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

Schreck et al. 2024

Machine Learning Journal Club
November 19, 2024

Introduction: Uncertainty Quantification

Ensemble Forecast

- Physics-Based Model
  - Change Initial Condition, Boundary conditions, Model Specifications
  - High computational cost
  - No uncertainty calibration
- Machine Learning Model
  - Cross Validation → Vary in data split
  - Deep Ensemble → Vary in weight initialization
  - Monte Carlo Dropout → Randomly deactivate weights, or using dropout layer
  - Still expensive
- Evidential Deep Learning

Motivation - Aleatoric and Epistemic Uncertainty

Consider two extreme cases:

1. Unpredictable output data (Aleatoric)

Input	3	4	3	4	3	4	3	4
Output	1	2	3	4	5	6	7	8

2. Noisy input data (Epistemic)

Input	1	2	3	4	5	6	7	8
Output	3	4	3	4	3	4	3	4

ML Ensemble Result:

Unpredictable output

→ Unreliable predictions for all models

→ **Large spread in ensemble result**

Noisy input

→ Each subset captures different dynamics

→ Every model makes different predictions

→ **Large spread in ensemble result**

Aleatoric and Epistemic Uncertainty

Aleatoric Uncertainty

- a.k.a. Stochastic Uncertainty, unexplained component, EV, etc.
- We cannot make accurate predictions with the input. The output "seems" stochastic with the given input.

$$\mathrm{E}(\mathrm{Var}(Y|X))$$

Epistemic Uncertainty

- a.k.a. Systematic Uncertainty, explained component, VE, etc.
- The input data is noisy, or some mislabelled samples (thus, systematic), so the model trained by those data cannot make accurate predictions.

$$\mathrm{Var}(\mathrm{E}(Y|X))$$

Solution - Uncertainty Examples

Consider two extreme cases:

1. Unpredictable output data (Aleatoric)

Input	3	4	3	4	3	4	3	4
Output	1	2	3	4	5	6	7	8

2. Noisy input data (Epistemic)

Input	1	2	3	4	5	6	7	8
Output	3	4	3	4	3	4	3	4

$$\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y|X)] + \mathrm{Var}(\mathbb{E}[Y|X])$$

Aleatoric Epistemic

How to improve model

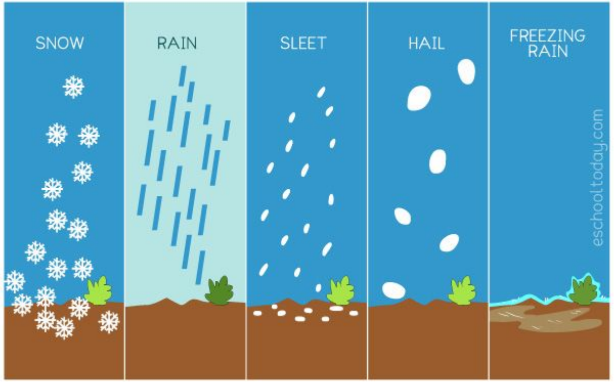
- Technically, we cannot remedy this problem
- Add other input that better describes the output

- Gather model data to make training independent from training split
- Reduce model complexity(?), so less overfitting

X: Training Process: data noise, model config…

Y: Model Output: Y = f(input)

Precipitation Type Prediction Example



Input: 4 meteorological variables (T, Td, u, v)

Output: Rain, Frozen Rain, Sleet, Snow

Problem: Inconsistent precipitation data

- The data and report are **crowd-sourced reports**. The outcome may vary due to subscale meteorological and societal factors
- **Mislabelled** sleet and freezing rain
- **Small Occurrences**: sleet and freezing rain

Therefore, we have probabilistic result for each label (multinomial distr.)

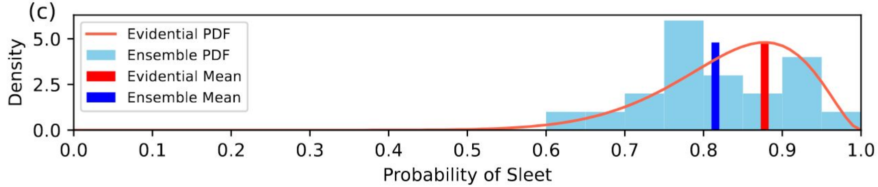
Evidential Neural Network

Traditional NN

- Loss: cross-entropy (logistic reg.)
- Output: Softmax, probability

Evidential NN

- Loss: MLE of Dirichlet Distribution
  - Mathematically convenience
- Output: probability density function



