Peeking Inside the Black Box:

Explainable AI Methods for a Precipitation-type Model

Belen Saavedra,

Dhamma Kimpara, David John Gagne II, John Schreck, Charlie Becker, Gabrielle Gantos

*NCAR **M**achine **I**ntegration and **L**earning of **E**arth **S**ystems*

August 2, 2023

This material is based upon work supported by the National Science Foundation under Grant No. ICER-2019758.

Why do we need to understand ML models?

XAI Pipeline

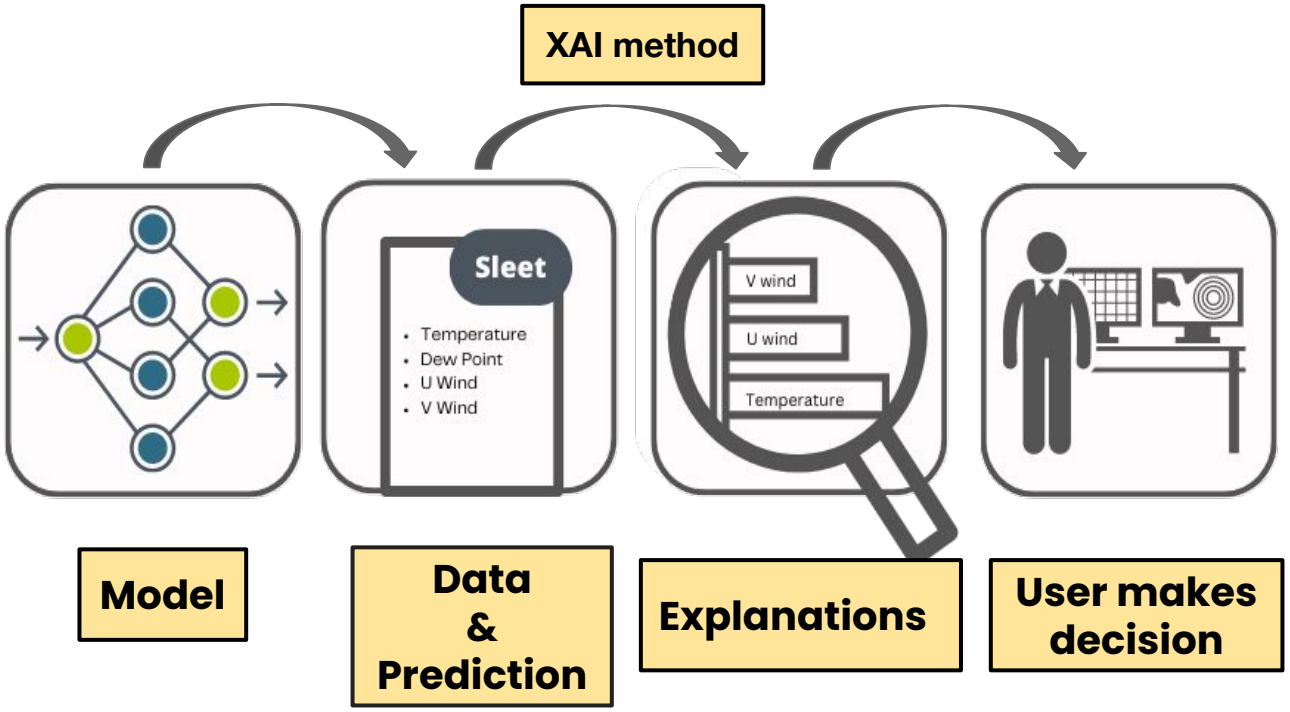
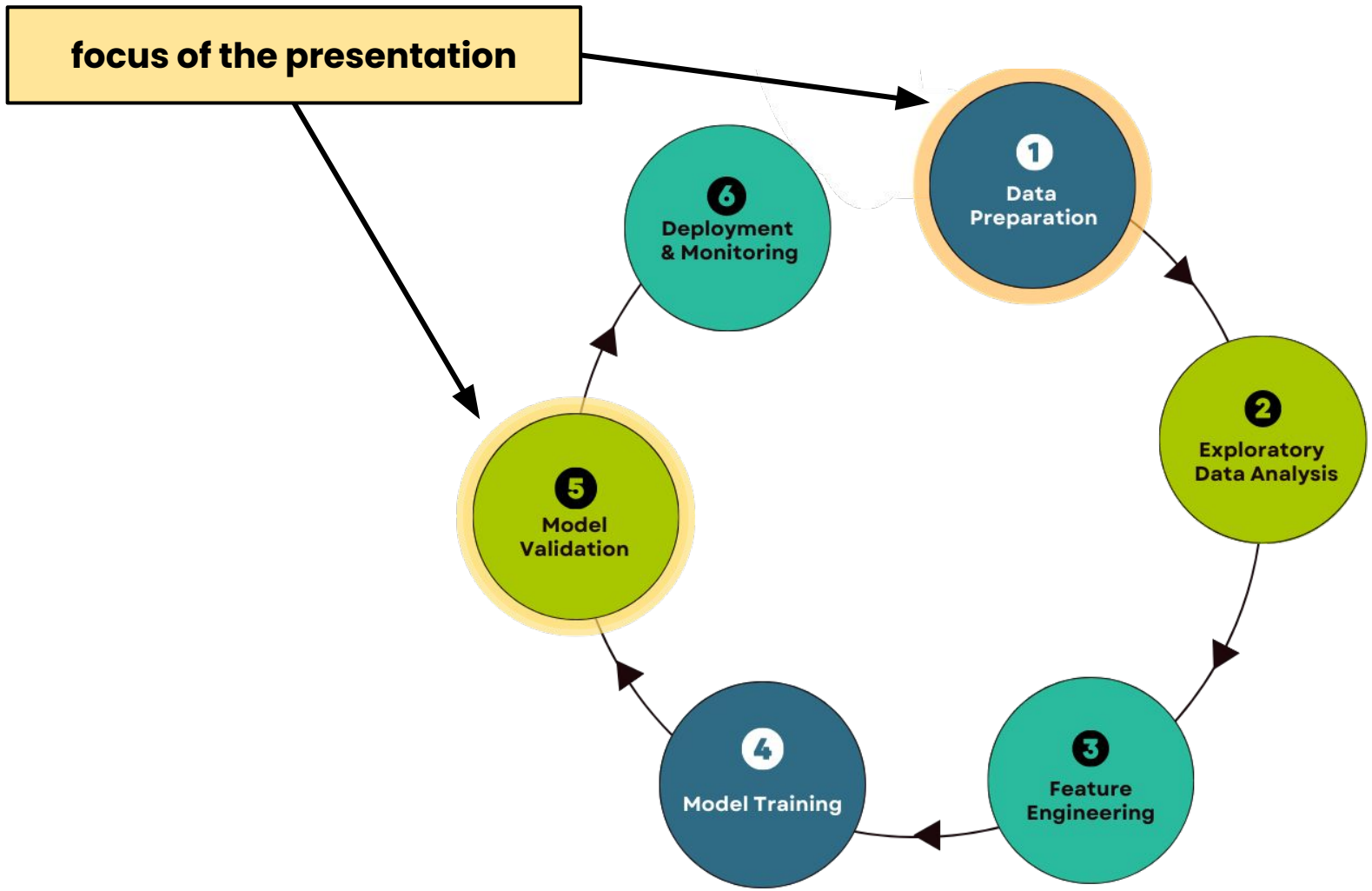


Figure 1. XAI pipeline

XAI methods aim to:

- Verify if the predictions of the ML models are consistent with the real-world
- Increase the credibility of machine learning models for both technical and non-technical users

