Combining Uncertainty Quantification and XAI to Understand the Sensitivities of Deep Learning Winter Precipitation Type Predictions

David John Gagne II
National Center for Atmospheric Research
Boulder, Colorado, USA

DoD Cloud Post-Processing and Verification Workshop
Sept. 13, 2023

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

https://avgeekery.com/ice-ice-baby-pilots-deal-wintry-mess/

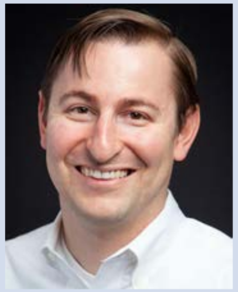
militarynews.com

- Transitions between liquid and frozen precipitation types can greatly impact transportation and logistics
- Forecasting p-type transitions is particularly challenging due to uncertainties in
  - thermodynamics
  - NWP models
  - observations
- ML methods with predictive uncertainty can help us understand and utilize uncertainty quantification (UQ) for more robust p-type forecasts
- Goals:
  - Introduce evidential deep learning
  - Connect uncertainty estimates with physical features
  - Link predictions to input features with XAI

Paper in prep: Evidential Deep Learning: Enhancing Predictive Uncertainty Estimation for Earth System Science Applications

The NCAR Machine Integration and Learning for Earth Systems (MILES) Group

MILES Core



David John Gagne
ML Scientist II
CISL/RAL

John Schreck
ML Scientist
CISL

Gabrielle Gantos
Associate Data Scientist II
CISL

Charlie Becker
Associate Data Scientist II
CISL

Kirsten Mayer
Project Scientist I
CGD

Will Chapman
Project Scientist I
CGD

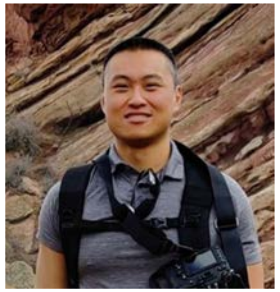
Mariana Cains
Project Scientist I
MMM

Chris Wirz
Project Scientist I
MMM

Jacob Radford
Research Associate
CIRA

Maria Molina
Professor
U. Maryland

Thomas Martin
Software Engineer
Unidata



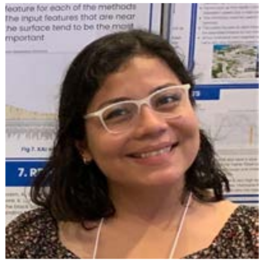
Wayne Chuang
Integration Engineer
LEAP

Julie Demuth
Project Scientist III
MMM

Jeff Anderson
Senior Scientist
CISL

Taysia Peterson
Admin
CISL

Dhamma Kimpara
Intern
CISL



Belen Saavedra Rios
Intern
CISL

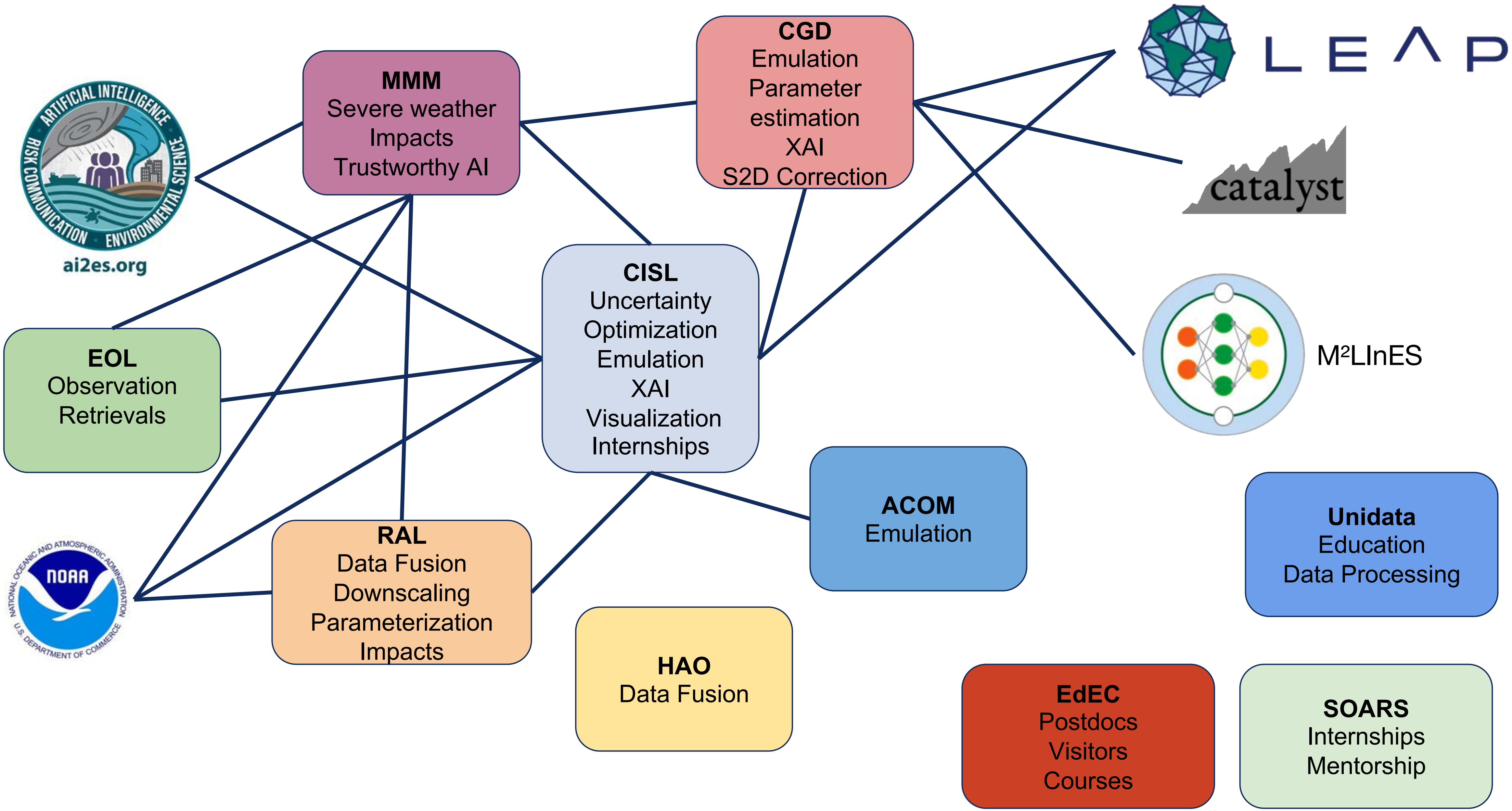
Hayden Outlaw
Intern
CISL

Da Fan
Visitor
Penn State

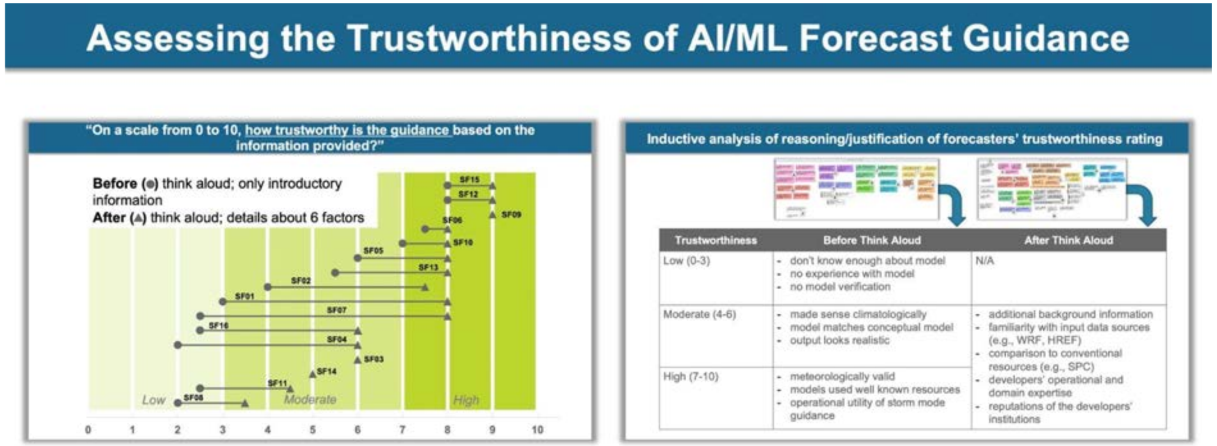
Bill Petzke
Software Engineer III, RAL

MILES+

The NCAR/UCAR AI Web

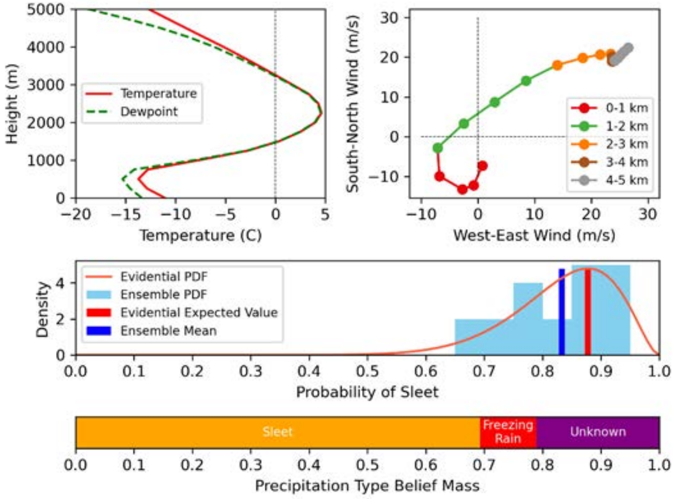


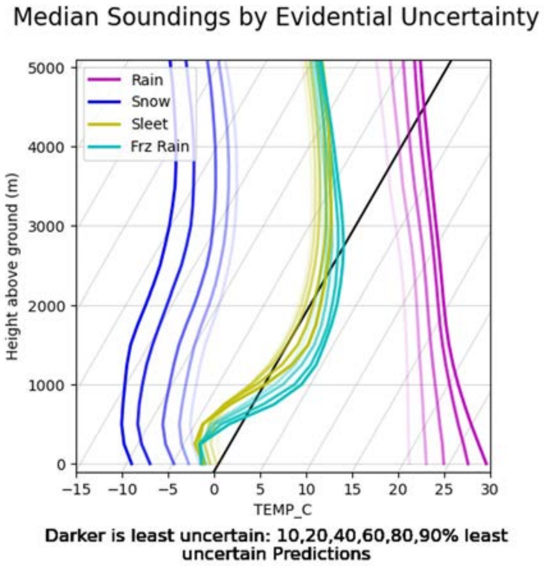
AI2ES: Developing Trustworthy AI Systems with User and Domain Expert Guidance



Forecasters need to personally use a model or piece of guidance over time to build trust in it.

