrated ice (VII) • Vertically integrated liquid (VIL) • **VIL Density**
Lightning – ENI	• **Total lightning flash rate (LtgFR)** • **Total lightning flash density (LtgFD)** • *d/dt*(LtgFR) (*dt* = 2 min) • Lightning jump algorithm (LJA) anomaly
NWP – RAP	• CAPE 0-3 km AGL • **CAPE -10°C to -30°C (hailCAPE)** • Downdraft CAPE (DCAPE) • Lapse rate 0-3 km AGL • Lapse rate 700-500 mb • Lifted condensation level (LCL) • Lowest height of 0°C • **Lowest height of wet bulb 0°C (wet bulb zero)** • Minimum average relative humidity 700-450 mb • Relative humidity at 0°C • **MLCAPE (0-90 mb AGL)** • **MUCAPE** • **MLCIN (0-90 mb AGL)** • SBCAPE • **Precipitable water (PWAT)** • θ_e difference between surface and min(θ_e) in 700-450 mb • **Effective bulk shear (EBShear)** • Bulk shear 0-1 km AGL

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

	• Bulk shear 0-3 km AGL • Bulk shear 0-6 km AGL • **Mean wind 1-3km AGL (MW 1-3km)** • **Storm-relative helicity 0-1 km AGL (SRH01)** • Storm-relative helicity 0-3 km AGL (SRH03) • Significant Tornado Parameter (fixed method)
Satellite – GOES16	• **Normalized satellite growth rate ($\Delta\varepsilon_{tot}$)** • Rate of change in cloud-top phase (Δice)

942

943 *Table 2: The predictors used in the ProbHail naïve Bayesian classifier.*

ProbHail predictors
• *a priori* = 0.03 • Max MESH / wet bulb zero • LtgFR / EBShear • hailCAPE / PWAT • Normalized satellite growth rate

944

945 *Table 3: The predictors used in the ProbWind naïve Bayesian classifier.*

Cellular ProbWind predictors	**Linear ProbWind Predictors**
• *a priori* = *f*(MUCAPE, EBShear) • Max MESH • LtgFR / EBShear • Normalized satellite growth rate	• *a priori* = *f*(MLCAPE, MW 1-3km) • Max VIL Density • 98th % LLAzShear / MW 1-3km • Flash rate / 98th % MLAzShear
Final ProbWind = *f*(cellular ProbWind, linear ProbWind)	

946

947 *Table 4: The predictors used in the ProbTor naïve Bayesian classifier.*

ProbTor predictors
• *a priori* = 0.01 • MLCAPE / MLCIN • Max LLAzShear • 98th % LLAzShear / SRH01 • 98th % MLAzShear / LtgFD • EBShear / MW 1-3km

948

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

Table 5: A contingency table defining the joint distribution of yes and no forecasts (f_{yes} and f_{no}) and yes and no observations (O_{yes} and O_{no}). The terms are defined as follows: A_e *is the number of warned events (i.e., reports),* A_w *is the number of verified warnings,* B *is the number of missed events (reports), and* C *is the number of false alarms (i.e., unverified warnings).*

	f_{yes}	f_{no}
O_{yes}	A_e, A_w	B
O_{no}	C	N/A

Table 6: Validation dates from 2018. Each day represents the "convective day" from 12 UTC of the given date to 11:59 UTC of the next date.

Month	**Dates**	**Count**
January	12, 21-23	4
February	6, 7, 10, 11, 15, 16, 20, 21, 24, 25	10
March	1, 5, 10, 16-20, 23-28	14
April	3, 4, 6, 7, 10, 13-15, 21-23, 29, 30	13
May	1-31	31
June	1-30	30
July	1-31	31
August	1-31	31
September	1-21, 24-27, 30	26
October	1-14, 20-23, 28, 29, 31	21
November	1, 2, 5-7, 24, 30	7
December	1, 2, 9, 14, 20, 21, 26, 27, 31	9

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

Figure Caption List

Figure 1: The probability of a non-severe storm (column 1), probability of a storm with severe hail (column 2), and the ratio of severe hail probability to non-severe probability (column 3), conditional on the wet bulb 0ºC height and MRMS MESH (row A), effective bulk shear and ENI flash rate (row B), and precipitable water and CAPE between -10ºC and -30ºC (row C). Columns 1 and 2 are lookup tables in ProbHail. The larger values in the ratio plots (column 3) indicate larger contributions to ProbHail.

Figure 2: The conditional probability of a severe storm (red) and a non-severe storm (blue) given a normalized satellite growth rate from GOES-16. A larger ratio of the severe and non-severe probabilities (black) indicates a larger contribution of this predictor in the naïve Bayesian models.

Figure 3: The conditional probability of any severe, given the MUCAPE and effective bulk shear. This is an update to Figure 2 in Cintineo et al. (2014).

Figure 4: The conditional probability of a severe storm (red) and a non-severe storm (blue) given a MRMS MESH value. A larger ratio of the severe and non-severe probabilities (black) indicates a larger contribution of this predictor in ProbWind (cellular).

Figure 5: The conditional probability of severe wind from a linear-type storm, given the MLCAPE and mean wind 1-3 km AGL.

Figure 6: The conditional probability of severe wind from a linear-type storm (red) and a non-severe storm (blue), given its maximum VIL density. A larger ratio of the severe and non-severe probabilities (black) indicates a larger contribution of this predictor in ProbWind (linear).