Machine Learning Pipeline



Precipitation-type model

DATA

INPUT ⇒ Rapid Refresh

TRUE LABELS ⇒ mPING

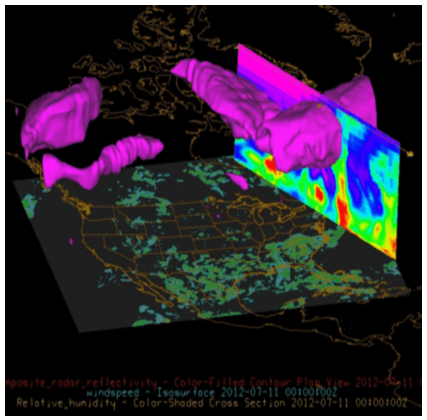
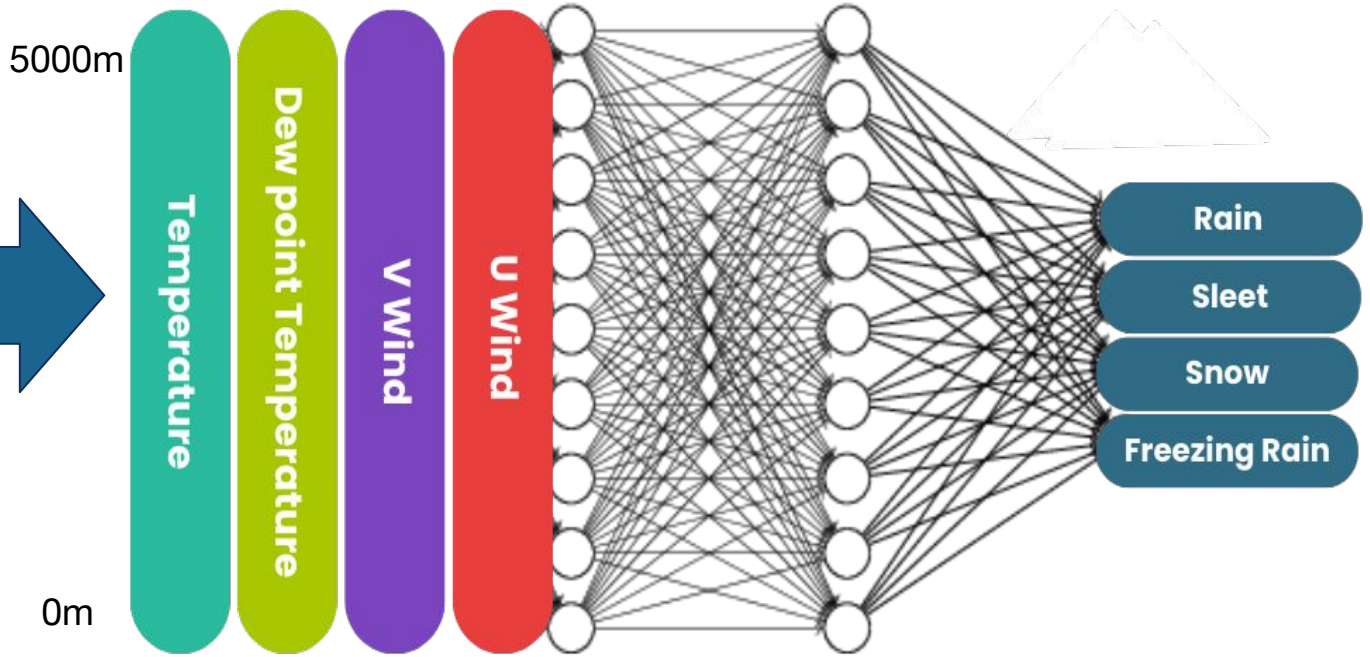


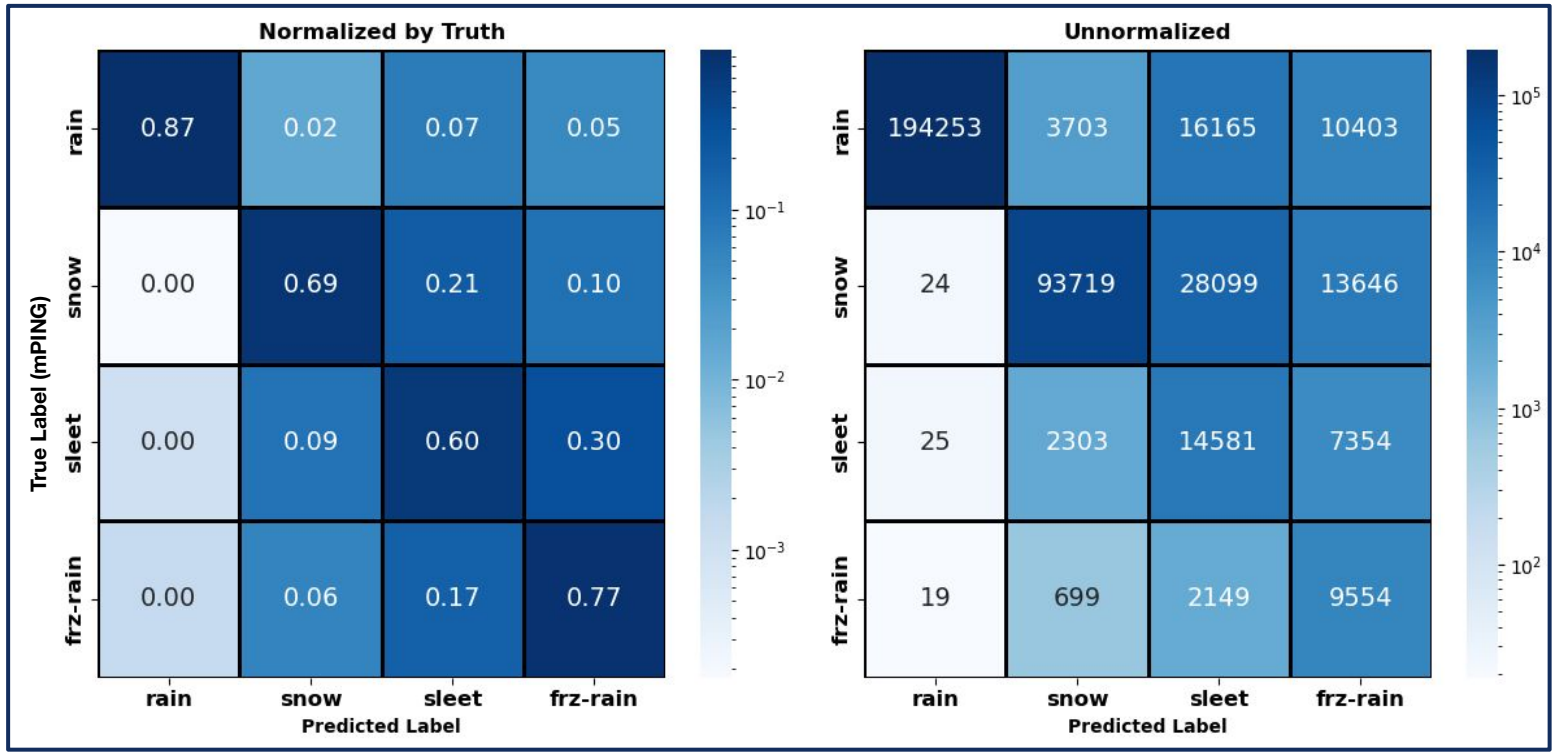
Fig 2. RAP data



Precipitation-type model

PERFORMANCE:

- mPING vs ML
- Overprediction of rain
- Under prediction of sleet and freezing rain



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?