mgcains@ucar.edu

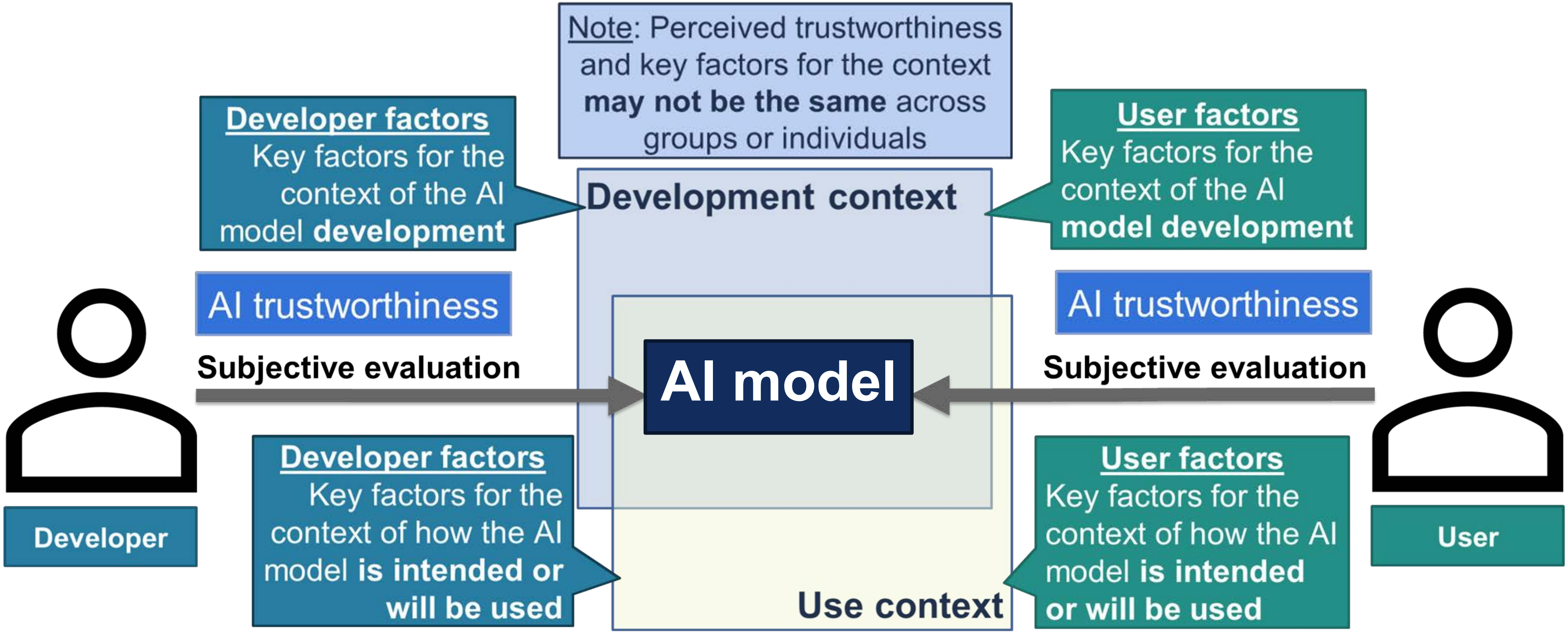




Vision: AI2ES is developing *novel*, *physically based* AI techniques that are demonstrated to be *trustworthy*, and will directly improve *prediction, understanding, and communication* of high-impact weather and climate hazards.

CISL: David John Gagne, John Schreck, Charlie Becker, Gabrielle Gantos
MMM: Julie Demuth, Chris Wirz, Mariana Cains
RAL: Bill Petzke
Unidata: Thomas Martin

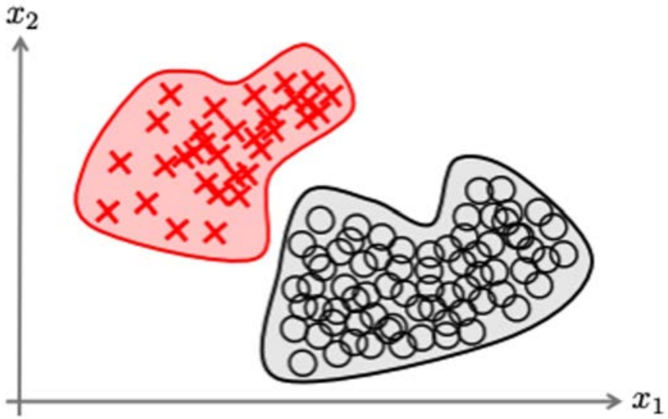
Our reconceptualization of trustworthy AI as **perceptual**



Wirz et al. 2023, (Re)Conceptualizing trustworthy AI: A foundation for change, In Prep.

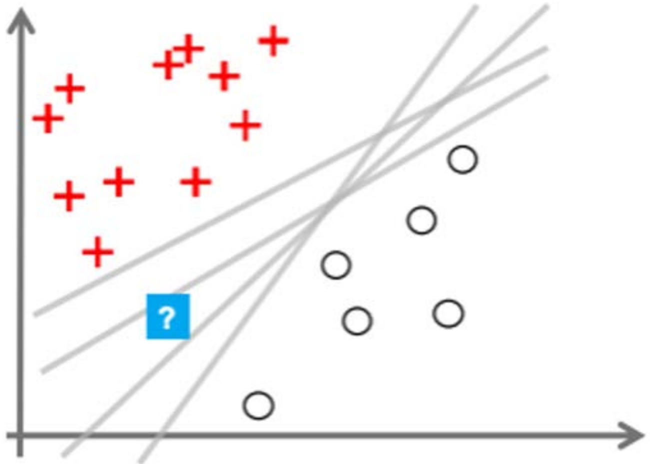

Decomposition of Uncertainty

Aleatoric Uncertainty



Definition: Uncertainty from unexplained variation in the data.
Estimated by: Single probabilistic AI model.

Epistemic Uncertainty



Definition: Uncertainty from variation in model predictions.
Estimated by: Ensemble of deterministic AI models.

Total Uncertainty

Definition: Combined aleatoric and epistemic uncertainty.
Estimated by:

1) Ensemble of probabilistic AI models
2) Single "evidential" (higher-order probabilistic) AI model
3) Bayesian AI models

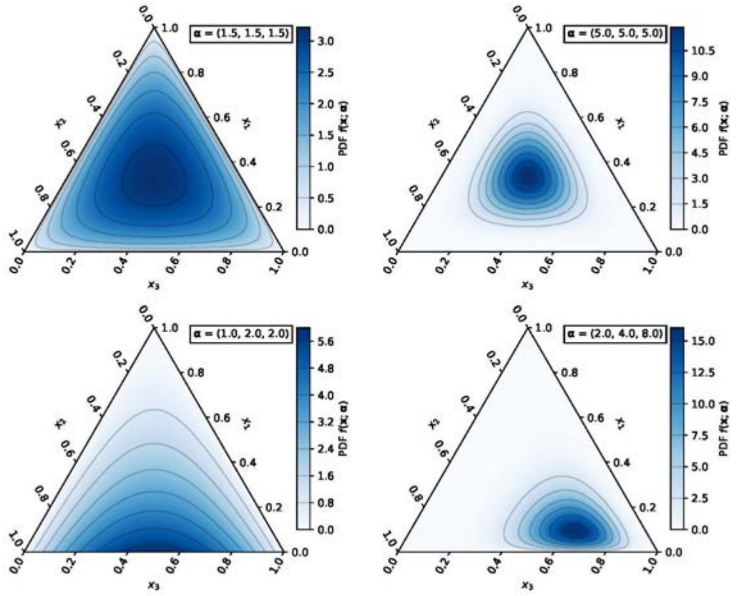
Collaborators
John Schreck, Charlie Becker, Gabrielle Gantos, Julie Demuth, Chris Wirz, Jacob Radford, Nick Bassil, Kara Sulia, Chris Thorncroft, Amy McGovern, Eliot Kim, Justin Willson, Kim Elmore, Maria Molina

Dirichlet Distribution: Model Classification Epistemic Uncertainty

$$
\begin{aligned}
\boldsymbol{\alpha} &= (\alpha_1, \ldots, \alpha_K) &&= \text{concentration hyperparameter} \\
\mathbf{p} \mid \boldsymbol{\alpha} &= (p_1, \ldots, p_K) &&\sim \operatorname{Dir}(K, \boldsymbol{\alpha}) \\
\mathbb{X} \mid \mathbf{p} &= (\mathbf{x}_1, \ldots, \mathbf{x}_K) &&\sim \operatorname{Cat}(K, \mathbf{p})
\end{aligned}
$$

then the following holds:

$$
\begin{aligned}
\mathbf{c} &= (c_1, \ldots, c_K) &&= \text{number of occurrences of category } i \\
\mathbf{p} \mid \mathbb{X}, \boldsymbol{\alpha} &\sim \operatorname{Dir}(K, \mathbf{c} + \boldsymbol{\alpha}) &&= \operatorname{Dir}(K, c_1 + \alpha_1, \ldots, c_K + \alpha_K)
\end{aligned}
$$



Source: Wikipedia

Theory of Evidence and Subjective Logic

How can we summarize epistemic uncertainty more effectively?

Classification probabilities must sum to 1, but what if we removed that restriction?

Subjective logic (SL) formulates *belief* b_k over K classes, plus u or "**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 belief mass, which is the normalized sum of evidence for an outcome.

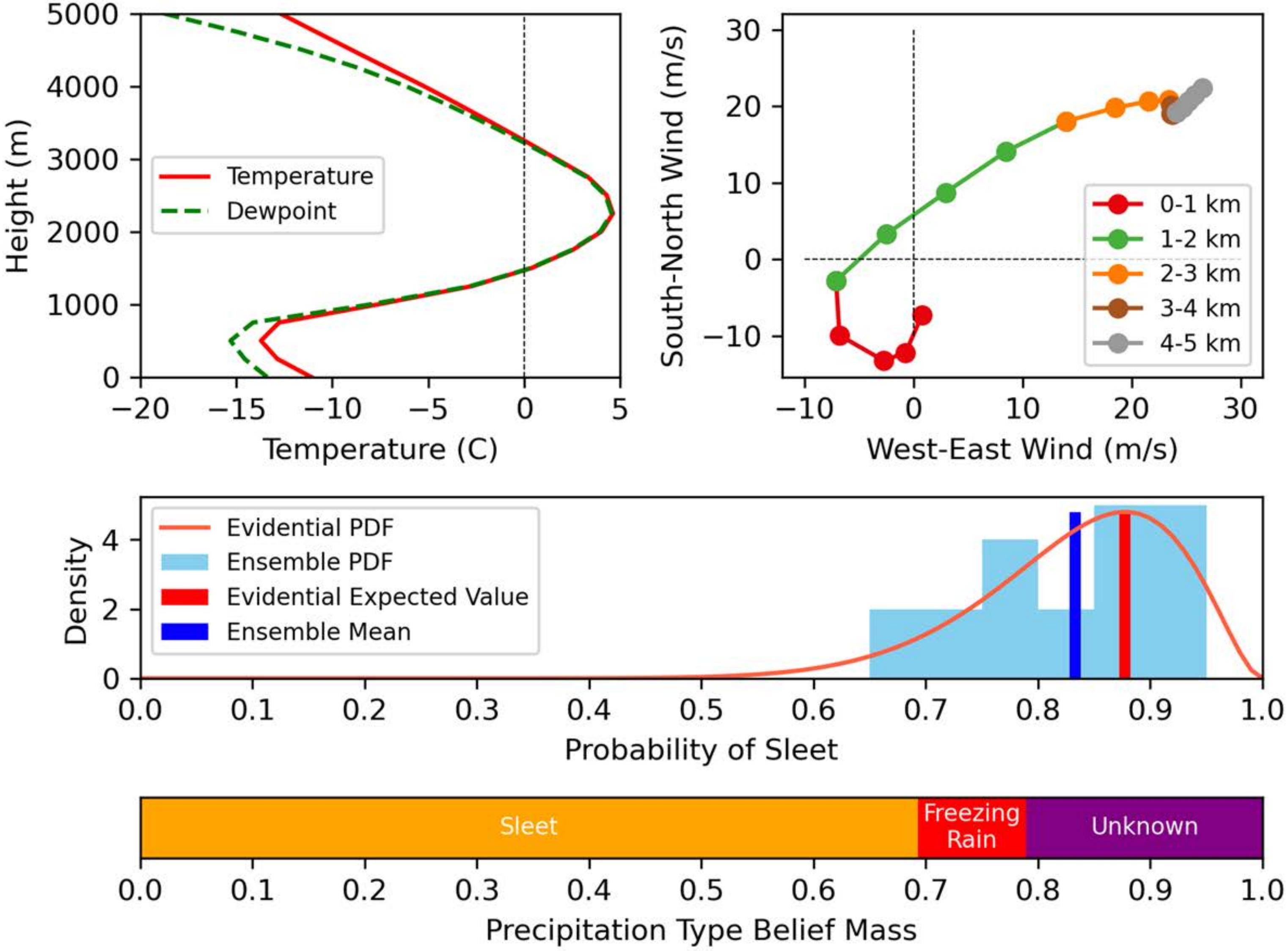
Each b_k is defined as

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

Dirichlet distributions can be updated based on adding new evidence to each outcome.

