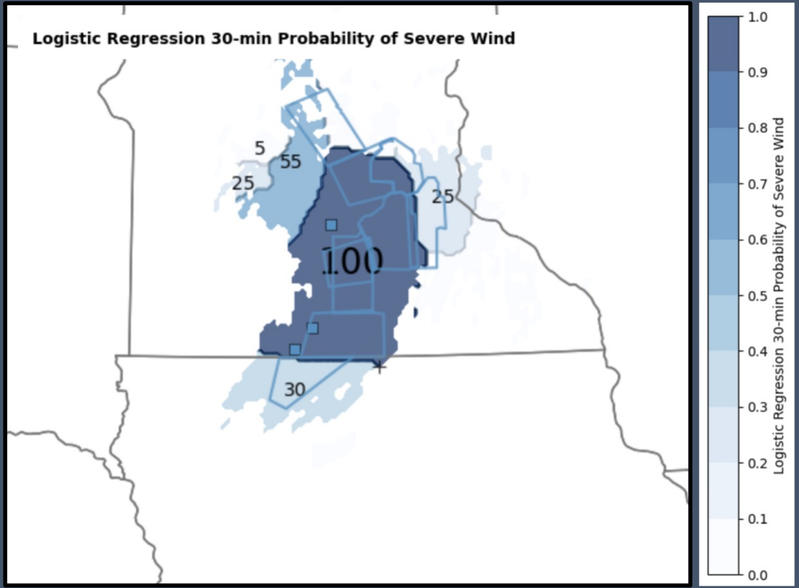


The Role of AI and machine-learning in the Hazardous Weather Testbed

Adam Clark and Corey Potvin, NSSL

AI4NWP Workshop - 28-29 November

***with important contributions from Thea Sandmael, Kristin Calhoun, and Monte Flora, and Eric Loken**

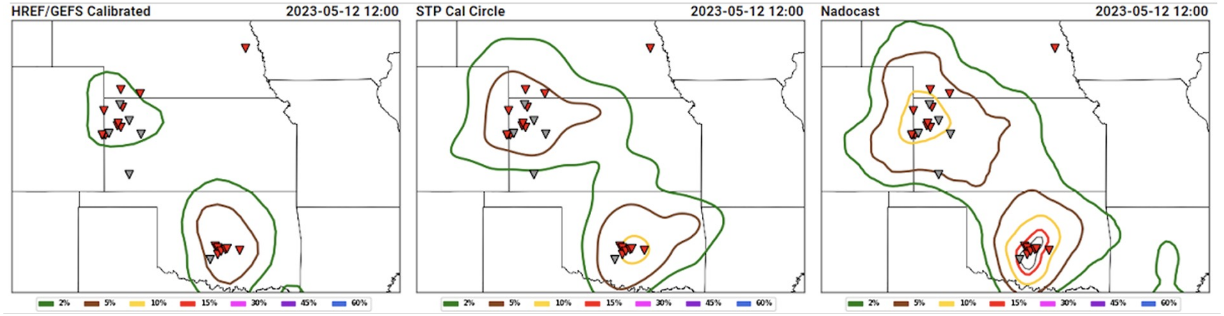
CIWRO
Storm Prediction Center
Norman, Oklahoma

Introduction & Motivation

- AI is a particularly powerful tool for severe weather applications because severe weather occurrence is NOT explicitly predicted by NWP models.

- Similar to human forecasters using ingredients-based methods for severe weather forecasting, AI can use the same ingredients (i.e., predictors) to produce skillful and reliable severe weather probabilities at a range of time and space scales.

- During SFE 2023, 12 different evaluation activities assessed a mix of AI-based & non-AI-based methods for producing calibrated hazard probabilities. In every activity, an AI-based product was the most skillful.



Tornado probabilities (contours) and observed tornadoes. An AI-based technique called "Nadocast" was the top rated tornado product.

History of AI in the SFE HWT

Initial attempts to use AI were not as good as traditional calibration techniques (e.g., look up tables, STP-based predictions, etc.). It took multiple R2O - O2R feedback cycles to get products that were useful to forecasters. HWT feedback extremely important because these products had never been seen before.

<u>SFE timeline of AI/ML assessments:</u>

- 2016-19 (1 project): ML-based hail predictions based on Gagne et al. (2017) & Burke et al. (2019).
- 2020 (4 projects): Gagne/Burke hail; Iowa State University (ISU) wind reports project; Loken random forest Day 1 hazard probs (Loken RFs); NCAR deterministic ML (RFs and NNs)
- 2021 (6 projects): WoFS-ML (Flora); ISU wind reports; Loken RFs; NCAR convective mode probabilities; NCAR HRRR-based ML probs; GEFS-ML (Colorado State University - Hill et al. 2023)
- 2022 (8 projects): WoFS-ML; ISU wind reports; Loken RFs; GEFS-ML; NCAR convective mode guidance; County-based watch guidance (HREF-based; SPC); Nadocast (Hempel @ SPC); flow-dependent ML (A. Johnson)
- 2023 (8 projects): WoFS-ML; WoFS-PHI (Loken); ISU wind reports; Loken RFs; GEFS-ML; GEFS-ML operational (Clark/Hoogewind); Nadocast; NCAR HRRR-based ML

- Rapid growth! Especially last 3 years.

How is AI used in the SFE?

- Generation of calibrated hazard guidance
    - **Grid-based:** Gridded sets of predictors input into algorithms that output grids of hazard probabilities. Input can come from deterministic, ensemble, CAM, and/or non-CAM systems.
    - **Object-based:** Storm-objects identified in CAMs, and properties of these objects and near-object environments are used to generate probabilities that these objects were produce a specified hazard.

- Generation of convective mode guidance
    - Storm objects identified & storm attribute fields and shape characteristics used to objectively assign mode (e.g., line, supercell, unorganized cluster, etc.).

- Enhancing local storm report (LSR) database: Project led by ISU used ML algorithms to find the probability that wind reports result from gusts exceeding severe criteria (i.e., ≥ 50 knots). LSRs from non-severe wind - very common in the east and southeast US - can be filtered out.

AI Success Stories: GEFS-ML

A New Paradigm for Medium-Range Severe Weather Forecasts: Probabilistic Random Forest–Based Predictions

Aaron J. Hill, Russ S. Schumacher, and Israel L. Jirak

- Hill et al. (2023) developed a random-forest model for generating severe weather probabilities from the GEFS for Days 1-8.
- GEFS reforecasts were used to train and test their model, which used operational GEFS forecasts as input.
- These forecasts run in real-time and are used by SPC.
- Performance was so good that it motivated more aggressive Days 3-8 convective outlooks starting in 2022.
- One limitation is that - due to computational limitations - the GEFS reforecasts only included 5 members.
- ***With GEFSv12 being operational since 2020, it may be possible to leverage all 31 GEFS members to get an improved result.***

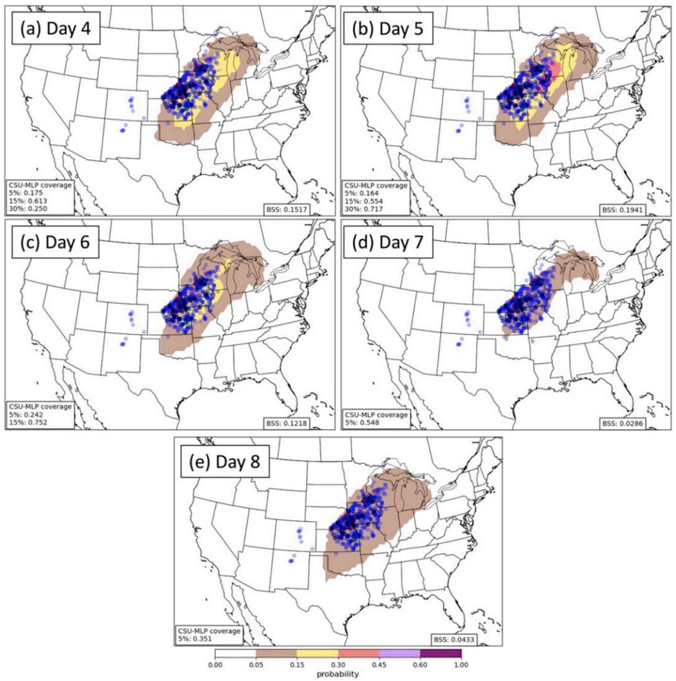


FIG. 14. Day-4–8 probabilistic CSU-MLP forecasts for any severe hazard valid at 1200 UTC 15 Dec 2021–1200 UTC 16 Dec 2021. NWS local storm reports for wind, hail, and tornadoes are included as blue, green, and red circles, respectively. Observation coverage and BSS are included in bottom-left and bottom-right corners of each panel, respectively.