Figure 7: The probability of a non-severe storm (column 1), probability of severe wind from a linear-type storm (column 2), and the ratio of severe wind probability to non-severe probability (column 3), conditional on 98th percentile LLAzShear and mean wind 1-3 km AGL (row A), and the ENI flash rate and 98th percentile MLAzShear (row B). Columns 1 and 2 are lookup tables in ProbWind (linear). The larger values in the ratio plots (column 3) indicate larger contributions in ProbWind

Figure 8: The probability of severe wind gusts for a storm conditional on the computed cellular and linear naïve Bayesian classifier probabilities. This is the final lookup table for ProbWind.

Figure 9: The a priori factor for ProbTor as a function of MLCIN (left) and MLCAPE (right) in a storm. The original a priori for ProbTor (0.01) is multiplied by the minimum of these two factors. The red horizontal "cutoff" lines denote the minimum value either function is allowed to attain (the value is 0.1). Where these lines intersect the blue lines show the values of MLCIN and MLCAPE where the minimum a priori factor occurs (-90 J kg^{-1} MLCIN and 150 J kg^{-1} MLCAPE). Please see the text for details on how these functions were created.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

Figure 10: The conditional probability of a tornadic storm (red) and severe, non-tornadic storm (blue), given its maximum 0-2 km AGL AzShear. A larger ratio of the severe and non-severe probabilities (black) indicates a larger contribution of this predictor in ProbTor.

Figure 11: The probability of a non-tornadic, severe storm (column 1), probability of a tornadic storm (column 2), and the ratio of tornadic probability to non-tornadic probability (column 3), conditional on 98th percentile 0-2km AGL AzShear and 0-1 km AGL storm-relative helicity (row A), ENI flash density and 98th percentile 3-6km AGL AzShear (row B), and effective bulk shear and mean wind 1-3 km AGL (row C). Columns 1 and 2 are lookup tables in ProbTor. The larger values in the ratio plots (column 3) indicate larger contributions in ProbTor.

Figure 12: Starting with the top left, clockwise: performance diagrams for PSv2 models constrained by 15 min, 30 min, 60 min, and 45 min storm longevity thresholds. Colored circles with labels denote certain forecast probability values. "probSevere" is the maximum hazard probability. National Weather Service (NWS) performance is denoted on the 45 min storm longevity performance diagram (lower left). Python code from Lagerquist and Gagne II (2019) was used to construct this plot.

Figure 13: PSv2 forecast calibration from 2017 and 2018 (left ordinate) and forecast probability frequency (purple bars; right ordinate).

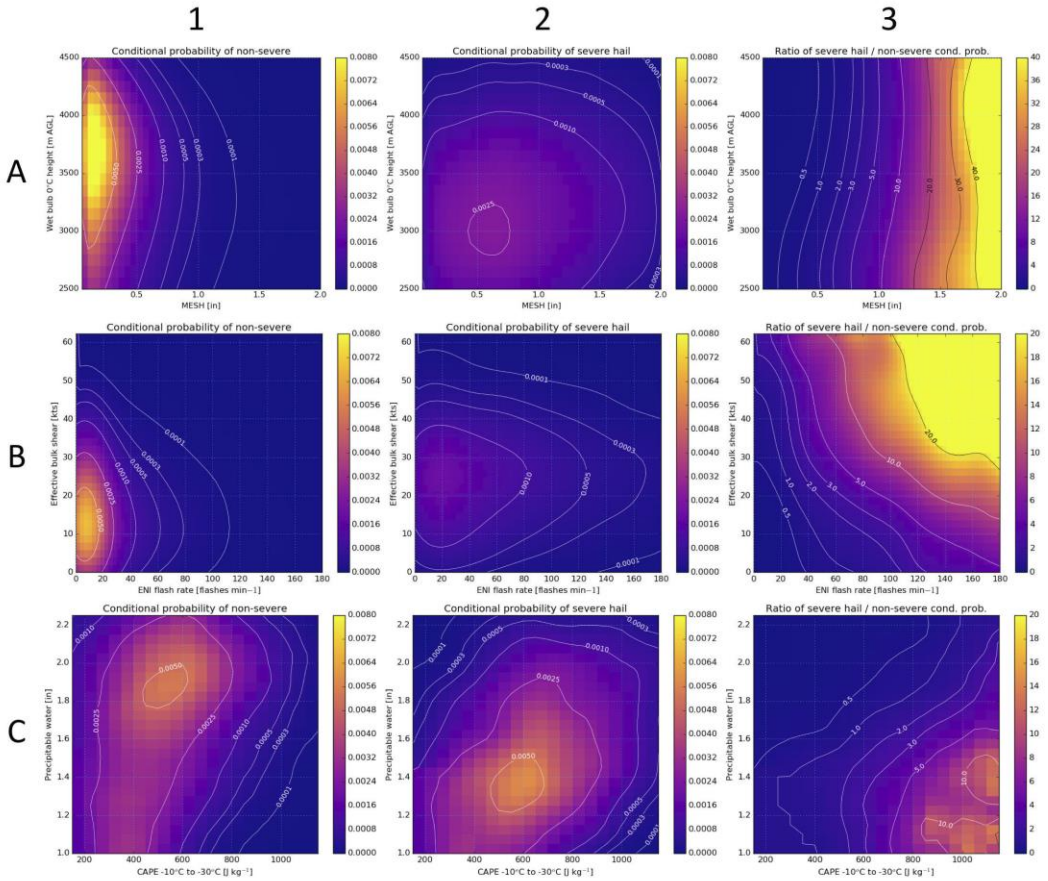
Figure 14: The maximum CSI by month for each PSv2 model and the National Weather Service (NWS) severe thunderstorm and tornado warnings (sev + tor) and tornado warnings only (tor only). The dates are from 2018 (see Table 6). Gray and purple bars denote the number of severe and tornadic storms, respectively, in the dataset for a given month. Note that "probSevere" refers to the maximum hazard probability, scored against any severe report type.