Source: Sensoy et al. 2018

Evidential Deep Learning



Full Classifier Evidential Loss

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

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

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

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

Pushes incorrect alphas toward 1 (uniform distribution)

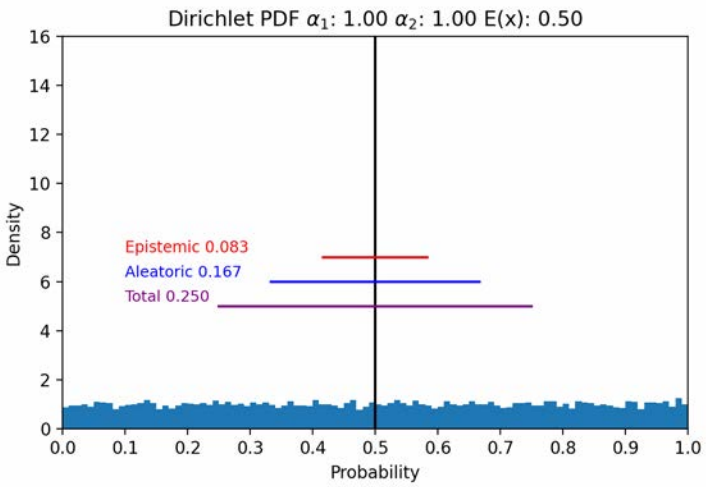
Sensoy, M., L. Kaplan, and M. Kandemir, 2018: Evidential deep learning to quantify classification uncertainty. *arXiv [cs.LG]*, https://arxiv.org/abs/1806.01768.

Dirichlet Aleatoric and Epistemic Uncertainties

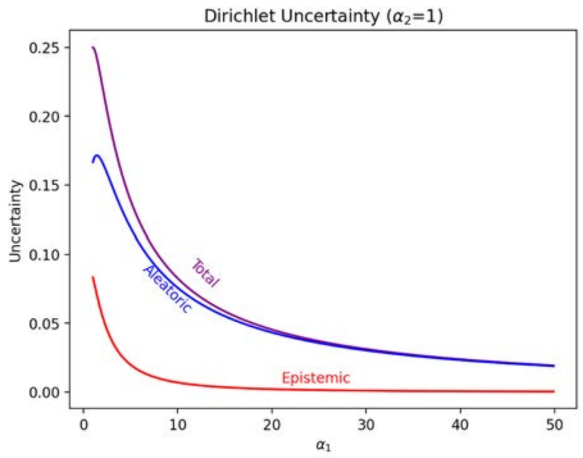
Law of total variance decomposes the total uncertainty into the sum of the unexplained variance plus the explained variance:

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

Aleatoric (unexplained) =
$$\begin{aligned}\mathbb{E}\left\{\mathrm{Var}(y_j|\boldsymbol{p})\right\} &= \mathbb{E}\left\{p_j(1-p_j)\right\} \\ &= \mathbb{E}(p_j) - \mathbb{E}(p_j^2) \\ &= \mathbb{E}(p_j) - \{\mathbb{E}(p_j)\}^2 - \mathrm{Var}(p_j) \\ &= \frac{\alpha_j}{S} - \left(\frac{\alpha_j}{S}\right)^2 - \frac{\frac{\alpha_j}{S}\left(1-\frac{\alpha_j}{S}\right)}{S+1}\end{aligned}$$



Epistemic (explained) =
$$\begin{aligned}\mathrm{Var}\left\{\mathbb{E}(y_j|\boldsymbol{p})\right\} &= \mathrm{Var}(p_j) \\ &= \frac{\frac{\alpha_j}{S}\left(1-\frac{\alpha_j}{S}\right)}{S+1}\end{aligned}$$



Total = Aleatoric + Epistemic

Probabilistic Forecast Example: Classifying Winter Precipitation Type

<u>Data</u>

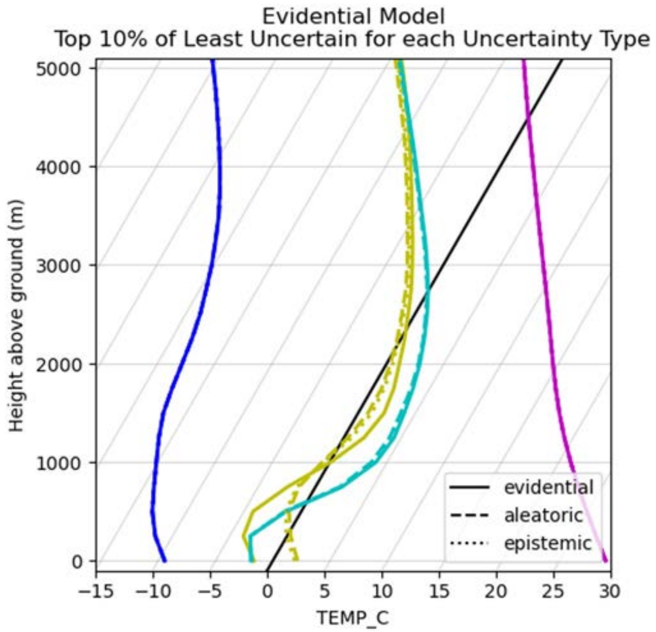
- ➢ **NOAA Rapid Refresh** Vertical Profiles
- ➢ Interpolate from pressure to height coords

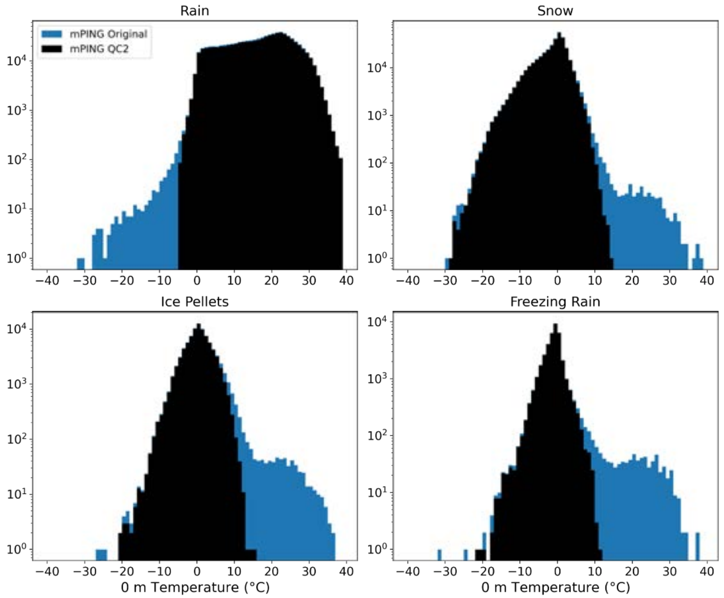
<u>Input</u> (0 - 5 km above surface, every 250 meters)

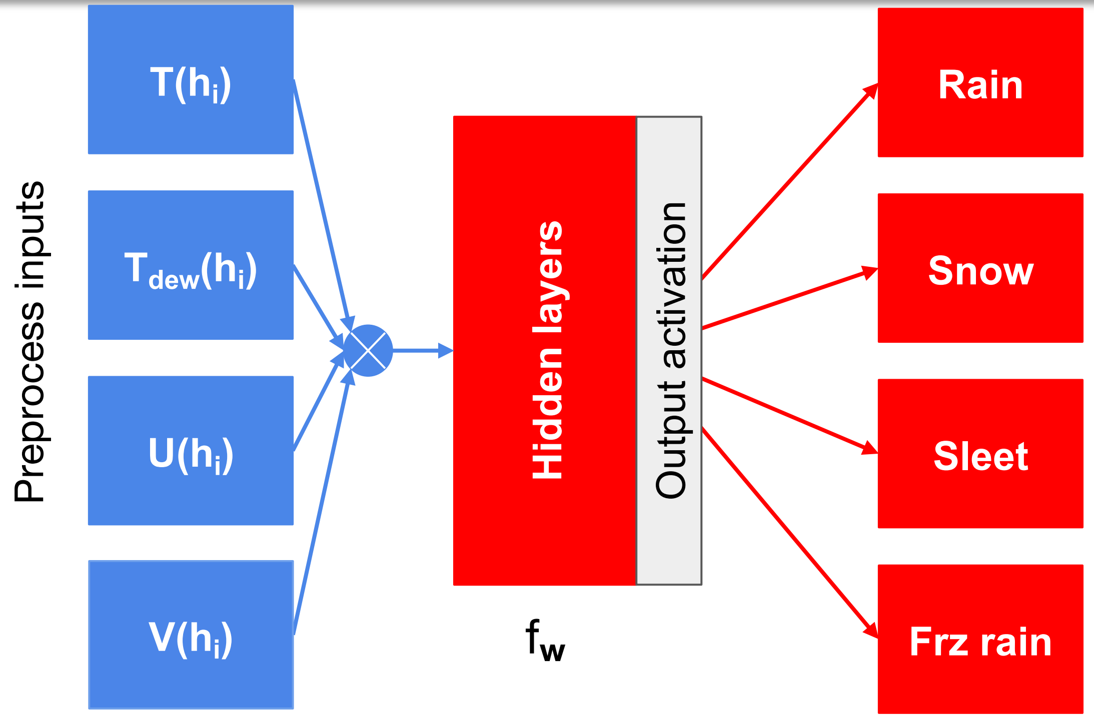
- ➢ Temperature, Dewpoint, U-Wind, V-Wind

<u>Target</u>

- ➢ mPING Crowd-sourced reports of winter precipitation types
  - ➢ *Rain, Snow, Sleet, Freezing Rain*







(i) Deterministic:

Predict probabilities for classes

Loss = Cross-entropy

p_k = Softmax($f_w(T,T_{dew},U,V))_k$

(ii) Evidential:

Predict evidence for classes

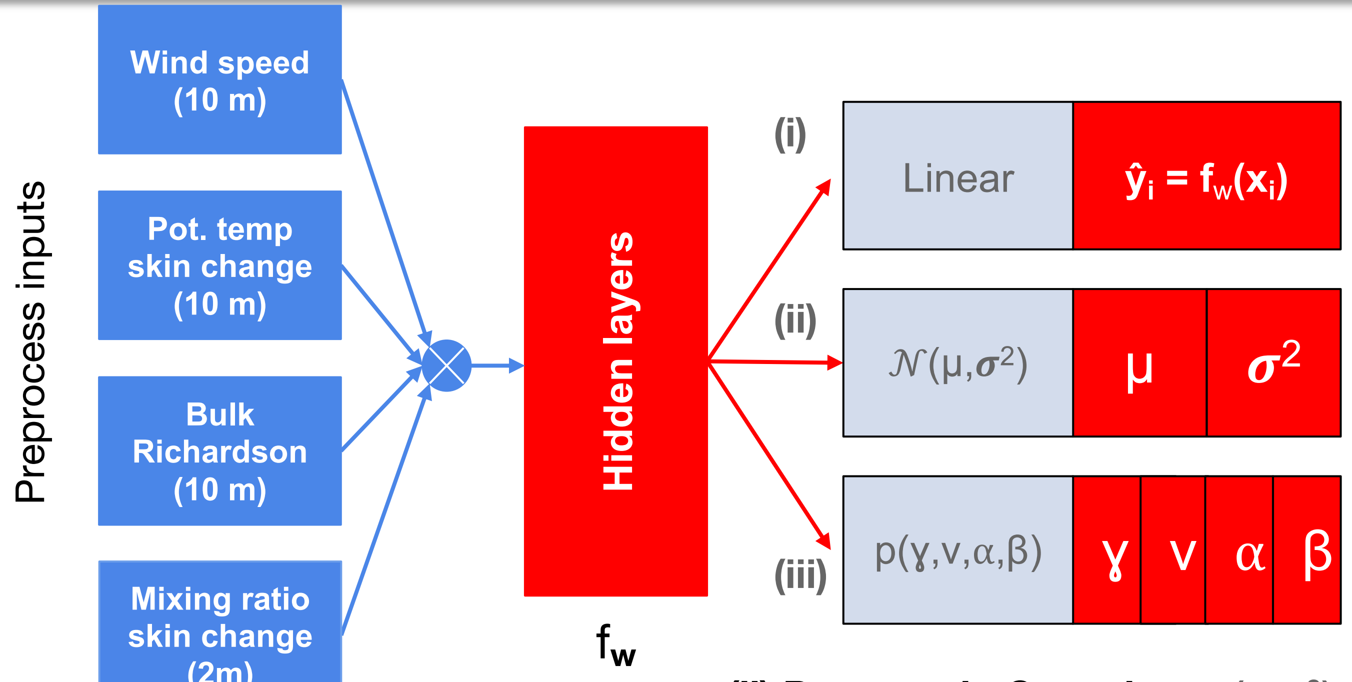
Loss = Evidential

e_k = ReLU($f_w(T,T_{dew},U,V))_k$

$\alpha_k = e_k + 1$

Compute S, evidential u, and the probabilities p_k

(a) P-type (categorical problem)