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

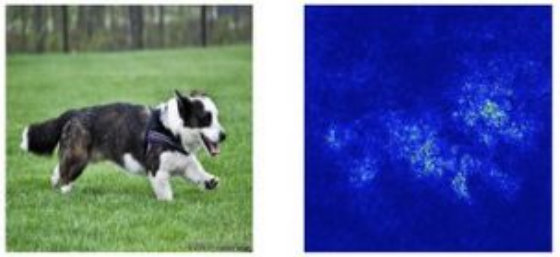
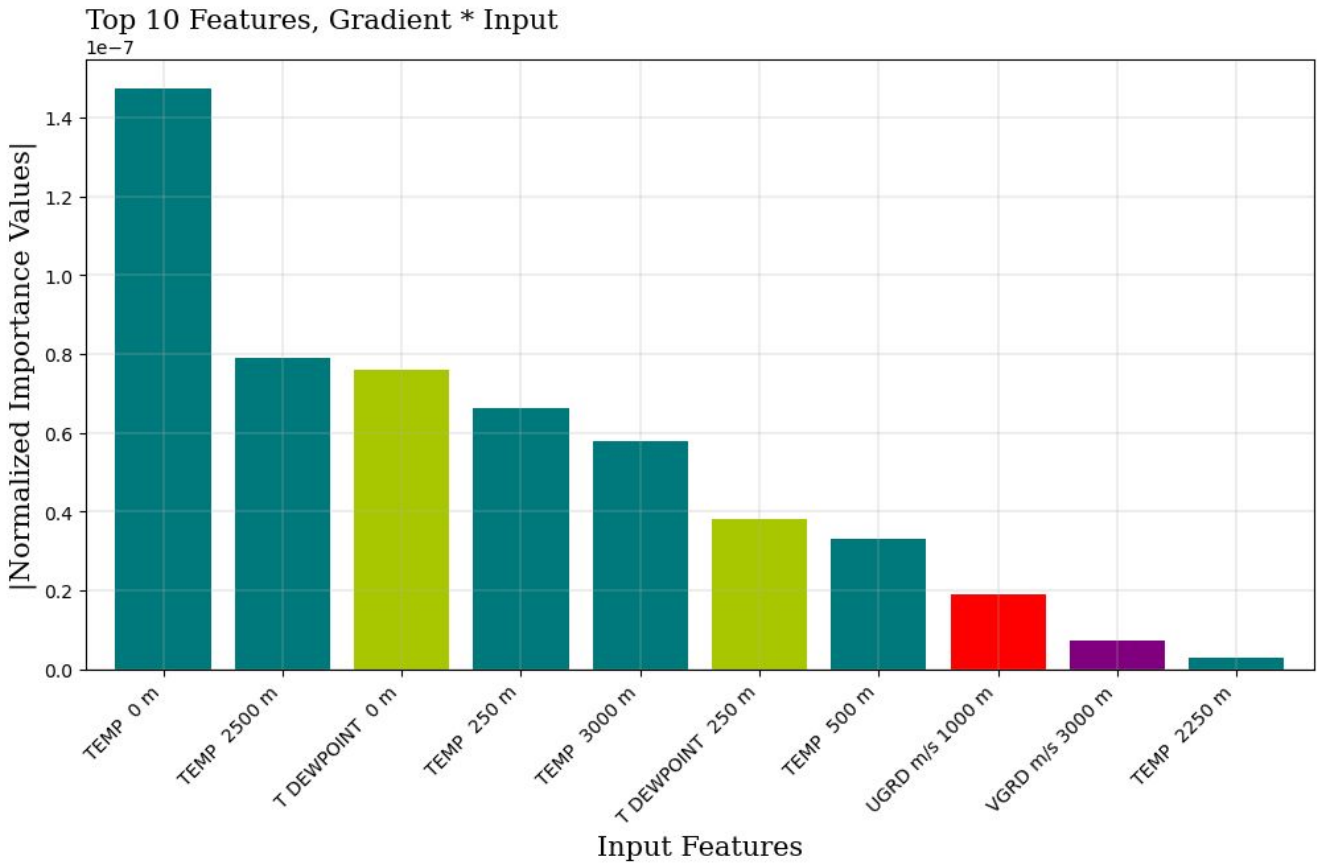


Fig. 3 Input * Gradient attribution method

Gradient * Input

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

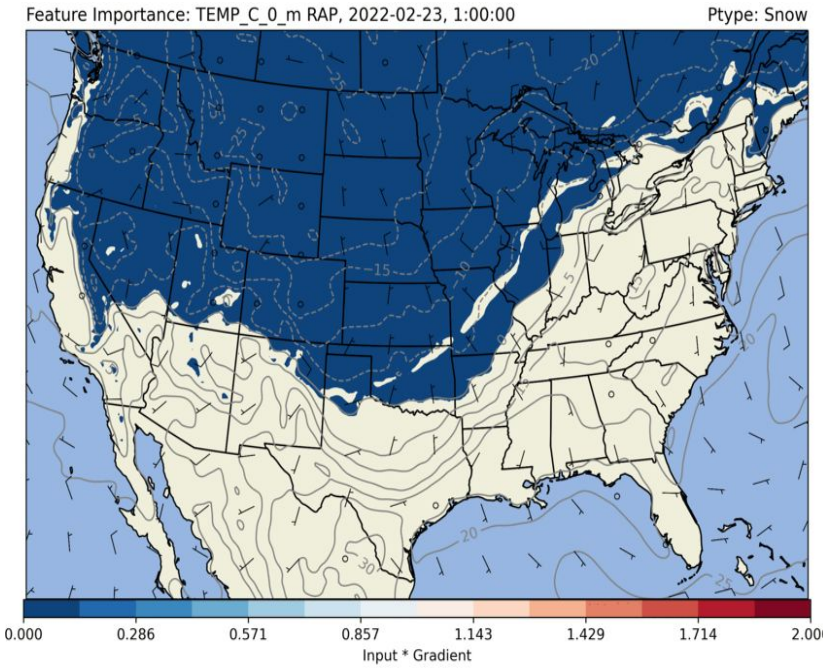


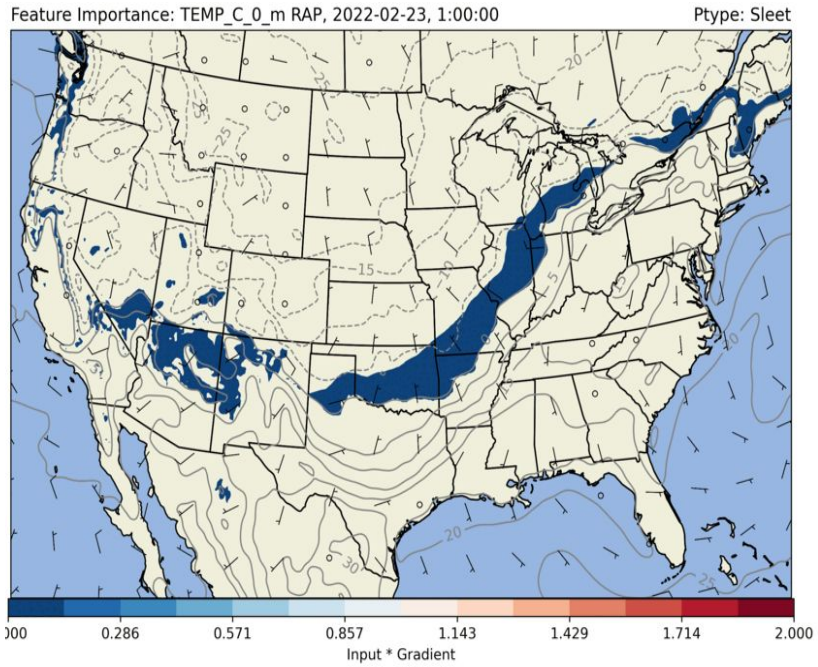
Gradient * Input works by multiplying the gradient of the model's output with the input features.

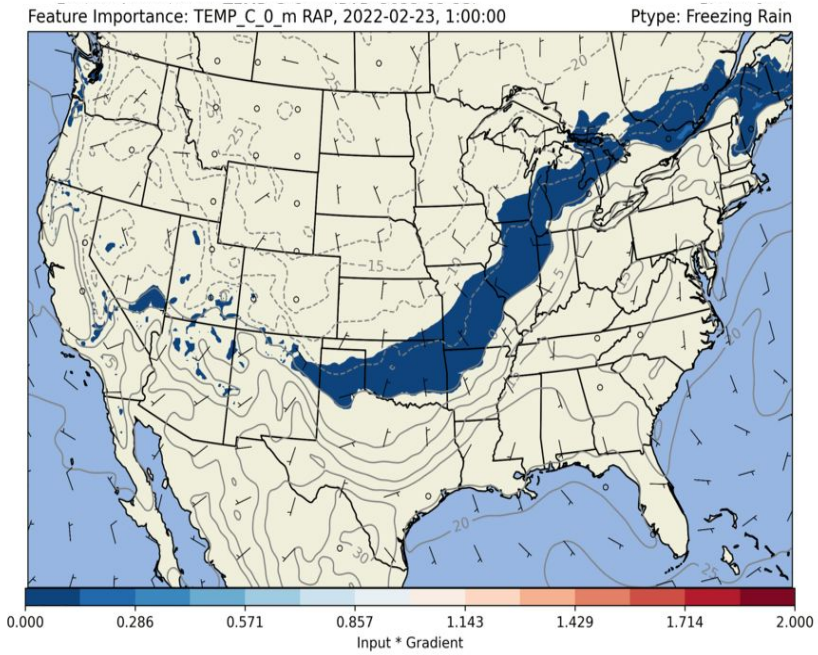
$$\mathbf{A}^{c}_{\text{Gradient}\odot\text{Input}} = \frac{\partial S_c(\mathbf{x})}{\partial \mathbf{x}} \odot \mathbf{x}.$$