Sample Output, Fig 1.c

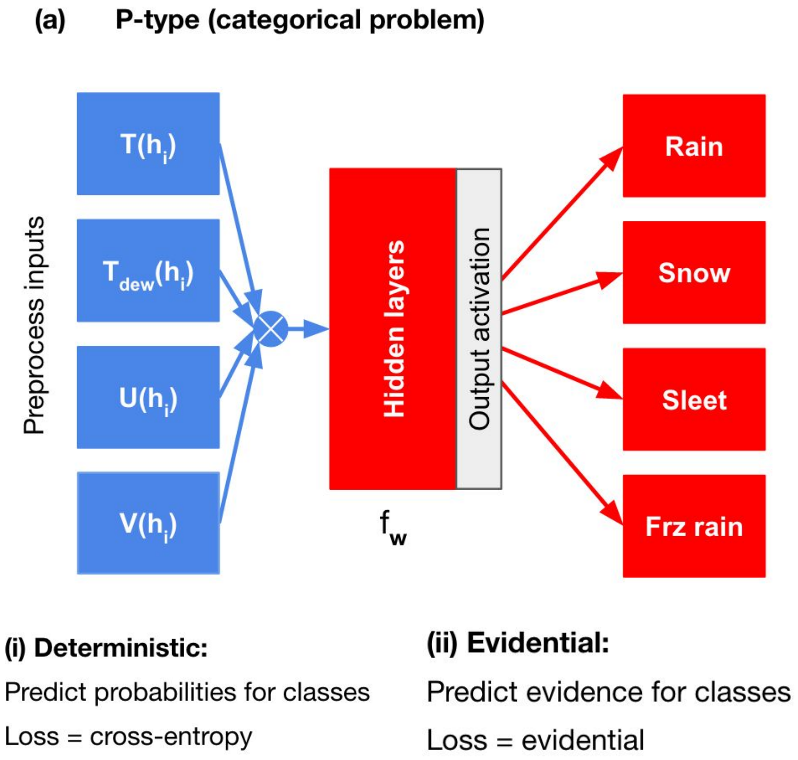


Fig 2.a

* For completeness, the real output equivalent in evidential NN is outputting normal-inverse gamma distribution

ENN Loss Function

- Minimize misclassification and uncertainty
- Kullback-Leibler Divergence (Eq.11)
    - Prevent premature convergence to uniform distribution due to misclassified samples.

Dirichlet Distribution

$$f(\mathbf{p}|\boldsymbol{\alpha}) = \begin{cases} \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} p_k^{\alpha_k - 1} & \text{for } \mathbf{p} \in S_K, B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)} \\ 0 & \text{otherwise,} \end{cases}$$

- ❖ Maximize the likelihood
- ❖ Conjugate prior integration, easy in Dir. Distr.

Misclassification Error | **KL Div., Regularizer**

$$\mathcal{L}_n(\mathbf{w}) = \sum_{k=1}^{K} (y_{n,k} - \hat{p}_{n,k})^2 + \frac{\hat{p}_{n,k}(1-\hat{p}_{n,k})}{S+1} + \nu_t \sum_{n=1}^{N} KL\left[D(\mathbf{p}_n|\widetilde{\alpha}_n)||D(\mathbf{p}_n|\mathbf{1})\right]$$

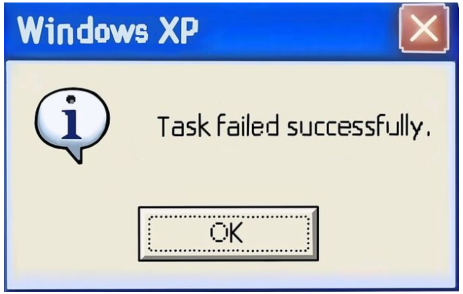
Uncertainty, Dirichlet Variance

Evaluation Metrics

Brier Score (BS)

- Basically feed label into MSE, ~~and trust the process works out, What a BS.~~

$$\mathrm{BS} = \frac{1}{NK}\sum_{n=1}^{N}\sum_{k=1}^{K}(y_{n,k} - p_{n,k})^2$$



- 0: Perfect Score, 1: Everything is Wrong (Perfectly Wrong?)

