Interpretation of Deep-learning Models for Prediction of Tornadogenesis

Ryan Lagerquist[1,2]; Amy McGovern[1]; David John Gagne II[3]; Cameron Homeyer[1]; Travis Smith[2]

[1] University of Oklahoma School of Meteorology; [2] Cooperative Institute for Mesoscale Meteorological Studies;
[3] National Center for Atmospheric Research

1. Introduction

- Machine learning (ML) is becoming widely used in weather research.
- ML is often faster and better than competing prediction methods.
- However, many are reluctant to adopt ML in operations, because it is a "black box" (does not explain decisions to user).

- Our work attempts to bridge this gap.
- We apply several interpretation methods to a convolutional neural network (CNN) trained to predict tornadogenesis.
- **Goal:** understand what CNN has learned, which has benefits in all three phases of ML (Selvaraju *et al.* 2017).

1. **Development phase**
   - Used for debugging (does the model learn relationships that make sense?)
2. **Operational phase**
   - Increases users' trust and understanding in the model
   - Highlights situations where model should (not) be trusted
3. **ML-superiority phase**
   - If ML ever vastly outperforms humans at forecasting, can be used to teach humans
   - Already being done for Chess (Johns *et al.* 2015) and Go (Silver *et al.* 2016)

- Also, ML interpretation can be used to form new scientific hypotheses (Wagstaff and Lee 2018).

2. Machine Learning

- **Prediction:** probability of tornadogenesis for each storm in next 60 minutes
- **Labels (ground truth):** NWS tornado reports
- **Predictors:** radar and soundings

- **Radar details:**
  - Storm-centered grid of 12 variables (Figure 1) every 5 minutes
  - 32 x 32, 1.5-km resolution, storm motion to the right
  - From GridRad dataset (Homeyer *et al.* 2017); currently experimenting with MYRORSS (Ortega *et al.* 2012)

- **Sounding details:**
  - From nearest grid cell in Rapid Refresh (RAP) or Rapid Update Cycle (RUC) analysis
  - RUC before 1 May 2012, RAP otherwise

- **Time period:** 145 days in 2011-17 (2011-13 for training, 2014-15 validation, 2016-17 testing)
- Performance on testing data shown in Figures 2-3