Gradient * Input: CONUS plots

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

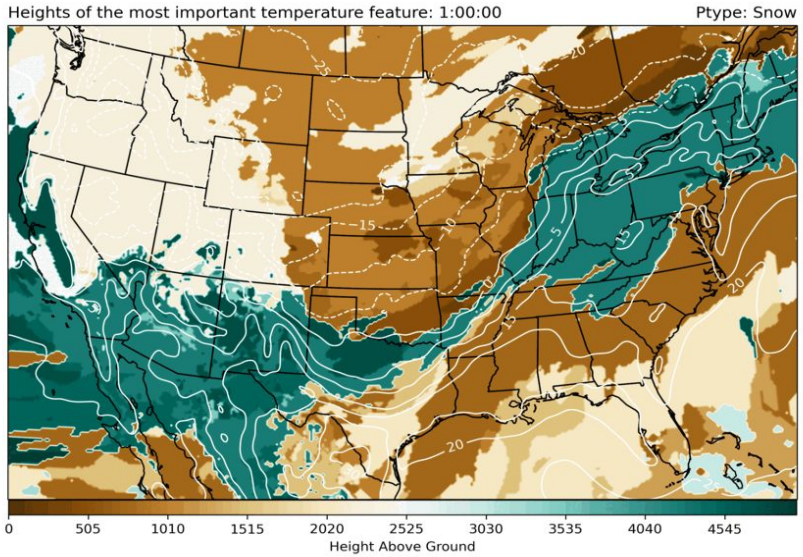


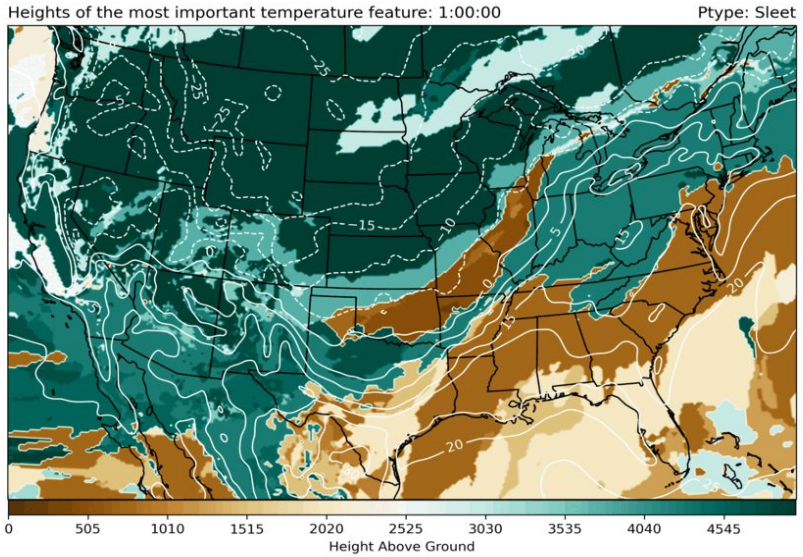


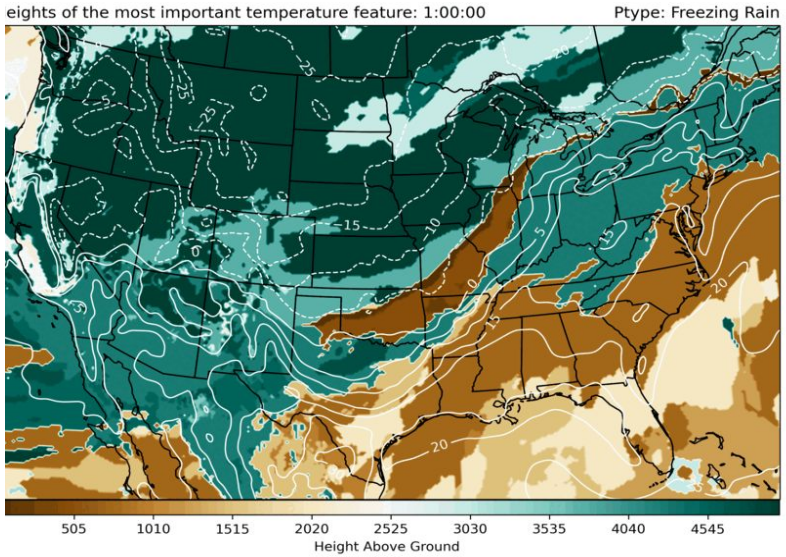


Gradient * Input: CONUS plots

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

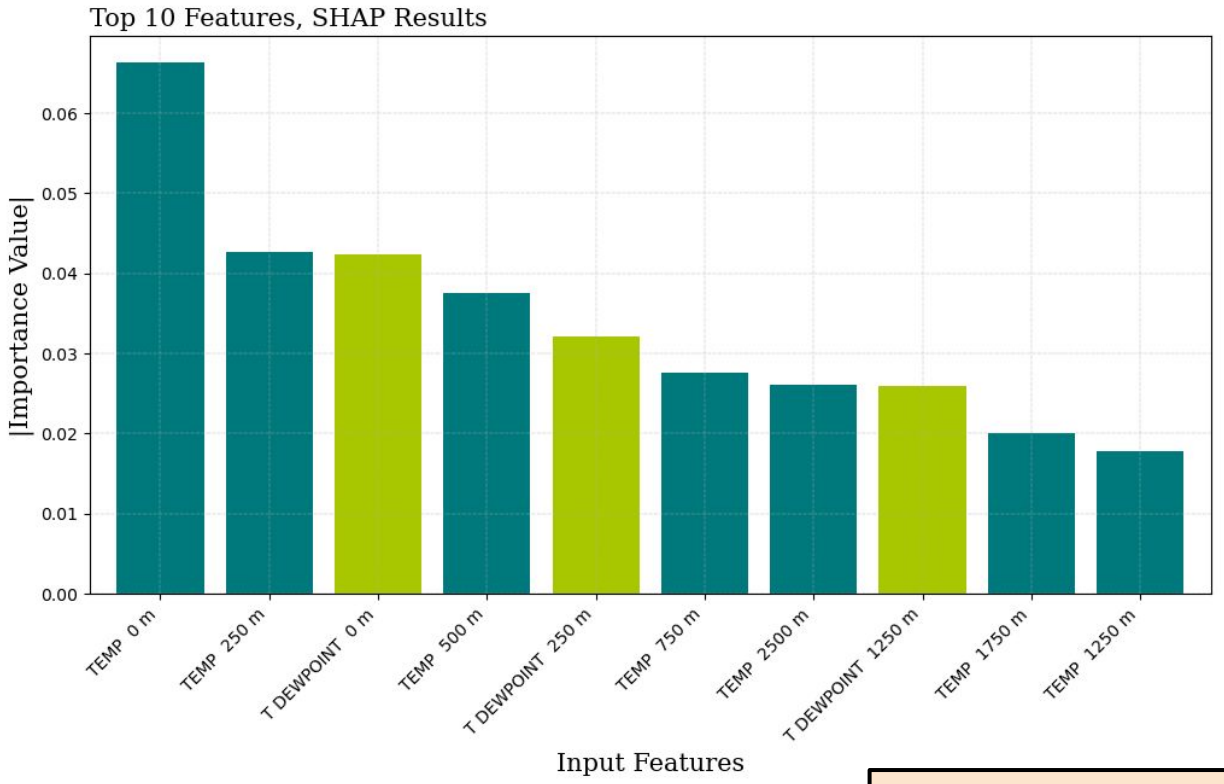






Shapley Additive Explanations (SHAP)