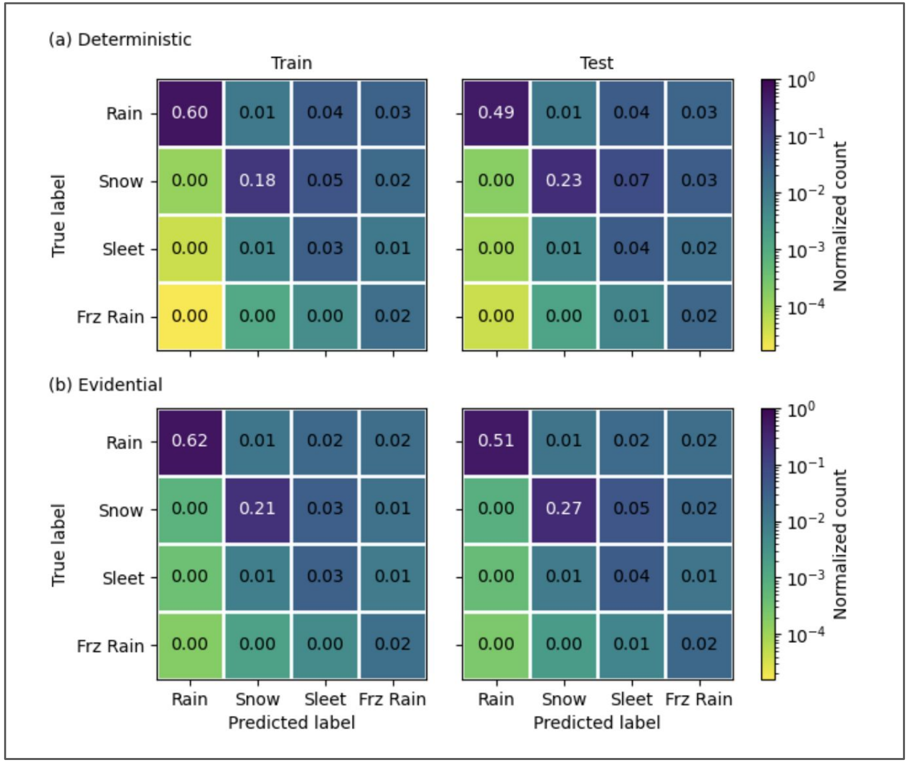
Brier Skill Score (BSS)

- Compare BS to Climatology

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}_{\mathrm{forecast}}}{\mathrm{BS}_{\mathrm{climatology}}}$$

- 1: Perfect Score, $-\infty$: Everything is Wrong including climatology

Results: Confusion Matrices (from supplemental of preprint…)