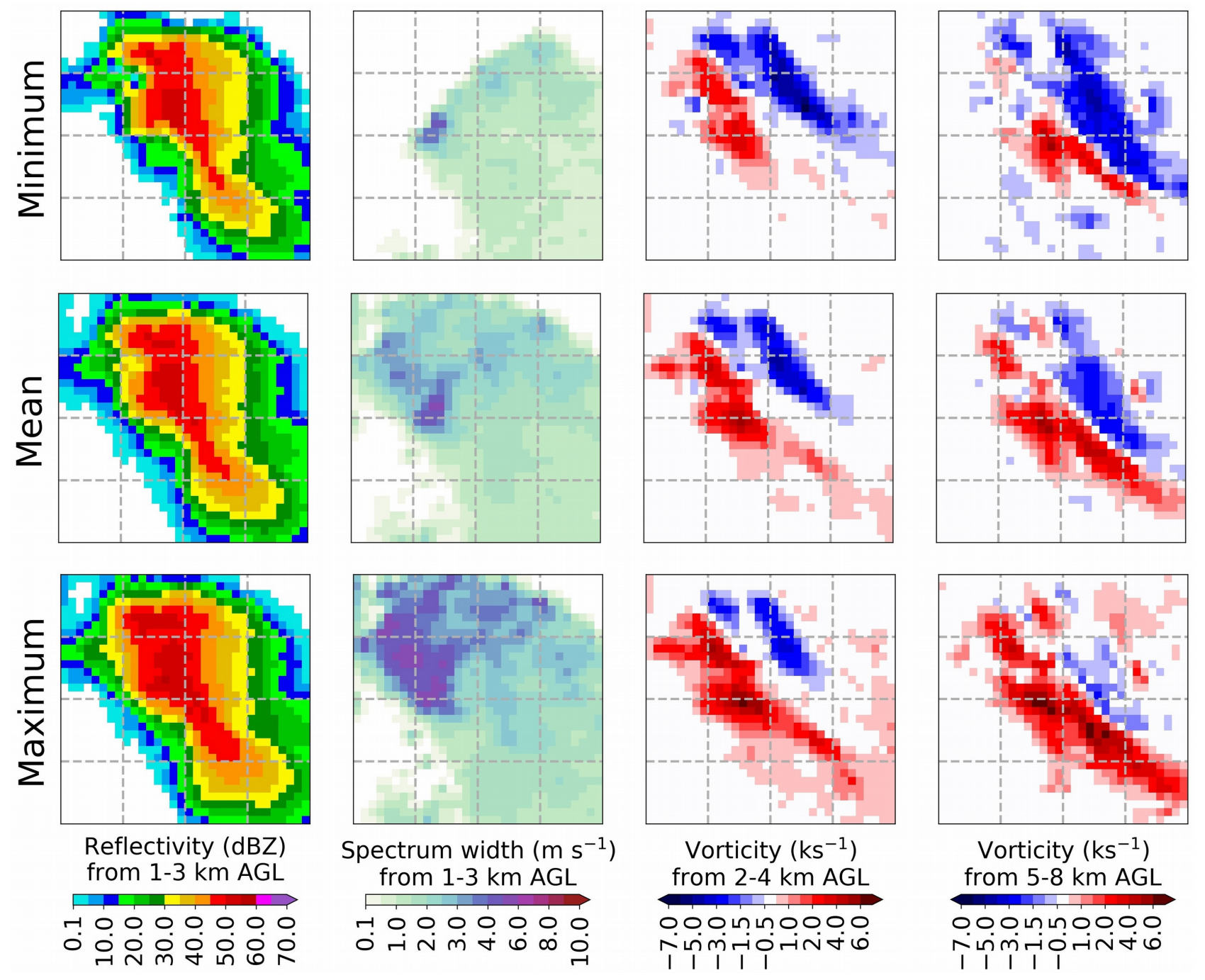

Figure 1: Radar predictors for one case (one storm cell at one time). CNN is trained to predict probability that storm will undergo tornadogenesis in the next hour.

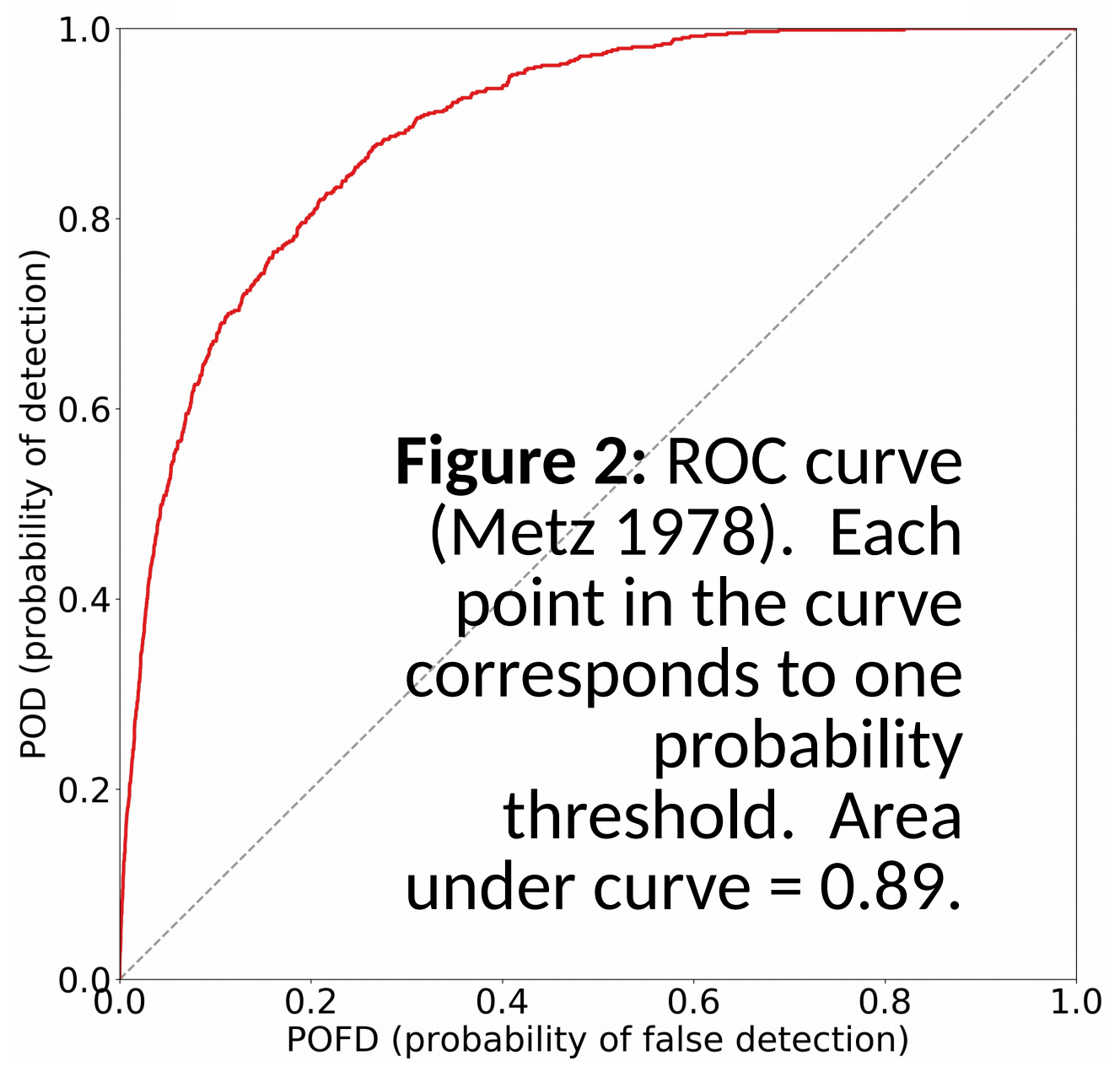

Figure 2: ROC curve (Metz 1978). Each point in the curve corresponds to one probability threshold. Area under curve = 0.89.