(i) Deterministic:

Predict values for the defined tasks

Loss = RMSE/MAE/etc

Number of outputs = number of tasks

(ii) Parametric Gaussian $\mathcal{N}(\mu,\sigma^2)$:

Predict the mean and variance for each task

Loss = NLL

Number of outputs = 2 * number of tasks

(iii) Parametric Normal-Inverse Gamma $p(\gamma,v,\alpha,\beta)$:

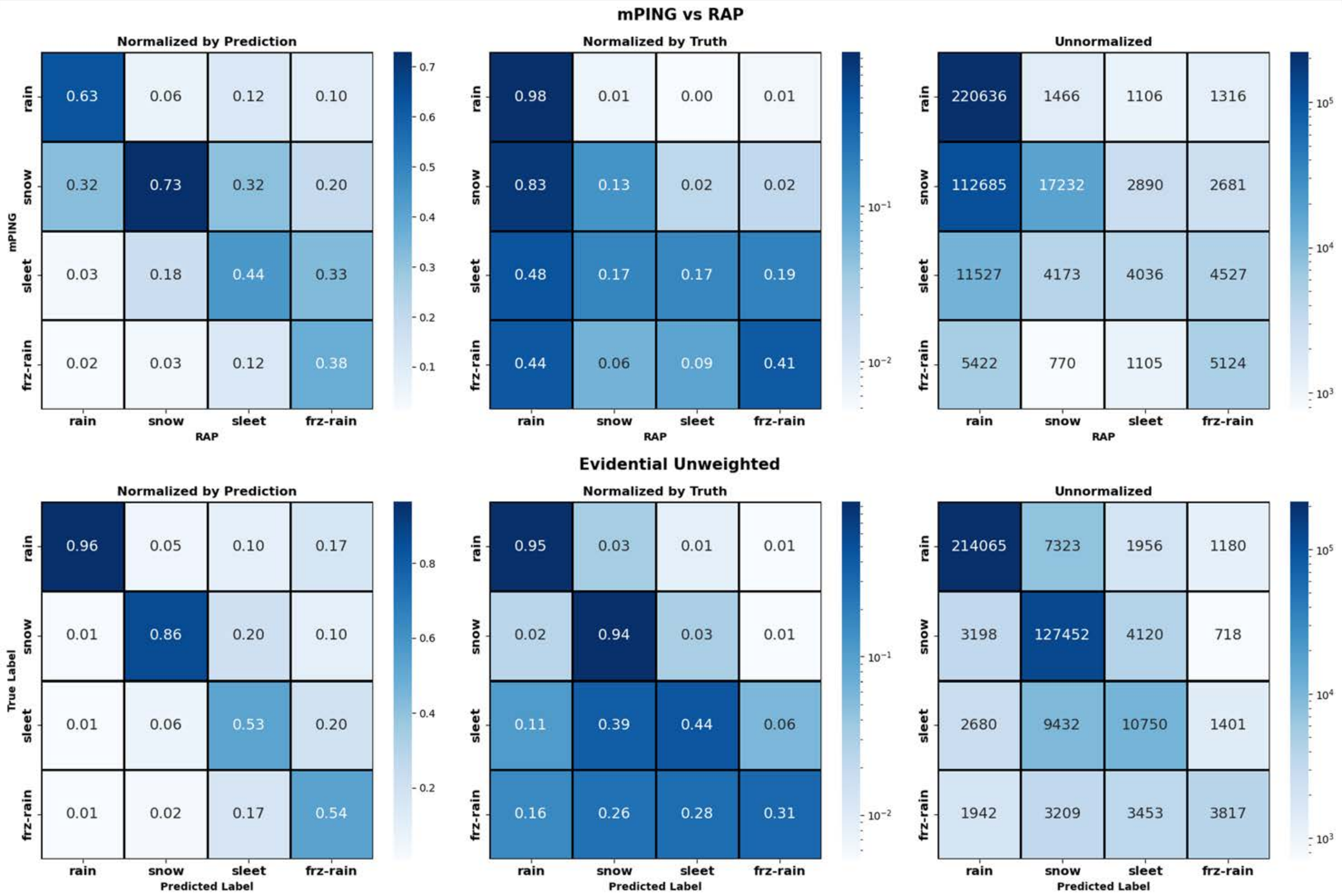
Predict evidence for parameters for each task

Loss = Evidential, Number of outputs = 4 * number of tasks

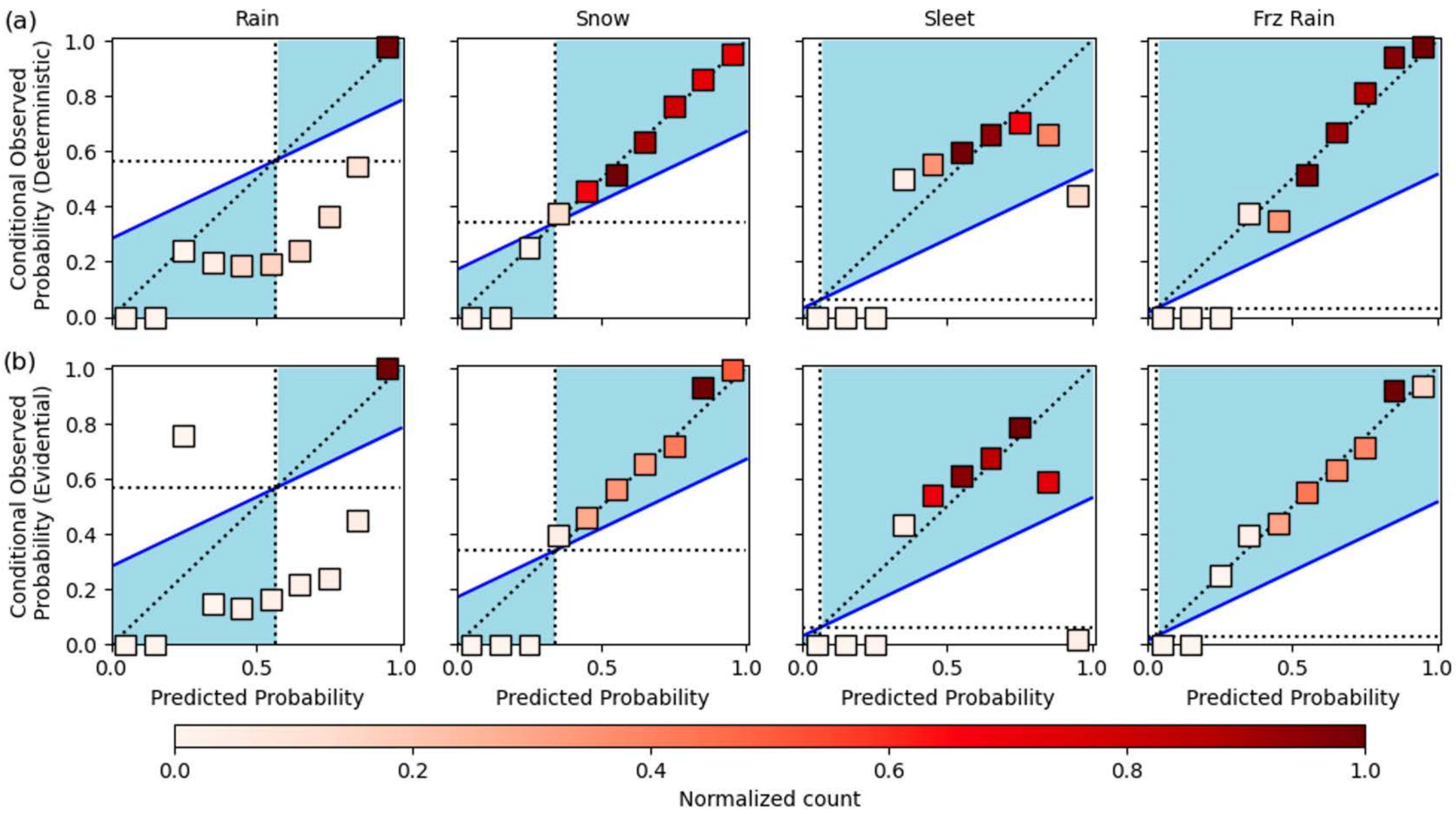
Post-prediction: Compute mean, aleatoric, and epistemic uncertainties

(b) Surface layer (regression problem)