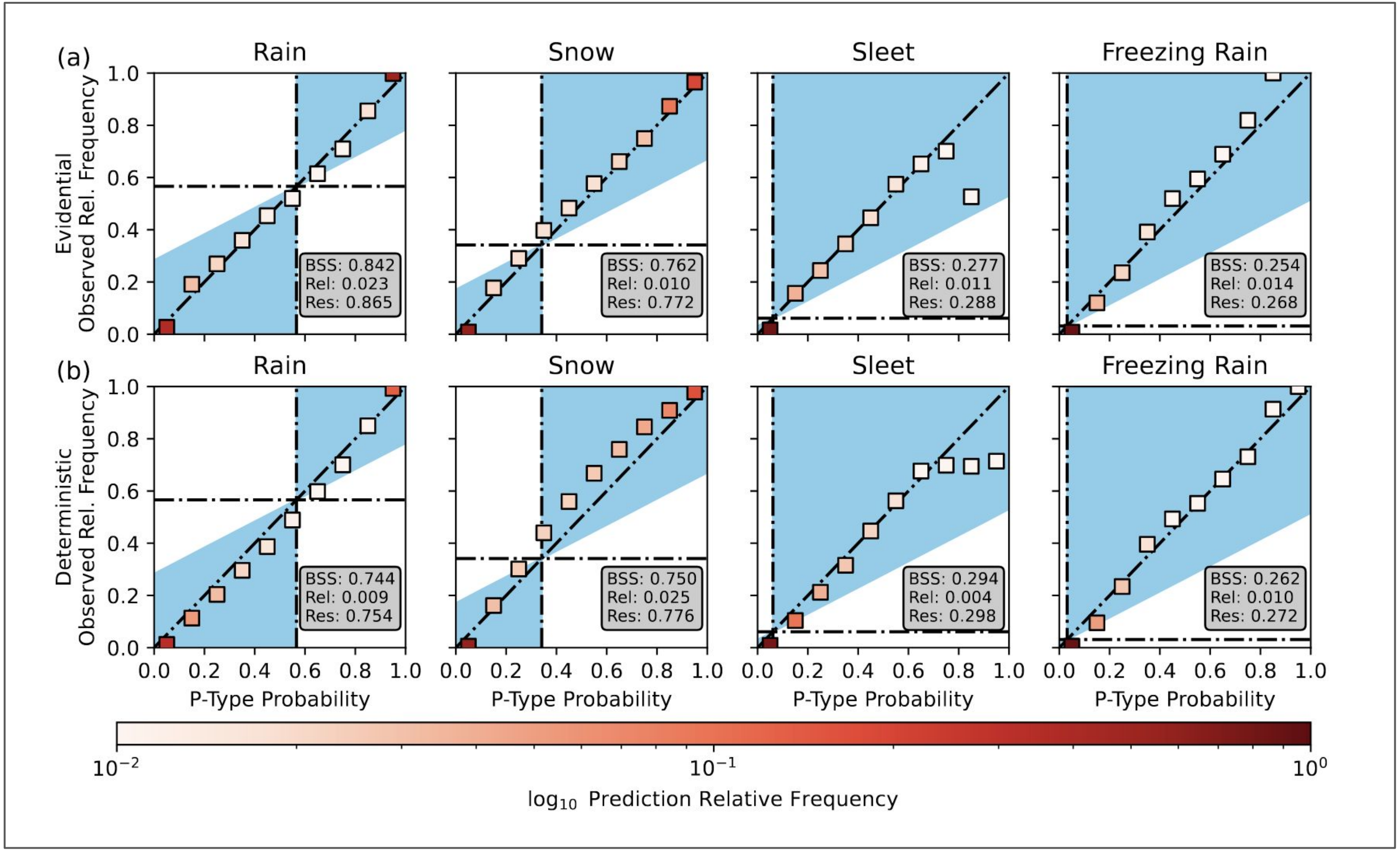


Key points:

1) Deterministic and evidential NNs have comparable performance

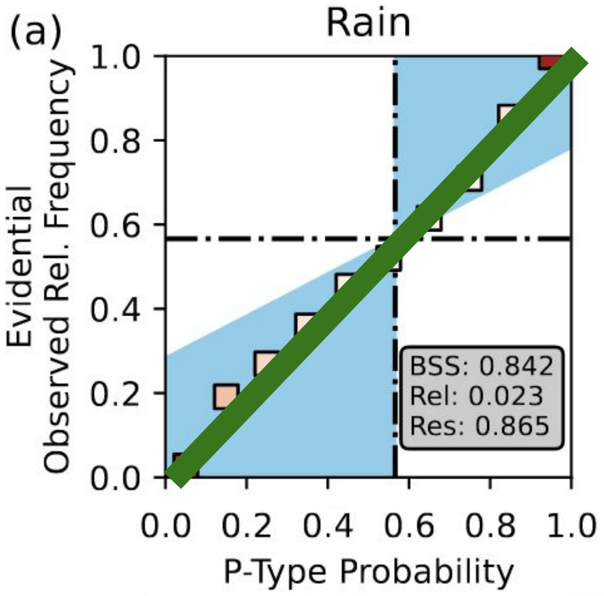
2) Very few sleet and freezing rain observations, as expected

Results



(a)
Rain
Snow
Sleet
Freezing Rain
Evidential
Observed Rel. Frequency
BSS: 0.842
Rel: 0.023
Res: 0.865
BSS: 0.762
Rel: 0.010
Res: 0.772
BSS: 0.277
Rel: 0.011
Res: 0.288
BSS: 0.254
Rel: 0.014
Res: 0.268
(b)
Rain
Snow
Sleet
Freezing Rain
Deterministic
Observed Rel. Frequency
BSS: 0.744
Rel: 0.009
Res: 0.754
BSS: 0.750
Rel: 0.025
Res: 0.776
BSS: 0.294
Rel: 0.004
Res: 0.298
BSS: 0.262
Rel: 0.010
Res: 0.272
P-Type Probability
10^{-2}
10^{-1}
10^{0}
$\log_{10}$ Prediction Relative Frequency