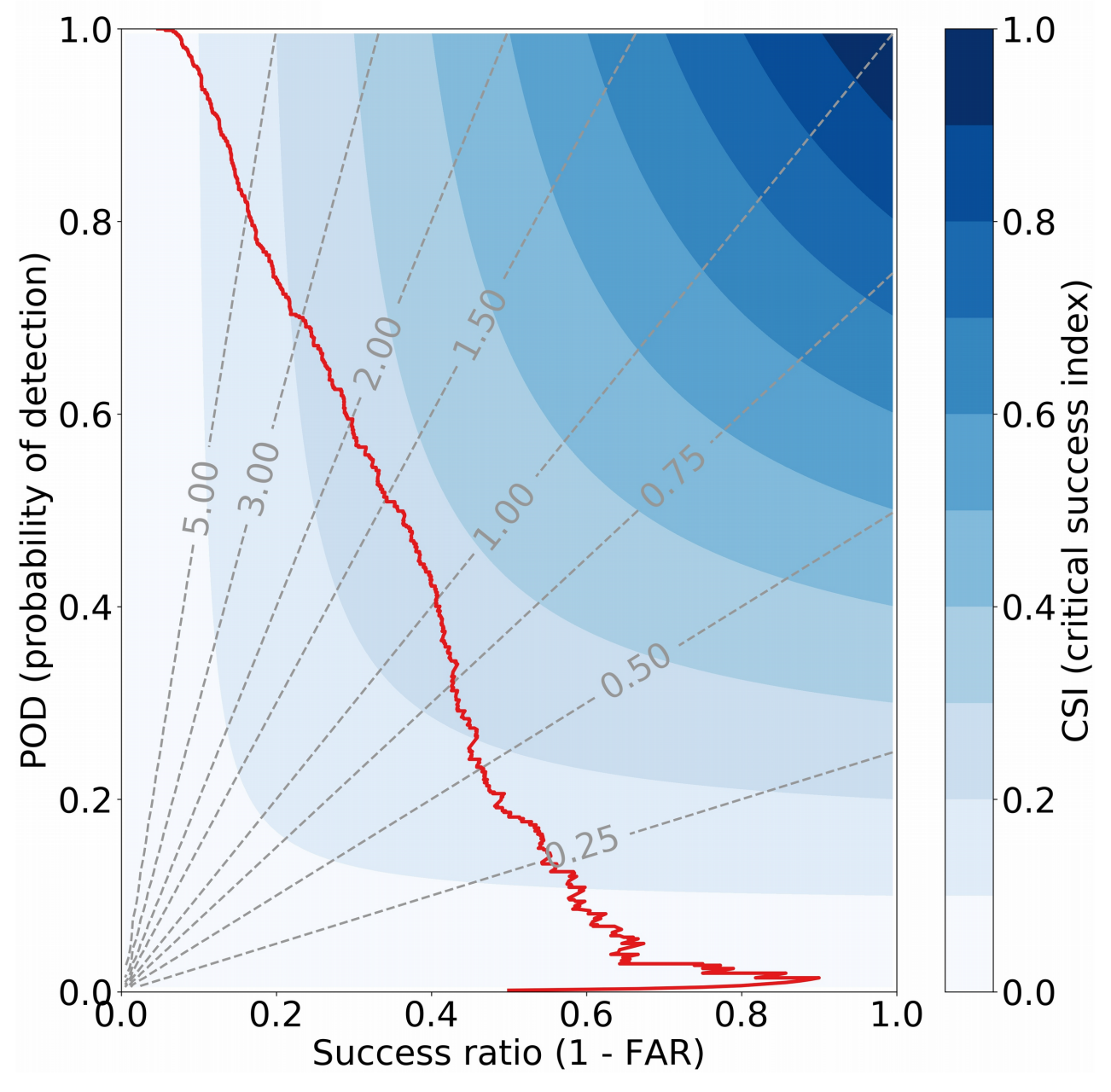

Figure 3: Performance diagram (Roebber 2009) for testing data. Dashed grey lines are frequency bias; each point in the red curve corresponds to one probability threshold. Maximum CSI = 0.27.

3. Permutation Importance

- Ranks importance of each predictor (x_j) by measuring how much performance declines when x_j is permuted (randomly shuffled over all cases).

- **Two versions:** single-pass (Breiman 2001) and multi-pass (Lakshmanan *et al.* 2015).

- **Single-pass:** only one predictor at a time is randomized.

- **Multi-pass:**
  - Find most important predictor and leave it randomized.
  - Find 2nd-most important and leave it randomized.
  - …Repeat until all predictors are randomized.

- Single-pass and multi-pass versions (Figure 4) agree on 4 of top 5 predictors:
  - v-wind
  - Max 1–3-km reflectivity
  - Max 5–8-km vorticity
  - Max 2–4-km vorticity

