Explainable Uncertainty in Machine Learning Weather Prediction

David John Gagne [1]
John Schreck [1], Charlie Becker [1], Gabrielle Gantos [1], Maria Molina [2], Yingkai Sha [1]
1. National Center for Atmospheric Research
2. University of Maryland

NCAR

ai2es.org

November 16, 2022

This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977.

Motivation

- Users want to know when and why machine learning predictions are uncertain due to
  - Inherently challenging predictions
  - The model is outside the realm of its training data
- Current machine learning methods provide multiple ways to characterize predictive uncertainty
  - Discrete probability distribution
  - Parameters of parametric distribution
  - Generate samples from stochastic model or ensemble
- **Problems**
  - Single-model approaches only account for uncertainty from variance in the training data
  - Sampling/ensemble approaches can be slow, computationally expensive, and still underdispersive
  - How to propagate model uncertainty into model explanations
- **Solutions**
  - Evidential deep learning: single model that can estimate aleatoric and epistemic uncertainty
  - Explainable uncertainty: Apply model-agnostic XAI to tie variations in inputs to variations in uncertainty

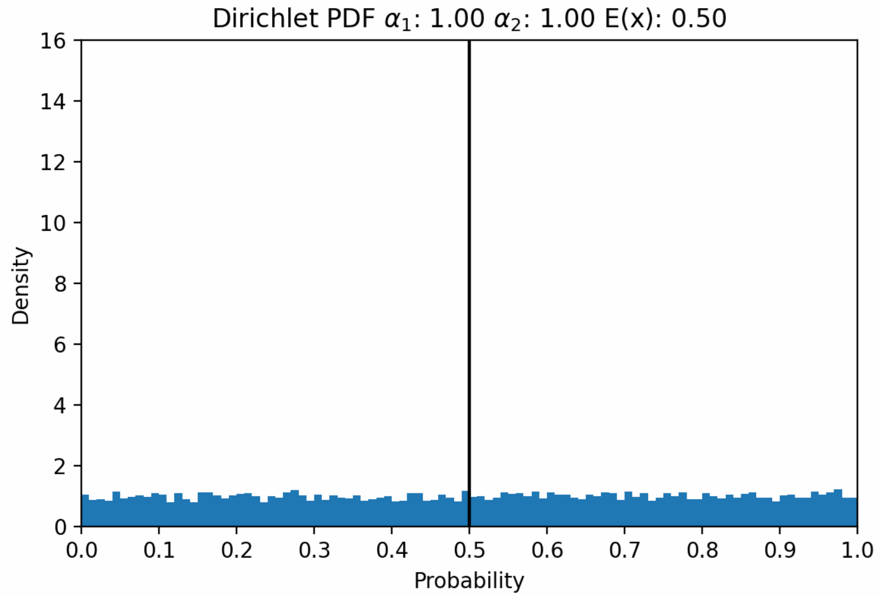
Background: Probabilities for classification problems

- Machine learning classification assumes the output is a categorical probability distribution with a fixed label
  - Example: p(Y) = (0.8, 0.1, 0.1); Y=(1, 0, 0)
- Neural networks estimate these probabilities by using a softmax activation function and categorical cross entropy
  - Ensures probabilities are between 0 and 1 and sum to 1 across classes
  - Categorical cross entropy maximizes the likelihood of the probability for the true class
- However, problems arise when interpreting the uncertainty from this approach
  - Probabilities generally overconfident
  - Loss only accounts for uncertainty in the data
- Instead: utilize a Bayesian approach to classification
  - Assume we are learning to build evidence for each class
  - Assume a prior distribution so that we can estimate the posterior distribution
  - Posterior distribution is a distribution of possible categorical distributions

Estimating the Uncertainty of a Categorical Distribution

We want to estimate the uncertainty in the predicted probabilities. The Dirichlet distribution is a good choice as the prior for Categorical(K) with free parameter vector **α** >=1:

$$D(p|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} p_i^{\alpha_i - 1} \quad \text{where} \quad \sum_{i=1}^{K} p_i = 1$$



The expected probabilities can be computed via MLE:

$$\mathbb{E}(p_i) = \frac{\alpha_i}{\sum_{j=1}^{K} \alpha_j}$$

The posterior (= prior * likelihood) is conjugate to the prior (e.g. it's also a Dirichlet distribution)