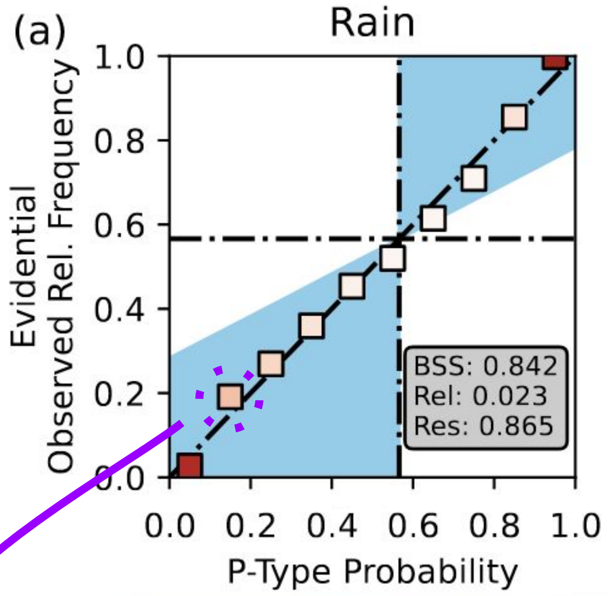
Results: Reliability Diagrams



Perfect Reliability

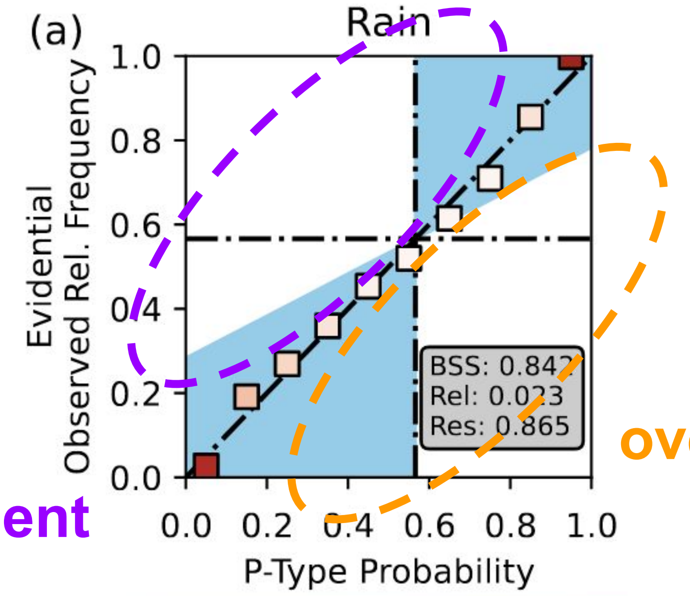
Reliability = Deviation of predicted <p-type> probability from relative frequency in observations

Results: Reliability Diagrams



For inputs with predicted P(rain)≈0.16, observed P(rain)≈0.19 → model underconfident