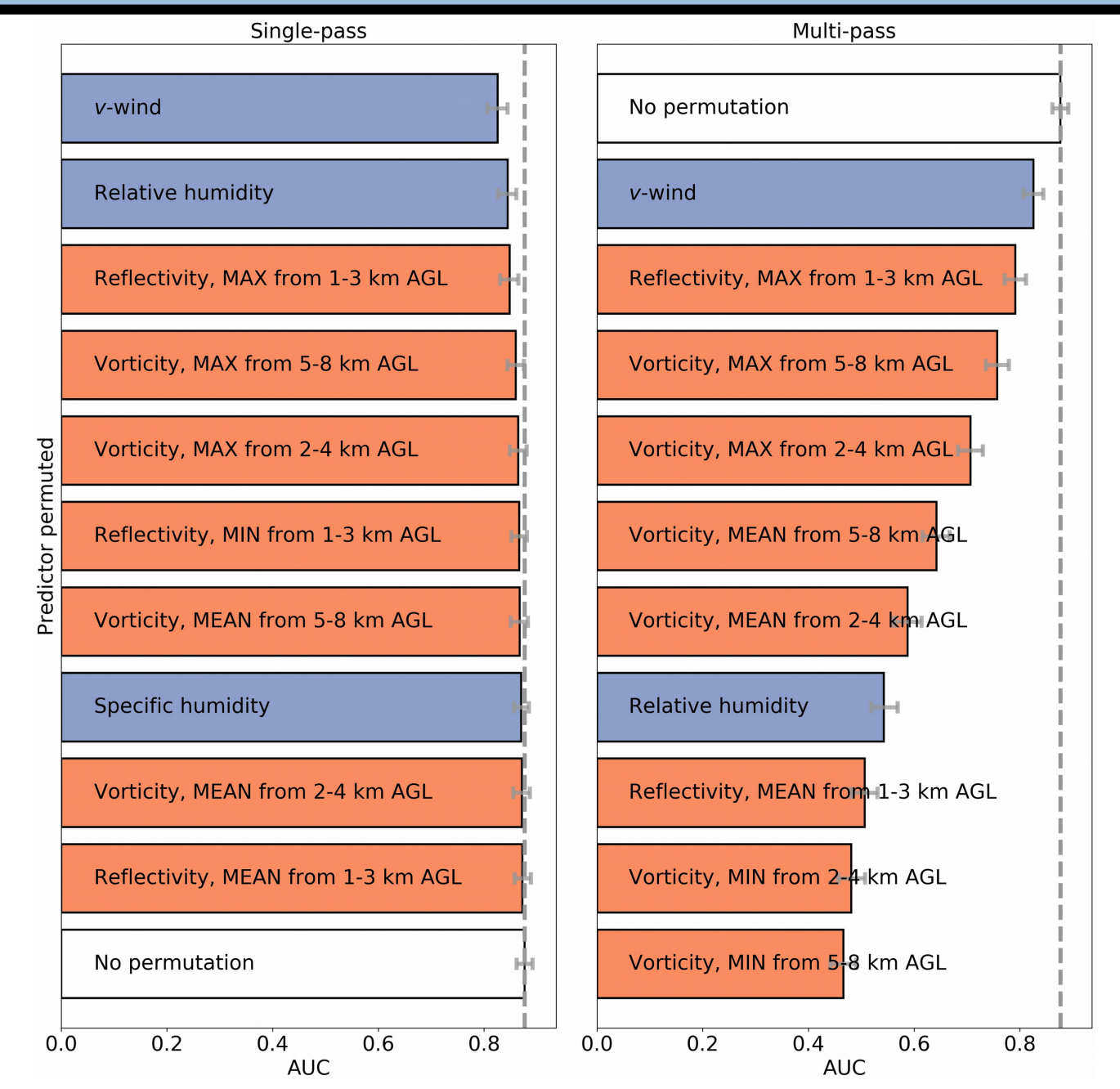

Figure 4: Results of permutation importance. Showing only top 10 (of 17) predictors, with most important at top. Radar predictors in orange; sounding predictors in purple; "AUC" is area under ROC curve for validation data.

4. Saliency Maps

- **Definition:** gradient of model prediction with respect to input value.

$$\text{saliency} = \left.\frac{\partial p}{\partial x}\right|_{x=x_0}$$

- p = model prediction (probability of tornadogenesis)
- x = input value (one predictor at one grid point)
- x_0 = actual value of x in dataset example

- Thus, saliency is linear approx to $\frac{\partial p}{\partial x}$ about $x = x_0$.

- Saliency tells us how prediction changes when x changes **a little bit**.

- Figure 6 shows composite saliency maps for 4 sets of storms:
  - **Best hits** = 100 tornadogenetic storms with highest forecast probabilities
  - **Worst false alarms** = 100 non-tornadogenetic storms with highest probs
  - **Worst misses** = 100 tornadogenetic storms with lowest probs
  - **Best correct nulls** = 100 non-tornadogenetic storms with lowest probs