AI Success Stories: GEFS-ML (methodology)

GEFS operational ML (developed by Clark & Hoogewind)

- Random Forest (RF) model trained with operational GEFS data; conceptually similar to Hill et al. (2023)
- Probabilistic forecasts for any severe weather report occurring within an 80-km grid-box are generated for Days 1-15.
- Separate RFs are configured for each lead time using 18 predictors extracted from the mean of 31 operational GEFS members (0000 UTC initializations) at 3-hourly output intervals, which are remapped to the 80-km NCEP 211 grid.
- For tuning & feature engineering (i.e., optimizing input format), *k*-fold cross validation was used to generate forecasts from 656 cases (7 folds; 84 cases/fold) covering the period 3 March 2021 to 1 February 2023.
- For real-time forecasting, the RF is trained from all 656 past cases.

Table 1 List of the 18 predictors used in GEFS operational ML.

GEFS Operational ML Predictors		
(1) Bulk Shear (0-1 km AGL)	(7) Lapse Rate (700-500 mb)	(13) Temperature (2-m AGL)
(2) Bulk Shear (0-3 km AGL)	(8) Surface-based LCL	(14) Precipitation (3-h accum.)
(3) Bulk Shear (0-6 km AGL)	(9) Sig. Tornado Parameter	(15) u-wind (10-m AGL)
(4) Surface-based CAPE	(10) Mean-sea-level pressure	(16) v-wind (10-m AGL)
(5) Surface-based CIN	(11) Precipitable water	(17) Wind magnitude (10-m AGL)
(6) Storm relative helicity (0-3 km AGL)	(12) Specific humidity (2-m AGL)	(18) Most unstable CAPE

AI Success Stories: GEFS-ML (results)

- During the 2023 SFE, participants rated (1-10 scale) the quality of severe weather guidance from **GEFS reforecast ML** and **GEFS operational ML** at each lead time from Days 3-7.
- In each evaluation, a single valid time is displayed to show the evolution of forecasts with lead time.

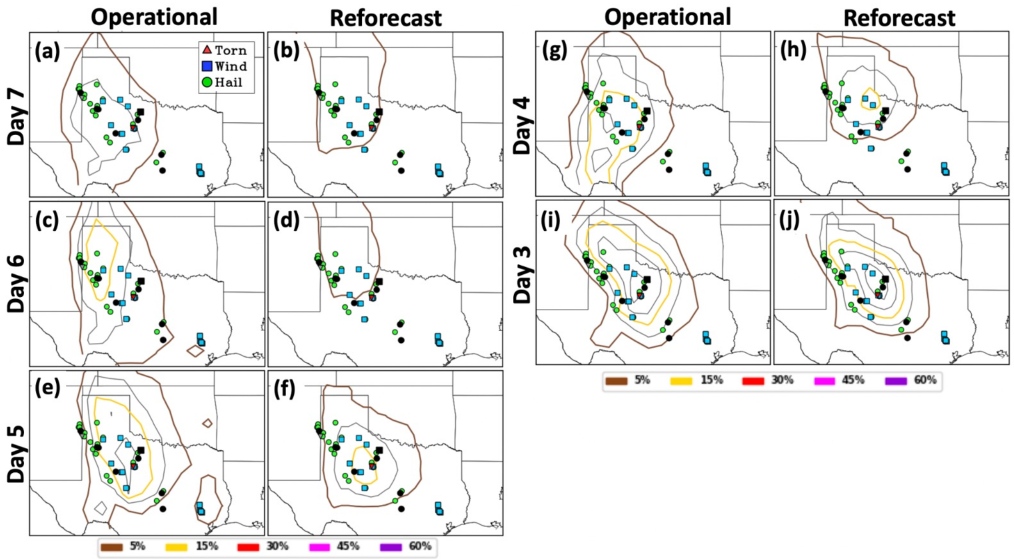
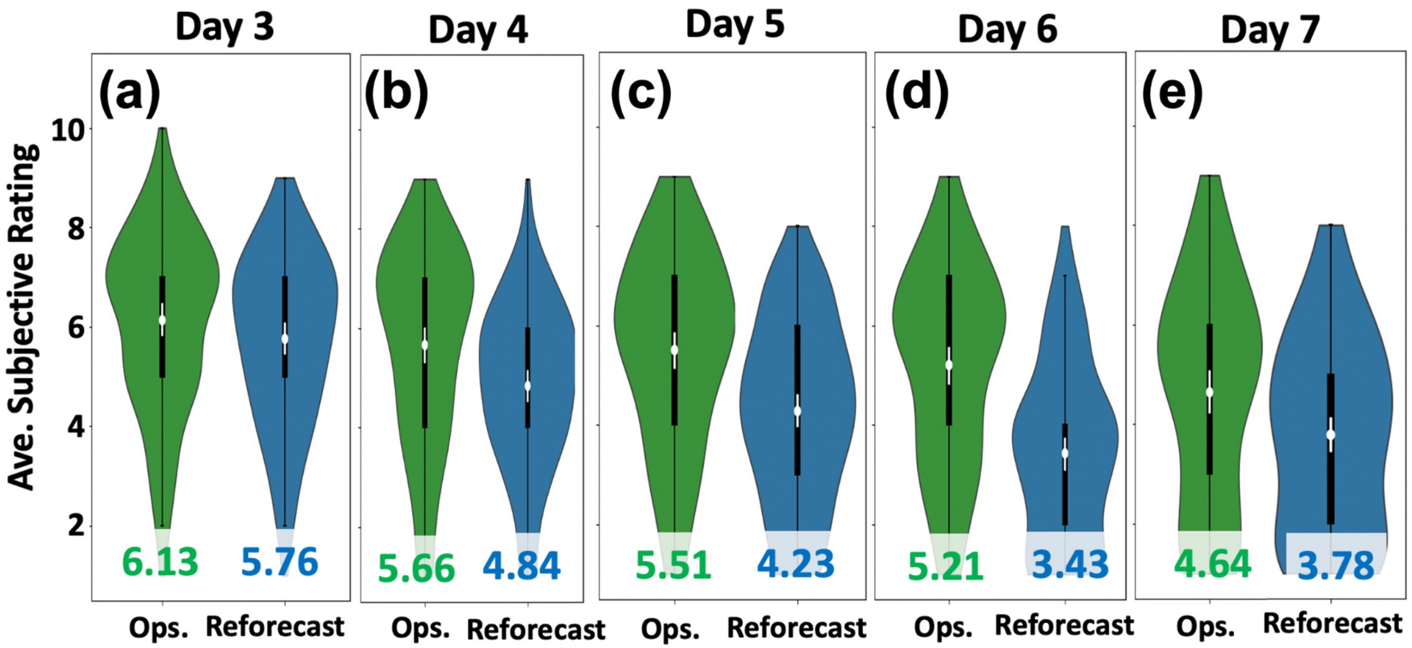


Figure 1 Severe weather probabilities at Day 7 lead time from (a) GEFS operational ML, & (b) GEFS reforecast ML. (c)-(d), (e)-(f), (g)-(h), and (i)-(j), same as (a)-(b), except for lead times of 6, 5, 4, & 3 days, respectively. Locations of observed storm reports are overlaid.

Example from 23 May 2023

- GEFS operational ML tends to generate higher probabilities at longer lead times that often correspond quite well to observed severe weather.