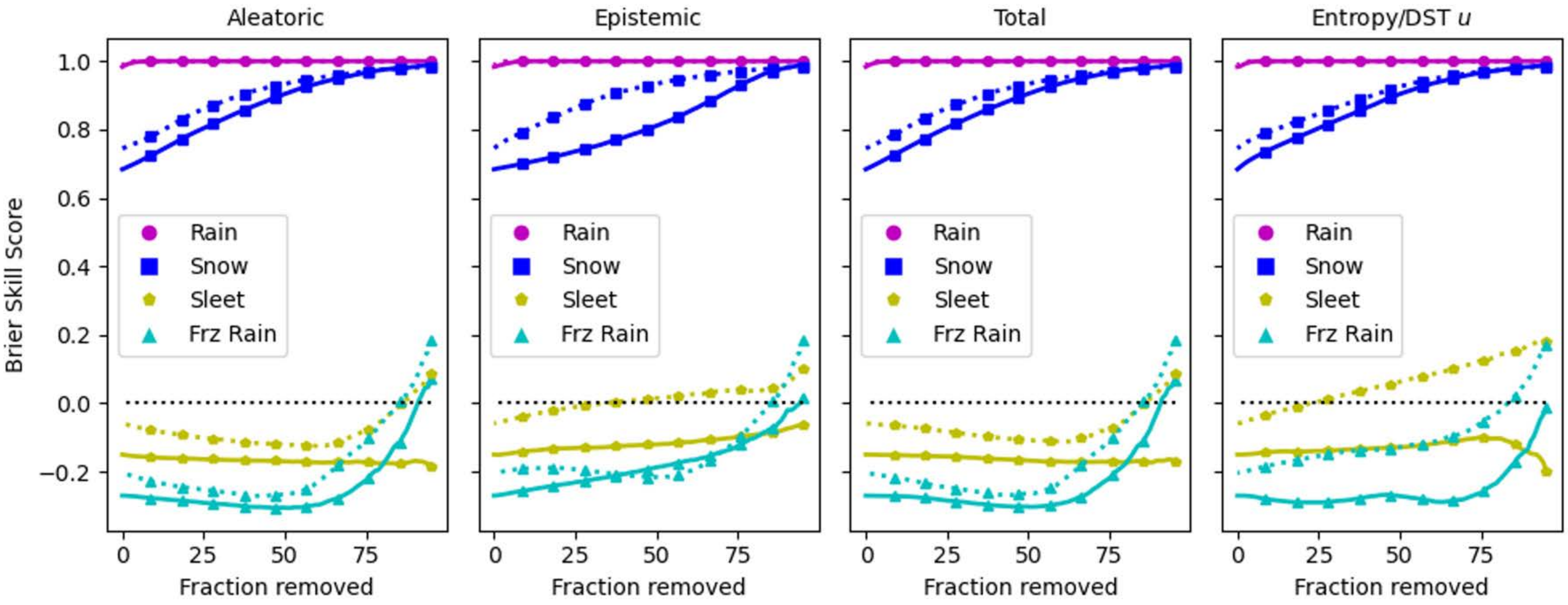
Precipitation-Type Validation



Probabilistic Verification

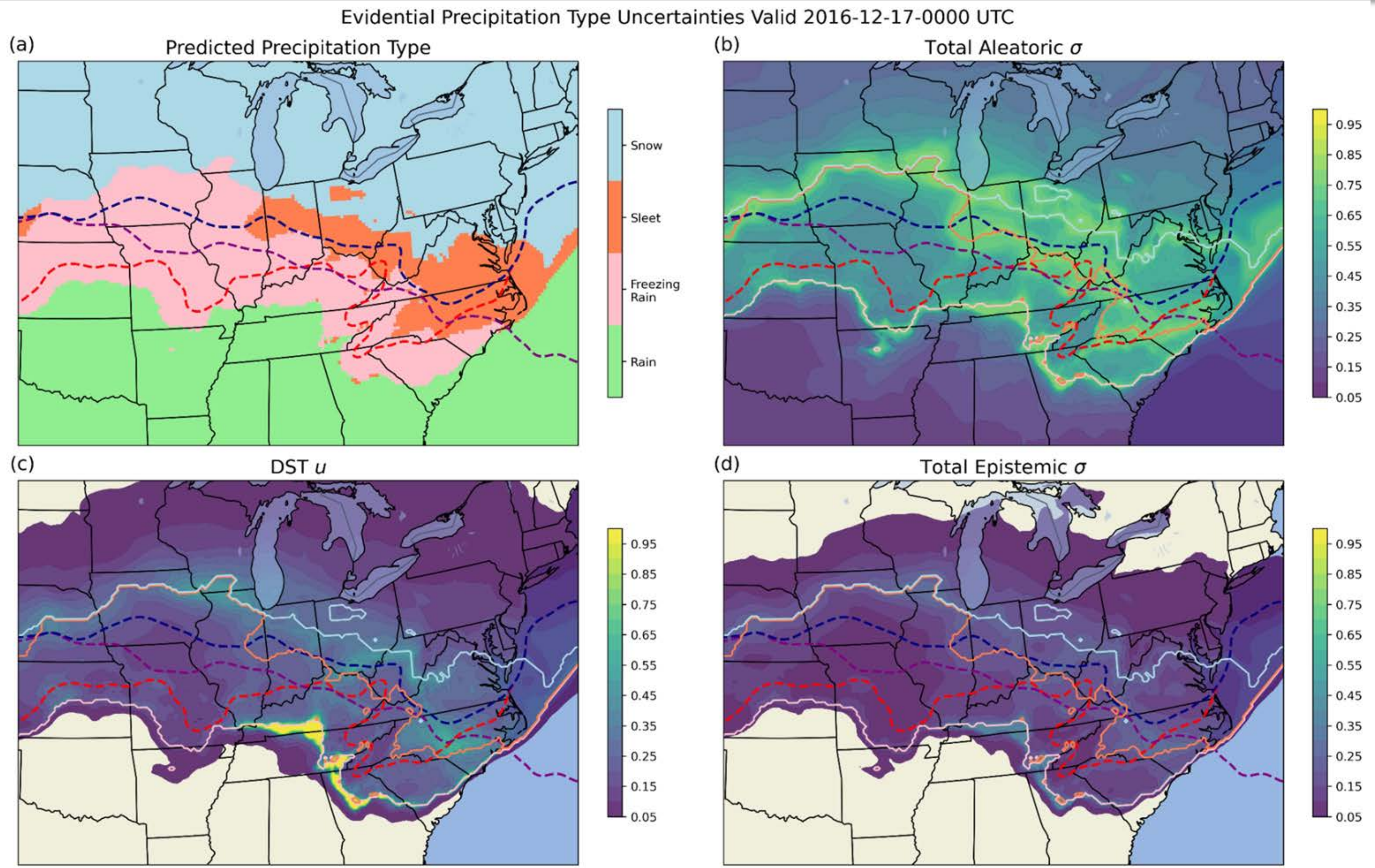


P-type Drop Fraction

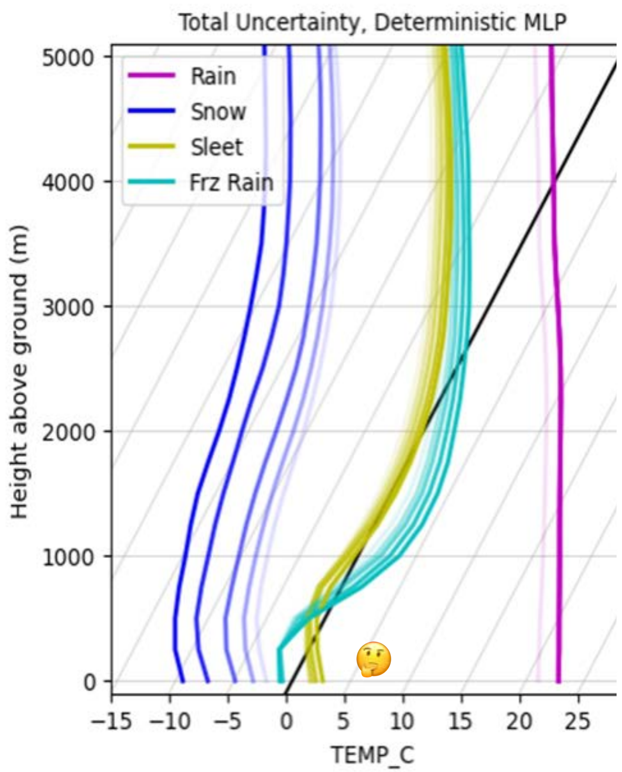


How well does each type of uncertainty discriminate between easier and harder to classify events?

