Using Machine Learning to Predict Convection-Allowing Ensemble Forecast Skill: Evaluation with the NSSL Warn-on-Forecast System

Corey K. Potvin[1,2], Montgomery L. Flora[3,1], Patrick S. Skinner[3,1,2], Anthony E. Reinhart[1], Brian C. Matilla[3,1]

[1]NOAA/OAR/National Severe Storms Laboratory; [2]School of Meteorology, University of Oklahoma; [3]Cooperative Institute for Severe and High-Impact Weather Research and Operations

OVERVIEW

- Ensembles are underdispersive and have weak spread-error correlations
- Machine learning (ML) could improve objective guidance for forecast skill by incorporating information beyond forecast spread
- We have trained ML models that predict forecast skill for the NSSL Warn-on-Forecast System (WoFS)
- The ML models outperform rigorous baselines and motivate similar methods for "forecasting forecast skill" of larger-scale ensembles

METHODS

Data: WoFS 0-3-h forecasts and Multi-Radar / Multi-Sensor (MRMS) composite reflectivity (CREF) from > 100 days during 2017-2021 HWT Spring Forecasting Experiments.

Prediction task: For each MRMS storm at *t*=0, generate storm-centered 'initial domain' and 'forecast domain' centered on estimated future storm position in 1, 2, or 3 h. Using features extracted from initial & forecast domains (Fig. 1), predict accuracy of the WoFS CREF within the forecast domain: POOR, FAIR, or GOOD (Fig. 2).

Labels: Per forecast domain: (1) compute Extended Fractions Skill Score (eFSS) for several CREF thresholds & neighborhoods, then take mean; (2) convert mean eFSS to percentile based on dataset-wide climo; (3) POOR: < 20th percentile; FAIR: 20th-80th percentiles; GOOD: > 80th percentile.

Learning algorithms: Ordinal logistic regression (OLR), ordinal random forest (ORF). Ordinal methods respect class ordering (POOR → GOOD).

Baselines: Single-feature OLR models. *Persistence* BL uses CREF eFSS within initial domain. *Spread* BL uses ensemble stdev of max CREF within forecast domain (best spread metric among many examined).

Features: Started with 323, then reduced to 10 or 15 with little skill loss.

RESULTS

- OLR and ORF perform similarly (Fig. 3); we focus on simpler, faster OLR
- ML models substantially outperform baselines (Fig. 3)
- POOR is most skillfully discriminated class, followed by GOOD (Fig. 4)
- ML models are generally reliable (Fig. 5)
- Egregious predictions (e.g., POOR classified as GOOD) are rare (Fig. 6)
- *ML is a promising framework for "forecasting forecast skill" in the WoFS, and we expect this is true for much larger-scale models*

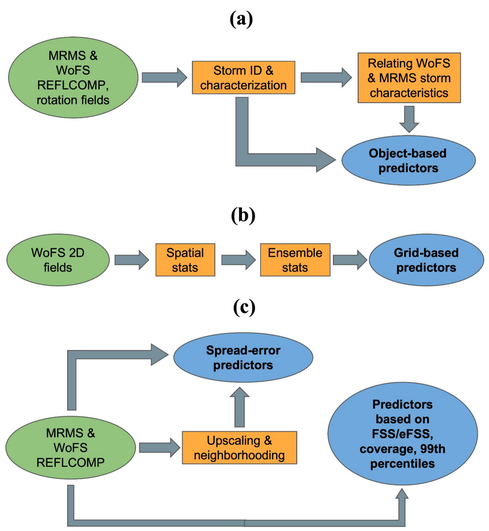

Fig. 1. Predictor generation process for (a) object-based predictors, (b) grid-based predictors, and (c) other predictors.

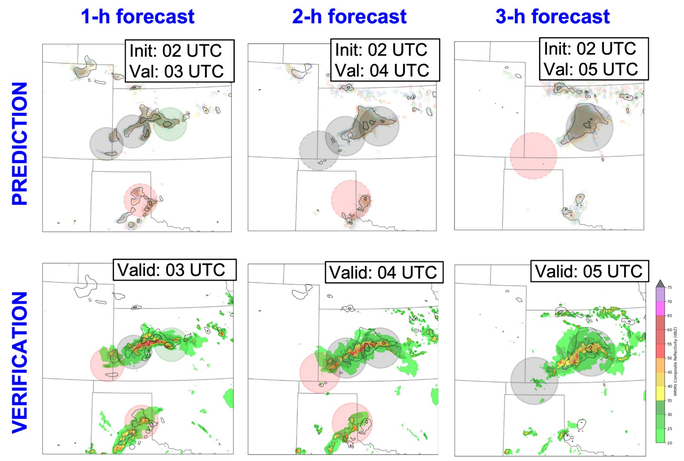

Fig. 2. ML predictions (top row) and verification (bottom row). Red = POOR, Gray = FAIR, Green = GOOD. WoFS member CREF > 45 dBZ paintballs shown in prediction panels; MRMS CREF in verification panels. WoFS probability-matched mean CREF contours of 30 dBZ, 50 dBZ shown in all panels.

