**NOAA ProbSevere 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*

Jordan Gerth

*Cooperative Institute of Meteorological Satellite Studies, University of Wisconsin—Madison*

Jason Brunner

*Cooperative Institute of Meteorological Satellite Studies, University of Wisconsin—Madison*

**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 trained 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. This paper overviews the ProbSevere system and details the construction and selection of model predictors. An evaluation of the four models demonstrated that v2 is more skillful than v1 for each severe hazard with higher critical success index scores. The discussion highlights PSv2 in NOAA’s Hazardous Weather Testbed (HWT) and current and future research for convective nowcasting.

1. Introduction

Severe storms are a common occurrence across the contiguous United States (CONUS) causing loss of life and over 1 billion USD in annual damages (e.g., Changnon 2011). Such storms occur in every season and may occur in nearly any locale. 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, 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 Rapid Refresh (HRRR); Benjamin et al. 2011), 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]; Goodman et al. 2013), terrestrial lightning arrays (e.g., Earth Networks Inc. [ENI] Total Lightning Network [ENTLN], Vaisala National Lightning Detection Network [NLDN]), and Multi-Radar Multi-Sensor products (MRMS; Smith et al. 2016). This development presents 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 new 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 from 2013 to 2016 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 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. This model has been in unofficial operational use by the NWS 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 NOAA Operations Proving Ground (OPG), 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 ≥ 1 in), 2) straight-line convective winds (gust ≥ 58 mph), and 3) tornadoes. Thus, ProbSevere version 2 (PSv2) was developed from 2016 to 2018 with three models specific to the aforementioned hazards (ProbHail, ProbWind, and ProbTor), as well as an “any severe” model (probSevere). PSv2 was evaluated at the HWT in 2017-2019 and is planned to become operational within MRMS in 2020. This paper gives some general background on the ProbSevere system of data integration and prediction, then describes the predictor evaluation and selection process, as well as the forecast performance for each model. The discussion section highlights future areas ripe for rapid development.

1. ProbSevere system overview

The real-time ProbSevere system has several aspects: 1) observation processing, 2) storm identification and tracking, 3) predictor extraction and probability computation.

1. Observation processing

There are four streams of data processing in ProbSevere from different observation sources: 1) geostationary satellites, 2) MRMS products, 3) ENI total lightning, and 4) Rapid Refresh (RAP) model output (used as near-storm environment “observations”). The processing of each stream is run in parallel, enabling more efficient updates to ProbSevere output products.

1. GOES-East

ProbSevere currently uses GOES-East data only (located at 75o W), as 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., CA, OR, WA), where the average pixel area ranges from 16-24 km2. As resources allow, GOES-West can be easily incorporated in ProbSevere, but for the time being, the increase in processing expense outweighed any expected gains in using GOES-West data. As GOES-East data stream in every 5 min for the CONUS sector, they are processed using radiative transfer software to create the 11-µm top-of-troposphere emissivity (εtot). This field is the emissivity a cloud would have if it were at the tropopause (Pavolonis 2010a) and helps normalize storm depth seasonally and latitudinally. The εtot is then remapped to a cylindrical equidistant projection over the CONUS, at 0.02o x 0.02o resolution (approximately 2 km x 2 km in the midlatitudes). The time rate of change of εtot, the “normalized satellite growth rate” (∆εtot) within satellite-identified storms is one of the predictors in ProbHail and ProbWind. The ∆εtot is analogous to decreases in the minimum observed 11-µm brightness temperature in the cloud object, which has been shown to help discern between severe and non-severe convection in the growth stage of cumulus development (Cintineo et al. 2013).

1. MRMS

Several MRMS products are used in the ProbSevere models. These 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 (VILD), merged composite reflectivity (COMPREF), 0-2 km AGL azimuthal shear (LLAzShear), and 3-6 km AGL azimuthal shear (MLAzShear). The COMPREF, VILD, and MESH arrive natively in a 0.01o x 0.01o cylindrical equidistant projection and are simply cropped to the ProbSevere domain, whereas LLAzShear and MLAzShear arrive in a 0.005o x 0.005o cylindrical equidistant projection and are remapped to the 0.01o x 0.01o projection and cropped to align with the reflectivity-based fields.

1. ENI total lightning

Total lightning from the ground-based ENTLN are ingested as binary files every minute in ProbSevere. They contain attributes of each flash recorded in the previous minute, 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 0.01o x 0.01o spatial resolution on the same domain as the processed satellite and MRMS gridded products. This total lightning flash density field (LtgFD) has units of fl min-1 km-2.

1. RAP NWP

ProbSevere pulls operational RAP analysis data, as well as 1-hour, 2-hour, and 3-hour forecast fields, approximately every hour from the National Centers for Environmental Prediction (NCEP) Central Operations (NCO). The fields extracted and computed are: MUCAPE, MLCAPE (0-90mb AGL), MLCIN (0-90mb AGL), 0-1 km AGL storm-relative helicity (SRH01), 1-3 km AGL mean wind (MEANWIND), effective bulk shear (EBS), lowest wet-bulb 0oC height (WBZ), CAPE in the -10oC to -30oC layer (CAPEM10M30), and precipitable water (PWAT). These field are remapped to a 0.04o x 0.04o 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 (t0), the previous hour analysis (t−1), and the 1-, 2-, and 3-h forecasts (tF1, tF2, and tF3, respectively). Thus, it is “centered” on tF1, which is the current time when the data become available. For most fields, the temporal compositing is a maximum operation of the 5 hours, but for WBZ and MLCIN, it is a minimum operation. The MLCIN and SRH01 use a 3-hour temporal compositing (still centered on tF1). The spatial smoothing uses a 5 × 5 gridpoint (~67.5 km × 67.5 km) Gaussian filter, with a smoothing radius equal to three standard deviations.