Aleatoric and Epistemic Uncertainties

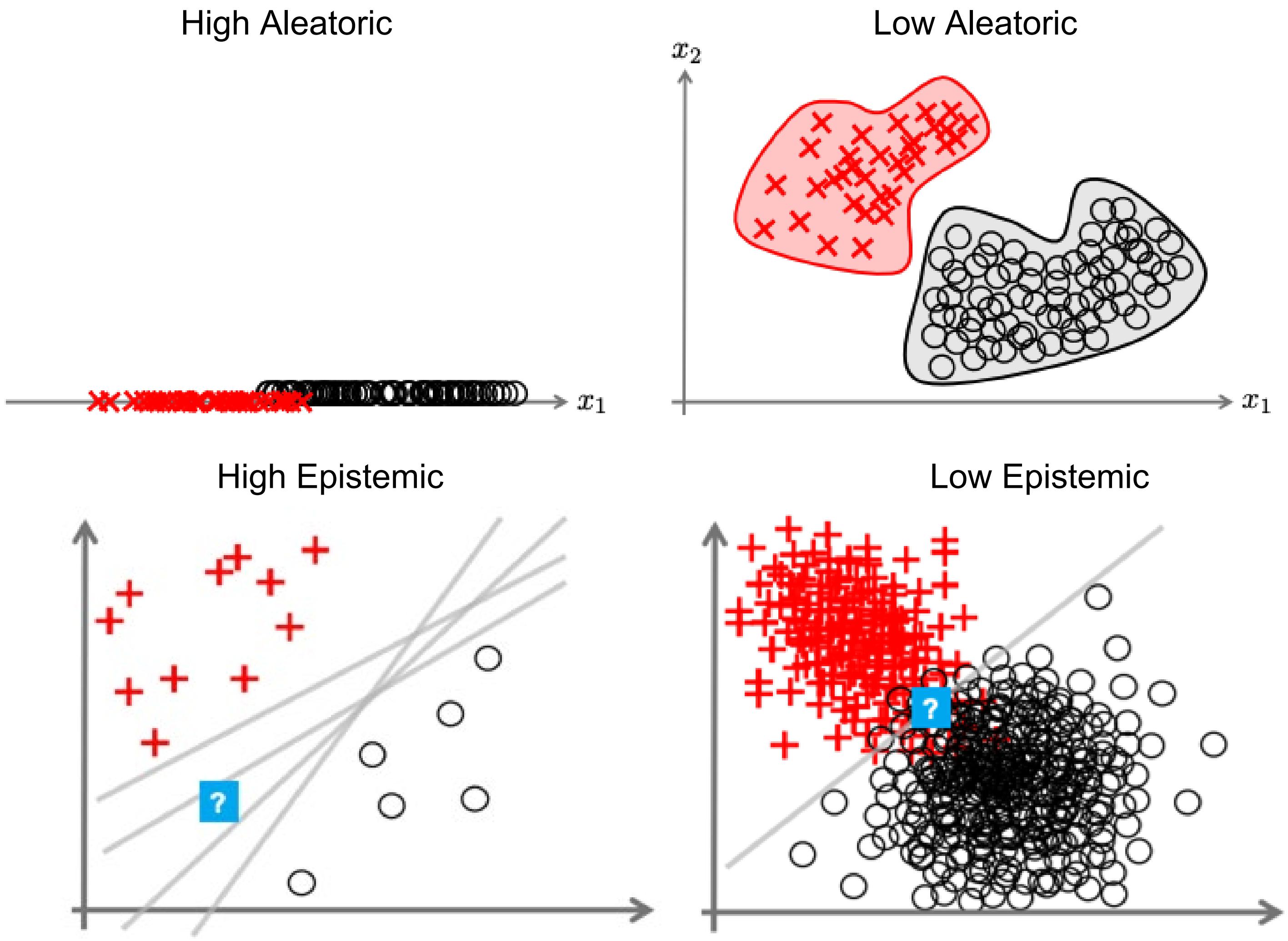
Aleatoric

- Uncertainty from data variance
- Irreducible: more examples do not help; more relevant features would be needed

Epistemic

- Uncertainty from model variance
- Reducible: more data examples can reduce this uncertainty

Hullermeier and Waegeman (2020)
arXiv:1910.09457v3



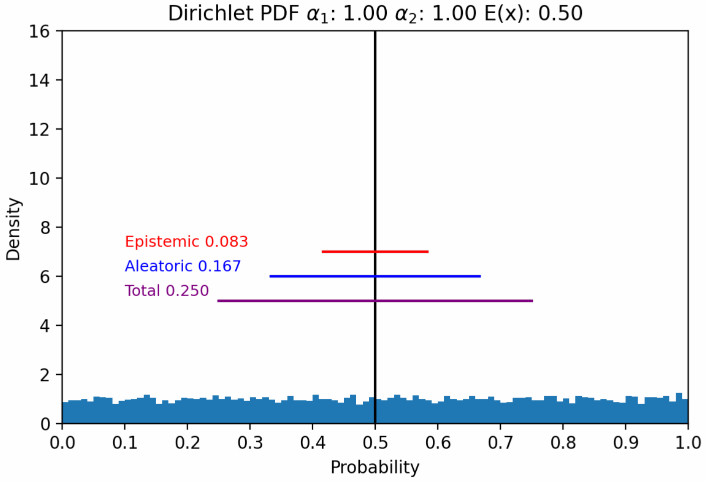
$$\mathrm{Var}(Y_j) = E(\mathrm{Var}(Y_j|\theta)) + \mathrm{Var}(E(Y_j|\theta))$$

Aleatoric — Epistemic

Dirichlet Aleatoric and Epistemic Uncertainties

Law of total uncertainty decomposes the total uncertainty into the sum of the unexplained variance plus the explained variance:

$$\mathrm{Var}(y_j) = \mathbb{E}\left(\mathrm{Var}(y_j|\boldsymbol{p})\right) + \mathrm{Var}\left(\mathbb{E}(y_j|\boldsymbol{p})\right)$$



Aleatoric (unexplained) $= \mathbb{E}\left\{\mathrm{Var}(y_j|\boldsymbol{p})\right\} = \mathbb{E}\left\{p_j(1-p_j)\right\}$

$$= \mathbb{E}(p_j) - \mathbb{E}(p_j^2)$$

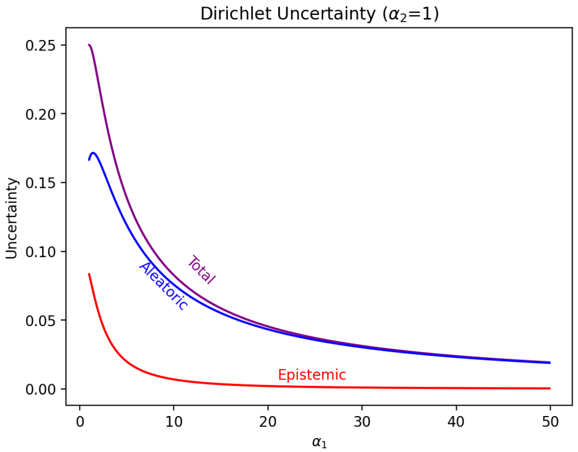
$$= \mathbb{E}(p_j) - \left\{\mathbb{E}(p_j)\right\}^2 - \mathrm{Var}(p_j)$$