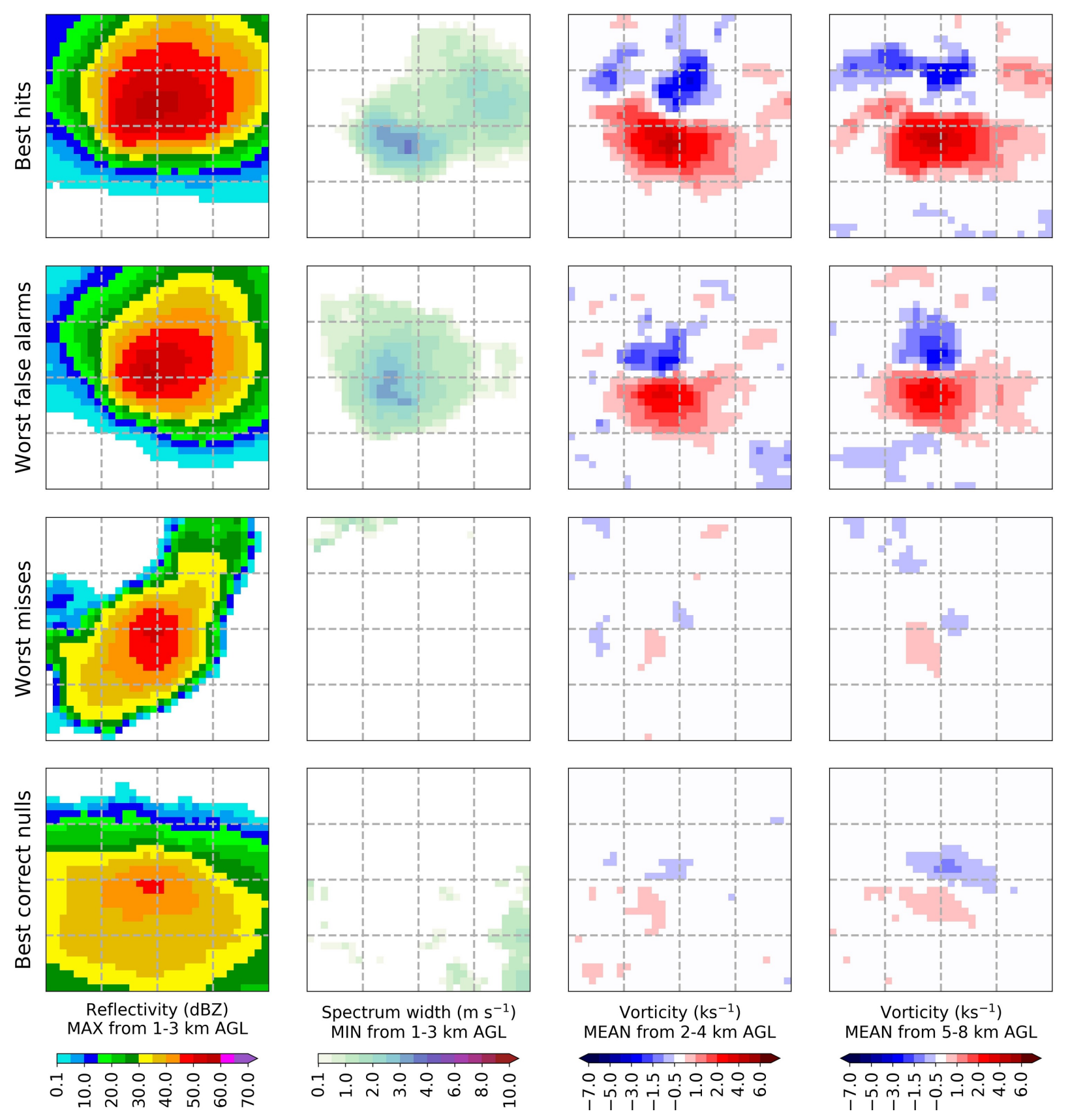

Figure 5: Composite map for each set of storms, showing only 4 of the 12 radar predictors.