Figure 15: Top left: probability of detection (POD); top right: false alarm ratio (FAR); bottom left: critical success index (CSI); bottom right: most skillful probability threshold for PSv2. The POD, FAR, and CSI correspond to the probability threshold maximizing CSI. These scores are aggregated for National Weather Service county warning area (CWA) boundaries and include adjacent CWAs (see text).

Figure 16: The percentage of Hazardous Weather Testbed forecaster 'yes' responses to daily end-of-operations questions, by year and model type (left panel: all ProbSevere v2 models; right panel: ProbTor model only). The two questions were, "does the product increase your confidence in warning decision making?" ("Increase confidence?"), and, "does the product increase your lead time to severe hazards?" ("Increase lead time?"). The ProbTor only questions referred to tornado warnings and lead time to tornado occurrences, whereas the ProbSevere v2 (all models) questions pertained to severe thunderstorm and tornado warnings and lead time to any National Weather Service-defined severe weather type.

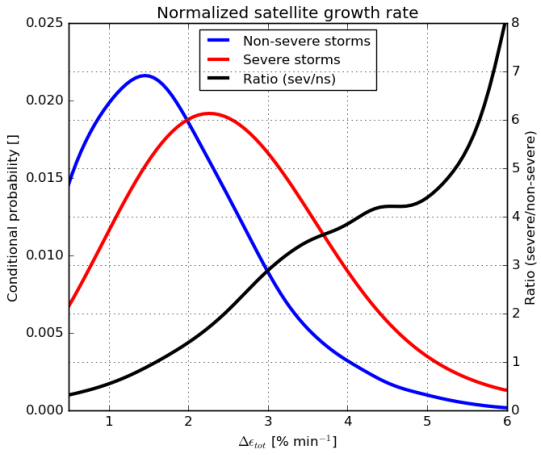
Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

1044 Figures



1045

1046 *Figure 1: The probability of a non-severe storm (column 1), probability of a storm with severe hail (column 2), and the ratio of*
1047 *severe hail probability to non-severe probability (column 3), conditional on the wet bulb 0°C height and MRMS MESH (row A),*
1048 *effective bulk shear and ENI flash rate (row B), and precipitable water and CAPE between -10°C and -30°C (row C). Columns 1 and*
1049 *2 are lookup tables in ProbHail. The larger values in the ratio plots (column 3) indicate larger contributions to ProbHail.*



1050

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

Figure 2: The conditional probability of a severe storm (red) and a non-severe storm (blue) given a normalized satellite growth rate from GOES-16. A larger ratio of the severe and non-severe probabilities (black) indicates a larger contribution of this predictor in the naïve Bayesian models.

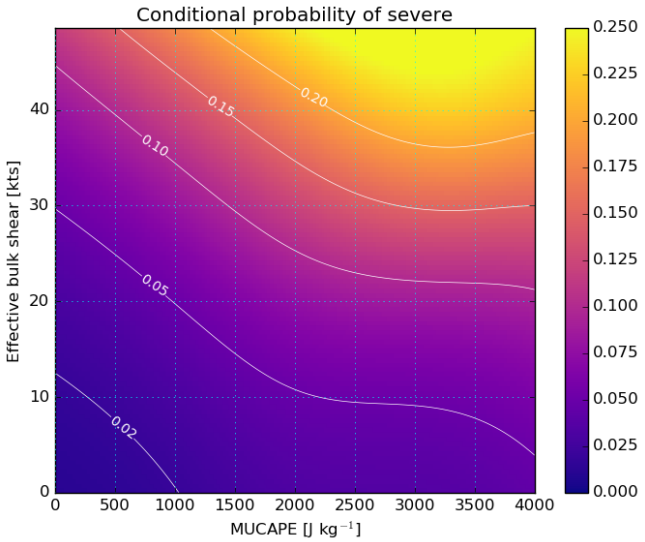


Figure 3: The conditional probability of any severe, given the MUCAPE and effective bulk shear. This is an update to Figure 2 in Cintineo et al. (2014).

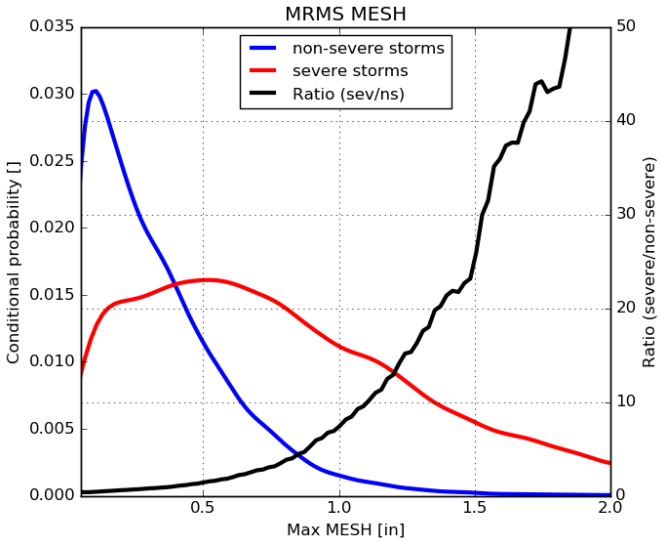
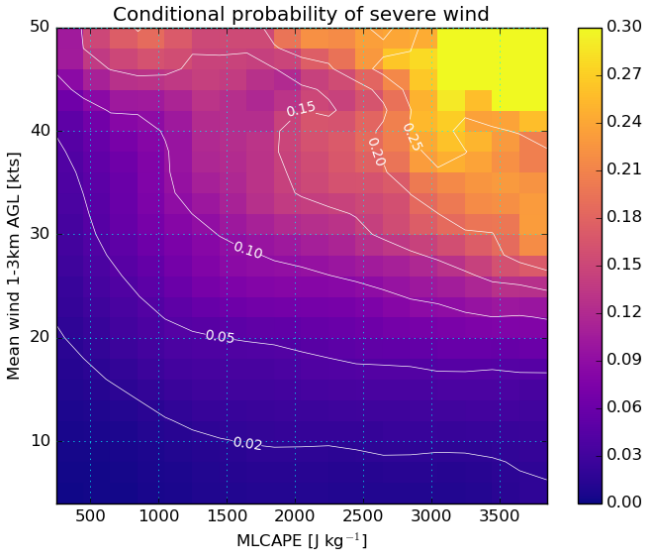