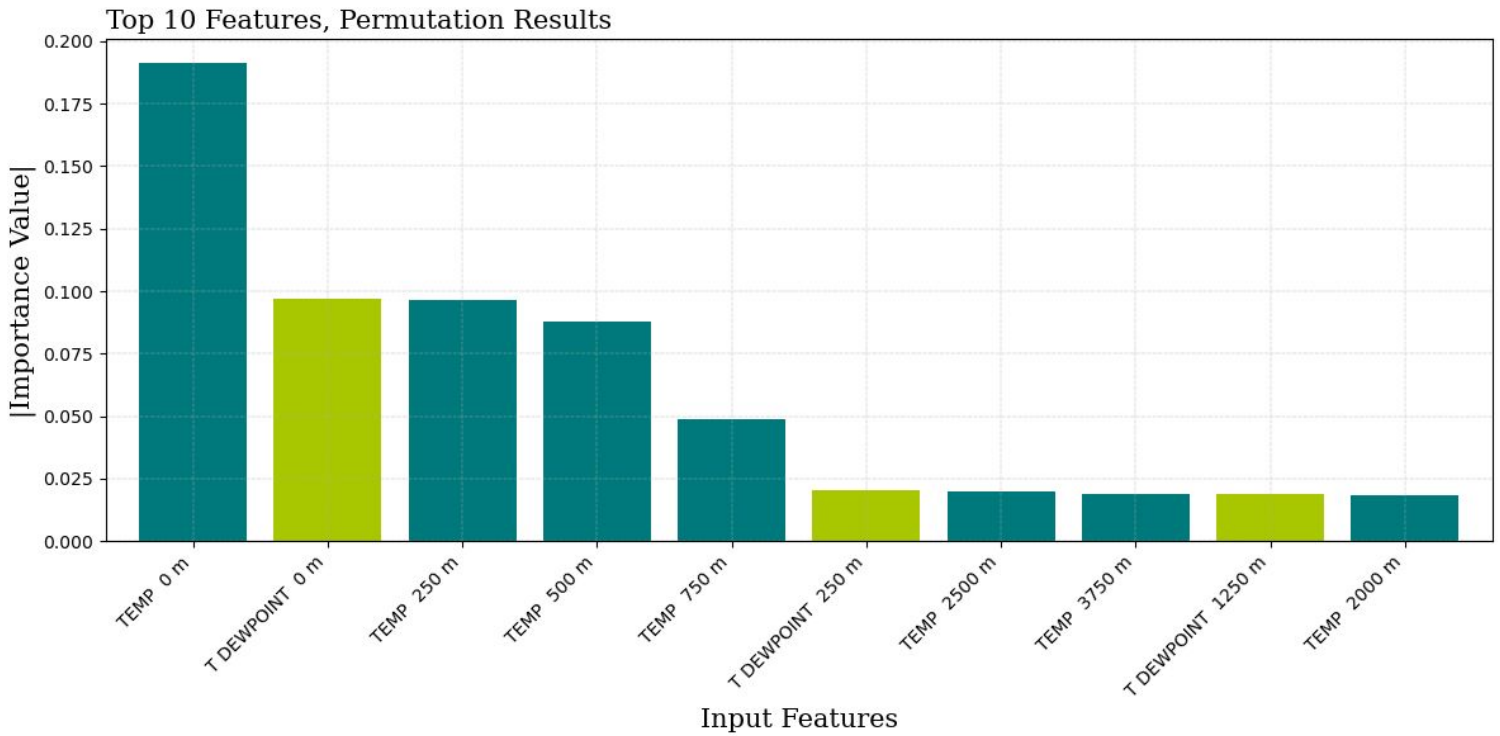
How much does **each feature contribute to the model's predictions**?



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

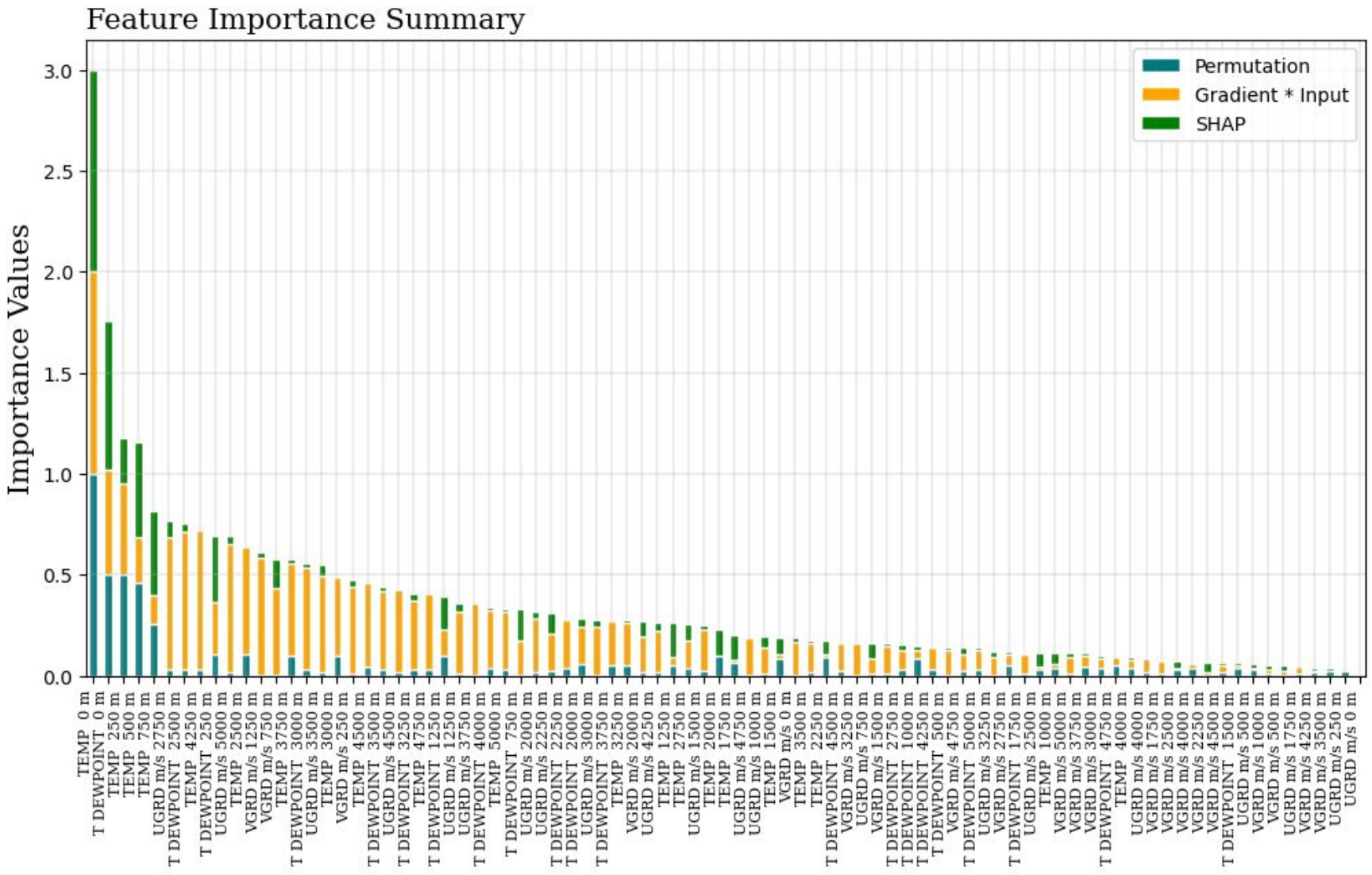
Permutation Feature Importance

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.