AI Success Stories: GEFS-ML (results)



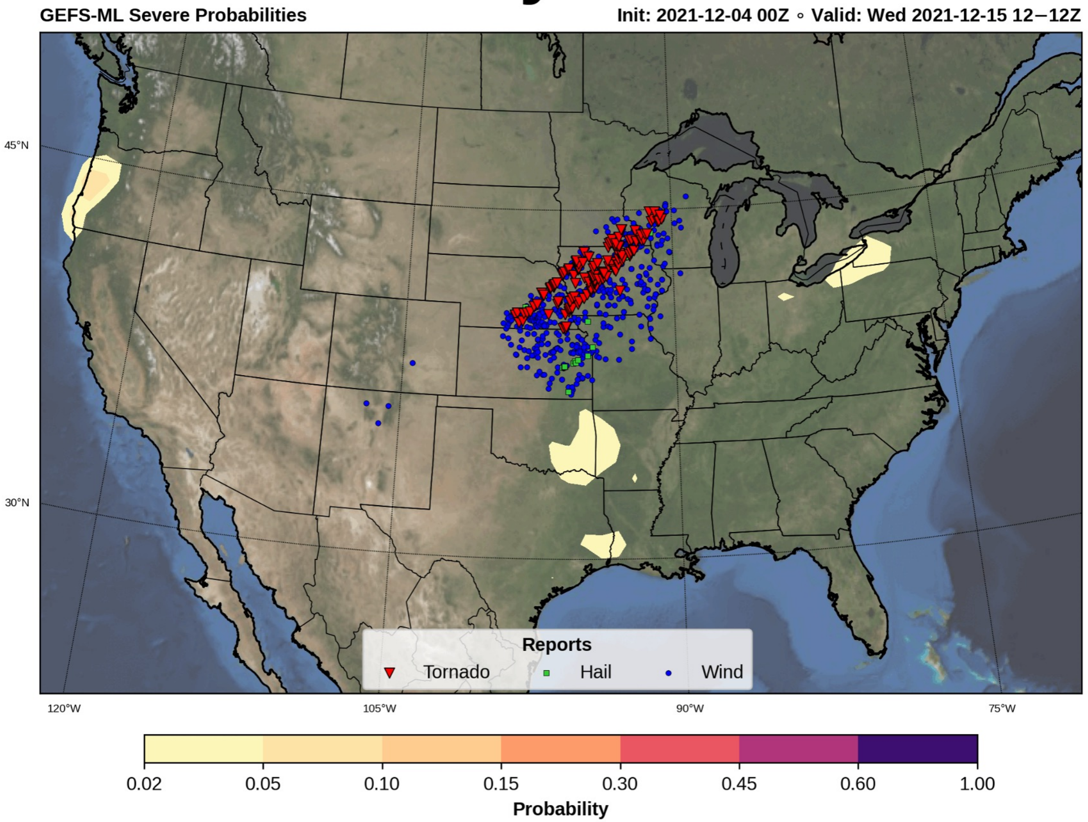
(a) Violin plots indicating the Day 3 lead time distributions of subjective ratings for GEFS Operational ML (green) & GEFS reforecast ML (blue). (b)-(e) same as (a), except for Day 4-7 lead times, respectively. The numbers at the bottom of each violin plot indicate the mean subjective ratings.

At each lead time, GEFS operational ML was clearly the best performing algorithm with statistically significant differences at every lead time examined.

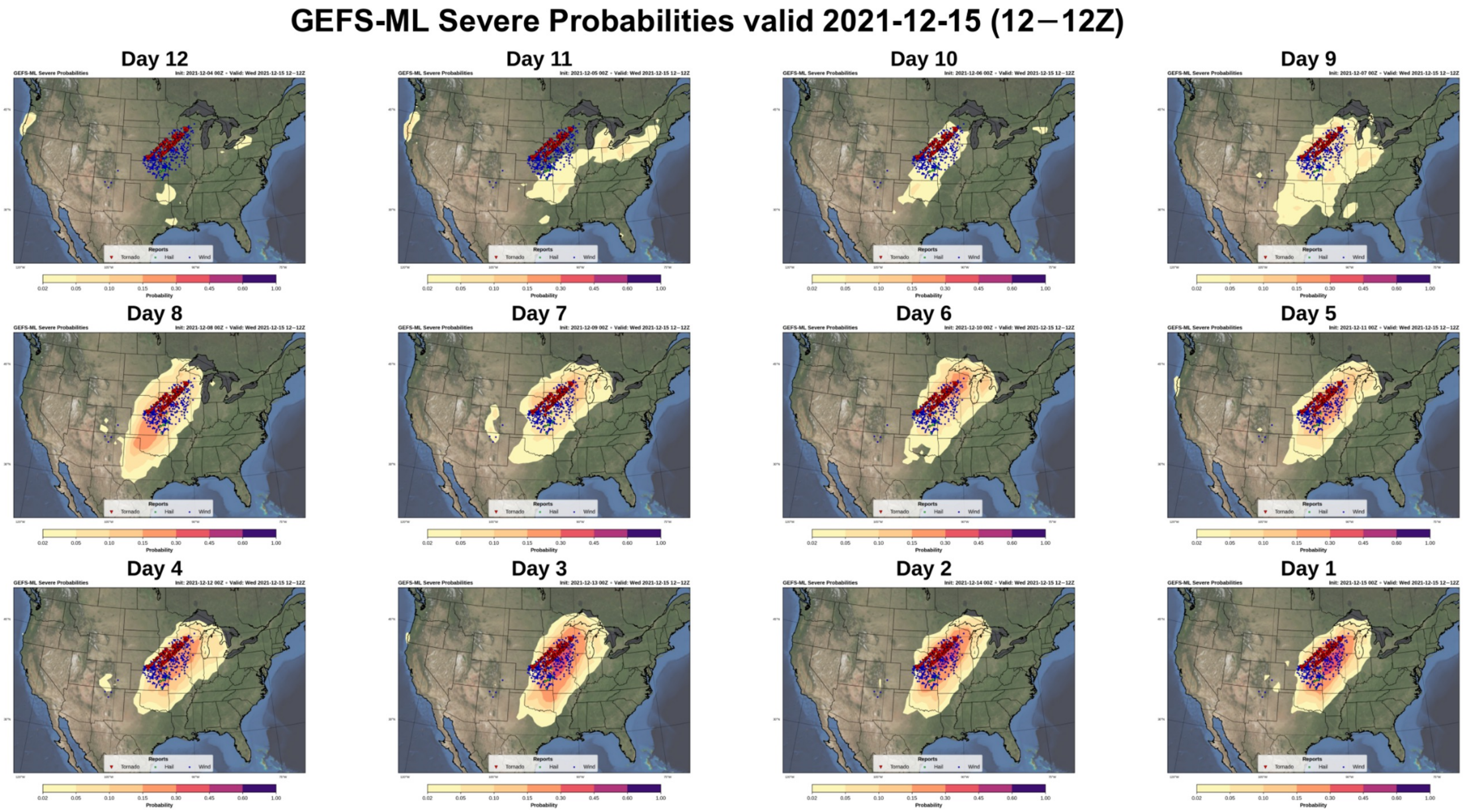
AI Success Stories: GEFS-ML (results)

Day 12

Example case:
15 December 2021



AI Success Stories: GEFS-ML (results)



Loken RF: HREFv3-based Day 1 and 2 hazard probabilities

Raw HREFv3 data → Period max/min → Upscale to 80 km grid → Ensemble mean → Predictors from multiple points, plus lat/lon