1. Storm identification and tracking

C14 and C18 summarize the tracking procedures in ProbSevere. Sieglaff et al. (2013) details the satellite tracking methodology, and both radar and satellite tracking use the WDSS-II algorithm w2segmotionll as the backbone object identification and tracking software (Lakshmanan et al. 2003). Some w2segmotionll configuration options have changed for radar object tracking, so we will summarize the changes here. The WDSS-II algorithm uses an enhanced watershed algorithm to create objects. In this case, the algorithm searches for local maxima of 40 dBZ ≤ COMPREF ≤ 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 km2). 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 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 footprint (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 the post-processing merging step 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.

1. 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 ∆εtot is computed for satellite objects when possible and shared with overlapping radar objects after a parallax correction. MRMS, lightning, and NWP attributes are extracted from the radar object footprint. Model predictors are then computed from the extracted observations and the probabilities are computed using a Naïve Bayesian classifier (NBC; 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. Using Bayes’ theorem, the probability of a storm producing a targeted hazard given a set of observed predictors **F** is defined by

. (1)

is the sample frequency of the hazard occurring (the *a priori*). Naturally, = 1 – . The “naïve” assumption of predictor independence allows for simplification of Bayes’ theorem by reduction of dimensionality. Thus, Eqn. (1) can be rewritten as

, (2)

with *Fi* denoting the value of the *i*thpredictor, and *N* the number of predictors. The denominator can be rewritten as

. (3)

The Π is the product operator, multiplying the probability of the *i*th predictor conditional on the storm being a member of . 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**. Please see Kossin and Sitkowski (2009) for details on dimensionality reduction of Bayes’ 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 NBC works well even while violating this assumption. However, the models’ performances do degrade when too many highly correlated predictors are used, as will be elaborated in the next section.

The final probSevere model (probability of any severe hazard) 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 take into account the dependent nature of hazards (e.g., a joint 3D lookup table of the three NBCs).

1. Predictor selection methodology

The NBCs were trained on 167 days of data from 2015 and 2016, encompassing the months of January through November. Preliminary local storm reports (LSRs) from NOAA’s Storm Prediction Center (SPC) rough log for each day were matched up to ProbSevere IDs in time and space in the same manner as C18 (i.e., finding the spatially closest centroid of a ProbSevere storm object to the report location within a +/- 2 min). 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 predictors explored. For the MRMS-based predictors, several percentile values were considered 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 flashes from within a storm, 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 drawn from the smoothed fields (see section 2aiv) were considered. And for the GOES-16 predictors, temporal maximum values of the satellite trends within a 2.5-hr time window were used. One advantage of the NBC 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.

As stated previously, the NBC 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, we generated a few rules of thumb when constructing models to test. For each model, we attempted 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 highly-correlated fields into a two-dimensional (2D) predictor, which helps reduce the negative impact of the correlation on the NBC. Both one-dimensional (1D) and 2D predictor distributions were smoothed with kernel density estimation (KDE) using a normal kernel function and optimally chosen bandwidths, unless otherwise indicated, 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 or 2D conditional probability vector or matrix, depending on the input shape of the data.

Models were thus constructed in an ad hoc manner, using the most favorable predictors from the training dataset, which were determined by looking at the maximum ratio between P(*Fi* | *C*yes) and P(*Fi* | *C*no) and the difference in means of each class. The models were then evaluated systematically on independent data from 2016, by iterating on previous model designs and tests. In this way, ProbSevere model construction is still somewhat of an art and most likely suboptimal. However, ProbSevere’s utility has been demonstrated in the NWS (C18) and section 7 details a validation based on data collected in real-time from 2018.

1. ProbHail

The probability of severe hail NBC (ProbHail) was trained using two classes: 1) storms that produced severe hail (diameter ≥ 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 four predictors summarized in Table 2. These include the 1) max MESH / WBZ (Figure 1), 2) LtgFR / EBS (Figure 2), 3) CAPEM10M30 / PWAT (Figure 3), and 4) ∆εtot (Figure 4). The most important field in ProbHail is the max MESH. The WBZ was found to generally and correctly increase probabilities for storms when it is low (< 3000 m AGL) and correctly decrease probabilities for storms when it is high (> 4000 m AGL). ProbHail still struggles with low-topped hailstorms as the MESH is generally low (≤ 0.5 in) compared to taller storms that produce 1 in diameter surface hail and the lower WBZ does not compensate enough. Furthermore, these storms also tend to exhibit low LtgFR. This may indicate a shortcoming in the training dataset whereby more storms are needed to populate this area of phase space.

1. ProbWind

The probability of severe wind NBC (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 help mitigate potential cross-contamination between classes.