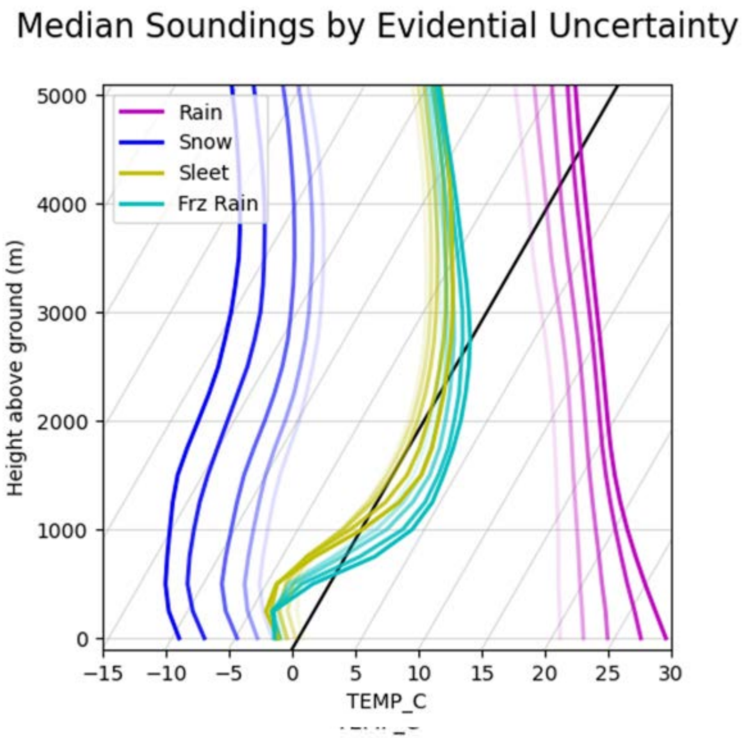
Regional Case Study



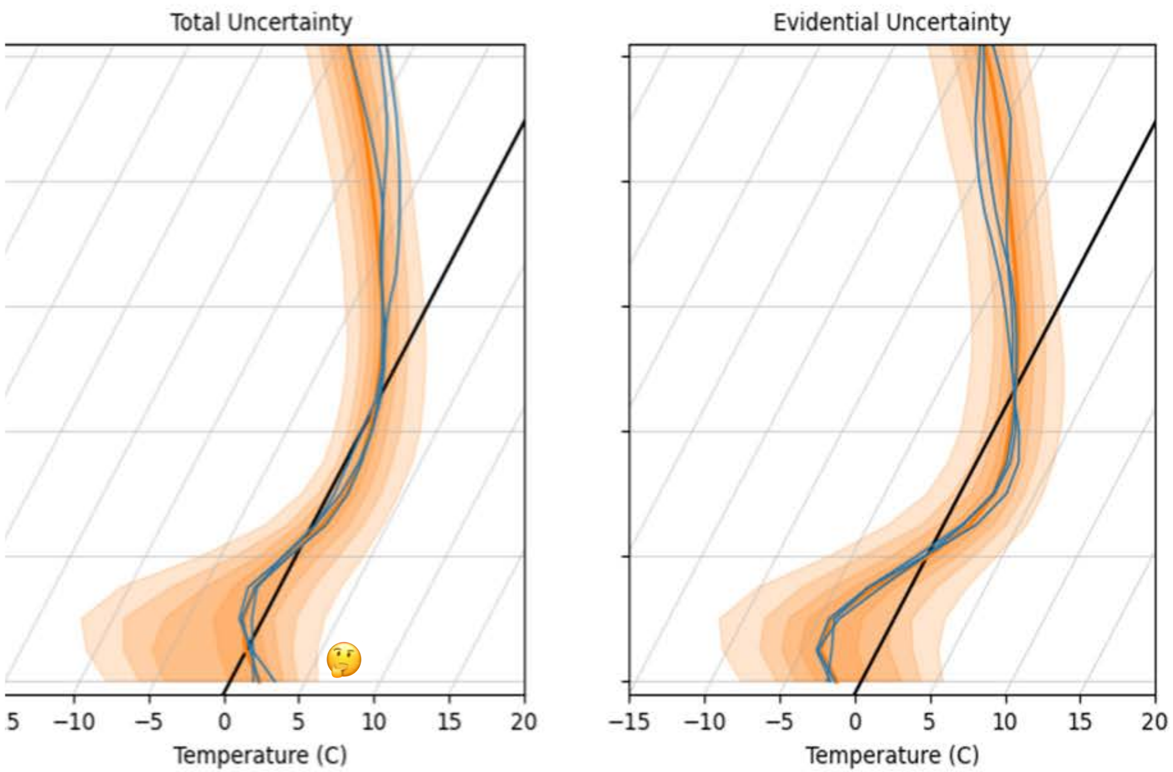
Evaluation: Binned by Uncertainty



MLP with Monte Carlo Dropout



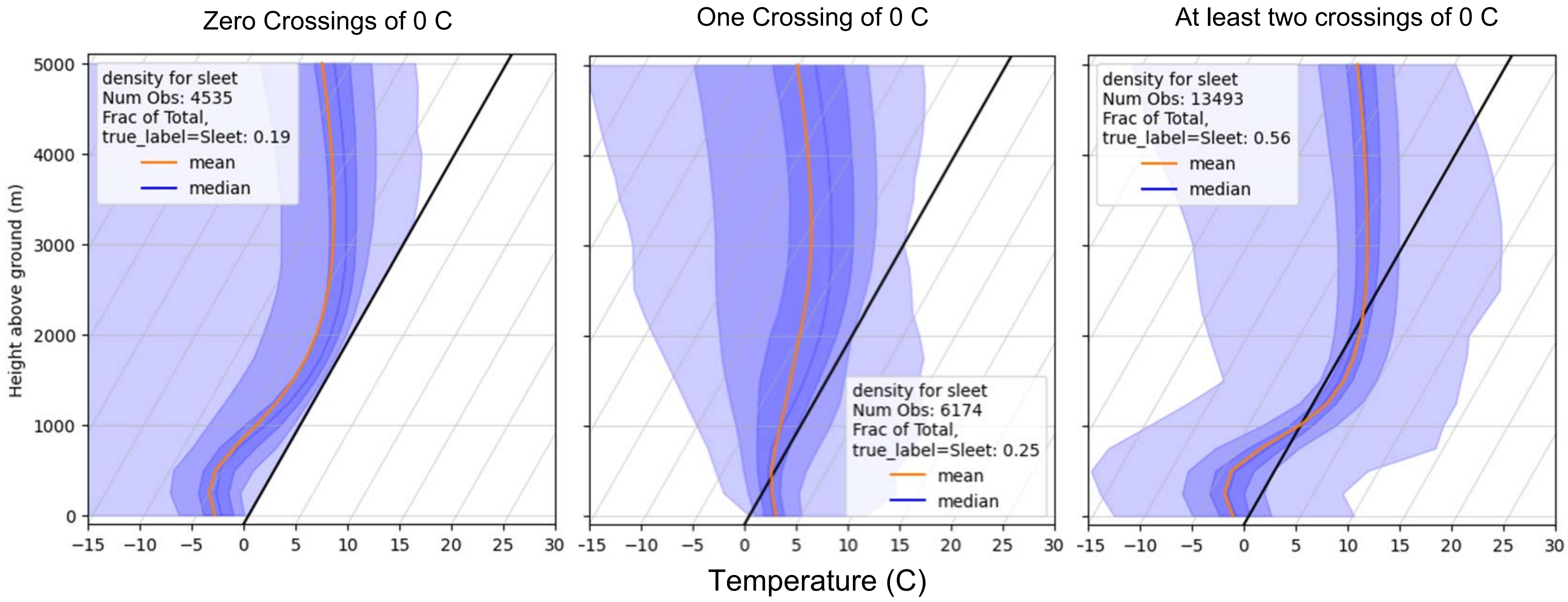
Evidential Model



Evidential Model, Sleet

Root cause: Data Quality

- "ground truth" labels are from crowdsourced observations
- some quality control done, but not enough:



Post hoc XAI methods

Gradient * Input	Which **features are most influential** in predicting the model's output?
Shapley Additive Explanations (SHAP)	How much does **each feature contribute to the model's predictions**?
Permutation Feature Importance	How does the **performance of the model change** when the information content of a feature is destroyed?

$$\mathbf{A}^{c}_{\text{Gradient}\odot\text{Input}} = \frac{\partial S_c(\mathbf{x})}{\partial \mathbf{x}} \odot \mathbf{x}.$$

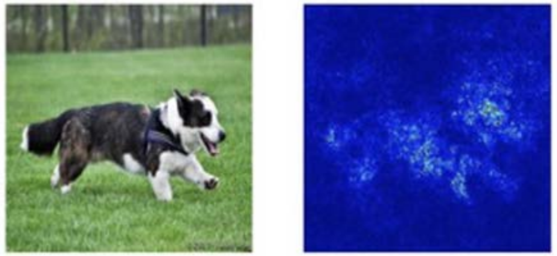
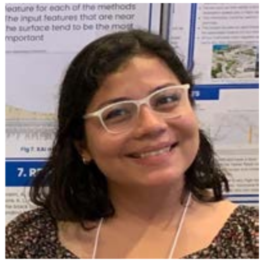
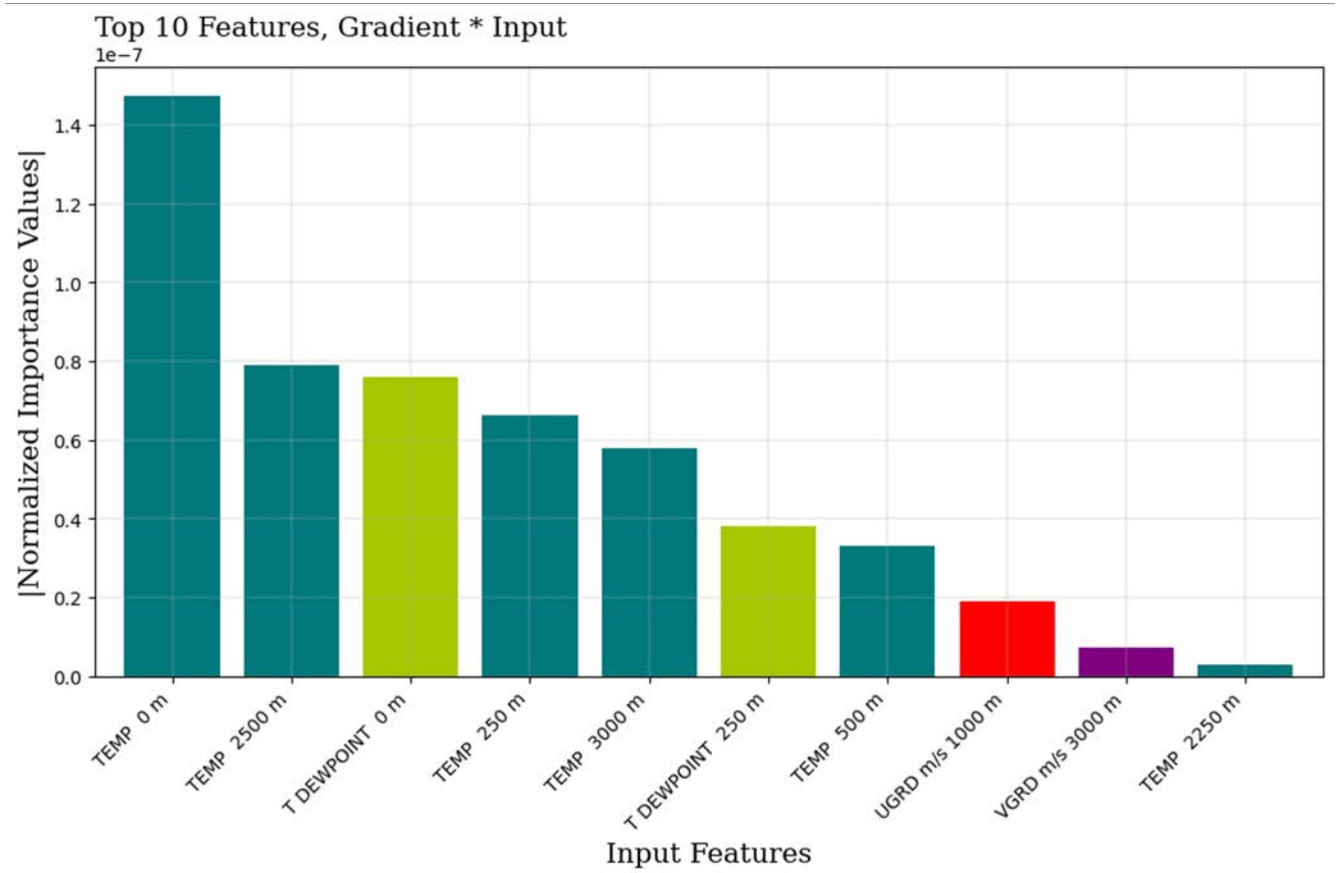


Fig. 3 Input * Gradient attribution method



Gradient * Input

Which **features are most influential** in predicting the model's output?

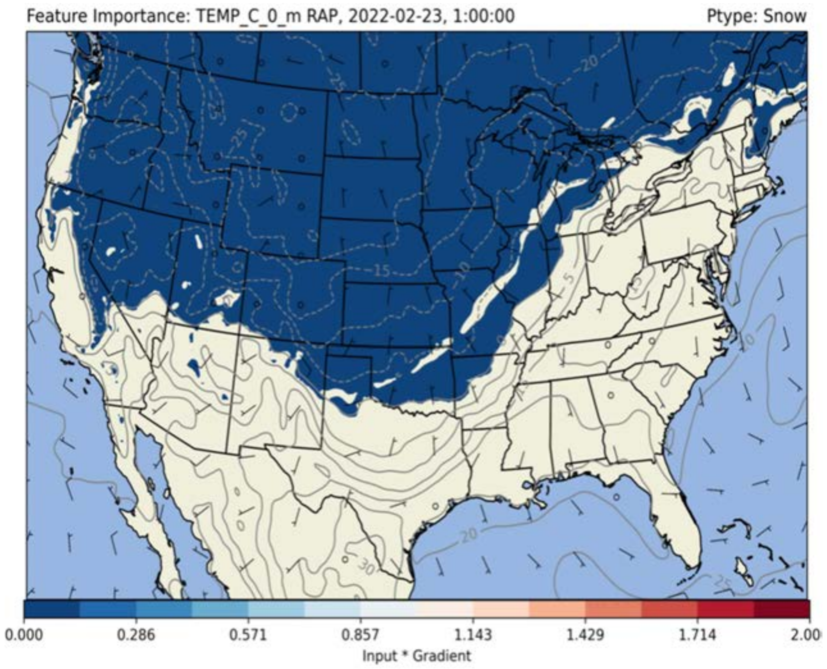


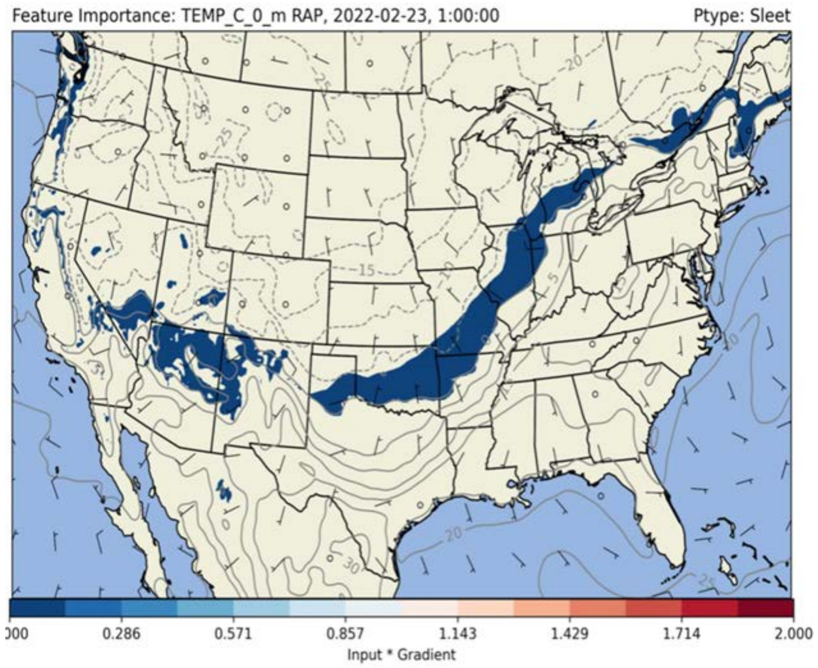
Gradient * Input works by multiplying the gradient of the model's output with the input features.