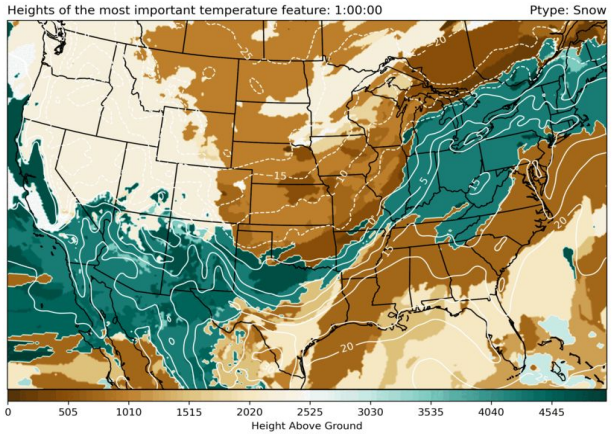
XAI Results Summary

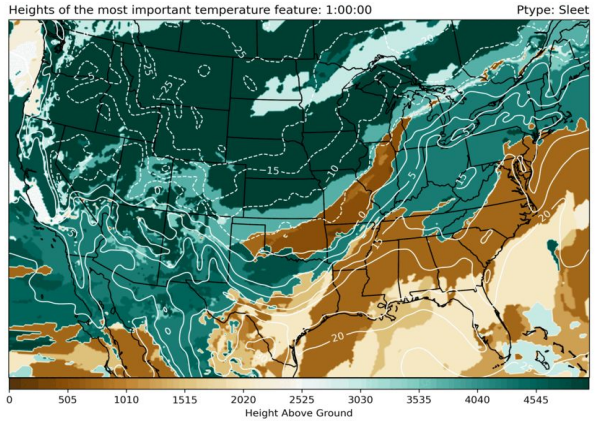


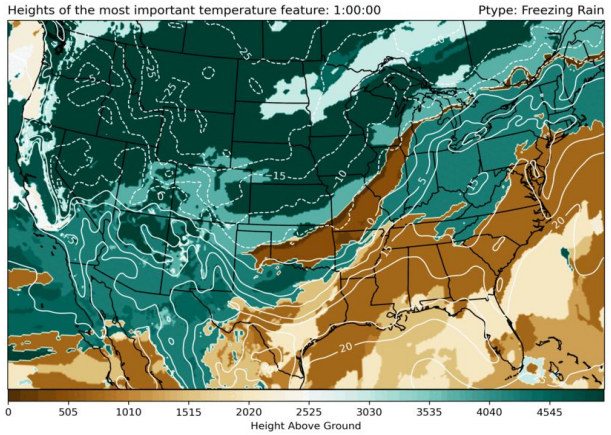
Limitations of XAI methods

- ➔ There is not a XAI method that works for every explainability task
  Some factors to consider:
  - ◆ model type
  - ◆ scope of the explanation
  - ◆ audience - who needs to understand the model?
- ➔ They can be computationally expensive
- ➔ XAI methods often rely on simplification techniques may not capture the nuances of the decision-making process
- ➔ The results of XAI might be hard to interpret

Main Takeaways and Future Steps







XAI Methods:

- ➔ Temperature at 0m is the top feature for across the three XAI methods
- ➔ The Input features that are near the surface tend to be the most important
- ➔ Each XAI method provides slightly different results.

Broader implications:

- ➔ Support the communication of the predictions of this model to a wide audience (decision makers, forecasters and general users)

References

[1] McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., & Smith, T. (2019). Making the black box more transparent: Understanding the physical implications of machine learning.

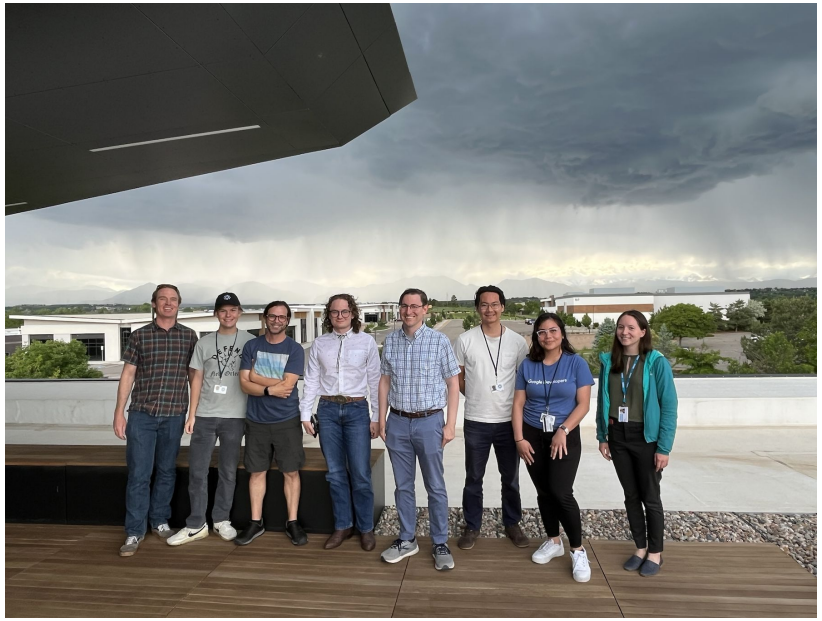
Figures:

Fig 1. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. arXiv preprint arXiv:1602.04938, 2016.

Fig 2. RAP, NOAA, Rapid Refresh/Rapid Update Cycle, https://www.ncei.noaa.gov/products/weather-climate-models/rapid-refresh-update#:~:text=The%20National%20Centers%20for%20Environmental,for%20smaller%20regions%20of%20interest.

Fig 3. https://i.stack.imgur.com/Nxhrr.jpg

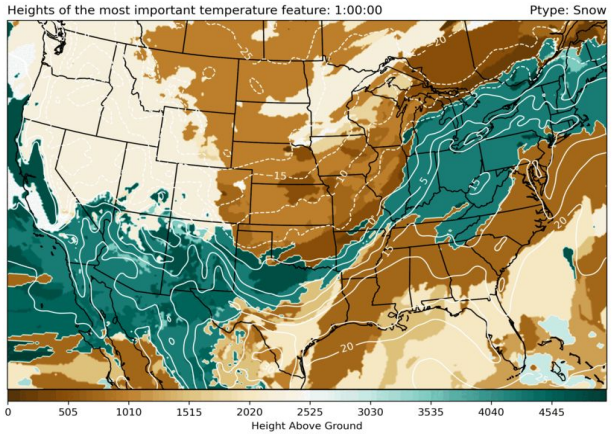
Acknowledgements

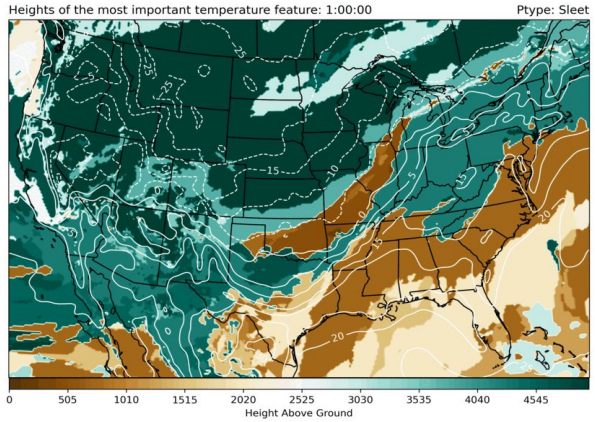


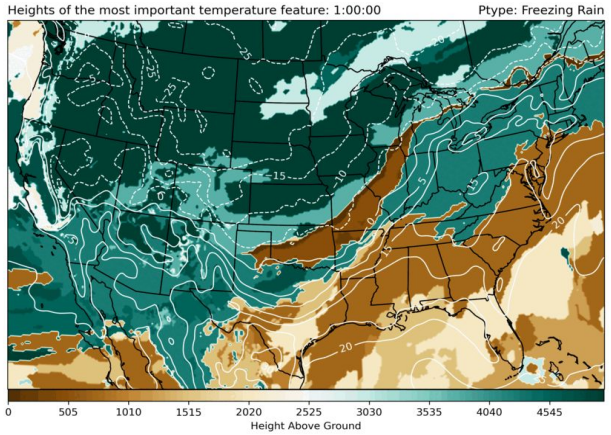
Part of the MILES group and Vaisala interns

- **MILES group:**
  David John Gagne II, John Schreck, Charlie Becker, Gabrielle Gantos, Dhamma Kimpara, Hayden Outlaw
- **SIParCS:**
  Virginia Do, Julius Owusu Afriyie, and the intern cohort
- **NESSI**:
  Jerry Cyccone, Ben Fellman, and the NESSI 2023 cohort
- **Funding:**
  National Science Foundation under Grant No. ICER-2019758.

Main Takeaways and Future Steps







XAI Methods:

- ➔ Temperature at 0m is the top feature for each of the methods
- ➔ The Input features that are below 1000 meters above ground tend to be the most important
- ➔ Each XAI method provides slightly different results.