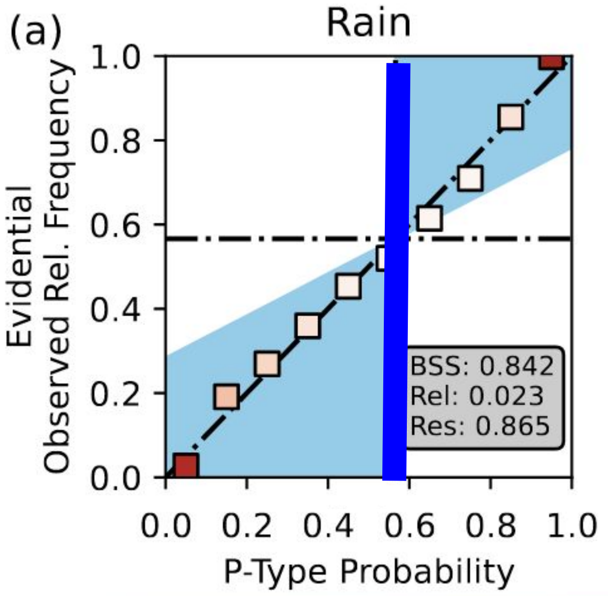
Results: Reliability Diagrams



Model
underconfident

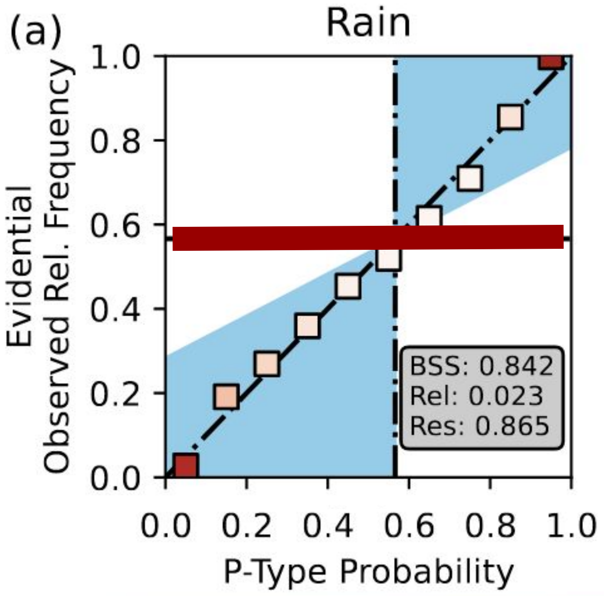
Model
overconfident

Results: Reliability Diagrams



Climatology

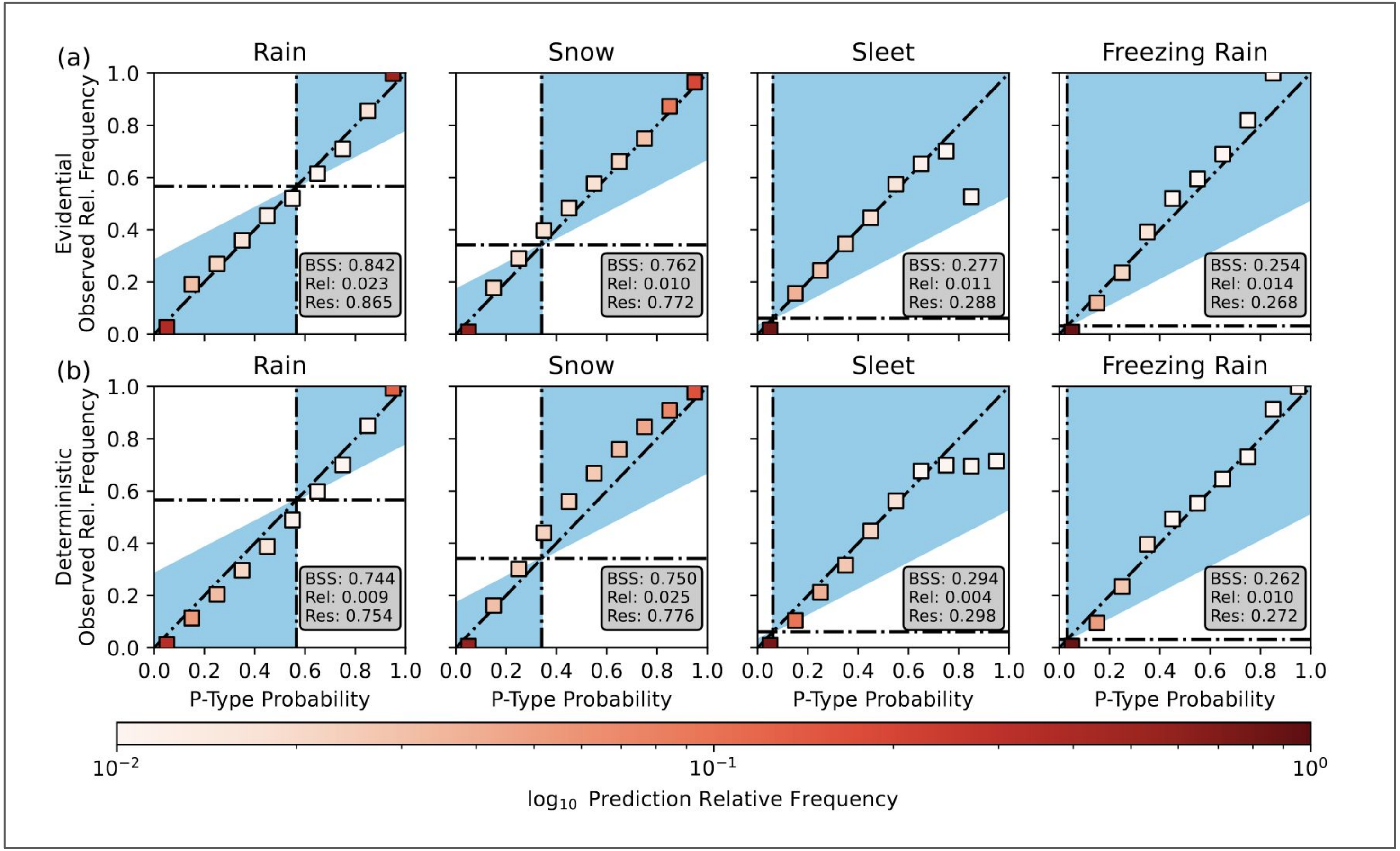
Results: Reliability Diagrams



No Resolution Line