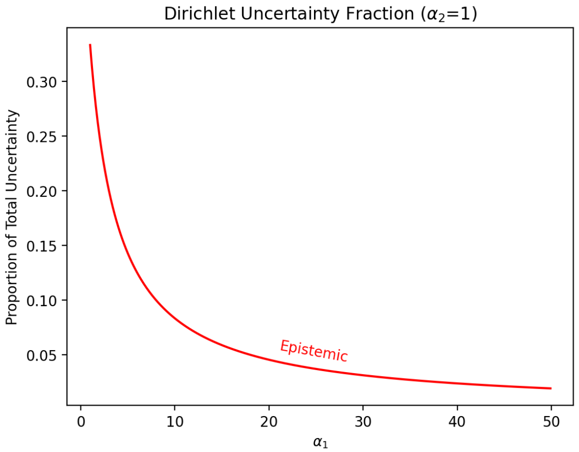
$$= \frac{\alpha_j}{S} - \left(\frac{\alpha_j}{S}\right)^2 - \frac{\frac{\alpha_j}{S}\left(1-\frac{\alpha_j}{S}\right)}{S+1}$$

Epistemic (explained) $= \mathrm{Var}\left\{\mathbb{E}(y_j|\boldsymbol{p})\right\} = \mathrm{Var}(p_j)$

$$= \frac{\frac{\alpha_j}{S}\left(1-\frac{\alpha_j}{S}\right)}{S+1}$$

Total = Aleatoric + Epistemic





Theory of Evidence

Dempster-Shafer Theory of Evidence (DST), a generalization of Bayesian theory of subjective probabilities, assigns ***belief masses*** to subsets of possible labels for an observation.

If belief masses for an observation are all equally likely ~ "*I do not know.* "

Subjective logic (SL) formulates *belief assignments* b_k over K classes, plus "I don't know ", as a Dirichlet distribution (prior). For a NN with K outputs

$$u + \sum_{k=1}^{K} b_k = 1$$

where b_k is the *kth* ReLU output, interpreted as the "*belief mass*" of the *kth* class, and u is the uncertainty mass of the K outputs.

Each b_k is defined as

$$b_k = \frac{e_k}{S} \quad \text{where} \quad S = \sum_{i=1}^{K}(e_k + 1) \quad \text{and thus} \quad u = \frac{K}{S}$$

Theory of Evidence

With $\boldsymbol{\alpha}_i = e_i + 1$, the Dirichlet probability density function is again

$$D(p|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} p_i^{\alpha_i - 1} \qquad \text{where} \qquad \sum_{i=1}^{K} p_i = 1$$

except now the average probabilities are computed using the evidence for each class:

$$p_i = \frac{\alpha_i}{\sum_{j=1}^{K} e_j + 1} = \frac{\alpha_i}{S_K}$$

Sensoy, M., L. Kaplan, and M. Kandemir, 2018: Evidential deep learning to quantify classification uncertainty. *arXiv [cs.LG]*, https://arxiv.org/abs/1806.01768.

Full Classifier Evidential Loss

$$\mathcal{L}(\Theta) = \sum_{i=1}^{N} \mathcal{L}_i(\Theta) + \lambda_t \sum_{i=1}^{N} KL[D(\mathbf{p_i}|\tilde{\boldsymbol{\alpha}}_i) \,||\, D(\mathbf{p}_i|\langle 1, \ldots, 1\rangle)],$$

MLE Loss

Distance from 0-evidence/uniform prior

Annealing coefficient $\lambda_t = \min(1.0, t/50)$ | $\tilde{\boldsymbol{\alpha}} = \boldsymbol{y}_i + (1 - \boldsymbol{y}_i) \odot \boldsymbol{\alpha}$ Alphas of misleading evidence

MLE Loss

$$\mathcal{L}_i(\Theta) = \int ||\boldsymbol{y}_i - \boldsymbol{p}_i||_2^2 \frac{1}{B(\alpha_i)} \prod_{j=1}^{K} p_{ij}^{\alpha_{ij}-1} d\boldsymbol{p}_i = \sum_{j=1}^{K} (y_{ij} - \hat{p}_{ij})^2 + \frac{\hat{p}_{ij}(1 - \hat{p}_{ij})}{(S_i + 1)}$$

MSE

Variance

Distance from 0-evidence prior

$$KL[D(\mathbf{p}_i|\tilde{\boldsymbol{\alpha}}_i) \,||\, D(\mathbf{p}_i|\mathbf{1})]$$

$$= \log\left(\frac{\Gamma(\sum_{k=1}^{K} \tilde{\alpha}_{ik})}{\Gamma(K)\prod_{k=1}^{K} \Gamma(\tilde{\alpha}_{ik})}\right) + \sum_{k=1}^{K} (\tilde{\alpha}_{ik} - 1)\left[\psi(\tilde{\alpha}_{ik}) - \psi\left(\sum_{j=1}^{K} \tilde{\alpha}_{ij}\right)\right],$$

Pushes incorrect alphas toward 1 (uniform distribution)

Sensoy, M., L. Kaplan, and M. Kandemir, 2018: Evidential deep learning to quantify classification uncertainty. *arXiv [cs.LG]*, https://arxiv.org/abs/1806.01768.