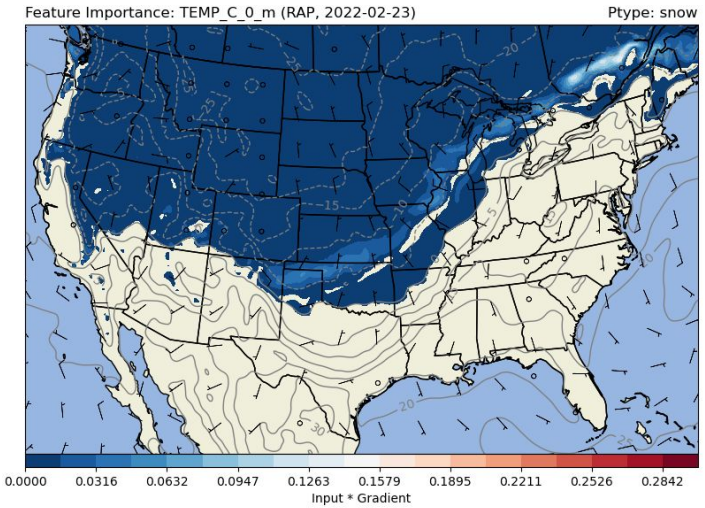
Broader implications:

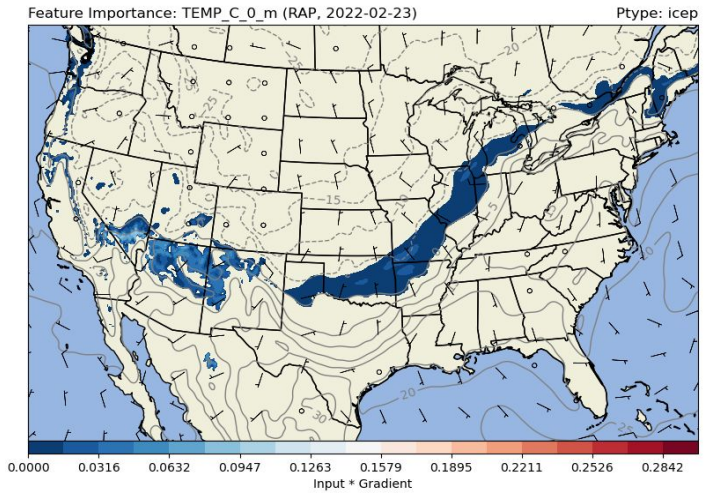
- ➔ Support the communication of the predictions of this model to a wide audience (decision makers, forecasters and general users)

Questions/Feedback?