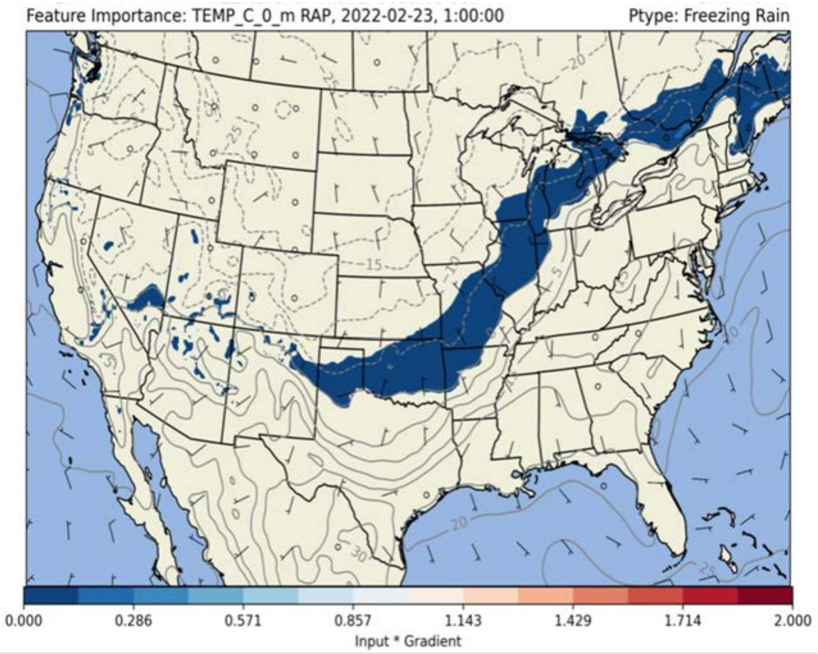
$$\mathbf{A}^c_{\text{Gradient}\odot\text{Input}} = \frac{\partial S_c(\mathbf{x})}{\partial \mathbf{x}} \odot \mathbf{x}.$$

Gradient * Input: CONUS plots

Which **features are most influential** in predicting the model's output?

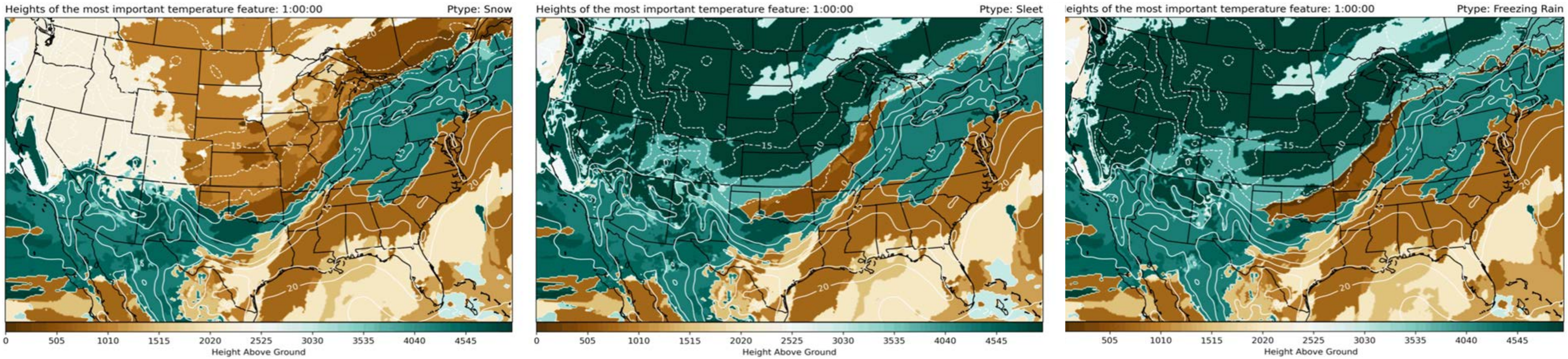






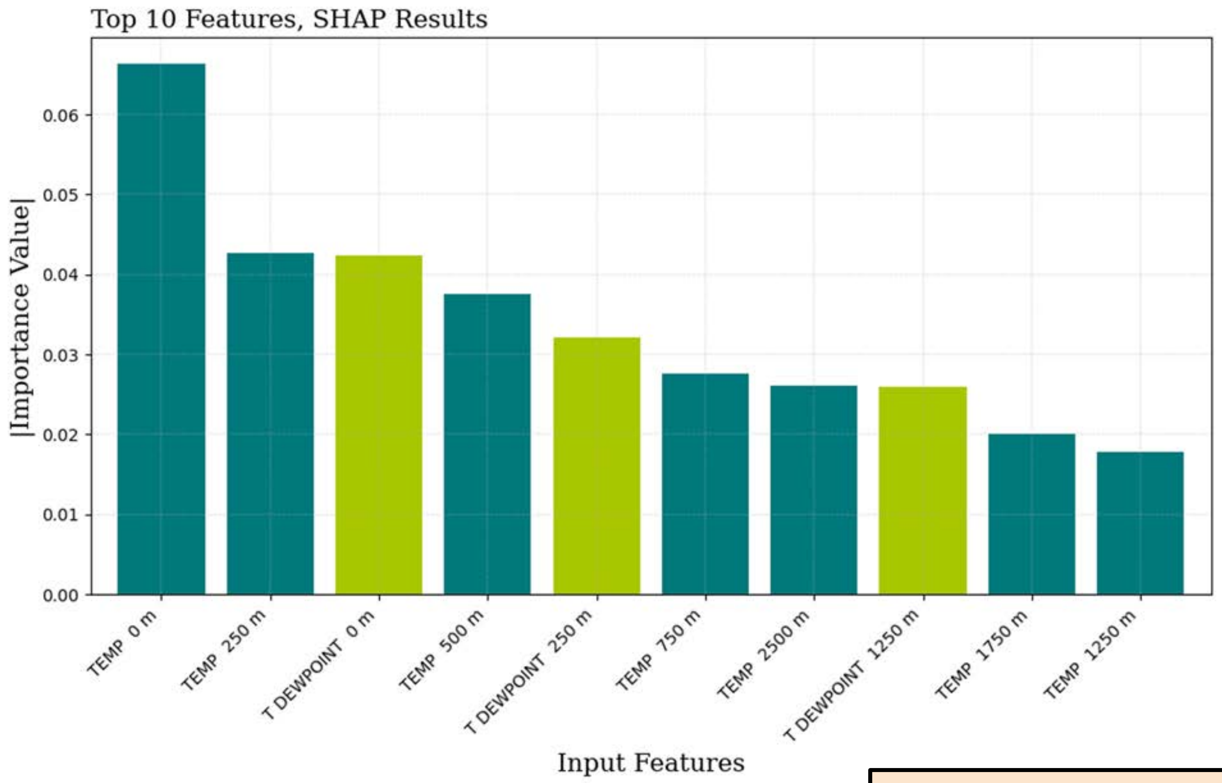
Gradient * Input: CONUS plots

Which **features are most influential** in predicting the model's output with respect to their height?



Shapley Additive Explanations (SHAP)

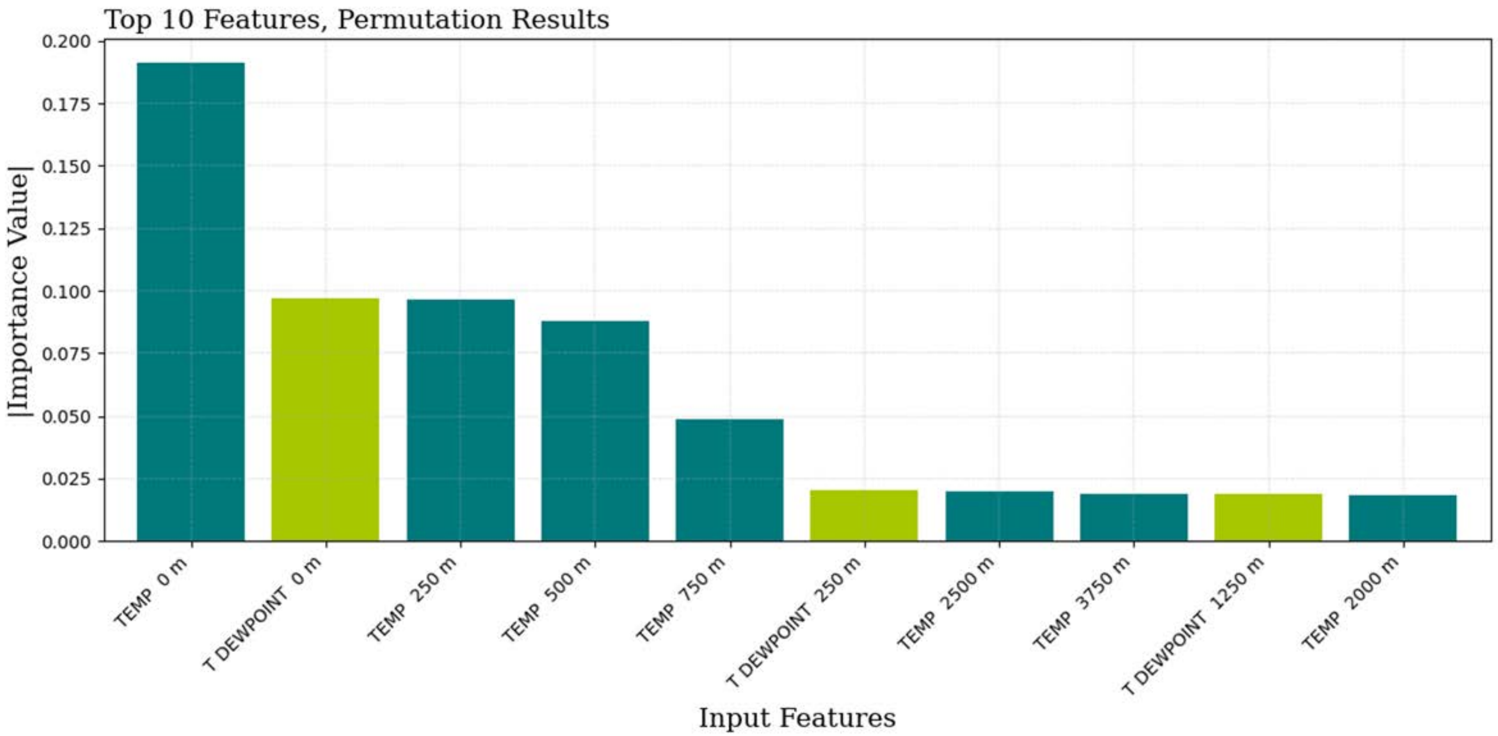
How much does **each feature contribute to the model's predictions**?



SHAP calculates the average contribution of each feature, representing how much each feature influences the model's prediction