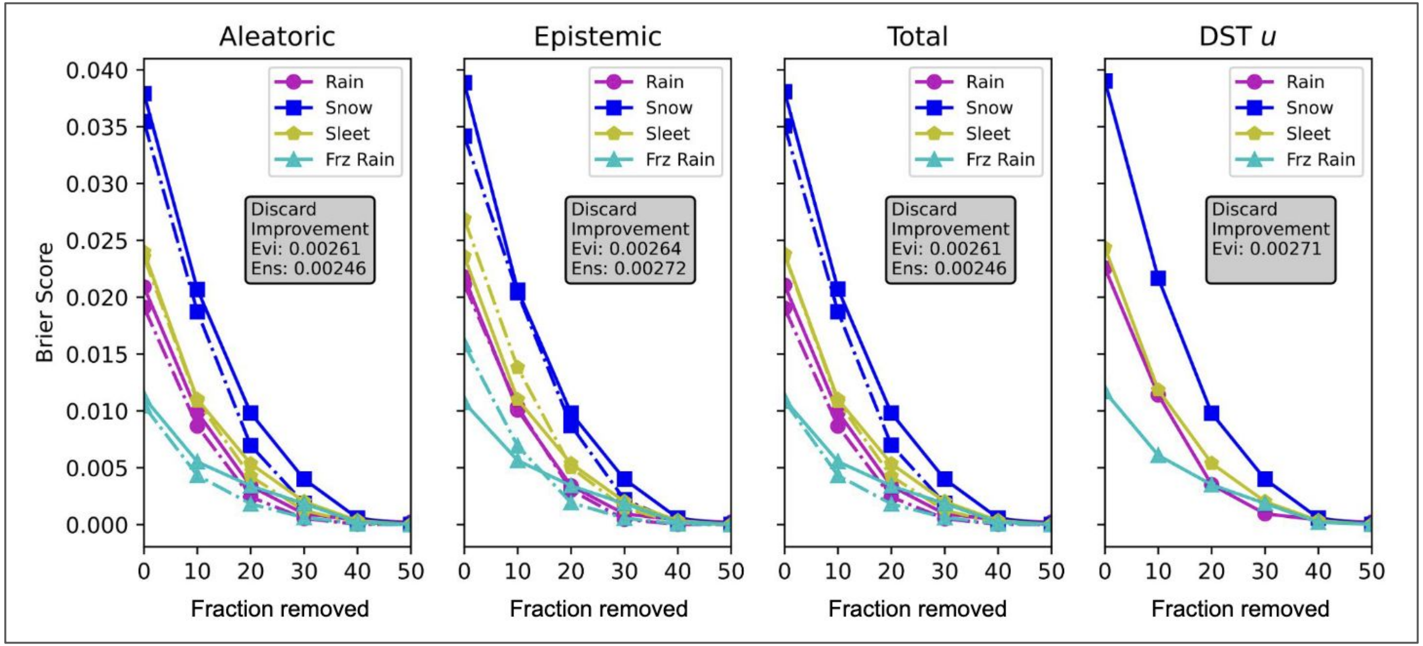
Resolution = Average difference of predicted <p-type> probabilities from climatology

Results



- <u>Rain and snow:</u> Evidential > Deterministic
- <u>Freezing rain and sleet:</u> Evidential < Deterministic

Results

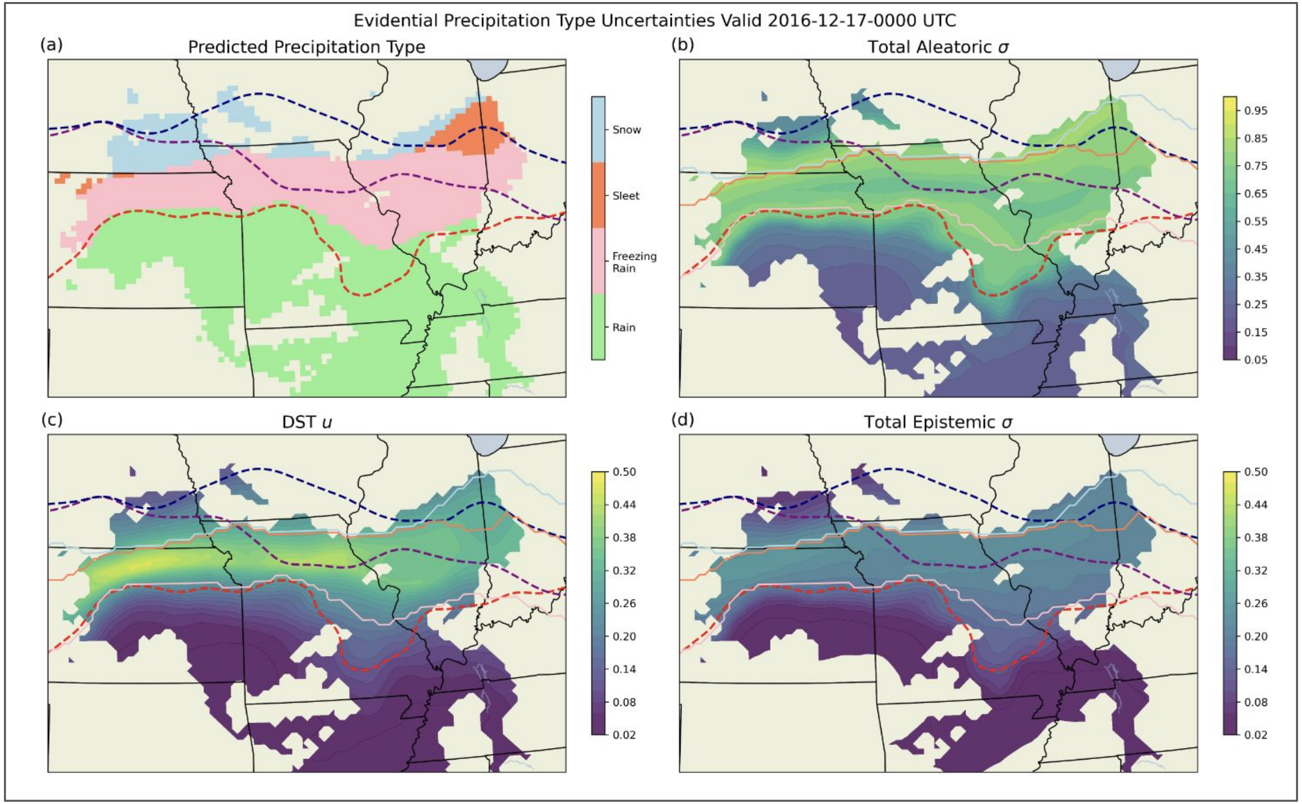


Discard test?