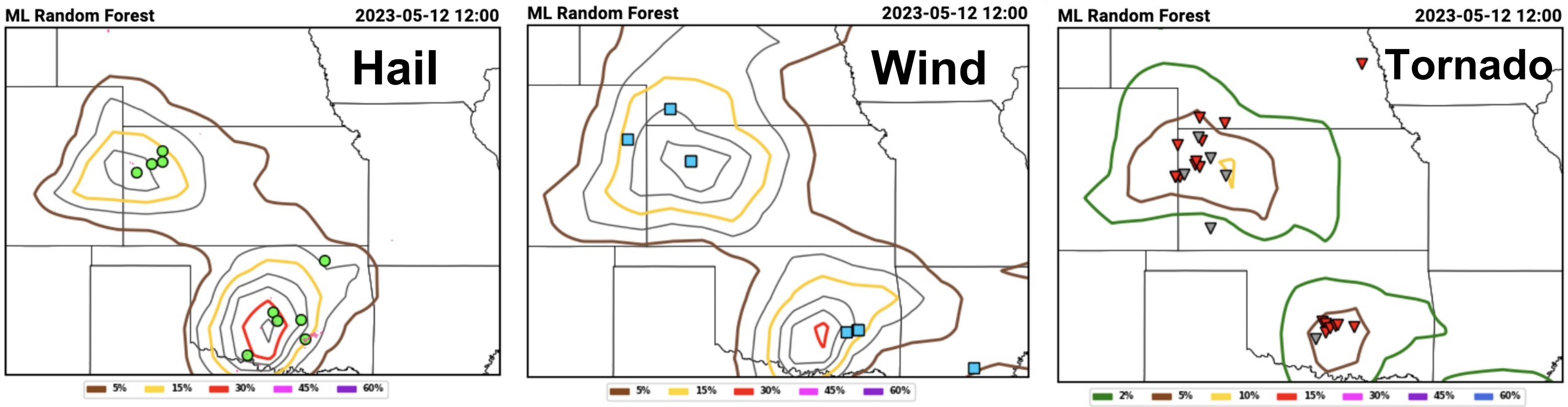
↓

Input to Random Forests

↓

Hazard probabilities

Example Real-time Day 1 Forecast: Valid 24h ending 12z 12 May 2023

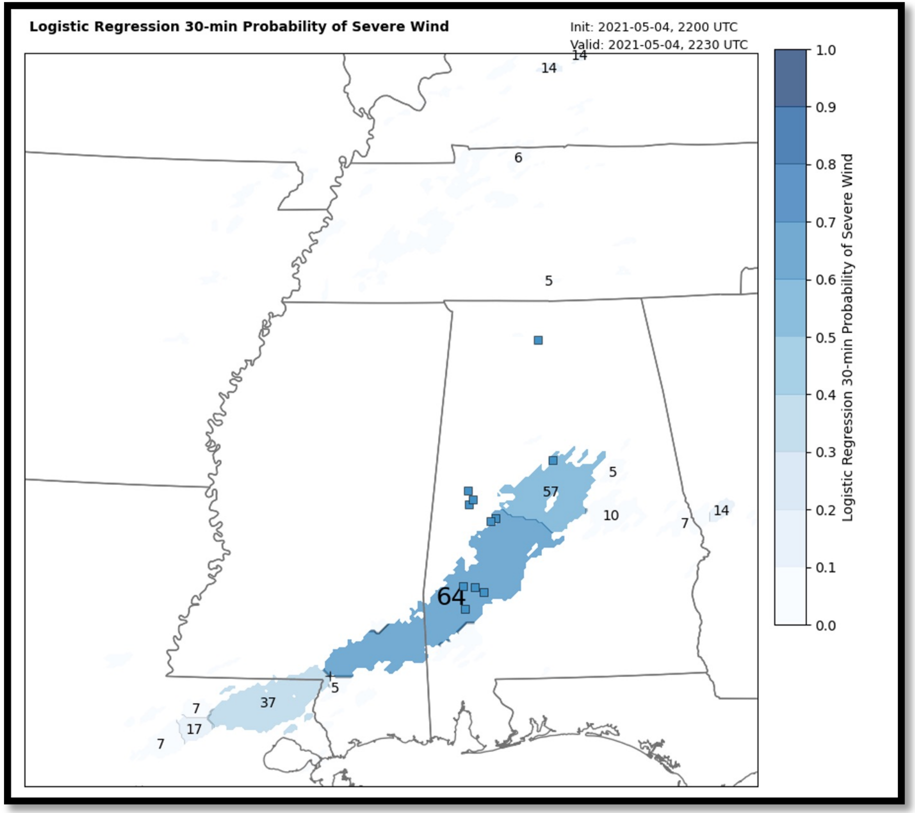


Object-based ML severe guidance for WoFS (Monte Flora)

Create objects from ensemble member forecasts of updraft tracks (30-min swaths)

Train logistic regression models to predict probability of severe report of each type (wind/hail/tor) occurring within each object

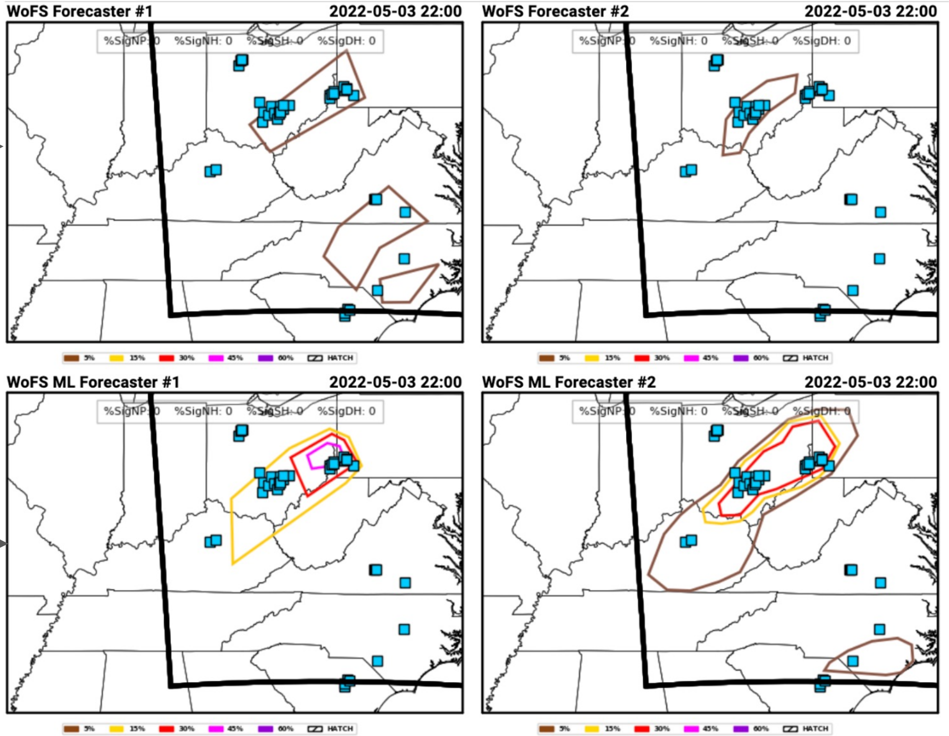
[Flora et al. (2021, *MWR*)](https://doi.org/)



Incorporating object-based ML led to subjectively better 1-h forecast outlooks in SFE 2022

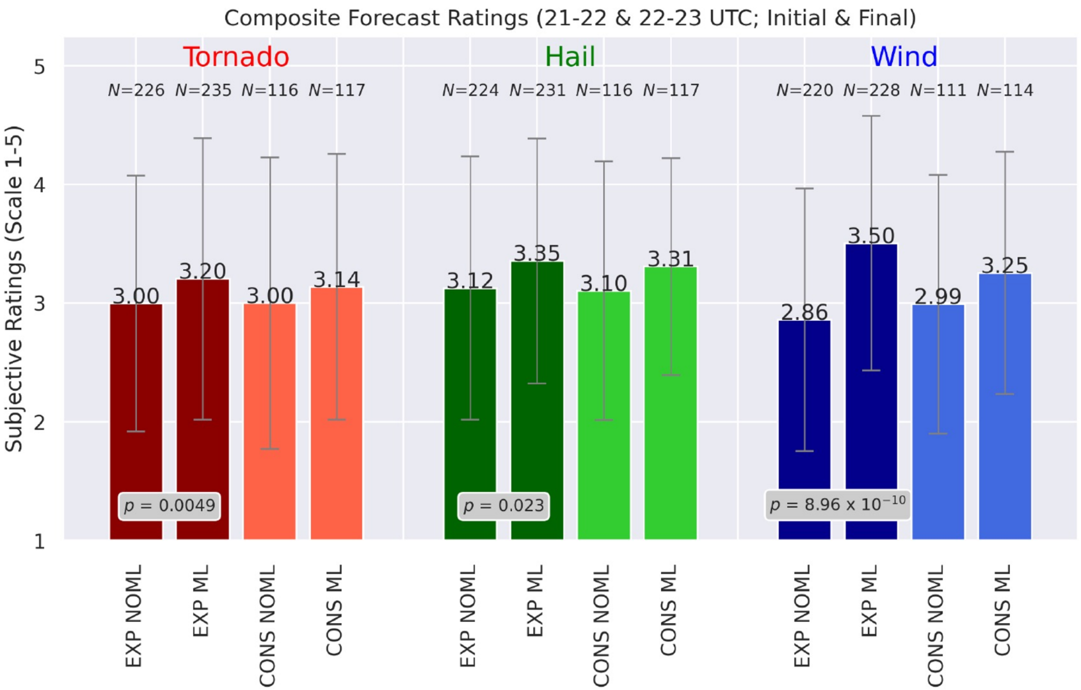
These expert forecasters had access to the full WoFS suite

These expert forecasters had access to the full WoFS suite + ML guidance



Flora et al. (2024, *WAF*, cond. accepted)

Incorporating object-based ML led to subjectively better 1-h forecast outlooks in SFE 2022 (cont.)