Benchmark Use Case: Estimating Storm Severity

Data: Simulated storm properties from Molina et al. (2021).

Storms are extracted from “High Resolution WRF Simulations of the Current and Future Climate of North America” dataset (https://rda.ucar.edu/datasets/ds612.0/)

Inputs

Pressure

Temperature

U wind

V wind

Water Vapor Mixing Ratio

Radar reflectivity (max)

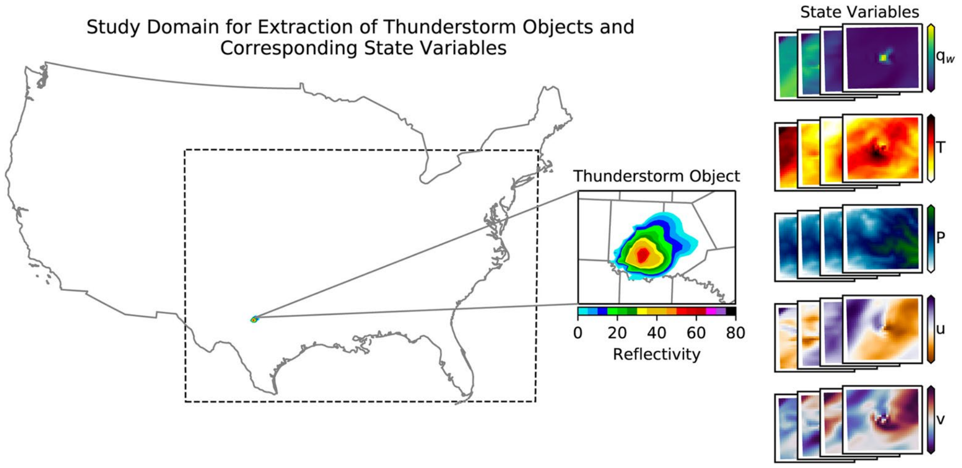
spatially averaged at 3 and 5 km AGL

Target

2-5 km Instantaneous Updraft Helicity > 75 $m^2 s^{-2}$

Training Years: 2001, 2002, 2003, 2005, 2006, 2008, 2009, 2010, 2012

Testing Years: 2000, 2004, 2007, 2011, 2013



Model

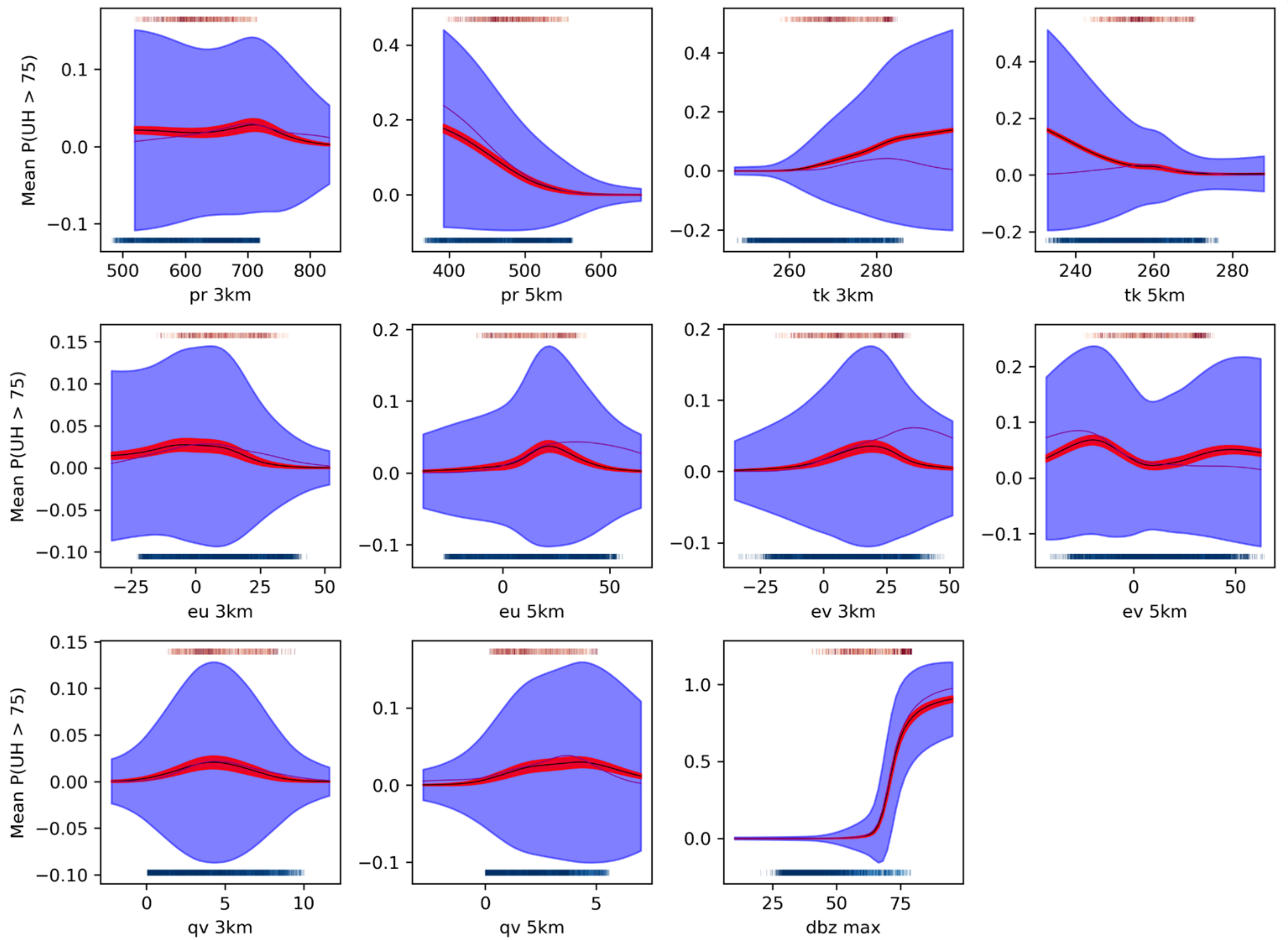
Dense neural network
Hidden layers: 2
Neurons/hidden layer: 100
ReLU activation function
Adam Optimizer
Learning rate: 0.01
Epochs: 30