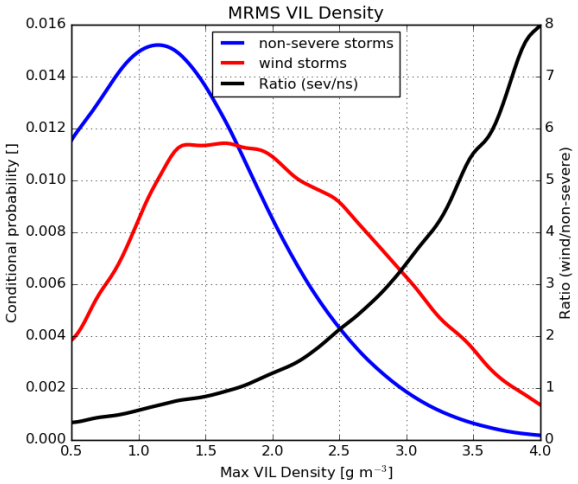
Figure 4: The conditional probability of a severe storm (red) and a non-severe storm (blue) given a MRMS MESH value. A larger ratio of the severe and non-severe probabilities (black) indicates a larger contribution of this predictor in ProbWind (cellular).

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.



1061

1062 *Figure 5: The conditional probability of severe wind from a linear-type storm, given the MLCAPE and mean wind 1-3 km AGL.*



1063
1064 *Figure 6: The conditional probability of severe wind from a linear-type storm (red) and a non-severe storm (blue), given its*
1065 *maximum VIL density. A larger ratio of the severe and non-severe probabilities (black) indicates a larger contribution of this*
1066 *predictor in ProbWind (linear).*

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

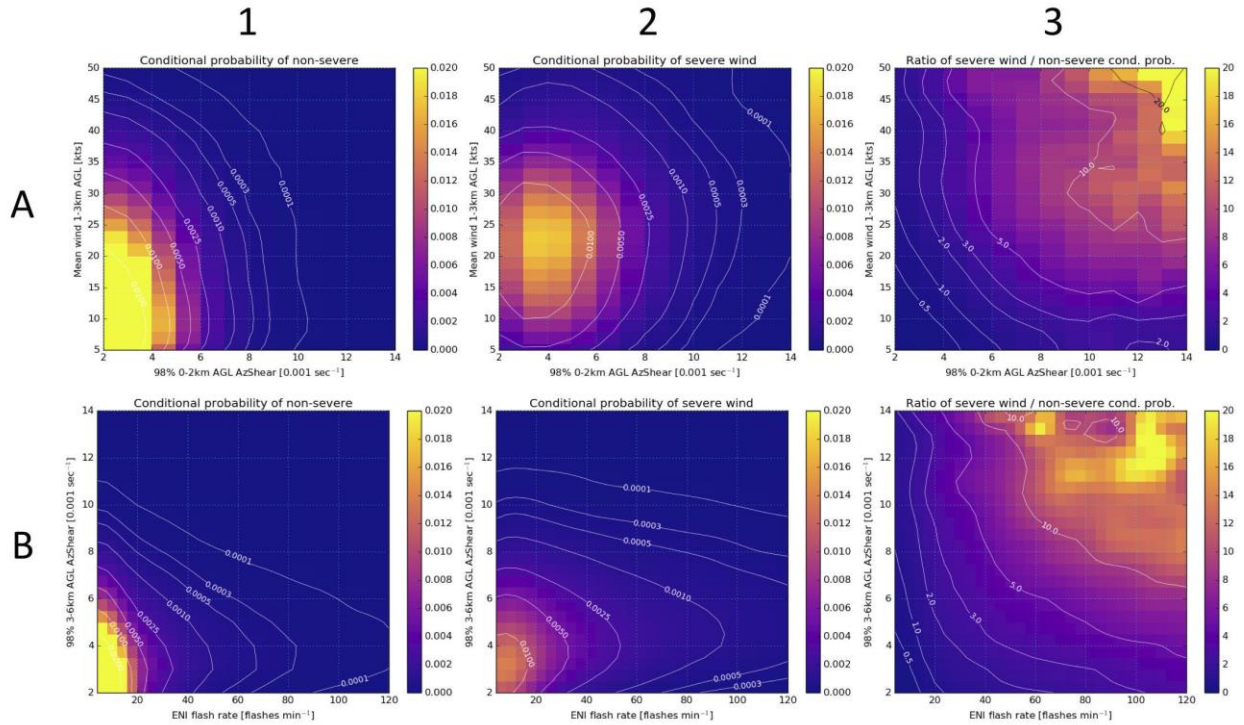


Figure 7: The probability of a non-severe storm (column 1), probability of severe wind from a linear-type storm (column 2), and the ratio of severe wind probability to non-severe probability (column 3), conditional on 98th percentile LLAzShear and mean wind 1-3 km AGL (row A), and the ENI flash rate and 98th percentile MLAzShear (row B). Columns 1 and 2 are lookup tables in ProbWind (linear). The larger values in the ratio plots (column 3) indicate larger contributions in ProbWind