Clark et al. (2023, *BAMS*), Flora et al. (2024, *WAF*, cond. accepted)

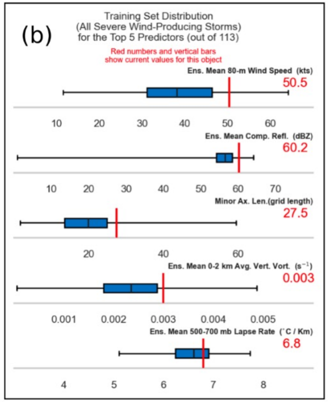
Object-based ML Explainability Products

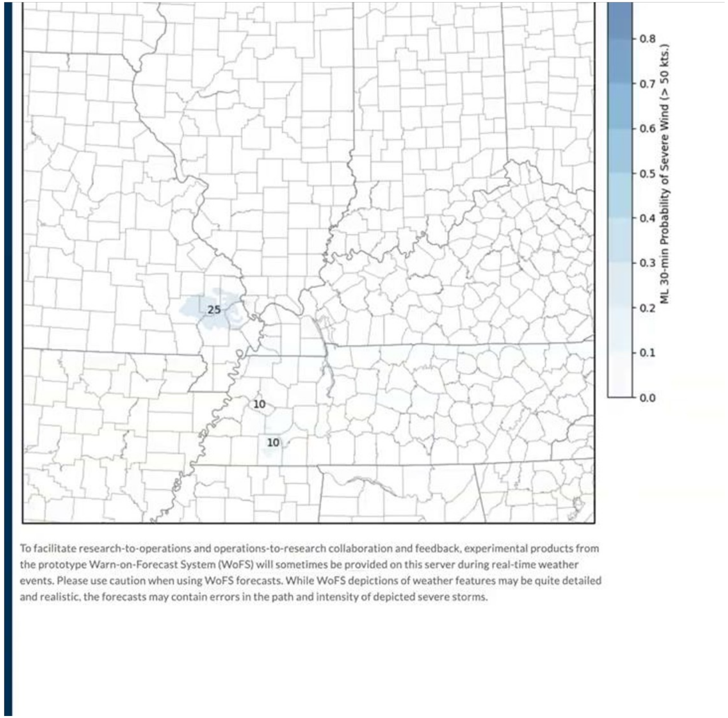
Clicking on storm object pops up explainability graphic

For each of most important 5 predictors, shows training set distribution for severe storms and the value for clicked storm

How does this storm compare to previous WoFS storms that overlapped severe reports?

Explainability Example



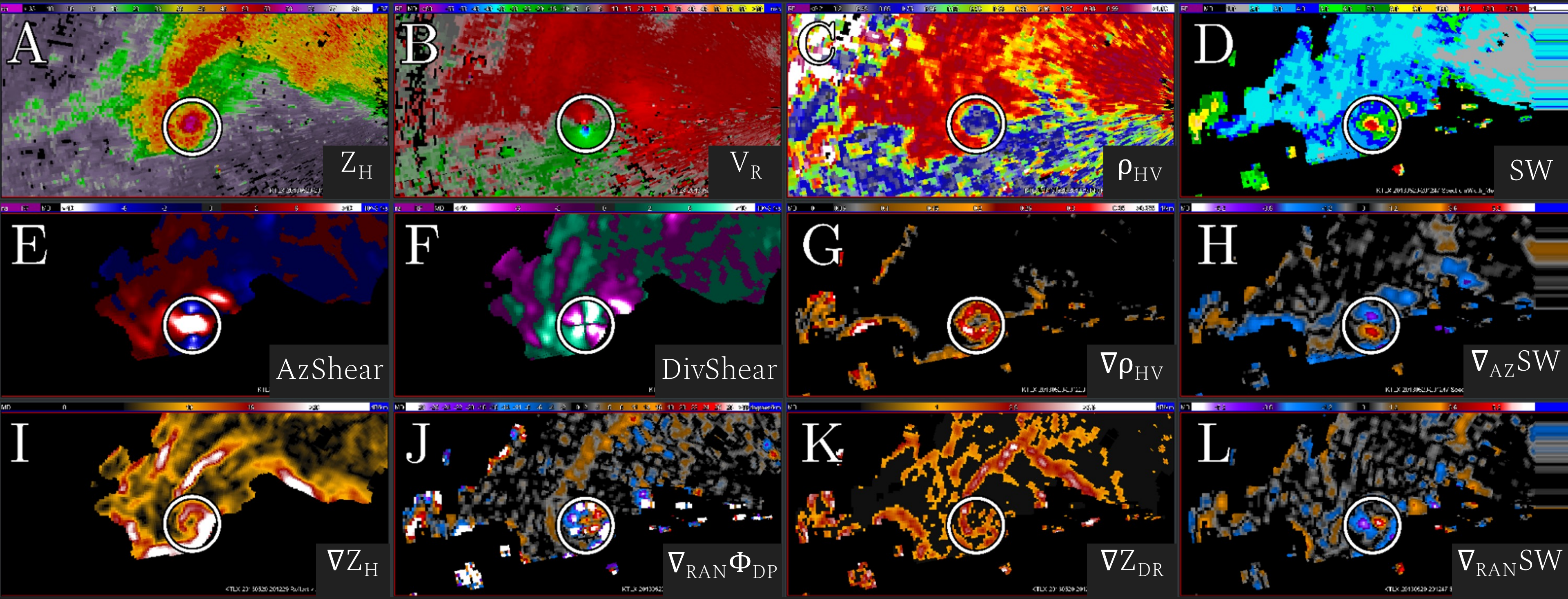


Flora et al. (2024, *AIES*) – “A Machine Learning Explainability Tutorial for Atmospheric Sciences”

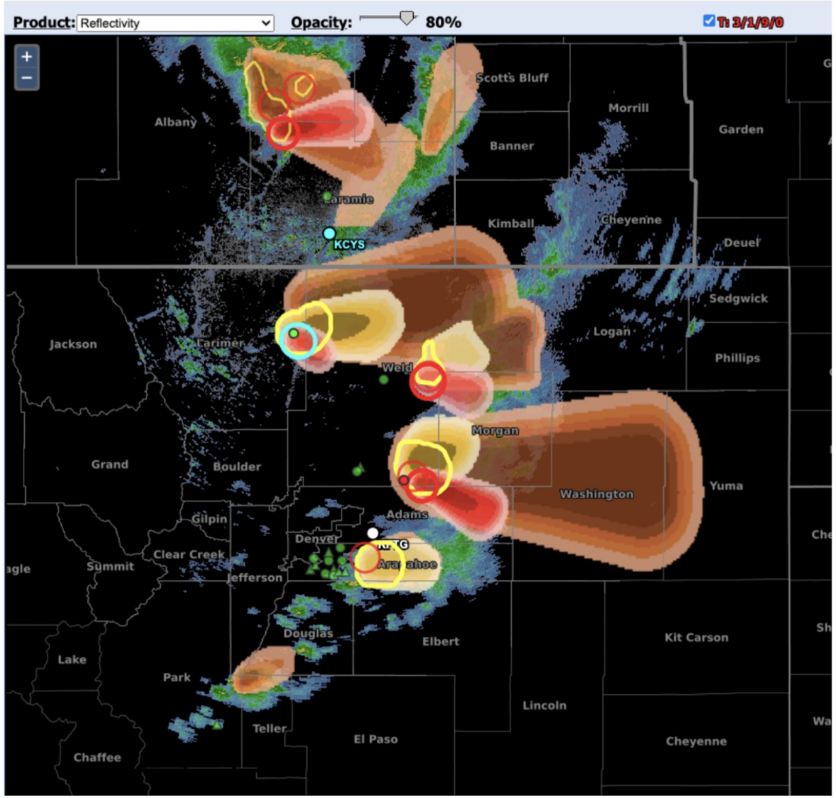
AI in the EWP

- Experimental Warning Program (EWP) focuses tools for issuing warnings (i.e., 0-1 h lead times).
- AI/ML algorithms use radar products combined with environmental analyses to generate probabilities that a given storm will produce a hazard (tornado, wind, or hail).
- Recent experiments have tested TORP (Tornado Probability Algorithm; Sandmael et al. 2023)