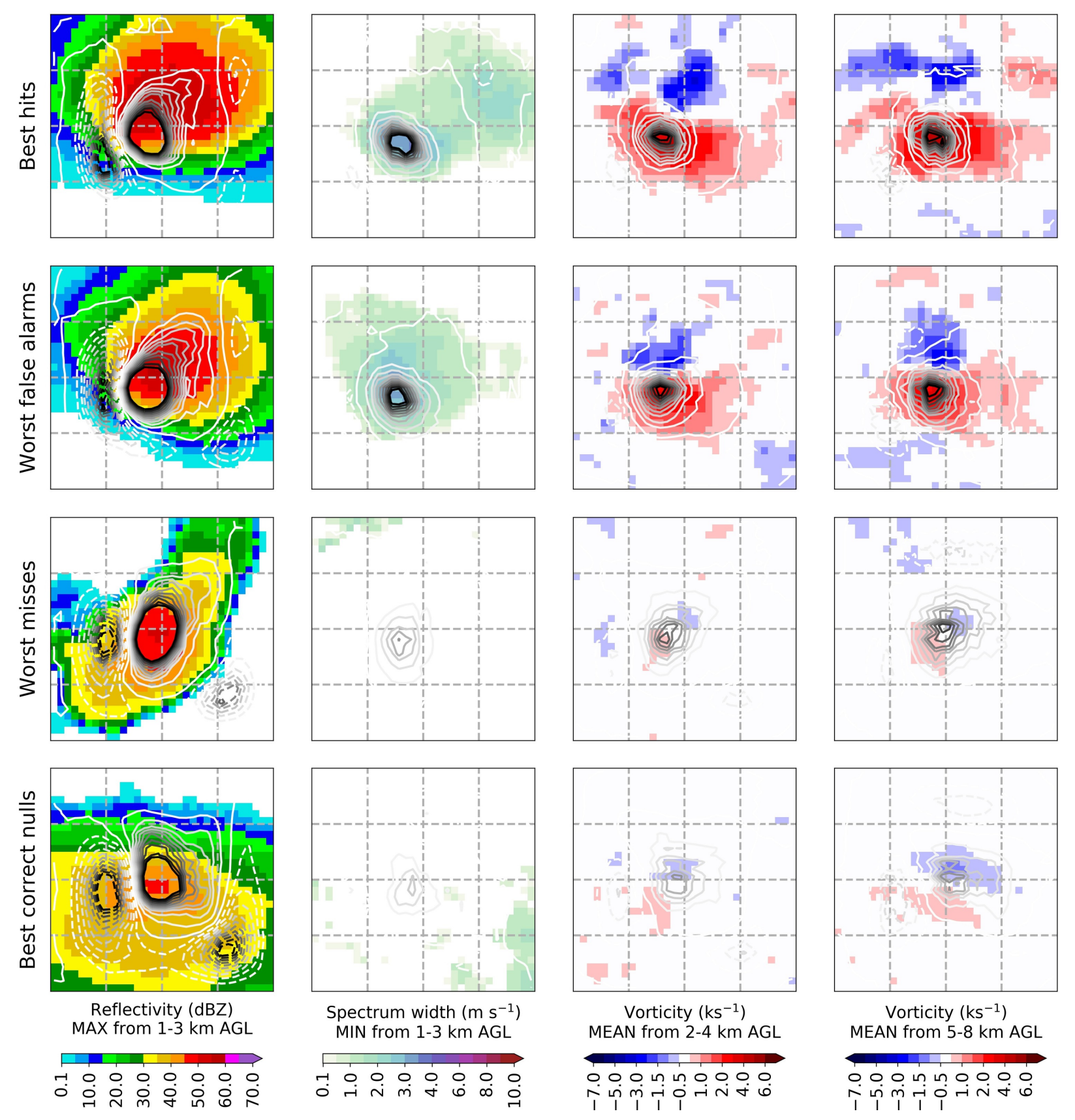

Figure 6: Composite saliency map for each set of storms. Heat maps represent input data (predictors). Solid contours are positive saliency (tornadogenesis probability increases with predictor values inside contour); dashed contours are negative saliency.

5. Backwards Optimization

- Also called "feature optimization" (Olah *et al.* 2017).
- **Goal:** create synthetic input that maximizes or minimizes model prediction.
- **Example:** create storm with tornadogenesis probability of 100% or 0%.

- Procedure involves gradient descent, which requires starting point. Examples:
  - Uniform image (all zeros)
  - Random image (Gaussian noise)
  - Dataset example

- We use dataset examples (Figures 7-8).
- Caveat: does not necessarily produce realistic data.