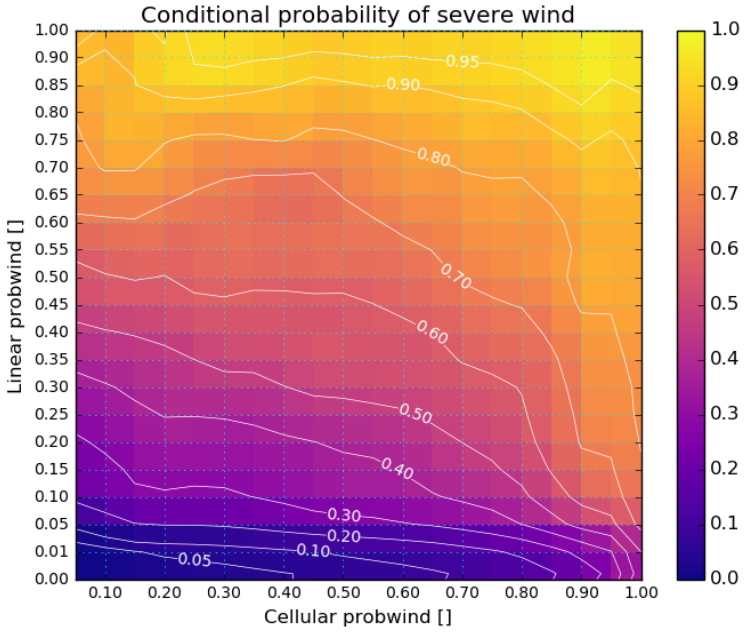


Figure 8: The probability of severe wind gusts for a storm conditional on the computed cellular and linear naïve Bayesian classifier probabilities. This is the final lookup table for ProbWind.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

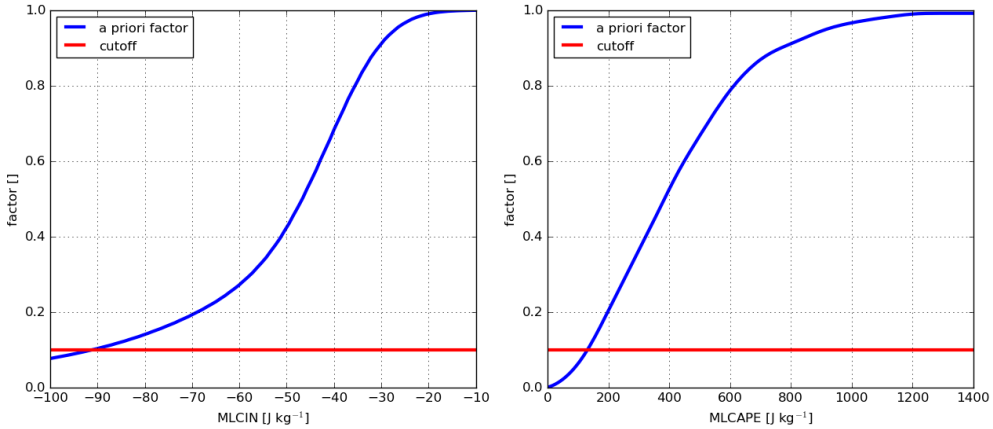


Figure 9: The a priori factor for ProbTor as a function of MLCIN (left) and MLCAPE (right) in a storm. The original a priori for ProbTor (0.01) is multiplied by the minimum of these two factors. The red horizontal "cutoff" lines denote the minimum value either function is allowed to attain (the value is 0.1). Where these lines intersect the blue lines show the values of MLCIN and MLCAPE where the minimum a priori factor occurs (-90 J kg^{-1} MLCIN and 150 J kg^{-1} MLCAPE). Please see the text for details on how these functions were created.

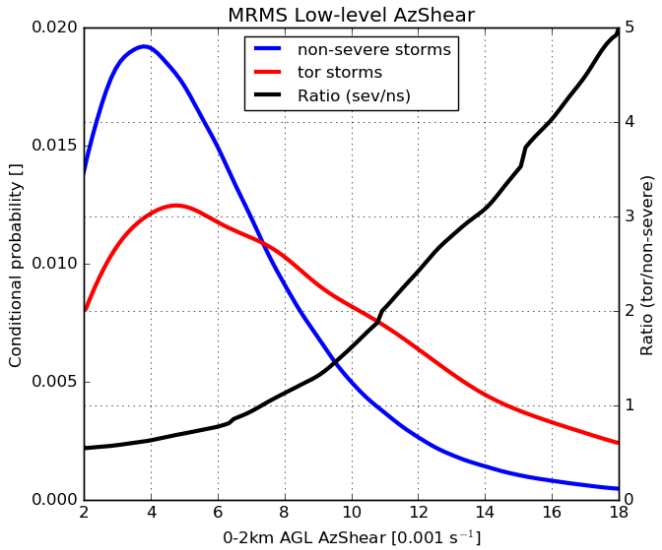


Figure 10: The conditional probability of a tornadic storm (red) and severe, non-tornadic storm (blue), given its maximum 0-2 km AGL AzShear. A larger ratio of the severe and non-severe probabilities (black) indicates a larger contribution of this predictor in ProbTor.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