It quickly became apparent that there are multiple conceptual models or mechanisms for severe wind gust producing thunderstorms (e.g., perturbation pressure forces, condensate loading, dry air entrainment into downdrafts and evaporative cooling [Wakimoto 2001]), which are difficult to account for in one NBC. We eventually created two NBCs for ProbWind—one for “cellular” windstorms and one for “linear” windstorms. While this may seem 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 storms such as supercells and dry and wet microbursts, while the linear model is appropriate for squall lines, bowing segments, quasi-linear convective systems (QLCS) and other mesoscale convective systems (MCS).

How the models are applied will be discussed shortly, but first their construction needs some discussion. 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 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., EBS, 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 of a given day were excluded if only one wind report was present or the wind type was not clear (e.g., different severe wind types in the same region during a given day). In general, regions of storms that were relatively small (≤ 20 km in diameter), and circular were considered cellular, while regions of storms that were relatively large or elongated ( > 20 km in one dimension) and had EBS ≥ 20 kts were considered linear. The cellular type could have very high EBS (e.g., supercell environments) or very low EBS (e.g., wet microbursts in “pulse” storms) The “typed” wind reports were then used as the “yes” class for the cellular and linear NBCs. This analysis was subjective, but it facilitated the training of the two NBCs in ProbWind in an expedient manner. In the future, a storm-typing algorithm may help create a less biased estimate of wind gust type for storms.

1. Cellular wind NBC

Despite numerous tests with predictors found to be useful for forecasting severity of cellular storms in previous studies such as the vertical θe difference (\*\*source\*\*), low-level lapse rates (\*\*source\*\*), and mid-level average humidity (\*\*source\*\*), the NBC from C18 (PSv1), or the “control”, performed the best on the wide range of cellular storms. This model has four predictors, listed in Table 3. The *a priori* is a function of MUCAPE and EBS (Figure 5), as in C14 (an update to their Figure 2), with a climatological frequency built in, so the model in essence has four predictors. It should be noted that this *a priori* takes the conditional probability grids of severe wind (for cellular storms) and all thunderstorms and multiplies each grid elementwise by the number of severe windstorms and all thunderstorms, respectively. The severe wind result is then divided by the all thunderstorm result to get a frequency, or probability.

The other predictors for this NBC are the spatial maximum MESH within an object (Figure 6), the LtgFR / EBS (Figure 2), and the ∆εtot (Figure 4). This NBC has been shown to predict well in a number of environments (see C18), but still struggles in lightning deficient storms as well as dry microbursts. Dry microbursts present a particularly difficult challenge to the current ProbSevere framework, as severe downbursts may occur in low-reflectivity storms, for which ProbSevere may not have identifiable objects. It is a challenge to create a model for a specific type of severe wind producing storm due to differing mechanisms of severe wind gust production.

1. Linear wind NBC

The linear wind NBC contains four predictors, found in Table 3. The MRMS LLAzShear and MLAzShear products and the MEANWIND proved to be very helpful in predicting severe wind gusts from linear systems. The four predictors are 1) the *a priori*, which is a function of MLCAPE and MEANWIND (Figure 7; computed in the same fashion as the *a priori* predictor in the cellular wind model), 2) maximum VILD (Figure 8), 3) 98th percentile LLAzShear / MEANWIND (Figure 9), and 4) 98th percentile MLAzShear / LtgFR (Figure 10). 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 (WINDINF) discriminated well between non-severe convection and convection producing severe wind gusts. The MEANWIND field in ProbWind is computed over 1-3 km AGL, which likely contains the top of the inflow layer of many storms. Thus, the MEANWIND is likely correlated with WINDINF and also demonstrated good discrimination between severe wind storms and non-severe storms.

Rather than having a physical basis for severe wind gust production, it was found that the AzShear products served as a proxy for the detection of strong velocity magnitudes in many storms, when the feature created an azimuthal gradient in the Doppler radar velocity field. When the feature does not create azimuthal gradients in the velocity field, strong velocity magnitude detection may be poor and ProbWind may exhibit lower probabilities. While the LLAzShear, MLAzShear, and MEANWIND help improve ProbWind with respect to PSv1 for severe wind storms, lightning deficient storms again pose a forecasting challenge to the model.

1. Final ProbWind

To create a final probability value for ProbWind, different thresholds of EBS and MEANWIND were evaluated, but we found no set of values that discriminated well enough for linear and cellular storms. This could be due to the fact that severe wind gusts occur in storms on a continuum of MEANWIND and EBS 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 determine when to execute each NBC.

In light of the challenge of automated discrimination between wind type, a 2D lookup table was created using the joint distributions of computed cellular and linear NBCs for thunderstorms (Figure 11). This was computed in the same fashion as the *a priori* predictor in the cellular wind model, except it is conditional on the NBC output from the cellular and linear models instead of physical quantities. Figure 11 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 (over 100 days from 2016; not shown). The increased performance is possibly due to the fact that some storms have both ‘cellular’ and ‘linear’ features.

1. ProbTor

The probability of tornado NBC (ProbTor) was trained using two classes: 1) storms that produced a tornado (EF0+) and 2) storms that produce a combination of severe hail and/or severe wind reports. Unlike ProbHail and ProbWind, the null class for this NBC 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—forecasters often ask the question, “is a severe storm likely to produce a tornado” rather than, “is a benign storm likely to produce a tornado”.