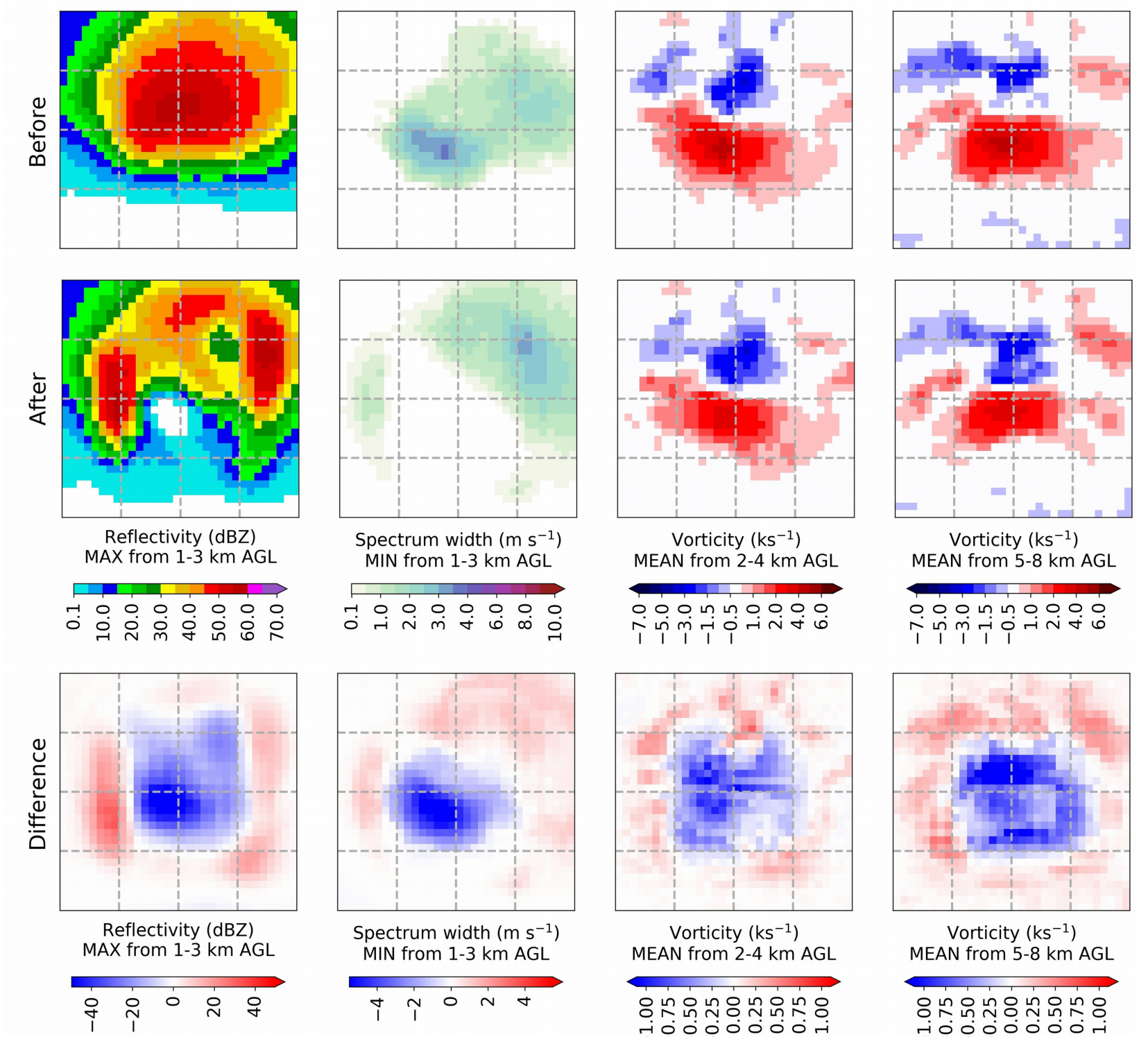

Figure 7: Results for 100 best hits. Backwards optimization applied to each storm separately, with goal of decreasing tornadogenesis probability to 0%.

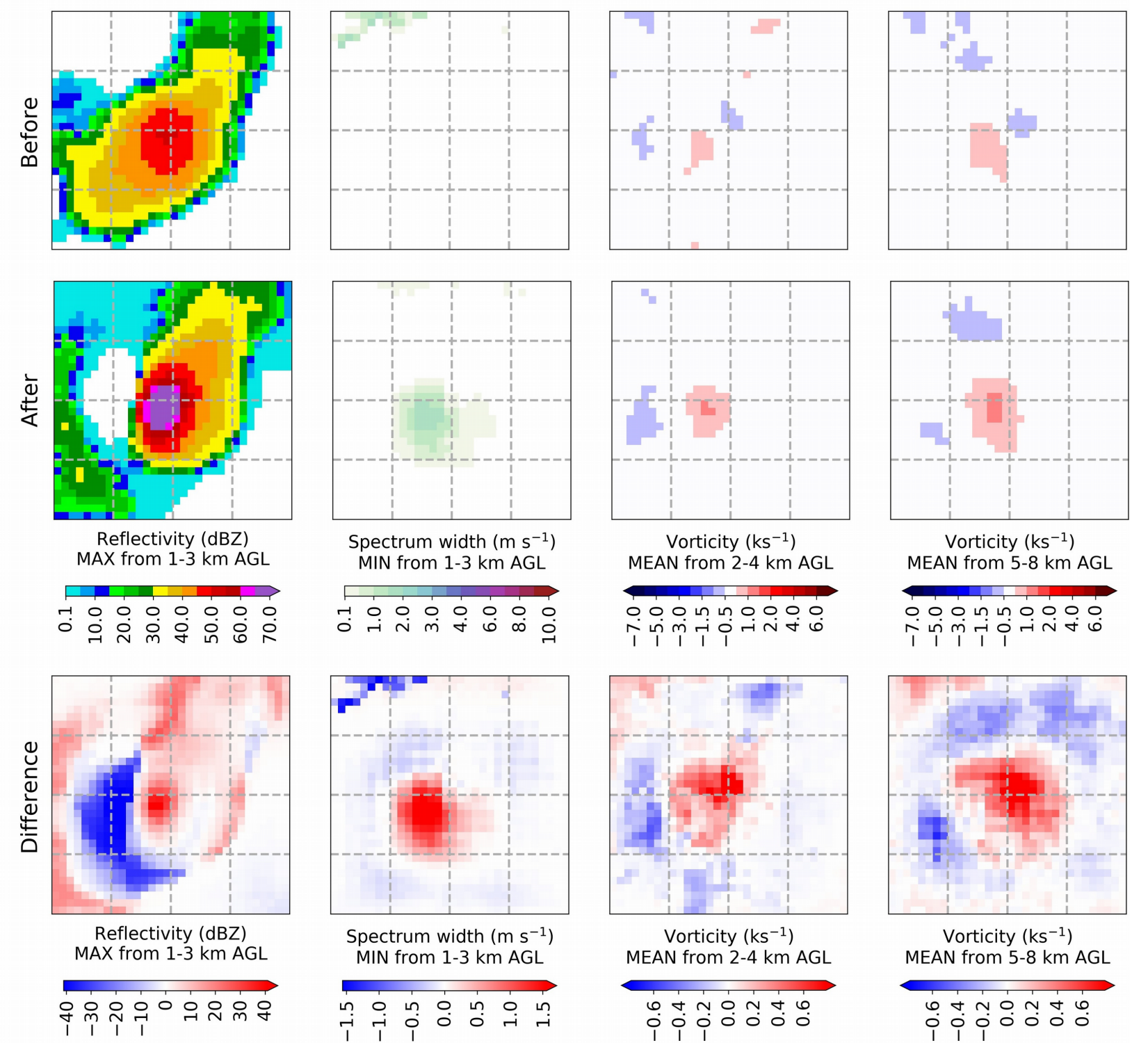

Figure 8: Results for 100 worst misses. Backwards optimization applied to each storm separately, with goal of increasing tornadogenesis probability to 100%.