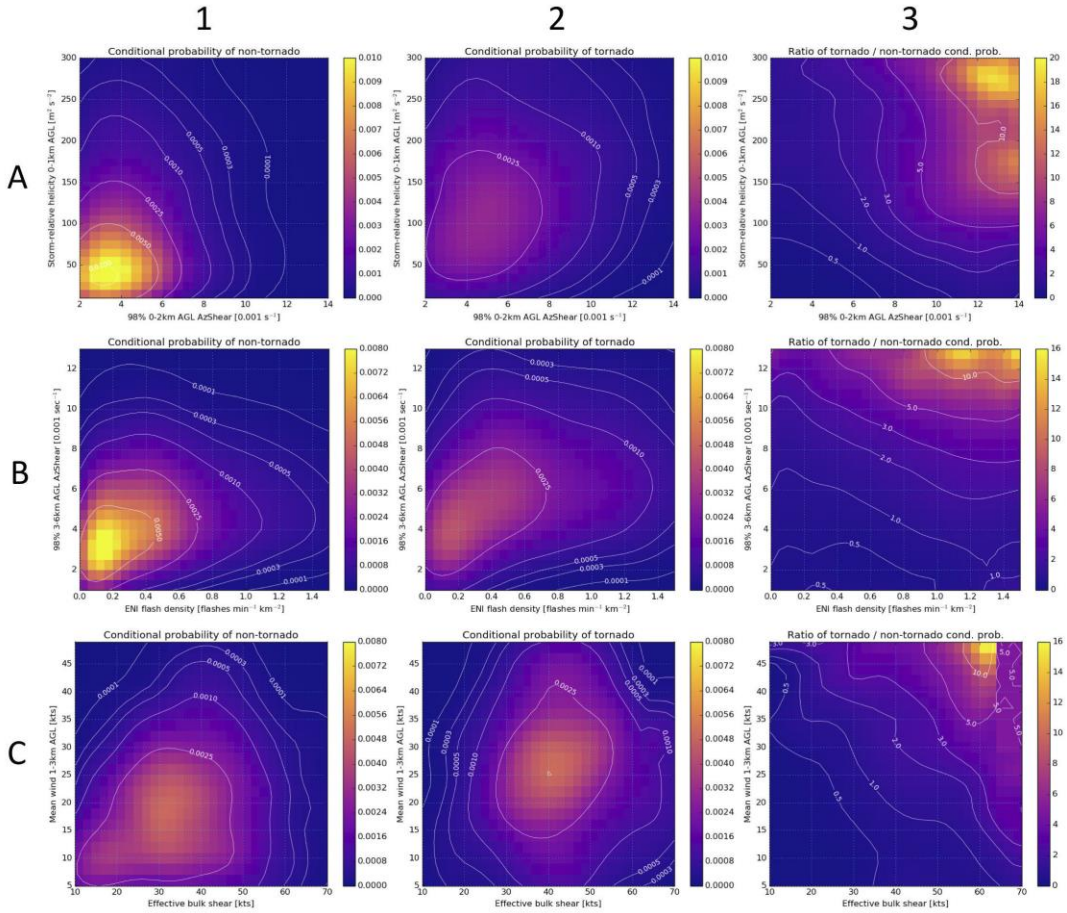


Figure 11: The probability of a non-tornadic, severe storm (column 1), probability of a tornadic storm (column 2), and the ratio of tornadic probability to non-tornadic probability (column 3), conditional on 98th percentile 0-2km AGL AzShear and 0-1 km AGL storm-relative helicity (row A), ENI flash density and 98th percentile 3-6km AGL AzShear (row B), and effective bulk shear and mean wind 1-3 km AGL (row C). Columns 1 and 2 are lookup tables in ProbTor. The larger values in the ratio plots (column 3) indicate larger contributions in ProbTor.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