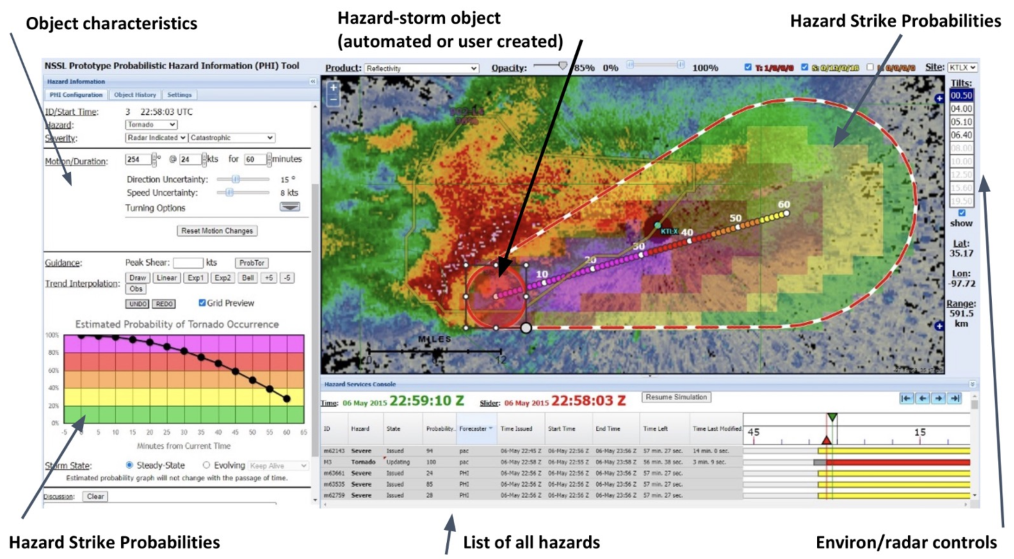
Predictors


A
Z_H
B
V_R
C
ρ_{HV}
D
SW
E
AzShear
F
DivShear
G
$\nabla\rho_{HV}$
H
$\nabla_{AZ}SW$
I
∇Z_H
J
$\nabla_{RAN}\Phi_{DP}$
K
∇Z_{DR}
L
$\nabla_{RAN}SW$

AI in the EWP

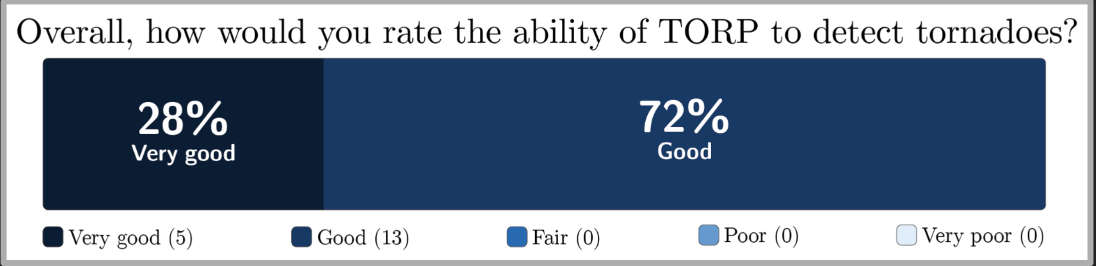


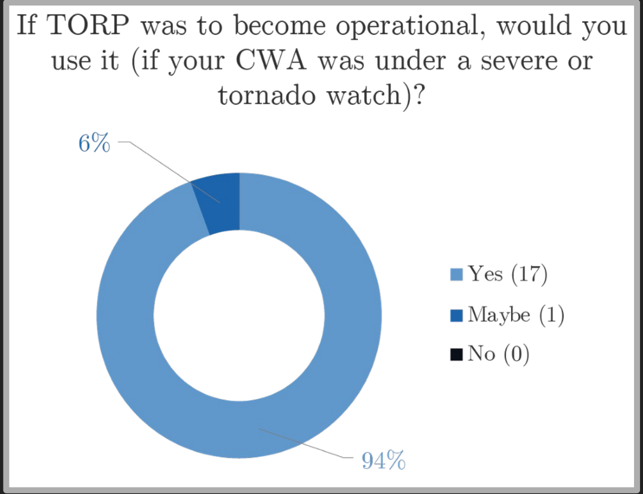
Automated “PHI plumes”



Forecaster Impressions

- “I came into this week a bit skeptical as I've found the legacy version not very useful. Well, after a week of using the [TORP] in various geographic regions and with different storm types, I am very impressed with this new version and it has exceeded my expectations.”
- “I have no reservations about the [TORP] becoming operational with only minor revisions.”
- “The probabilities generated through the random forest analysis were very useful, but the false alarm detections could be distracting in cases with several storms.”
- “Maybe”: A more organized readout to better utilize the information and additional filters for non-meteorological detections.





Future AI Applications in the HWT

- We will continue to use AI as a post-processing/calibration tool.

- Need to find ways that forecasters can understand the AI tools so they aren't a "black box". Recent work has explored use of "explainability" graphics. Forecasters like this concept, but have not found it useful in their forecasting.

- Find the right way to design ML algorithms is an art. "Feature engineering" refers to how predictors are configured, of which there are endless ways. Need a mix of computer science and meteorology knowledge.

- Quality of observations is a huge limitation for ML algorithms focused on severe storms. Need ways to enhance/supplement storm report database.

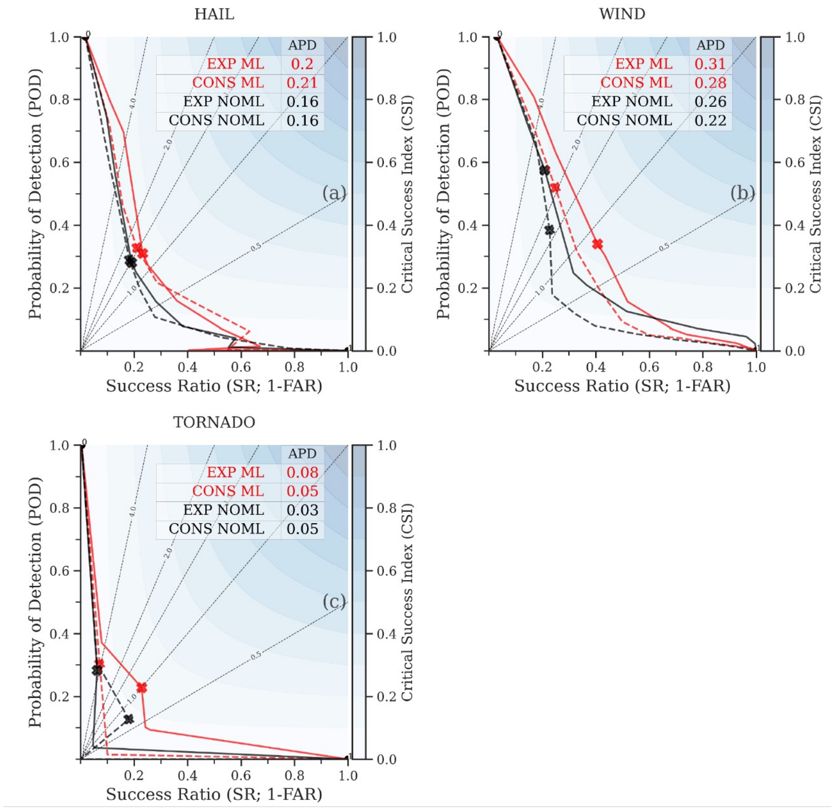
- We haven't yet explored the new frontier of pure AI-based forecasts (i.e., Pangu-Weather). Skill metrics show these forecasts outperform the best global models, but ***we need to get these products in front of forecasters to measure their "true" value! The HWT is ideal for this testing.***

Extra slides

Incorporating object-based ML led to *objectively* better 1-h forecast outlooks in SFE 2022

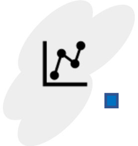
Participants with access to ML products ("ML") issued objectively better forecasts than participants w/o access to ML ("NOML")

True for both expert forecasters ("EXP") and non-experts ("CONS")

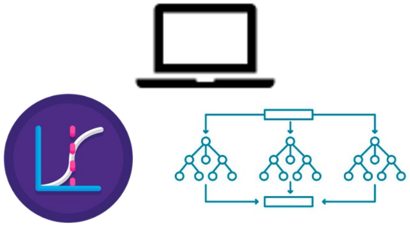


Object-based Severe Weather Guidance (Monte Flora)