ProbTor uses six predictors summarized in Table 4. These include the 1) MLCAPE / MLCIN (Figure 12), 2) max LLAzShear (Figure 13), 3) 98th percentile LLAzShear / SRH01 (Figure 14), 4) 98th percentile MLAzShear / LtgFD (Figure 15), and 5) EBS / MEANWIND (Figure 16). The *a priori* of 0.01 is approximately the number of tornadic storms divided by the total number thunderstorms in the training dataset.

The MLCAPE / MLCIN predictor was constructed in a unique way, compared to the other predictors in PSv2. It is generally well-known that tornadoes occur less frequently when surface-based CAPE is absent or is located above a deep layer of CIN (Davies, 2003). Using the MLCIN and MLCAPE distributions for the tornadic class of storms, cumulative distribution functions were created for each field (CDFs). The CDFs describe the fraction of tornadic storms that exhibited a value less than or equal to the given value of MLCIN or MLCAPE (see Figure 12). MLCIN helps reduce FAs.

The observational fields in ProbTor are the most essential to the NBC (particularly the LLAzShear) owing to their ability to depict rotation within a storm, but despite quality-control measures within the MRMS system (e.g., Miller et al. 2013), the AzShear fields can still be very noisy, on account of their derivation from Doppler radial velocity measurements from NEXRAD. This 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 against or not in accordance with a storm’s 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, MRMS AzShear has proven to be a skillful predictor in classifying between tornadic and non-tornadic storms.

The environmental EBS and SRH01 have been shown to help discriminate between tornadic and non-tornadic storms (Thompson et al. 2007). While the MEANWIND 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. This may be due to the fact that MEANWIND helps to capture strong low-level maxima (jets), which increases storm-relative inflow and supercell potential (Markowski and Richardson 2010) and has been associated with strong tornadoes (Broyles et al. 2018). Another possibility is that MEANWIND hints at strong mid-level storm-relative flow which can help strengthen the storm’s mesocyclone. Regardless, it is a useful field in ProbTor and is coupled with the EBS, with which it shares only a weak correlation (Pearson correlation = 0.17 for tornadic storms).

1. Validation
2. Method

After the initial training and validation of PSv2 using SPC storm reports from 2015 and 2016, PSv2 models were validated with data from 2018 from *Storm Data*, the National Center for Environmental Intelligence publication (NCEI) 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. 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 rate (FAR), and critical success index (CSI) for the different models could be computed. From the contingency Table 5, we can define the skill metrics thusly:

, (4)

, (5)

, (6)

where *Ae* is the number of warned events (e.g., hail, wind, or tornado reports), *Aw* is the number of verified warnings, *B* is the number of unwarned events (“misses”), and *C* is the number of unverified warnings (“false alarms)”.

To better conform to NWS warning and verification practices, “valid times” for a ProbSevere object were artificially issued once it attained a certain probability threshold. Valid times of 45 min for ProbHail, ProbWind and 30 min for ProbTor were chosen, since these values are the midpoints for the valid time ranges specified in the NWS WFO Severe Weather Products Specification document (NWS 2018). After a valid time expires for a “warning” of a given probability threshold, it can then be reissued if the probability threshold is met subsequently. Thus, a single storm can generate multiple verified and unverified warnings. Probability thresholds were binned in increments of 10% from 0% to 90%.

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

To mitigate storm object mergers and splits and better link together broken storm tracks, the python library “besttrack” was employed (Harrison 2017; Lakshmanan et al. 2015). This library utilizes the Theil-Sen estimator, which fits a line to storm centroid points (i.e., a storm track) by choosing the median of the slopes of all lines through pairs of points. This automated process helps mitigate but does not eliminate broken tracks. Thus, a longevity threshold of 45 min was placed on ProbSevere storm objects in order to ignore segments of storms that change object ID frequently. Applying the longevity threshold and a lightning activity threshold of 2 fl min-1 yielded 111,000 non-severe thunderstorms in the dataset.

1. Results

The left image in Figure 17 shows the maximum CSI and associated FAR, POD, and probability threshold for ProbHail, ProbWind, ProbTor, and PSv1 validated against the appropriate metric. PSv2 improves upon the skill of PSv1 for each report type. ProbHail has generally decreased FAR for hail-producing storms with respect to PSv1, while ProbWind has increased POD for wind-producing storms with respect to PSv1, and ProbTor has both increased POD and decreased FAR for tornadic storms. These improvements are due to more appropriate observations and predictors being incorporated to the hazard-specific models. The right image of Figure 17 compares probSevere of PSv2 at the most skillful (i.e., maximizing CSI) threshold (80%) to the NWS scores for all report types in the validation dataset. Unofficial NWS validation skill scores was obtained from the Iowa State Mesonet for all CONUS NWS offices for the dates in Table 6 (IEM, 2019). At the 80% probability threshold, PSv2 has slightly less FAR compared to the NWS (0.47 vs. 0.53), but much less POD (0.48 vs. 0.76), yielding a CSI of 0.34 compared to 0.41 for the NWS.

The reliability diagram for PSv2 shows an overforecasting bias above the 40% forecast probability threshold (Figure 18). While the bias certainly exists, it may be overestimated in this validation dataset, based on reliability found in previous research. This is likely due to differences in the validation methodology. Previous research used manual analysis to string together ProbSevere objects, which is very accurate for storms that split numerous times, but more expensive to perform. This research used the automated besttrack method to link object IDs, which may still struggle with proper storm track associations in long-lived storms with frequent ID changes (e.g., convective lines). Using the manual analysis approach on a 2017 dataset (spanning the entire year) yielded better forecast calibration. Nevertheless, the calibration of PSv2 for the 2018 dataset confirms that forecasters should be aware of the overforecasting bias at medium to larger forecast probabilities. Efforts are ongoing to make future versions of ProbSevere better calibrated.