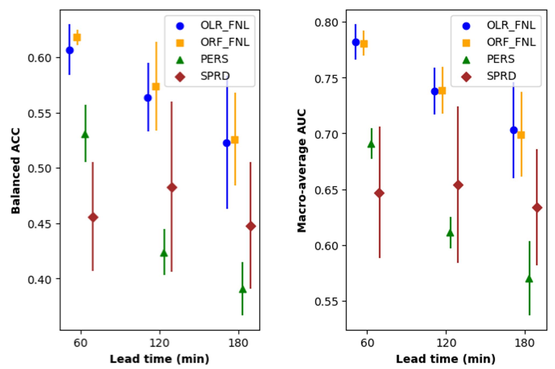

Fig. 3. Verification of the OLR and ORF models and of the PERS and SPRD baselines: (left) balanced classification accuracy and (right) macro-average AUC. Markers and bars represent the all-folds medians and standard deviations.

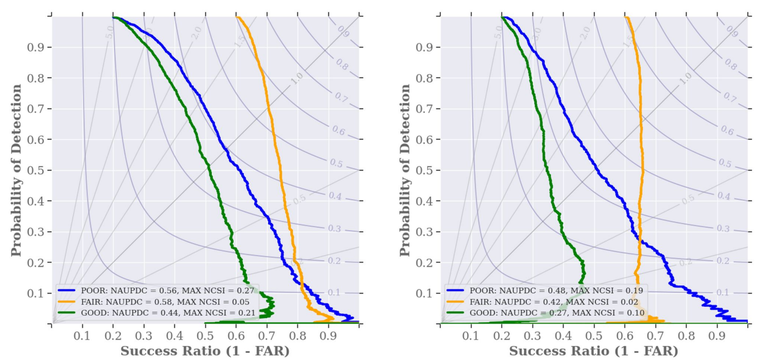

Fig. 4. Performance curves for OLR model at lead times of (left) 1 h and (right) 3 h. The normalized AUPC and maximum normalized CSI are listed.

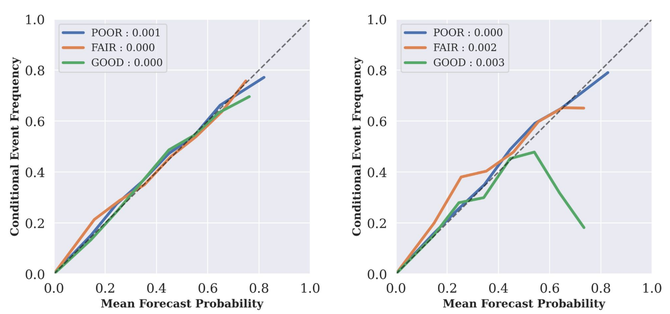

Fig. 5. Reliability curves at lead times of (left) 1 h and (right) 3 h. The reliability component of the BSS is listed for each class.

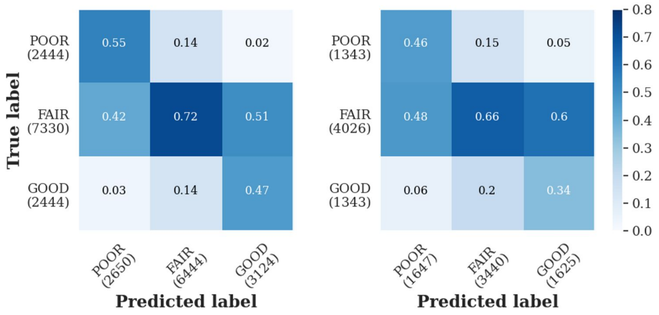

Fig. 6. Column-normalized confusion matrices at forecast times of (left) 1 h and (right) 3 h. Sample sizes are listed for each label.

Funding was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA21OAR4320204, U.S. Department of Commerce.