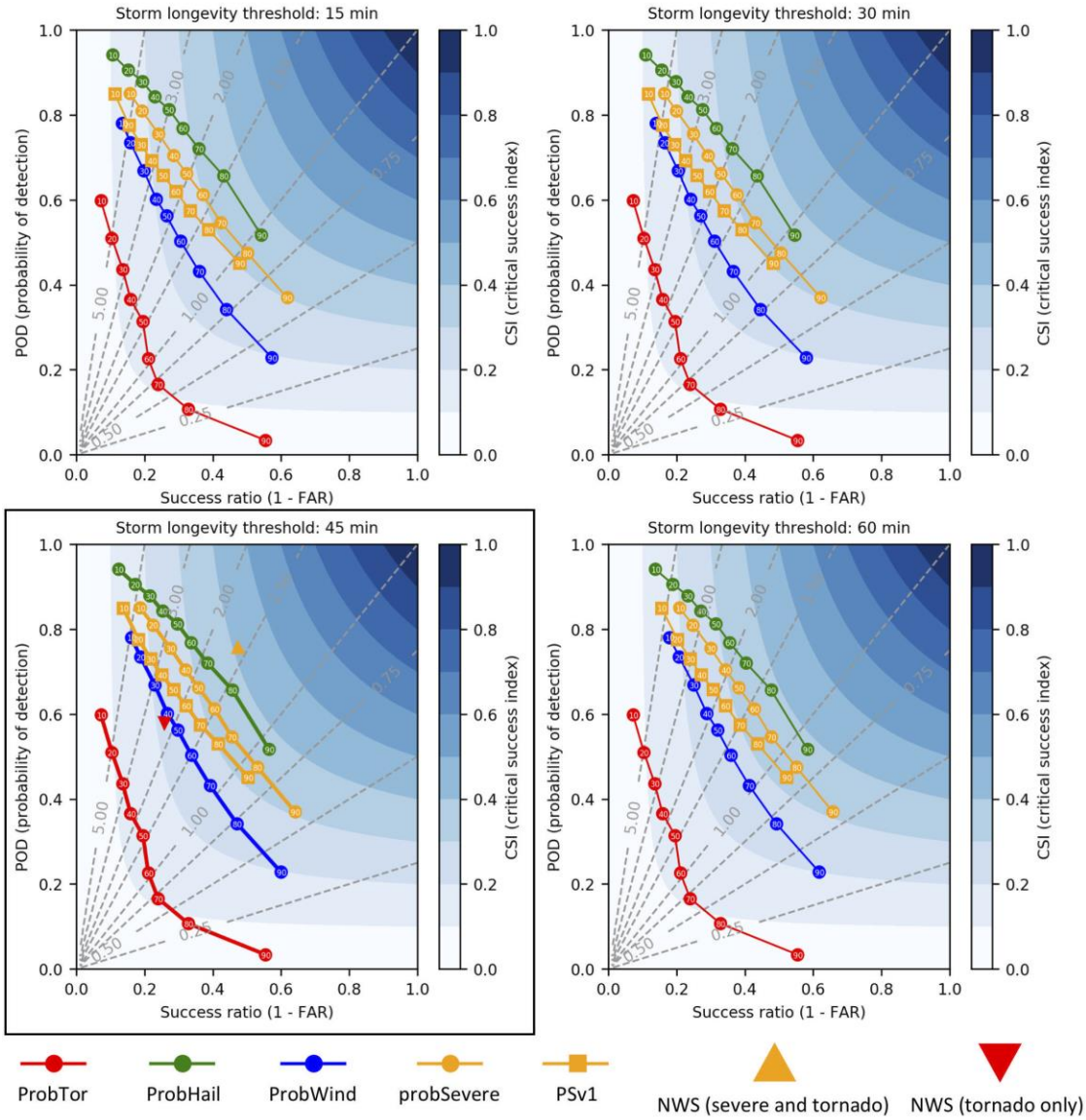


Figure 12: Starting with the top left, clockwise: performance diagrams for PSv2 models constrained by 15 min, 30 min, 60 min, and 45 min storm longevity thresholds. Colored circles with labels denote certain forecast probability values. "probSevere" is the maximum hazard probability. National Weather Service (NWS) performance is denoted on the 45 min storm longevity performance diagram (lower left). Python code from Lagerquist and Gagne II (2019) was used to construct this plot.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

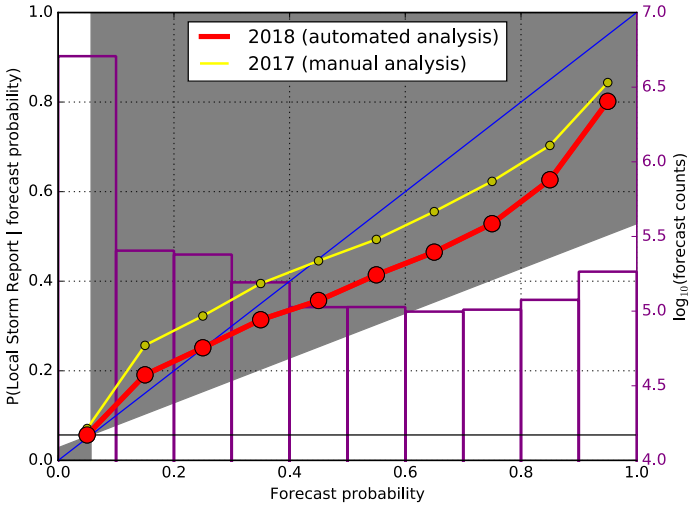


Figure 13: PSv2 forecast calibration from 2017 and 2018 (left ordinate) and forecast probability frequency (purple bars; right ordinate).

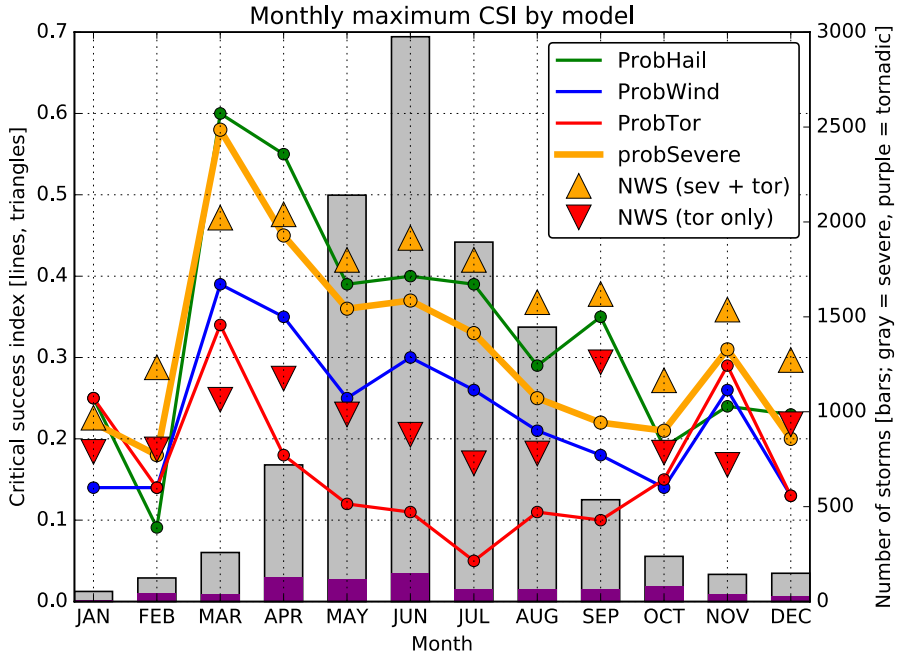


Figure 14: The maximum CSI by month for each PSv2 model and the National Weather Service (NWS) severe thunderstorm and tornado warnings (sev + tor) and tornado warnings only (tor only). The dates are from 2018 (see Table 6). Gray and purple bars denote the number of severe and tornadic storms, respectively, in the dataset for a given month. Note that "probSevere" refers to the maximum hazard probability, scored against any severe report type.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.