The annual cycle of ProbSevere skill (Figure 19) shows that probSevere (the maximum of ProbHail, ProbWind, and ProbTor) skill peaks in the spring months of March and April, likely due to the predominant hazard type being large hail and predominant storm type being cellular in nature. May, June, and July contain many more storms and different storm morphologies, with probSevere exhibiting CSI of about 0.35. August, September and October see a dip in skill to about 0.2 – 0.25 CSI, while there is a notable increase in CSI to 0.3 for November, possibly due to a stronger jet stream producing more sheared environments, which tend to be more predictable for severe weather. December, January, and February see another drop in CSI to about 0.2. Instability is low during these months and storms are relatively few. Low instability may indicate storms with low lightning activity, which ProbSevere has been shown to struggle with (C18). Notably, ProbTor, which has less dependence on total lightning, exhibits average or above average skill during the winter months (~0.13 – 0.25 CSI). ProbTor exhibits its lowest CSI in July (CSI = 0.05) but increases to nearly 0.3 in November, likely due to the presence of more dynamic synoptic systems.

Figure 20 demonstrates skill of PSv2 in a spatial sense across the country, as a function of NWS county warning areas (CWA). For each storm, the mean centroid position over its lifetime was used to place it within a CWA. For each CWA, storms were then aggregated using the given CWA and all adjacent CWAs. For example, The Milwaukee, WI CWA aggregation would include the CWAs of Milwaukee WI, La Crosse WI, Green Bay WI, Quad Cities IA/IL, and Chicago IL. This aggregation gives a more regional measure of skill for a given locale. Skill scores were then calculated as described previously. The bottom right image shows the most skillful forecast probability thresholds by region. We see quite a variance, from the 30-40% threshold in the Northeast, to 40-70% threshold in the Southeast, then gradually increasing from 60% to 90% proceeding westward from west of the Appalachian Mountains to the Great Plains. The Intermountain West maximizes CSI generally at the 80% threshold, whereas the West Coast varies greatly. However, the sample size diminishes significantly west of the Rocky Mountains. CSI is maximized from 0.45 to 0.5 in the Northern Plains, with the rest of the Plains region exhibiting CSI of about 0.35 to 0.4. The Northeast exhibits similar maximum CSI, but at lower probability thresholds. The Southeast exhibits lower CSI, approximately 0.25 to 0.3, owing to a higher FAR. This could be due to a number of reasons, including tall storms that grow quickly from a satellite perspective, lightning-rich storms, or high MESH storms that don’t verify as well as storms in other parts of the country with similar attributes. The Intermountain West and West Coast have notably lower CSI likely due to differences in storm type (e.g., relatively more dry microbursts), poorer radar coverage, and possibly due to a smaller sample size.

1. Discussion & Conclusion

Acknowledgements

The authors acknowledge the NOAA GOES-R Risk Reduction Program for support of this research, as well as David Harrison for advice on running the besttrack code. 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 *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](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 or 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.

Appendix A

w2segmotionll configuration options:

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

Table 1: Tested predictors in the ProbSevere models and their data sources.

|  |  |
| --- | --- |
| **Data source** | **Predictors** |
| **Radar – MRMS[[1]](#footnote-1)** | * EchoTop\_30 * EchoTop\_50 * H50Above0C * H50AboveM20C * H60Above0C * H60AboveM20C * LLAzShear * MLAzShear * COMPREF * MESH * POSH * ReflectivityM10C * ReflectivityM20C * VII * VIL * VILD |
| **Lightning – ENI** | * Total lightning flash rate (LtgFR) * Total lightning flash density (LtgFD) * d/dt(LtgFR) * Lightning jump algorithm (LJA) anomaly |
| **NWP – RAP** | * CAPE 0-3km AGL * CAPE -10oC to -30oC * DCAPE * Lapse rate 0-3km AGL * Lapse rate 700-500mb * LCL * Lowest height of 0oC * Lowest height of wet-bulb 0oC (WBZ) * Minimum average relative humidity 700-450mb * Relative humidity at 0oC * MLCAPE (0-90mb AGL) * MUCAPE * MLCIN (0-90mb AGL) * SBCAPE * Precipitable water (PWAT) * θe difference between surface and min(θe) in 700-450mb * Effective bulk shear (EBS) * Bulk shear 0-1 km AGL * Bulk shear 0-3 km AGL * Bulk shear 0-6 km AGL * Mean wind 1-3km AGL (MEANWIND) * Storm-relative helicity 0-1 km AGL (SRH01) * Storm-relative helicity 0-3 km AGL (SRH03) * Significant Tornado Parameter (fixed) |
| **Satellite – GOES16** | * Normalized satellite growth rate (∆εtot) * Rate of change in cloud-top phase (∆*ice*) |

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

|  |
| --- |
| **ProbHail predictors** |
| * *a priori* = 0.03 * Max MESH / WBZ * LtgFR / EBS * CAPEM10M30 / PWAT * Normalized satellite growth rate |