Testing Results

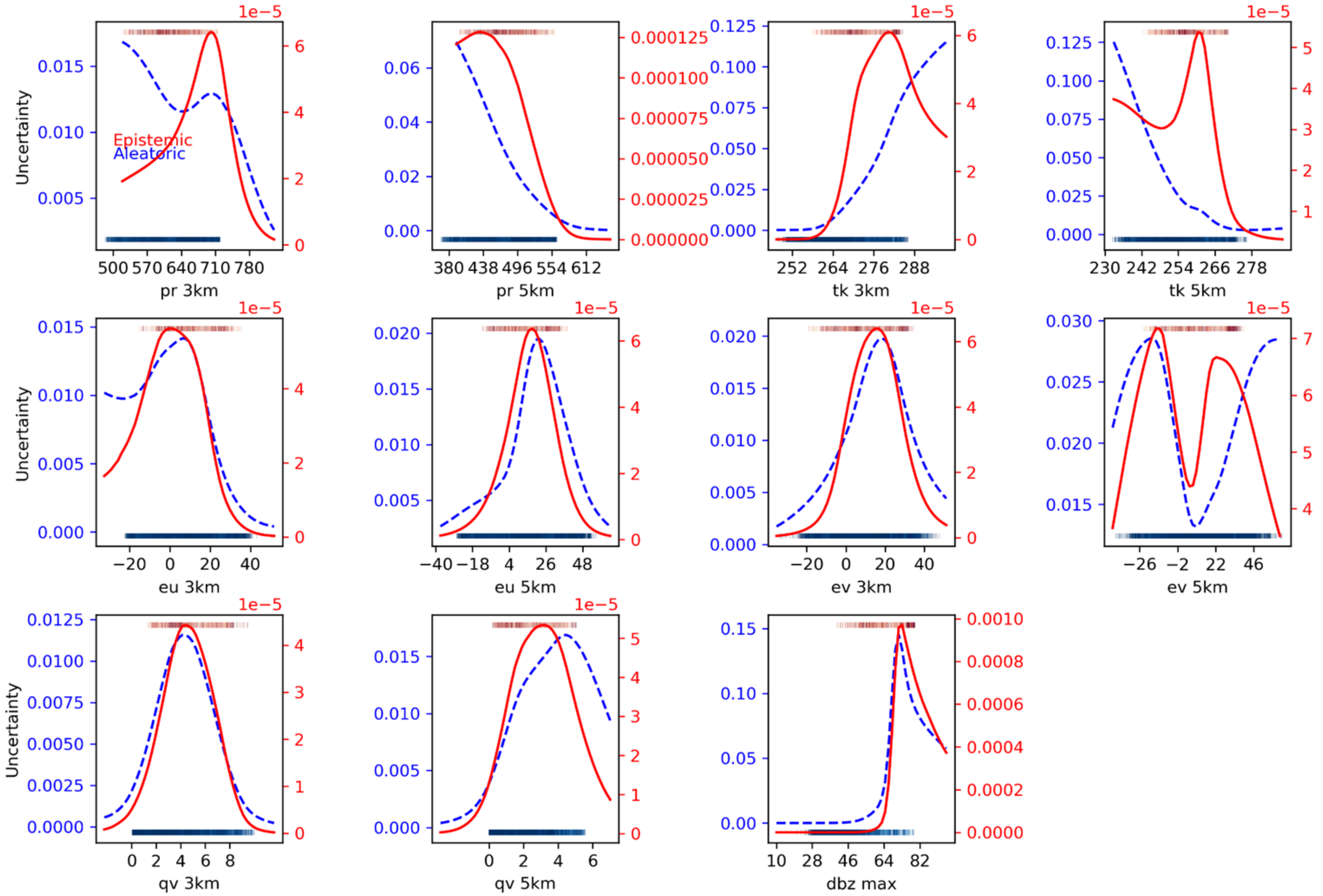
Model	AUC	Brier Score
Evidential	0.958	0.0180
Baseline	0.966	0.0179

Molina, M. J., D. J. Gagne, and A. F. Prein, 2021: A benchmark to test generalization capabilities of deep learning methods to classify severe convective storms in a changing climate. *Earth Space Sci.*, **8**, https://doi.org/10.1029/2020ea001490.