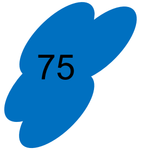
Extract WoFS data and observed severe wx reports from storm tracks



Train ML models (e.g., Log. Regression and Random Forest)

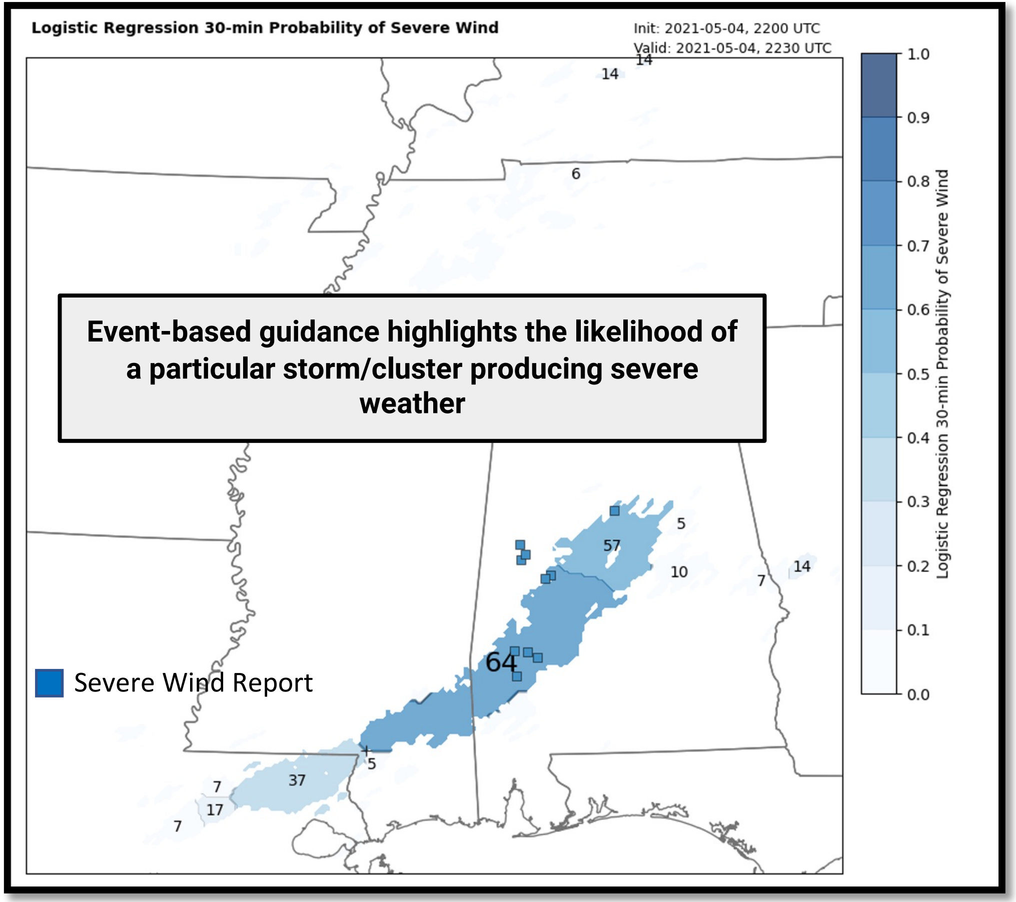


Assign future tracks a probability of a report occurring within it

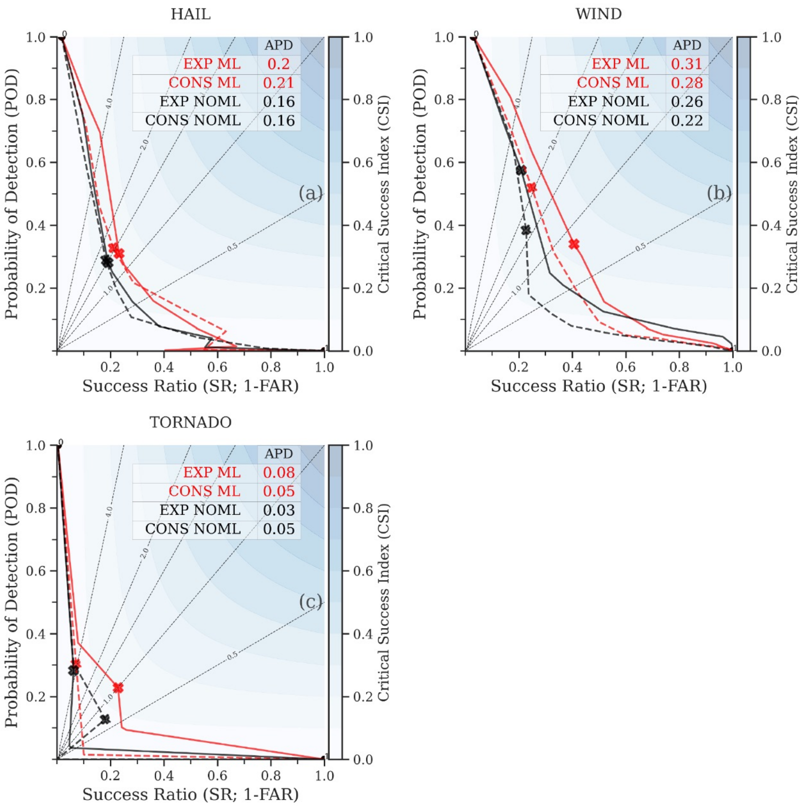


Evaluate in real-time and retrospective cases

Flora et al. (2021)



Flora et al. (2024, *WAF*, cond. accepted)



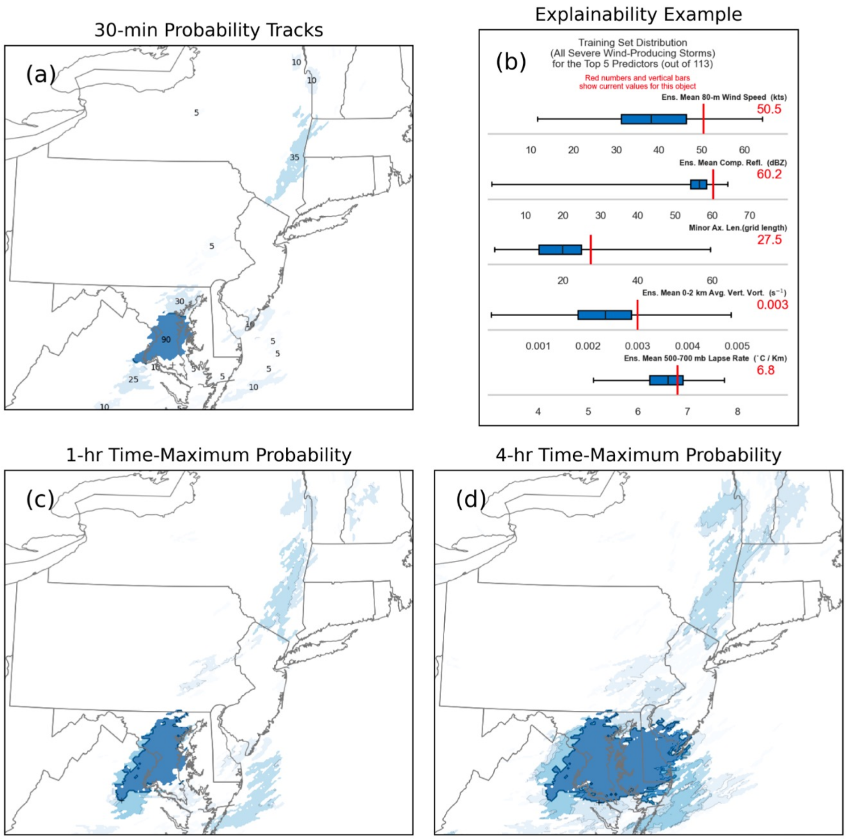
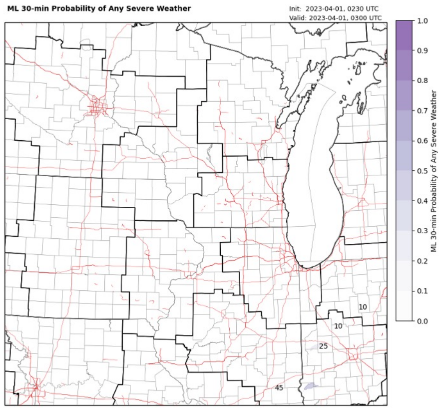