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

|  |  |
| --- | --- |
| **Cellular ProbWind predictors** | **Linear ProbWind Predictors** |
| * *a priori* = *f*(MUCAPE, EBS) * Max MESH * LtgFR / EBS * Normalized satellite growth rate | * *a priori* = *f*(MLCAPE, MEANWIND) * Max VILD * 98th % LLAzShear / MEANWIND * Flash rate / 98th % MLAzShear |
| **Final ProbWind = *f*(cellular ProbWind, linear ProbWind)** | |

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 * EBS / MEANWIND |

Table 5: A contingency table defining the joint distribution of yes and no forecasts (fyes and fno) and yes and no observations (Oyes and Ono). The terms are defined as follows: Ae is the number of warned events (i.e., reports), Aw 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).

|  |  |  |
| --- | --- | --- |
|  | *fyes* | *fno* |
| *Oyes* | *Ae, Aw* | *B* |
| *Ono* | *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 |

Figures

A screenshot of a cell phone

Description automatically generated

Figure 1: The probability of a non-severe storm (left), probability of a storm with severe hail (center), and the ratio of severe hail probability to non-severe probability, conditional on the wet bulb 0oC height and MRMS MESH. The left and center images are lookup tables in ProbHail. The larger values in the ratio plot (right) indicate a higher contribution to ProbHail.

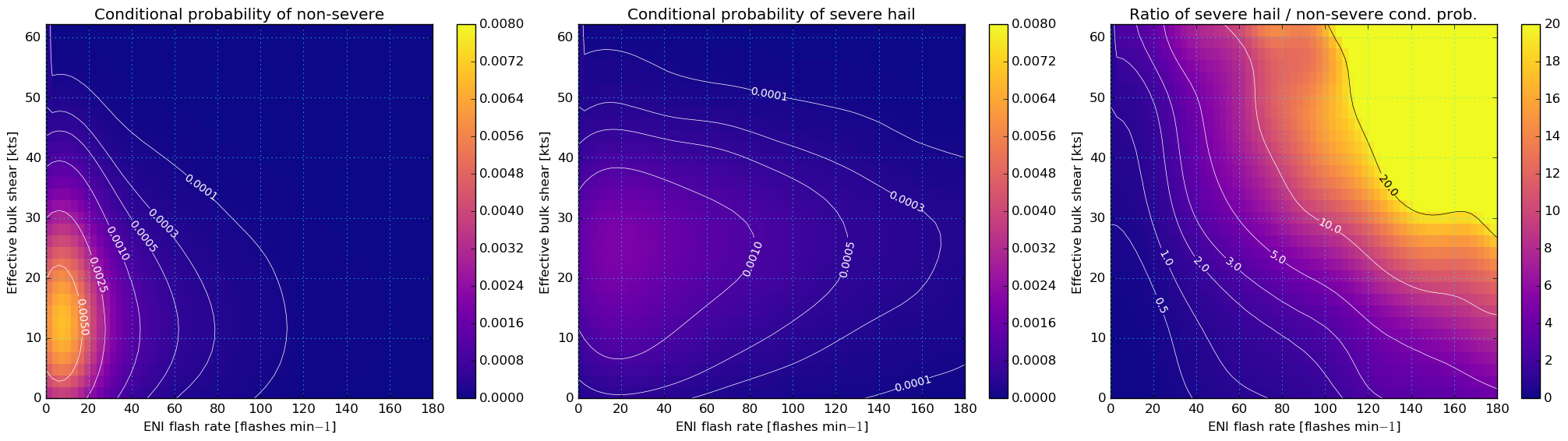


Figure 2: The probability of a non-severe storm (left), probability of a storm with severe hail (center), and the ratio of severe hail probability to non-severe probability, conditional on the effective bulk shear and ENI flash rate. The left and center images are lookup tables in ProbHail. The larger values in the ratio plot (right) indicate a higher contribution to ProbHail.