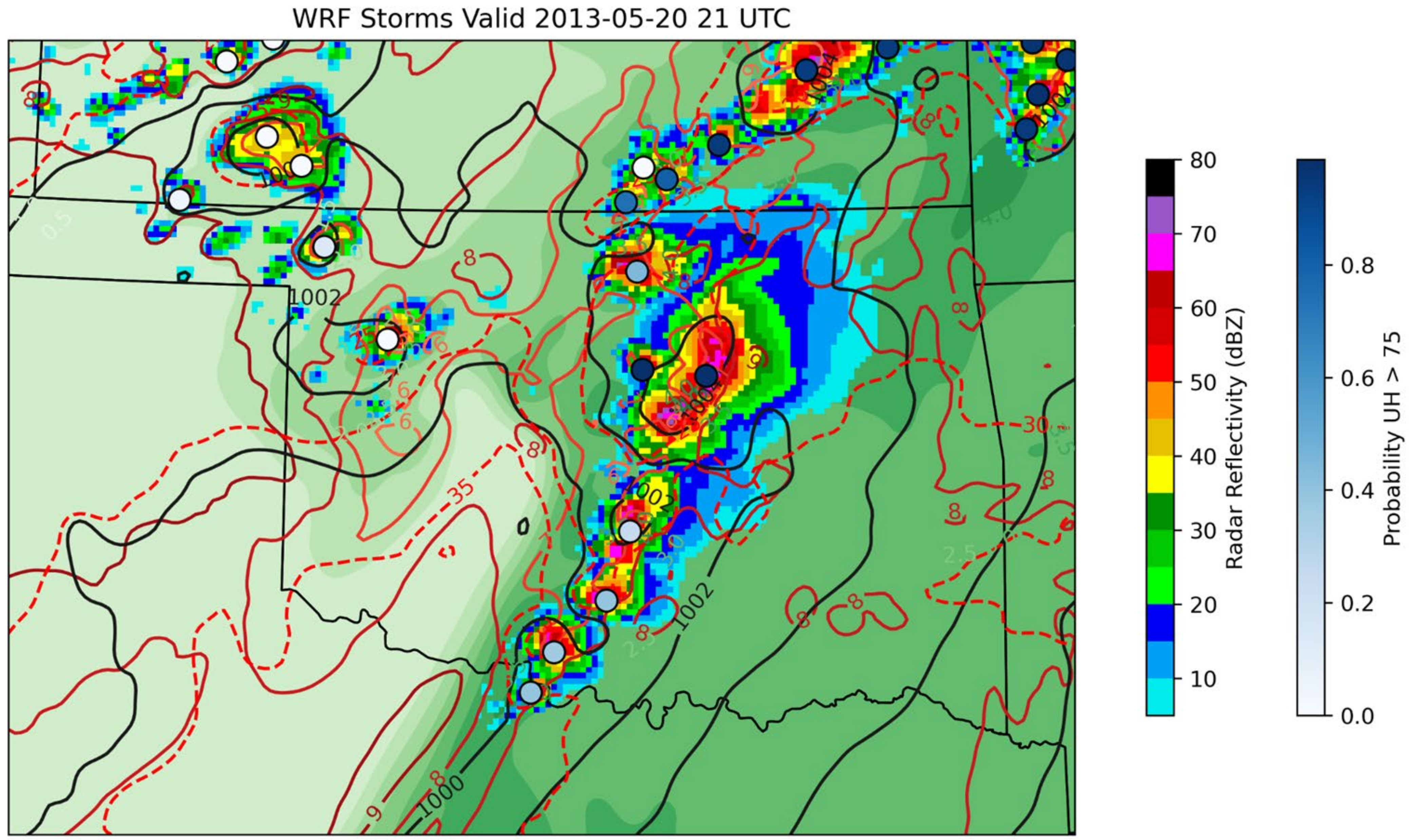
Partial Dependence with Uncertainty



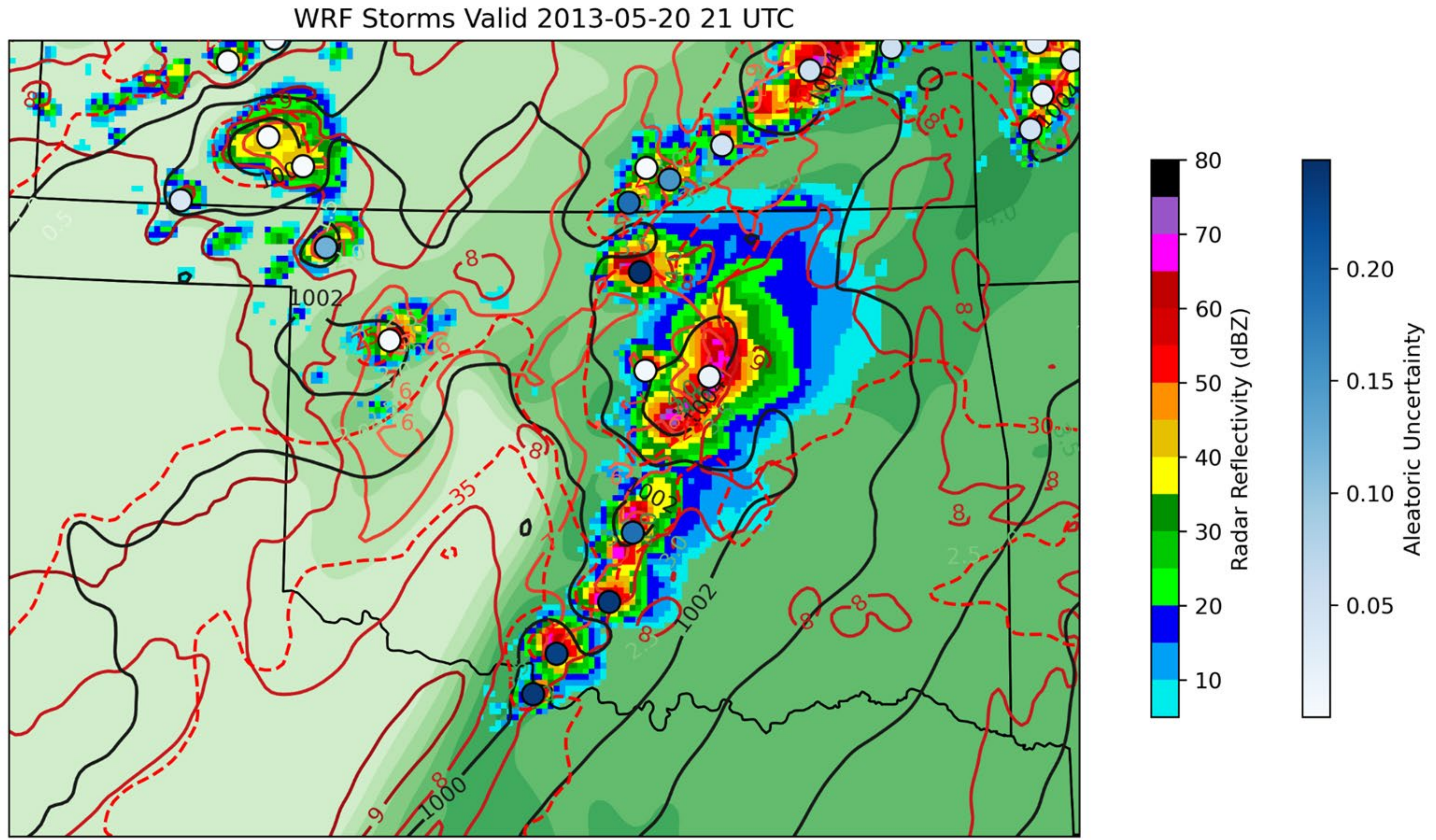
Aleatoric and Epistemic Partial Dependence