6. Novelty Detection

- **Goal:** find most novel image in trial set with respect to baseline set.
- Novelty detection also determines **which parts of novel image make it novel.**
- Used to flag Mars Rover images for further investigation (Wagstaff and Lee 2018).

- We apply novelty detection to a different CNN:
  - **Prediction:** probability that simulated storm will develop strong rotation (future vorticity ≥ 0.0054 s^{-1} anywhere in storm)
  - **Predictors:** storm-centered grids of reflectivity, temperature, wind
  - **Source:** one member of NCAR convection-allowing ensemble (Schwartz *et al.* 2015)

- **Baseline set** = 100 random (mostly weak) storms in 2015
- **Trial set** = 100 storms in 2015 with strongest future rotation

- **Novelty maps (Figure 9) highlight the following properties:**
  - Strong reflectivity core
  - Rotational winds near reflectivity core
  - Lack of surrounding deep convection

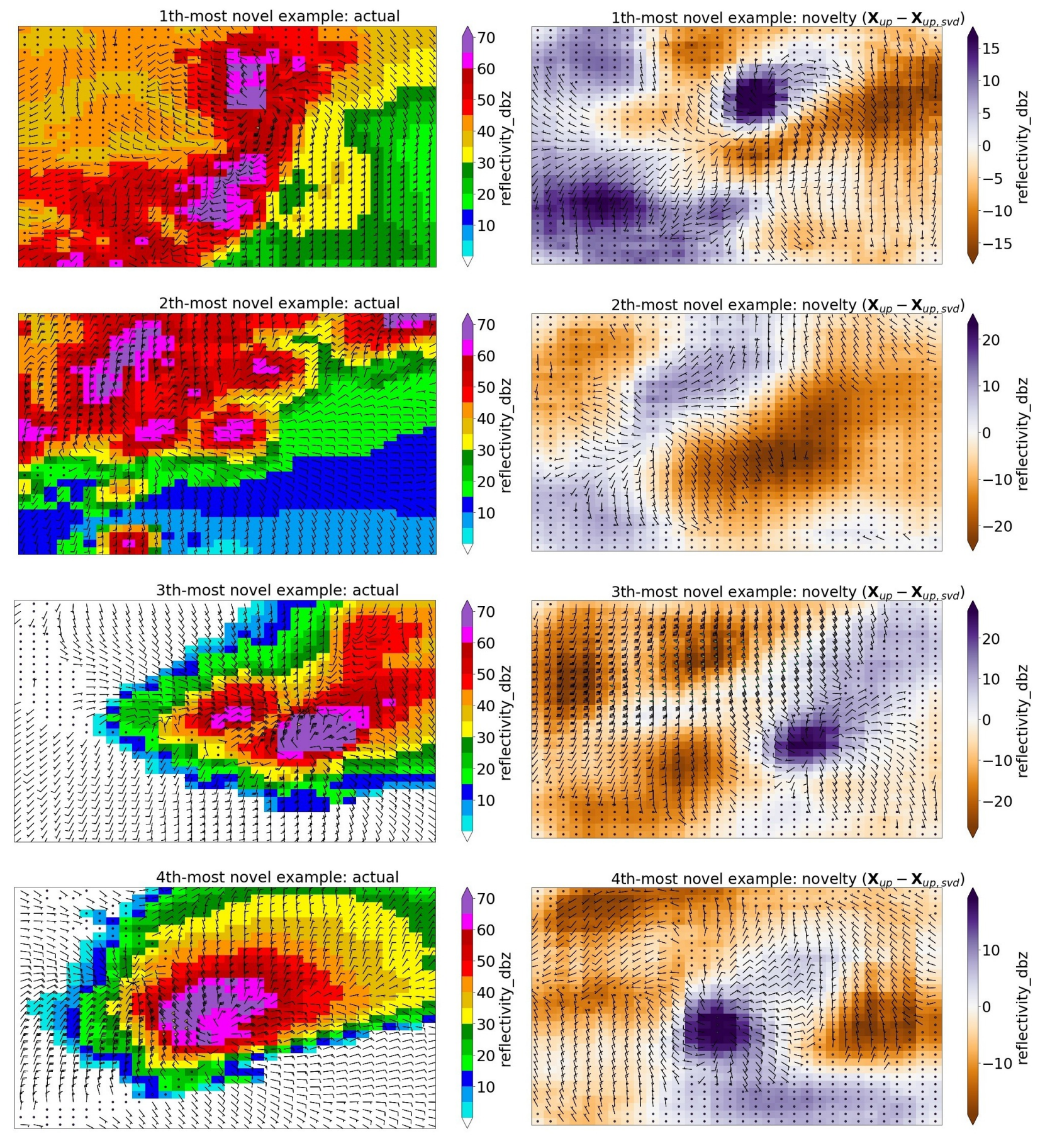

Figure 9: Most novel storms in the trial set (left column); most novel part of each storm (right column).

7. Future Work

- **BAMS publication** under review (McGovern *et al.* 2019).

- Develop ways to test **statistical significance** of interpretation results.
- Apply interpretation methods to **multiscale data** (*e.g.*, radar image + mesoscale and synoptic-scale environments).
- Use ML interpretation to help create **new scientific hypotheses.**