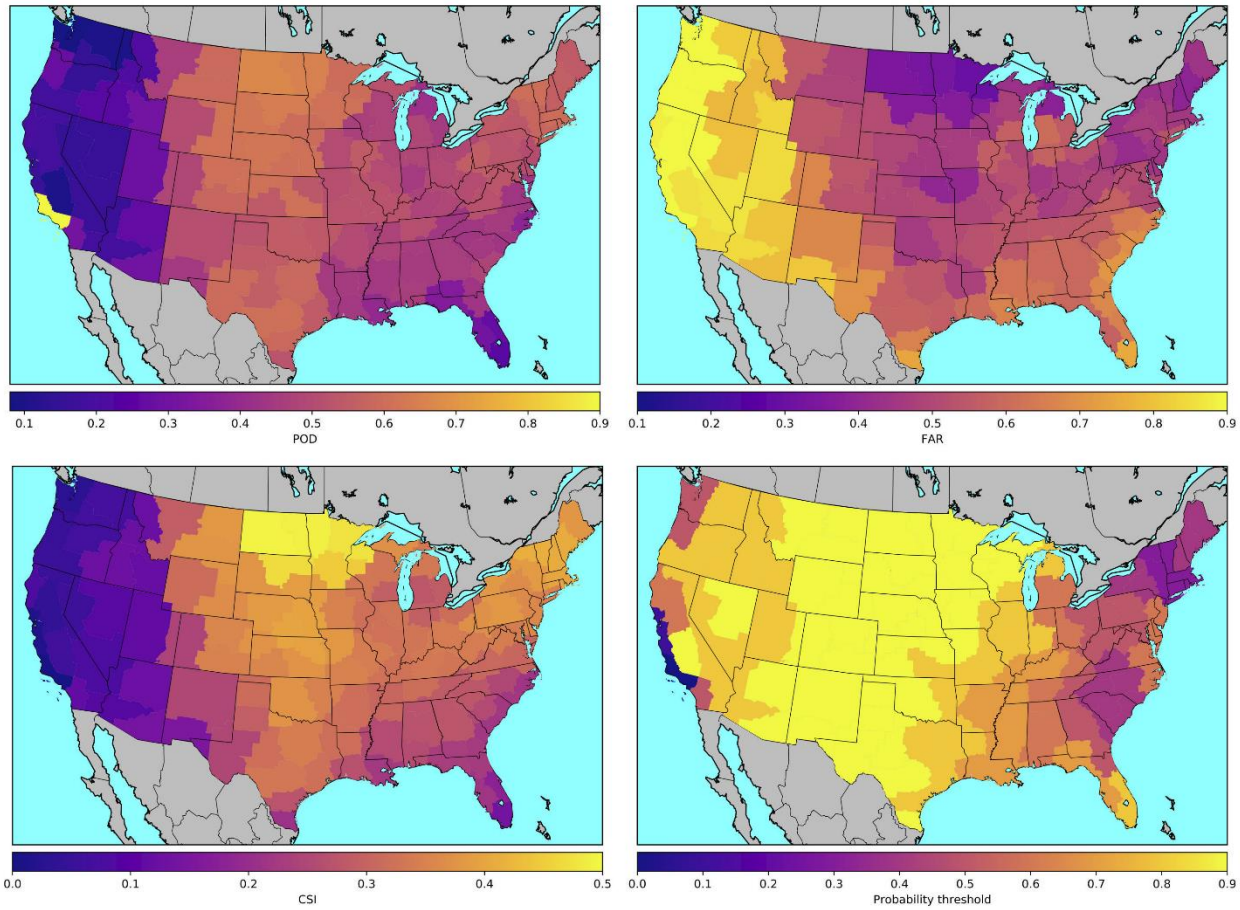


Figure 15: Top left: probability of detection (POD); top right: false alarm ratio (FAR); bottom left: critical success index (CSI); bottom right: most skillful probability threshold for PSv2. The POD, FAR, and CSI correspond to the probability threshold maximizing CSI. These scores are aggregated for National Weather Service county warning area (CWA) boundaries and include adjacent CWAs (see text).

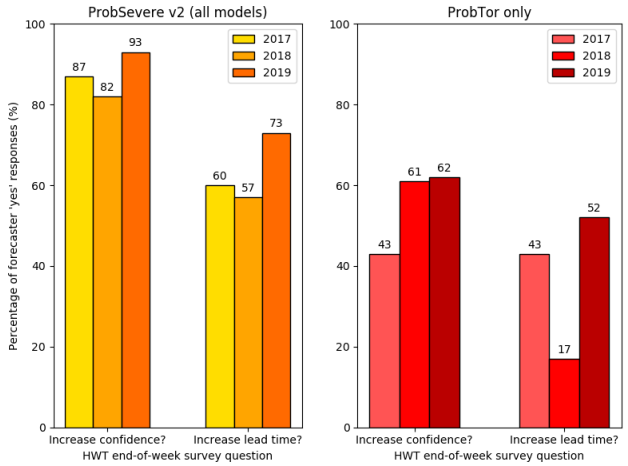


Figure 16: The percentage of Hazardous Weather Testbed forecaster 'yes' responses to daily end-of-operations questions, by year and model type (left panel: all ProbSevere v2 models; right panel: ProbTor model only). The two questions were, "does the product increase your confidence in warning decision making?" ("Increase confidence?"), and, "does the product increase your lead time to severe hazards?" ("Increase lead time?"). The ProbTor only questions referred to tornado warnings and lead time to tornado occurrences, whereas the ProbSevere v2 (all models) questions pertained to severe thunderstorm and tornado warnings and lead time to any National Weather Service-defined severe weather type.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-19-0242.1.