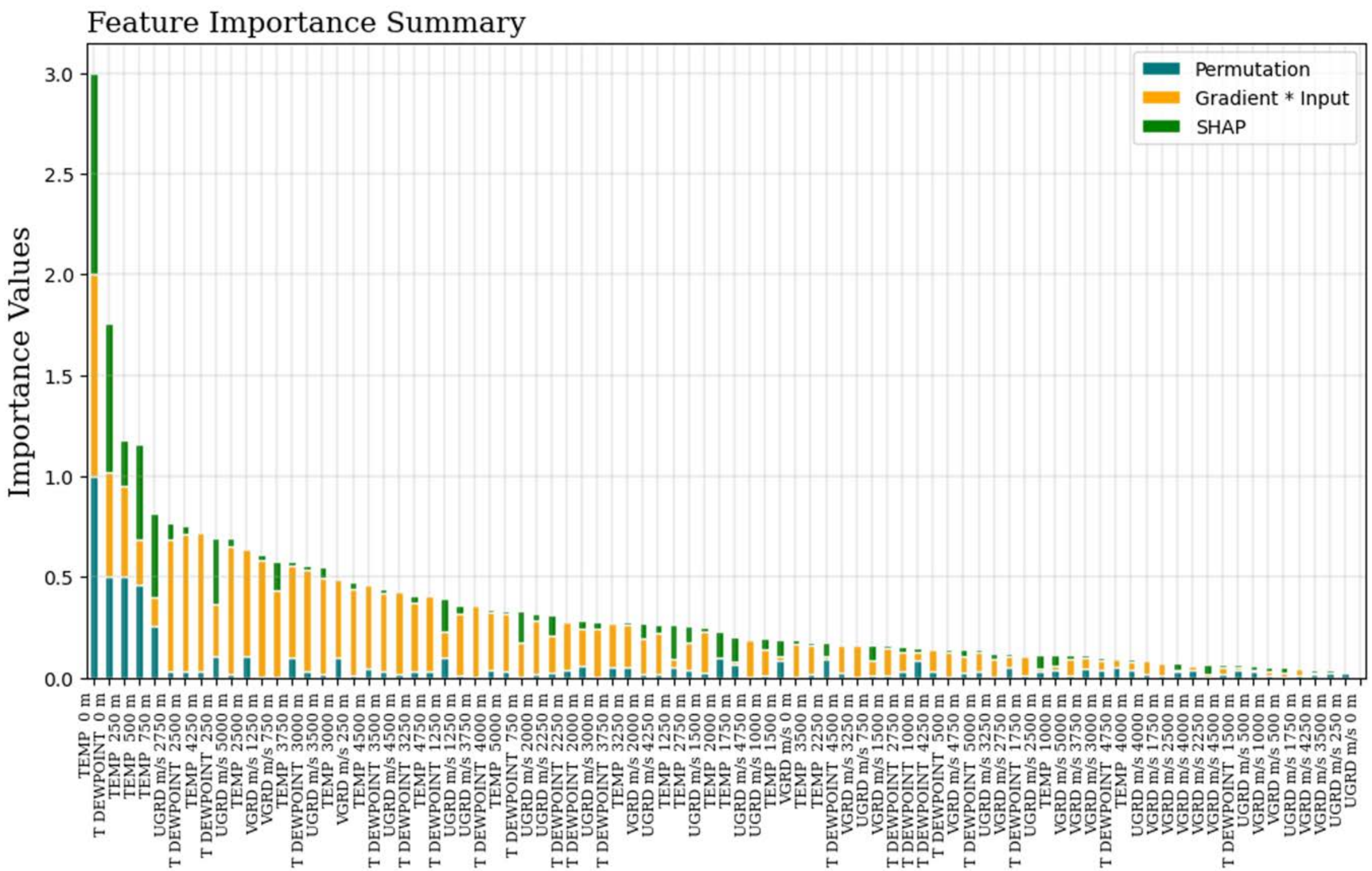
Permutation Feature Importance

What is the importance of each feature in predicting the model's output when the feature values are randomly shuffled?

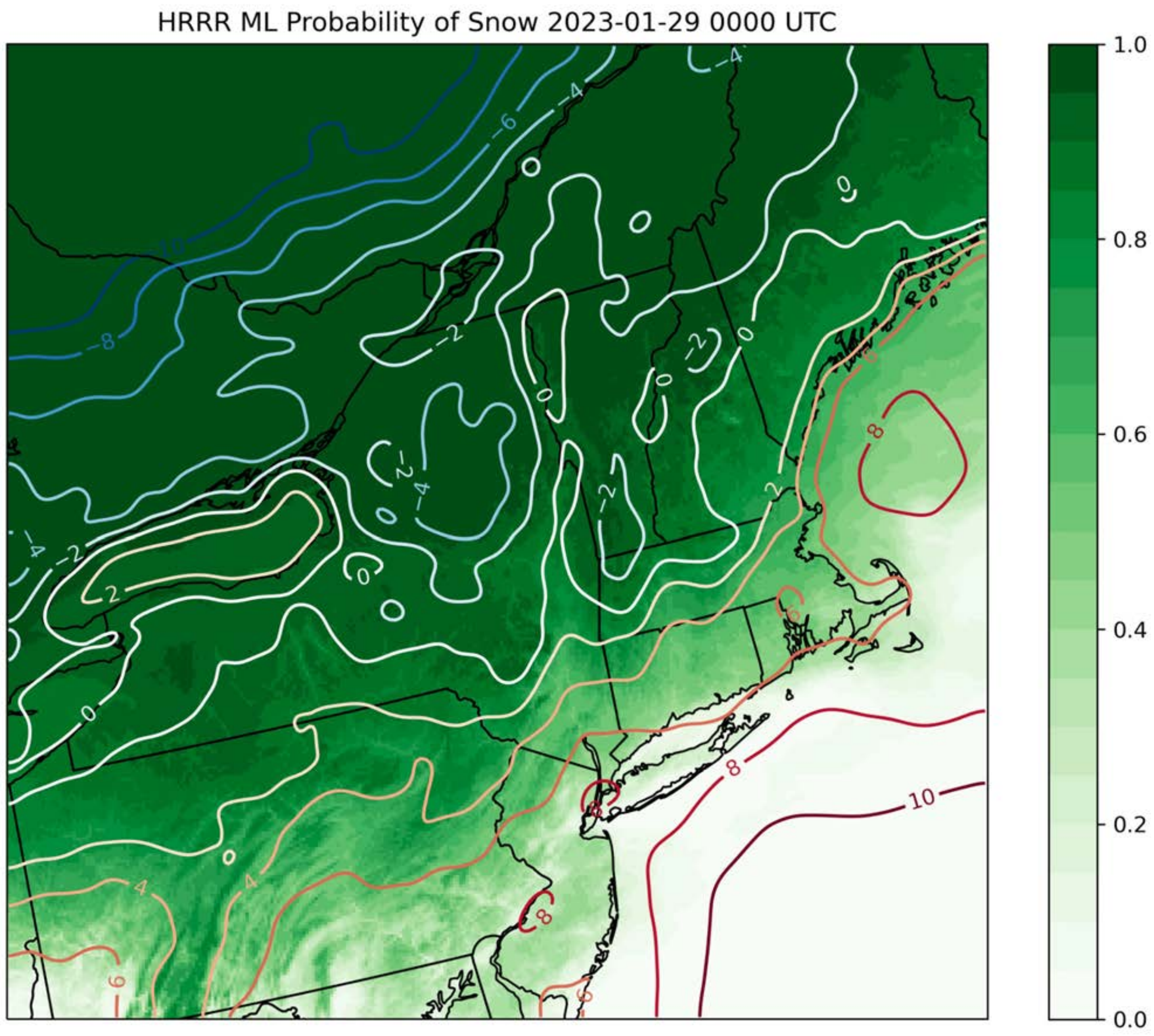


Permutation feature importance works by randomly shuffling the values of a single feature and measuring the resulting change in the model's performance. The feature with the largest change in performance is considered to be the most important feature.

XAI Results Summary



Transfer to Real-Time



- Planning to run in model in real-time on cloud this winter
- Working with risk communication team to perform interviews and/or experiments with stakeholders
- Have successfully run ML model on RAP, HRRR, and GFS in archival mode
- Partnered with Vaisala to test effect of ML p-type predictions on their road weather model

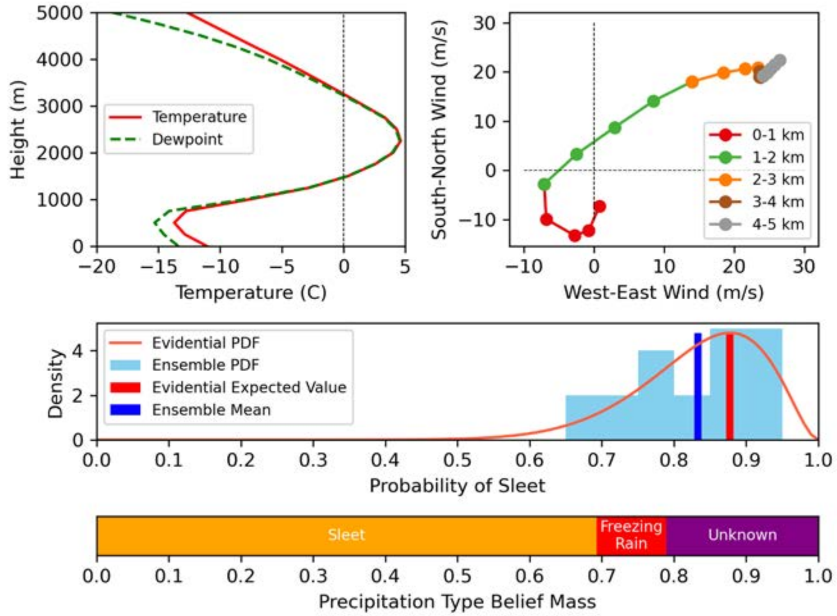
Limitations of Evidential Methods

- Requires calibration dataset to tune evidential regularizer coefficient
- Does not account for uncertainty in the inputs
- Uncertainty estimates will be underdispersive if the model is used outside its training context
  - e.g. train on observations/analysis but apply to forecast
  - transfer to different models
- No evidence prior may not be appropriate for rare events

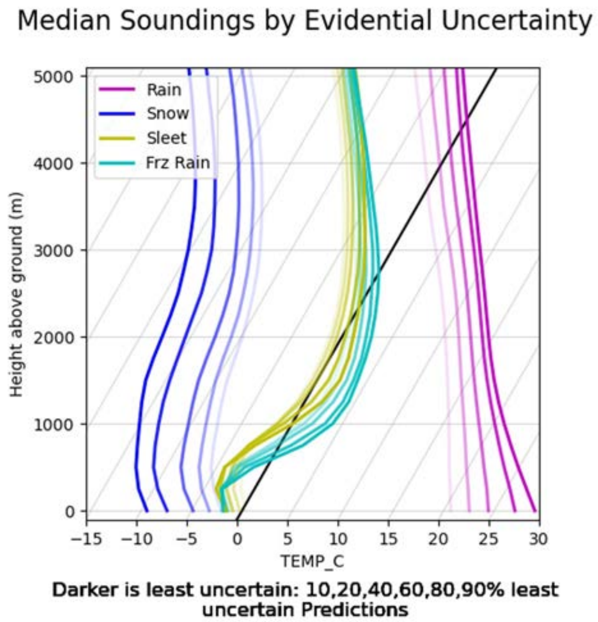
MILES Group Python Packages

- **miles-guess** (github.com/ai2es/miles-guess):
  - Implementations of evidential neural networks, deep ensembles, and Monte Carlo dropout
- **echo-opt** (github.com/NCAR/echo-opt):
  - Distributed hyperparameter optimization on HPC systems
  - Supports GPU allocation, XAI visualization for hyperparameter settings
- **hagelslag** (github.com/djgagne/hagelslag):
  - Object segmentation, tracking, and data extraction for convection-allowing model output
  - verification scores and plots
- **bridgescaler** (github.com/NCAR/bridgescaler):
  - Reproducible saving/loading of sklearn preprocessing scalers and transforms
  - Custom scalers for groups of variables and image patches
- **mlinwrf** (github.com/NCAR/mlinwrf):
  - Neural network and random forest implementations in Fortran
- **mlmicrophysics** (github.com/NCAR/mlmicrophysics):
  - Bin microphysics emulator for CAM/CESM

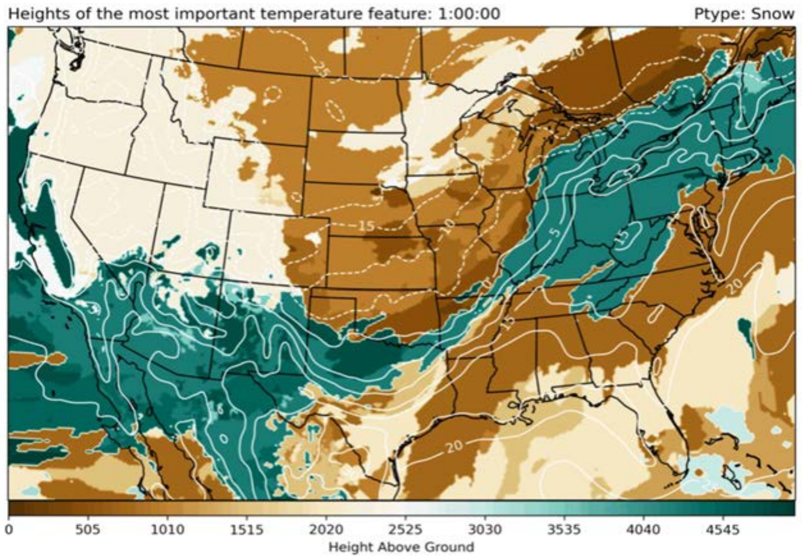
Summary



Evidential deep learning provides more comprehensive predictive uncertainty quantification.



Can composite soundings by uncertainty and get meaningful features



XAI diagnostics help connect predictions with atmospheric features.

Email: dgagne@ucar.edu