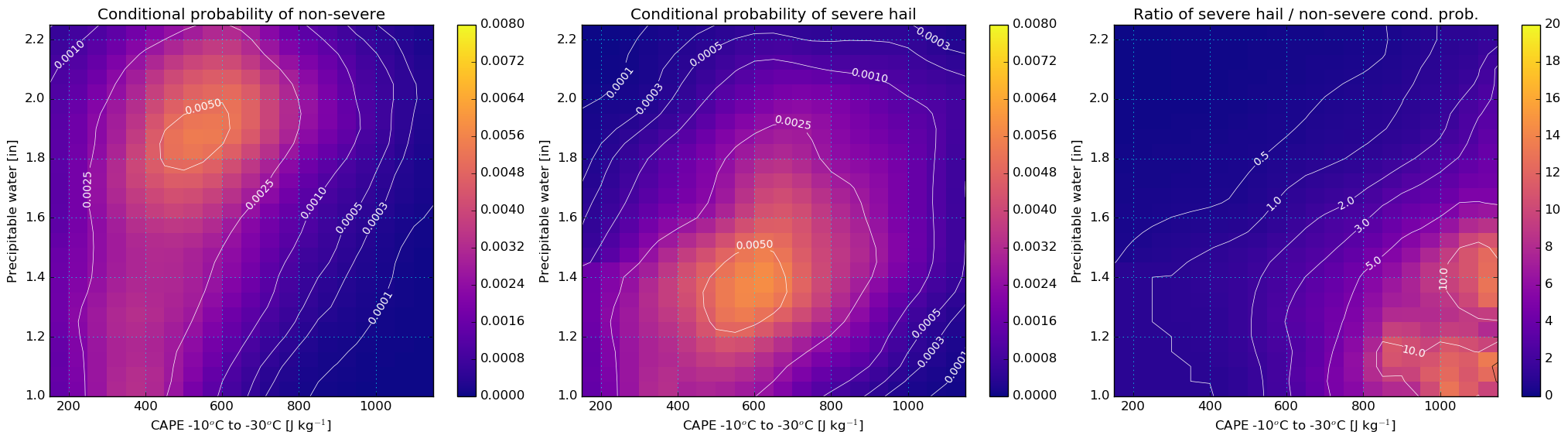


Figure 3: The probability of a non-severe storm (left), probability of a storm with severe hail (center), and the ratio of severe hail probability to non-severe probability, conditional on precipitable water and CAPE between -10oC and -30oC. The left and center images are lookup tables in ProbHail. The larger values in the ratio plot (right) indicate a higher contribution ProbHail.

A close up of a map

Description automatically generated

Figure 4: The conditional probability of a severe storm (red) and a non-severe storm (blue) given a normalized satellite growth rate from GOES-16. The ratio of the severe and non-severe probabilities (black) indicates the contribution of this predictor in the naïve Bayesian models.

A screenshot of a cell phone

Description automatically generated

Figure 5: The conditional probability of any severe, given the MUCAPE and EBS. This is an update to Figure 2 in Cintineo et al. (2014).

A close up of a map

Description automatically generated

Figure 6: The conditional probability of a severe storm (red) and a non-severe storm (blue) given a MRMS MESH value. The ratio of the severe and non-severe probabilities (black) indicates the contribution of this predictor in ProbWind (cellular).

A close up of a map

Description automatically generated

Figure 7: The conditional probability of severe wind from a linear-type storm, given the MLCAPE and mean wind 1-3 km AGL.

A close up of a map

Description automatically generated

Figure 8: The conditional probability of severe wind from a linear-type storm (red) and a non-severe storm (blue), given its maximum VIL density. The ratio of the severe and non-severe probabilities (black) indicates the contribution of this predictor in ProbWind (linear).

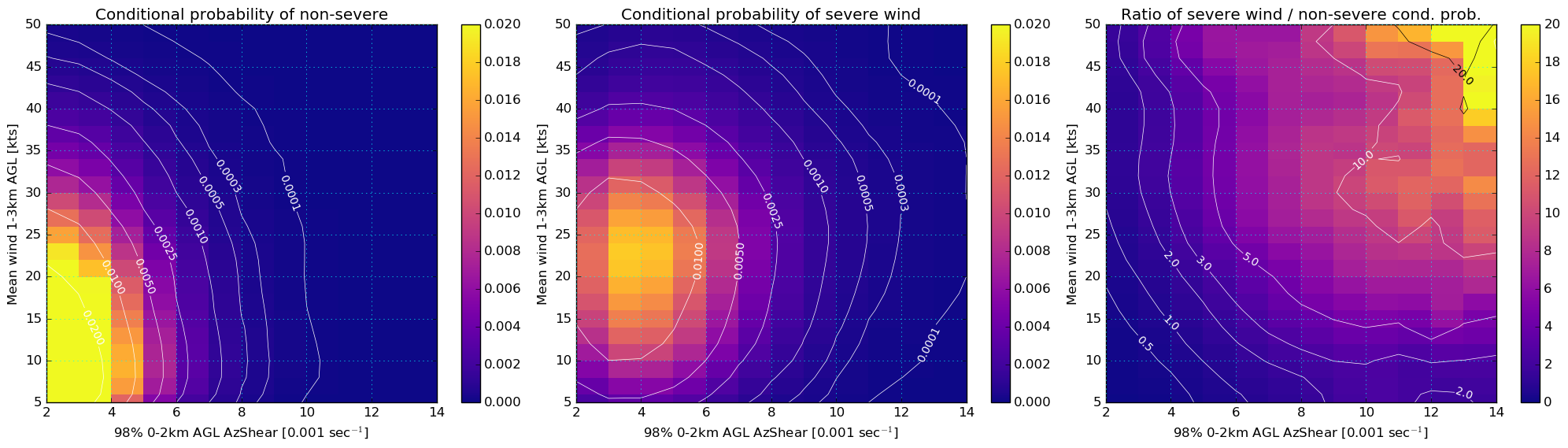


Figure 9: The probability of a non-severe storm (left), probability of severe wind from a linear-type storm (center), and the ratio of severe wind probability to non-severe probability, conditional on 98th percentile LLAzShear and mean wind 1-3 km AGL. The left and center images are lookup tables in ProbWind (linear). The larger values in the ratio plot (right) indicate a higher contribution in ProbWind.

A screenshot of a computer

Description automatically generated

Figure 10: The probability of a non-severe storm (left), probability of severe wind from a linear-type storm (center), and the ratio of severe wind probability to non-severe probability, conditional on ENI flash rate and 98th percentile MLAzShear. The left and center images are lookup tables in ProbWind (linear). The larger values in the ratio plot (right) indicate a higher contribution in ProbWind.