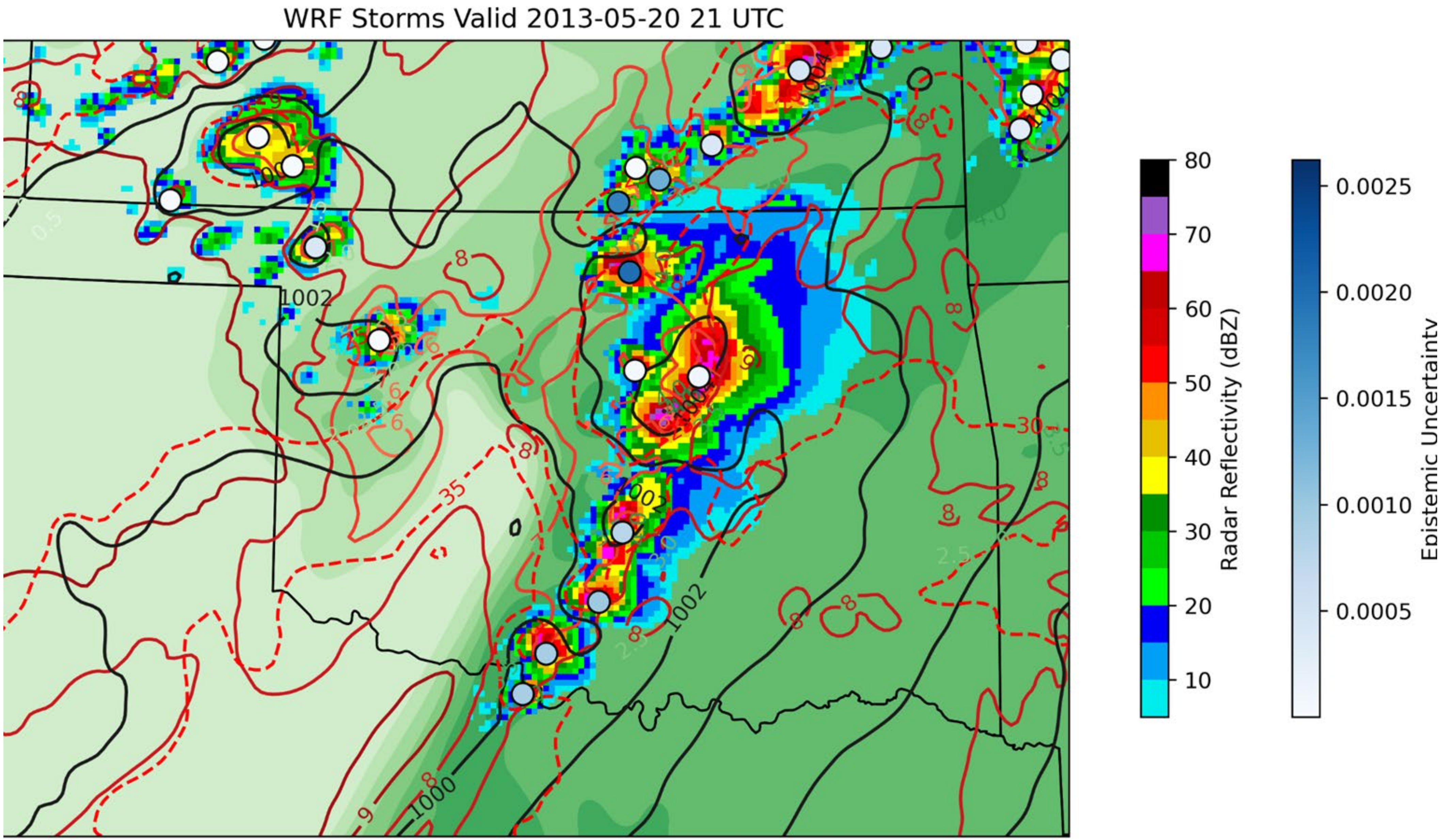
Case: UH Probability



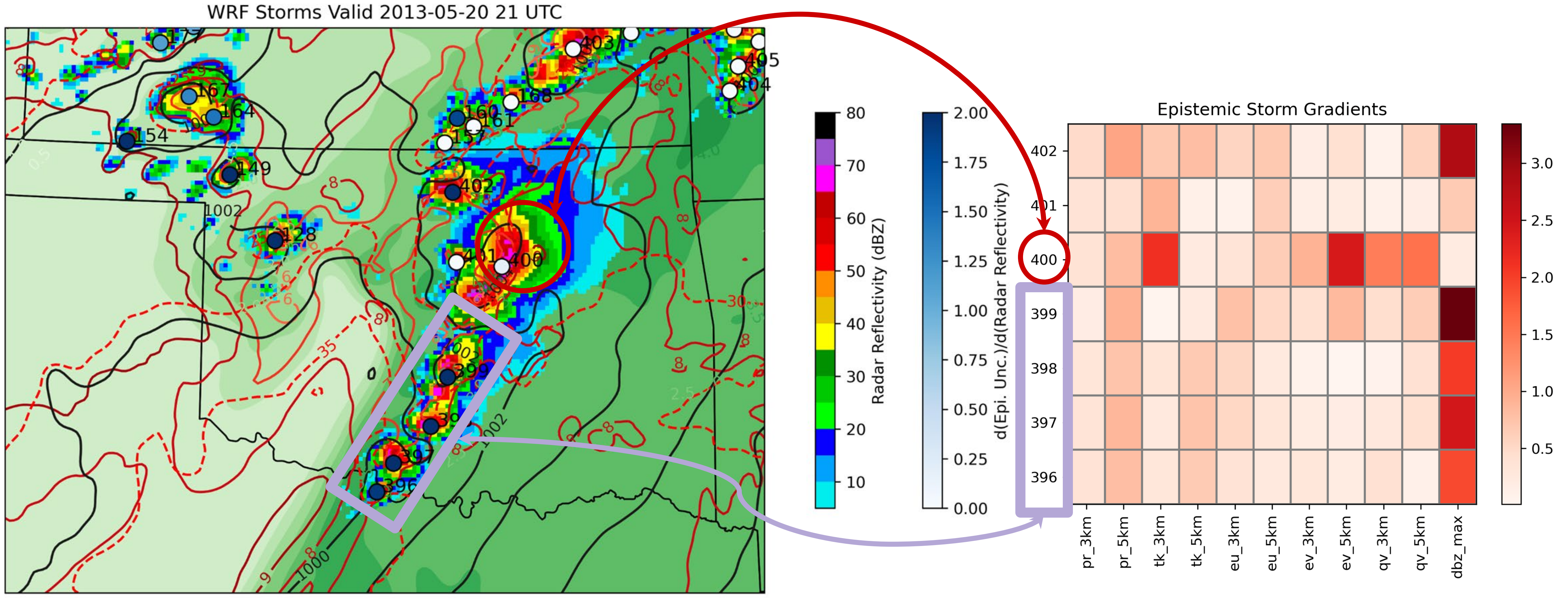
Case: Aleatoric Uncertainty



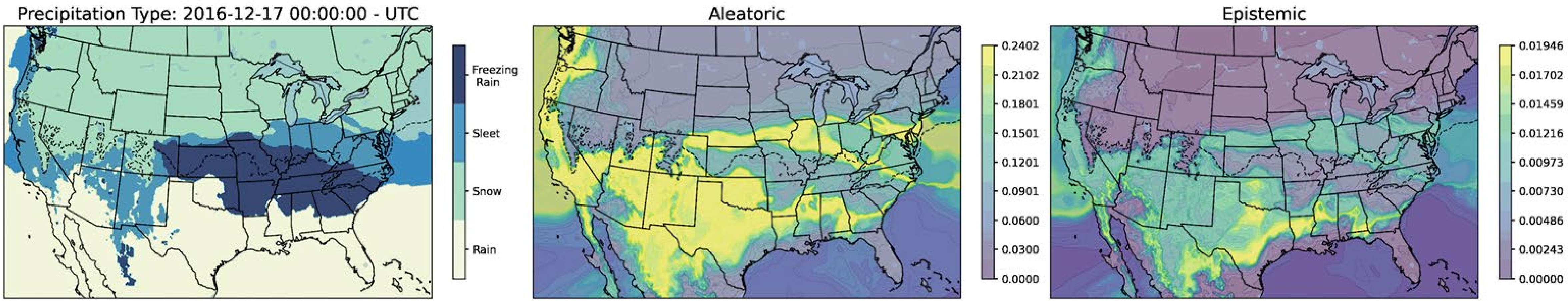
Case: Epistemic Uncertainty



Case: Neural Network Gradients



Second Benchmark: Precipitation Type Analysis



- Problem: analyze precipitation type
- Input: NOAA Rapid Refresh vertical profile
- Output: mPING precipitation type (rain, snow, sleet, freezing rain)
- Evidential model estimates aleatoric and epistemic uncertainties at all grid cell
- Aleatoric uncertainty high over broad precipitation transition region
- Epistemic uncertainty highest along strong temperature gradients

Summary and Challenges

Summary

- We implemented an evidential neural network for classification of updraft helicity and precipitation type
- Evidential models produce more robust estimates of uncertainty with a computationally inexpensive model and minimal architecture changes
- XAI on evidential models can reveal linkages between inputs and predicted uncertainty

Challenges

- Developing guidelines for more robust hyperparameter settings
- Evaluating the quality of the epistemic uncertainty estimate
- How best to use aleatoric and epistemic uncertainty in operations and research

Contact David John Gagne
Email: dgagne@ucar.edu
Twitter: @DJGagneDos