FIG. 2. Examples of the available ML products on the real-time WoFS web viewer. These include (a) 30-min guidance, (b) the interactive ML explainability product, (c) 1-hr guidance, and (d) the 4-hr guidance. For the 30-min guidance (a), probabilities rounded to the nearest 5% are overlaid. The 1-hr and 4-hr products are time-maximum composites of the 30-min guidance. These graphics come from the cloud-based WoFS webpage

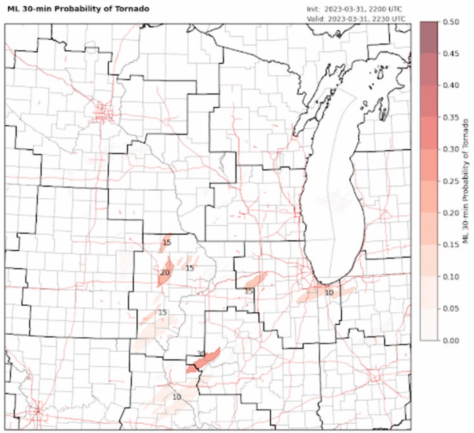
Forecasters share the ML guidance in real-time operations



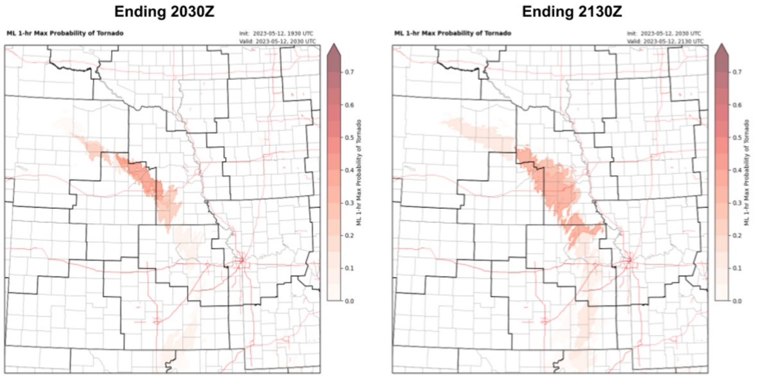
NWS Forecaster

"Up in IWX's neck of the woods, WoFS ML probabilities suggest that the potential for any severe weather the rest of the night is very low…"

"Machine learning tornado probs do bring high values towards the Quad Cities *themselves, while our mode may be changing, the QLCS environment and observed mesovorticies line up well with these signals"*



1830Z WoFS Machine Learning Max 1 Hour Tornado Probability



"The WoFS Machine Learning Tor Probs continue to remain elevated (graphic generated by forecaster)"

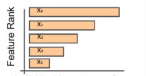
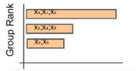
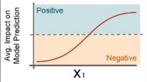
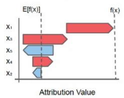
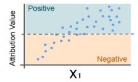
These graphics and feedback come from a NWS-WoFS google chat room & Southern Region Remote Mesoanalysis google chat rooms

ML Explainability

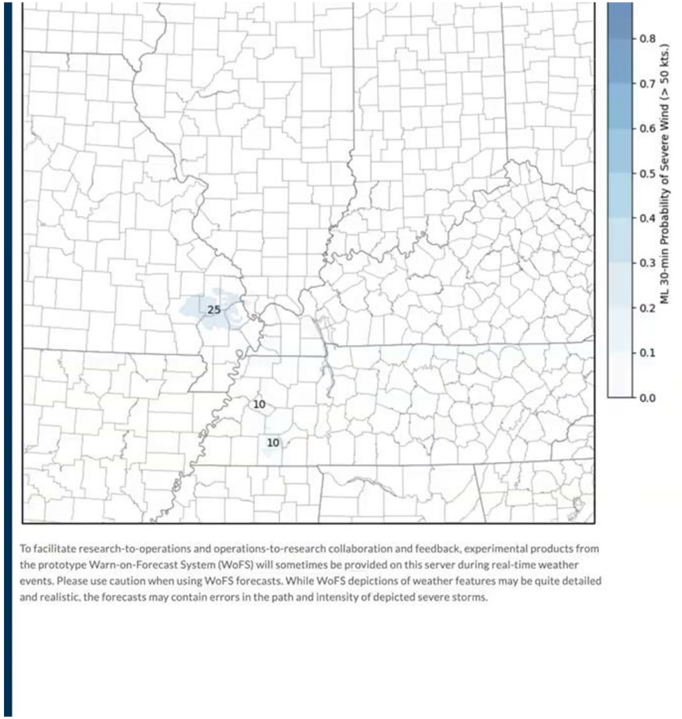
A Machine Learning Explainability Tutorial for Atmospheric Sciences

Montgomery L. Flora[a,b,e], Corey K. Potvin,[b,c,e], Amy McGovern[c,d,e] Shawn Handler[a,b,*],

https://doi.org/10.1175/AIES-D-23-0018.1

Grouping	Explainability Methods	Key Ideas	Visualizations	Scope
Feature Importance	Single-Pass Permutation Importance	**Measures:** feature importance by permuting (backward)/unpermuting (forward) features one at a time **Pros:** Quick to compute; parallelizable; model-agnostic **Cons:** Highly sensitive to correlated features and does not account for multivariate relationships between features.	 Individual Importance Scores	Global Explainability
Feature Importance	Grouped Permutation Importance	**Measures:** feature importance by permuting /unpermuting multiple features at a time **Pros:** parallelizable; model-agnostic; manually defined groups are highly understandable; for mutually exclusive groups grouped importance is quick to compute; includes feature co-dependencies when computing importance. **Cons:** Automatically defining feature groups is difficult; does not replace single-pass permutation importance	 Grouped Importance Scores	Global Explainability
Feature Importance	Shapley Additive Global Importance (SAGE)	**Measures:** feature importance using Shapley theory; unifies single-pass and grouped permutation importance **Pros:** model-agnostic; global-based version of SHAP; computationally quicker than computing SHAP; unifies global feature importance methods **Cons:** SAGE is limited to loss-based metrics; it's a new method and package so documentation is lacking and knowledge of sensitivities is unknown.		Global Explainability
Feature Effects	Accumulated Local Effects (ALE) and Partial Dependence (PD)	**Measures:** global model sensitivity to a feature across the full range of its values. **Pros:** quick to compute; parallelizable; model-agnostic; ALE is less sensitive to correlated features than PD; both can be used for functional decomposition; both can be computed for higher-order interactions **Cons:** PD is sensitive to correlated features; ALE can be noisy or biased when sample size is low		Global Explainability
Feature Relevance / Feature Attributions	SHapley Additive Explanations (SHAP)	**Measures:** feature attributions using an approximate version of Shapely values **Pros:** model-agnostic; only method that assigns attributions fairly and satisfies certain desirable properties (e.g., additivity, missingness, etc); exact Shapely values for tree models (ignore decision paths with missing features) **Cons:** slower compute time for a large set of examples or features	Single Example 	Local Explainability
Feature Relevance / Feature Attributions	Local Interpretable Model-agnostic Explanations (LIME)	**Measures:** feature attributions using the coefficients of a local linear model **Pros:** model-agnostic; fast compute time **Cons:** attributions do not add to the model's prediction; sensitive to the accuracy of the local model approximation; assumes feature independence	Multiple Examples 	Local Explainability
Feature Relevance / Feature Attributions	Tree Interpreter	**Measures:** feature attributions using the path of a decision tree or forest **Pros:** quick to compute ; attributions add to the model prediction **Cons:** model-specific; can assign lower attributions to features higher in the tree; new method (sensitivities are relatively unexplored)		Local Explainability

WoFS Viewer explainability product demo



To facilitate research-to-operations and operations-to-research collaboration and feedback, experimental products from the prototype Warn-on-Forecast System (WoFS) will sometimes be provided on this server during real-time weather events. Please use caution when using WoFS forecasts. While WoFS depictions of weather features may be quite detailed and realistic, the forecasts may contain errors in the path and intensity of depicted severe storms.

Scikit-Explain

A user-friendly, open source Python package for traditional ML model explainability.

Scikit-Explain GitHub https://github.com/monte-flora/scikit-explain

Like to contribute to scikit-explain or have questions?
monte.flora@noaa.gov