A close up of a map

Description automatically generated

Figure 11: The probability of severe wind gusts for a storm conditional on the computed cellular and linear NBC probabilities. This is the final lookup table for ProbWind.

A close up of a mans face

Description automatically generated

Figure 12: 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.

A close up of a map

Description automatically generated

Figure 13: The conditional probability of a tornadic storm (red) and severe, non-tornadic storm (blue), given its maximum 0-2 km AGL AzShear. The ratio of the severe and non-severe probabilities (black) indicates the contribution of this predictor in ProbTor.

A screenshot of a cell phone

Description automatically generated

Figure 14: The probability of a non-tornadic, severe storm (left), probability of a tornadic storm (center), and the ratio of tornadic probability to non-tornadic probability, conditional on 98th percentile LLAzShear and 0-1 km AGL storm-relative helicity. The left and center images are lookup tables in ProbTor. The larger values in the ratio plot (right) indicate a higher contribution in ProbTor.

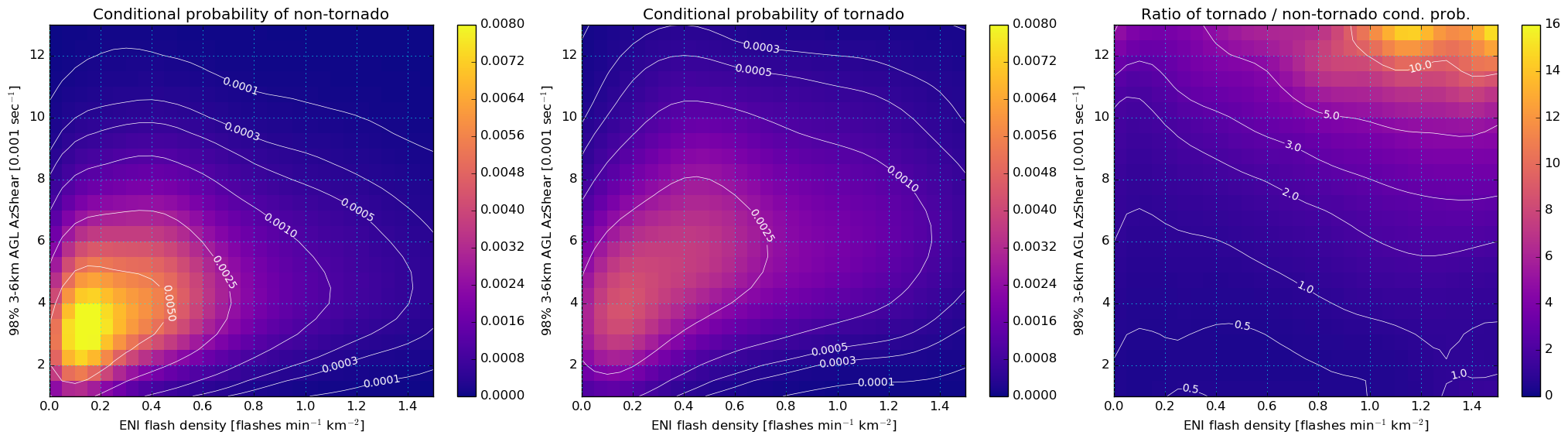


Figure 15: The probability of a non-tornadic, severe storm (left), probability of a tornadic storm (center), and the ratio of tornadic probability to non-tornadic probability, conditional on ENI flash density and 98th percentile MLAzShear. The left and center images are lookup tables in ProbTor. The larger values in the ratio plot (right) indicate a higher contribution in ProbTor.

A close up of a logo

Description automatically generated

Figure 16: The probability of a non-tornadic, severe storm (left), probability of a tornadic storm (center), and the ratio of tornadic probability to non-tornadic probability, conditional on effective bulk shear and mean wind 1-3 km AGL. The left and center images are lookup tables in ProbTor. The larger values in the ratio plot (right) indicate a higher contribution in ProbTor.



Figure 17: Left: The maximum CSI for ProbHail (PSv2 – hail), ProbWind (PSv2 – wind), ProbTor (PSv2 – tornado), and PSv1 scored against the hazard on the x-axis. The associated probability threshold, POD, and FAR are denoted above the bars. Right: A comparison of POD, FAR, and CSI for PSv2 (at 80% threshold) and NWS.



Figure 18: PSv2 forecast calibration from 2017 and 2018 (left y-axis) and forecast probability frequency (orange bars; right y-axis).



Figure 19: The maximum CSI per model and per month of 2018. Gray bars denote the number of severe storms in the dataset for a given month. Note that “probSevere” refers to the PSv2 model used for any severe hazard.

A close up of a map

Description automatically generated

Figure 20: Top left: probability of detection (POD); top right: false alarm ratio (FAR); bottom left: critical success index; bottom right: most skillful probability threshold for PSv2. The POD, FAR, and CSI correspond to this threshold. These scores are aggregated for NWS county warning area (CWA) boundaries and include adjacent CWAs (see text).

References

<https://www.weather.gov/lmk/supercell/dynamics>

<https://journals.ametsoc.org/doi/pdf/10.1175/1520-0434%282004%29019%3C0714%3AEOCALA%3E2.0.CO%3B2>

1. See the following link for a description of these MRMS fields. <https://www.nssl.noaa.gov/projects/mrms/operational/tables.php> [↑](#footnote-ref-1)