1) Sort test set from least certain to most certain.
2) Iteratively remove N% of least certain data and re-run trained model.

Key points: Test set performance improves as inputs become more “certain”. Similar results for ensemble vs. evidential UQ.

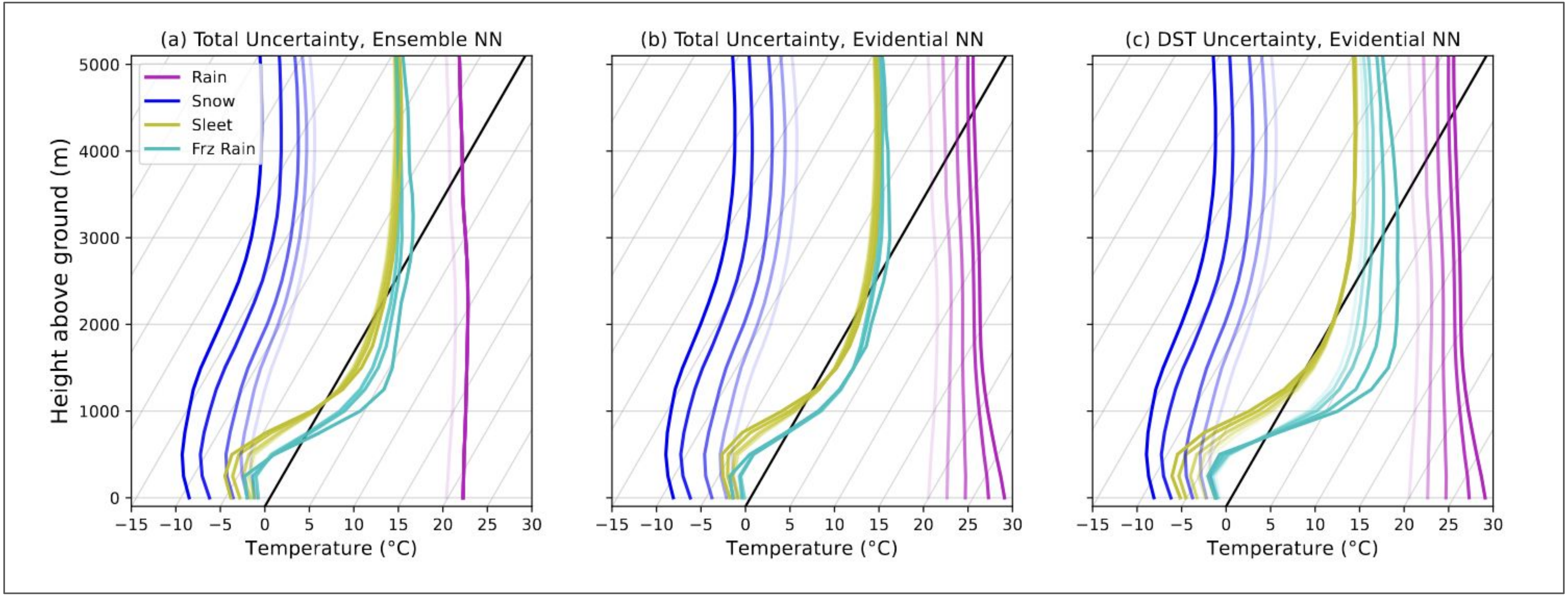
Results



Key points

1) Aleatoric uncertainty high near p-type transition boundaries (data insufficiently constrains the output)
2) Epistemic uncertainty high in freezing rain zone (insufficient # of examples)
3) Aleatoric uncertainty > Epistemic uncertainty

Results



Key point: Uncertainty trend is physically consistent.

Main Points

- Evidential NN provides UQ without computational cost ensemble-based NN UQ or physics-based UQ methods
- Evidential NN and deterministic NN have comparable p-type prediction performance (accuracy and calibration!)
- Aleatoric and epistemic uncertainties are easily computed from evidential NN, and make physical sense!

Discussion Questions

1. Should reducing one type of uncertainty be prioritized over another (if so, in what scenarios)?
2. What are some potential applications of uncertainty quantification in your research?
3. How much utility does this work have for forecasters? How “trustworthy” is this method?
4. Maybe Combine ENN with other neural networks? Like GAN-CERNN
5. EVVE, lol