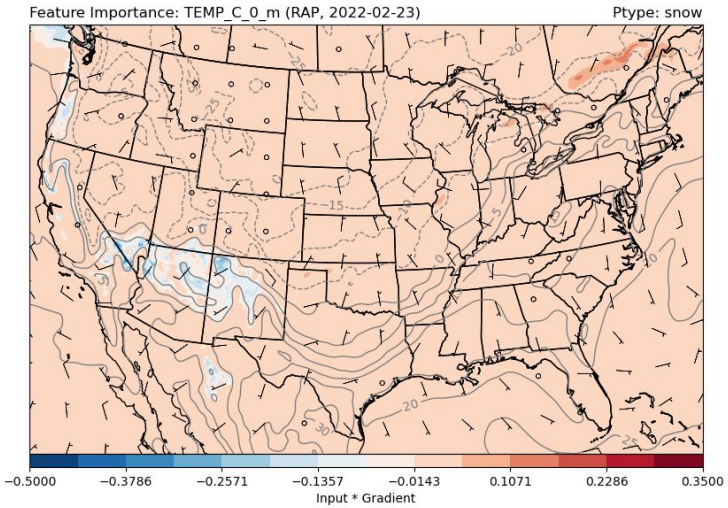
Appendix: Input * Gradient CONUS plots

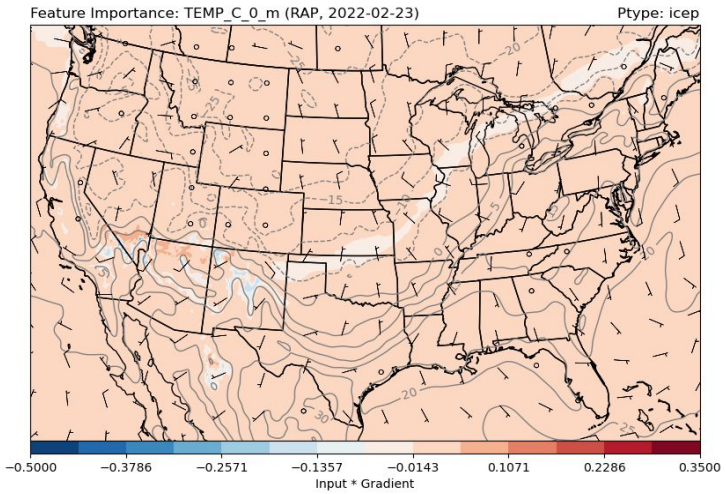
Absolute Values



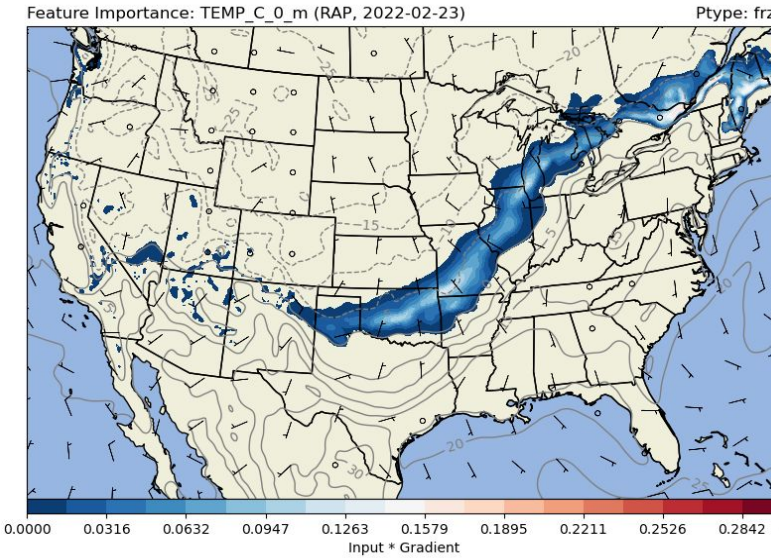


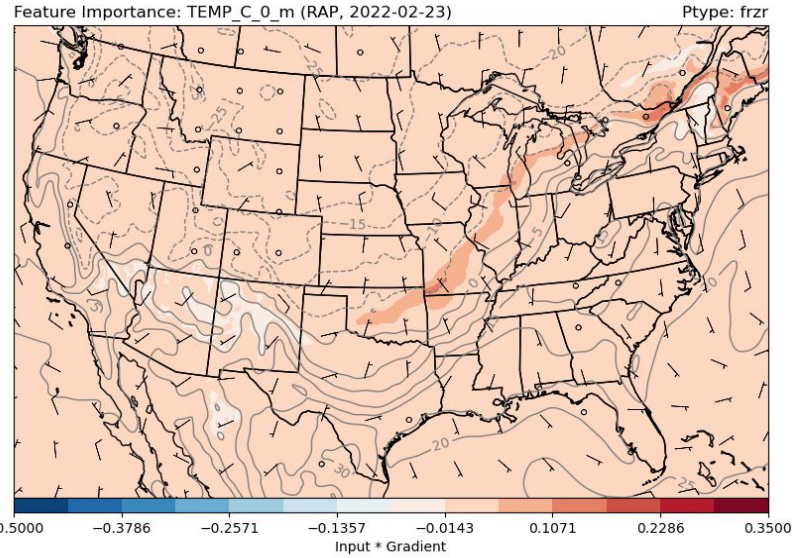
Raw Values

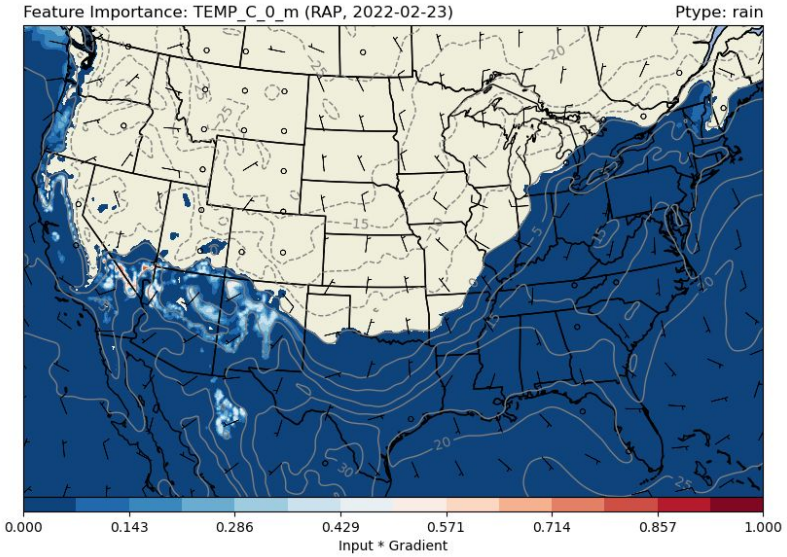


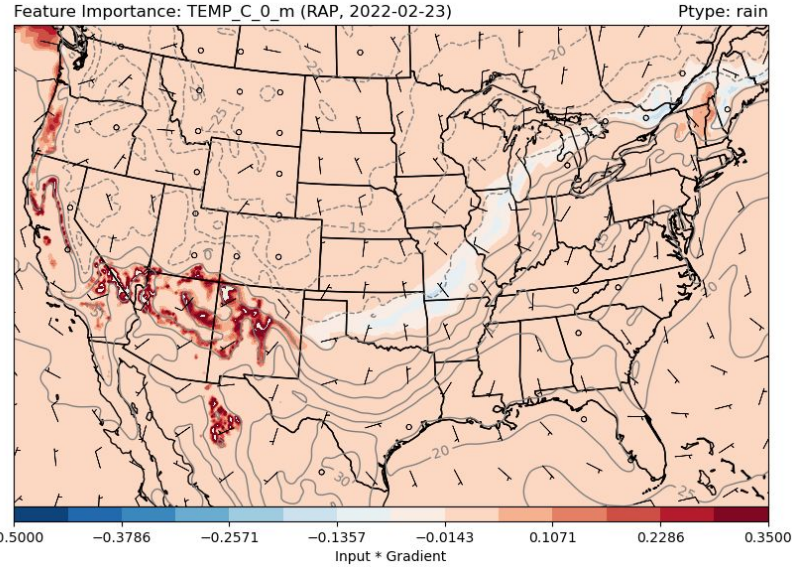


Feature Importance by precipitation type: Temperature

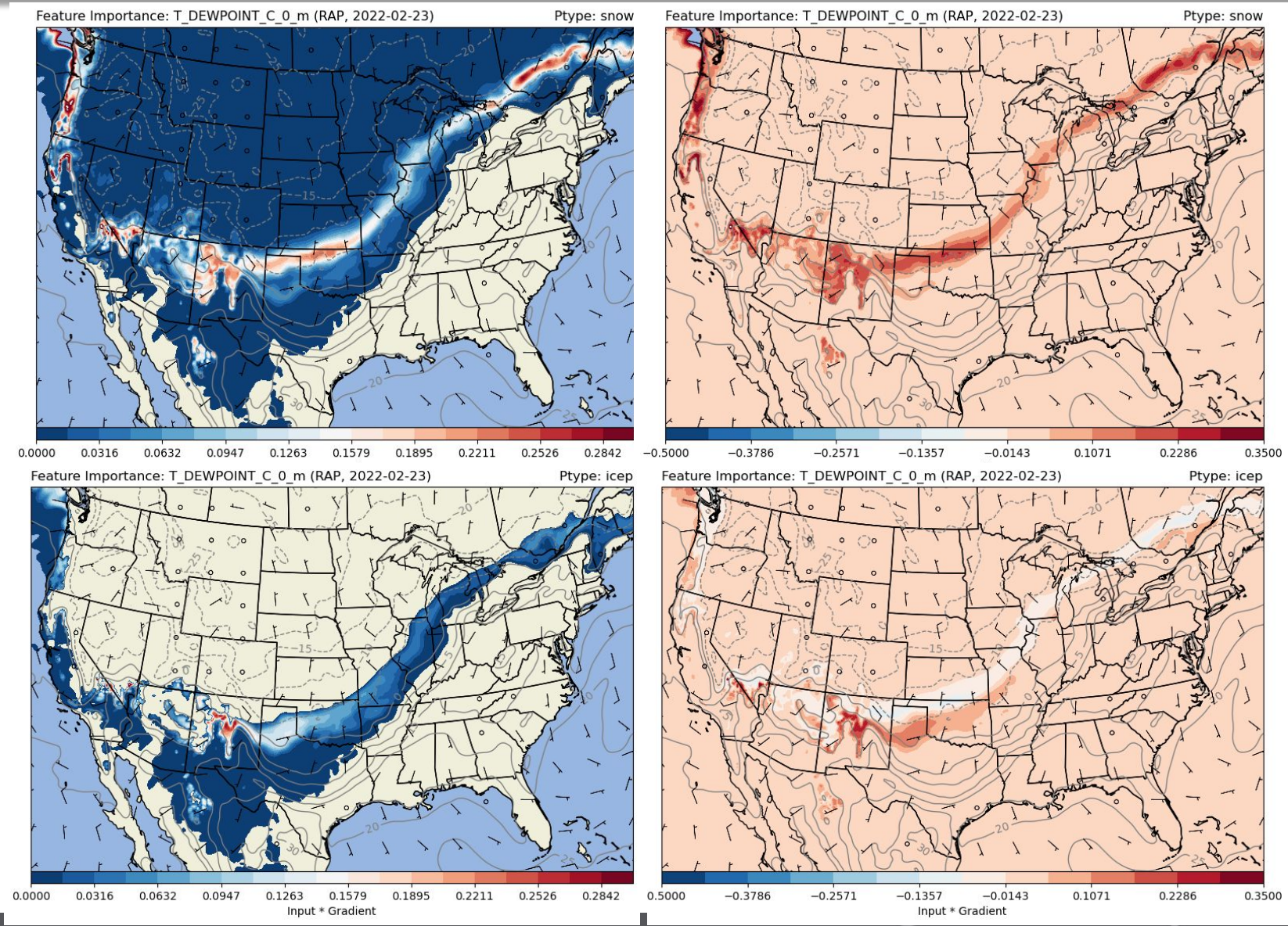




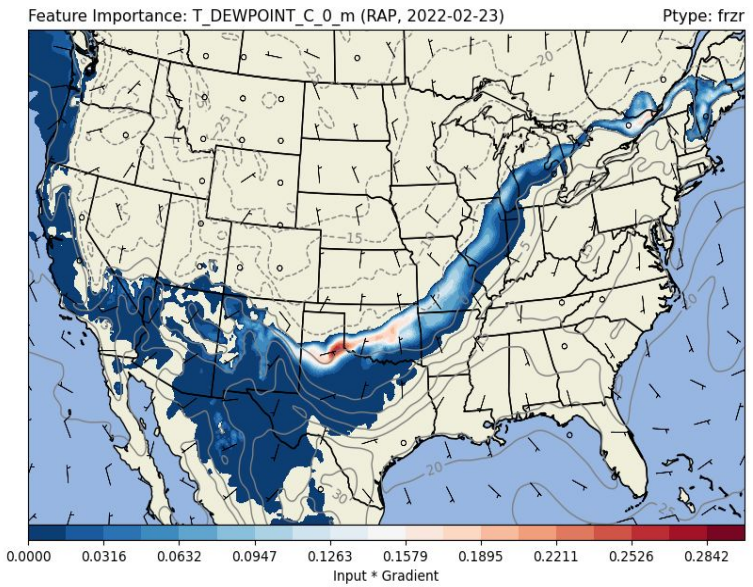


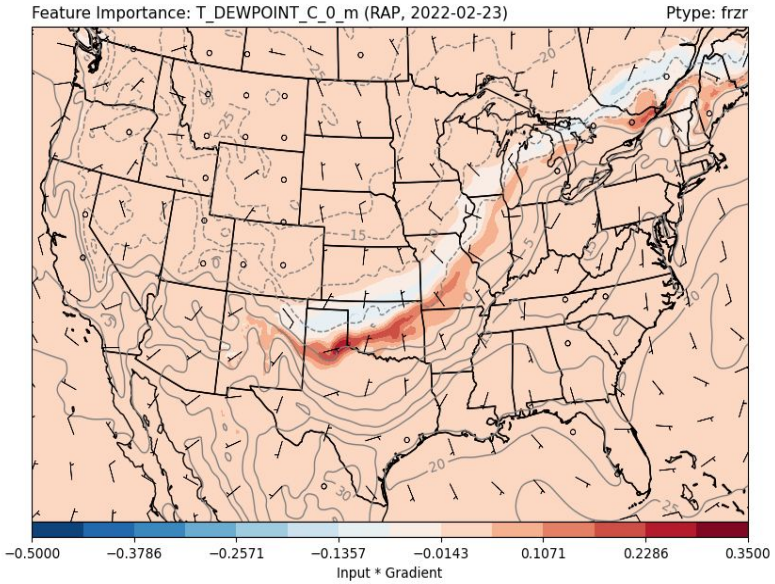


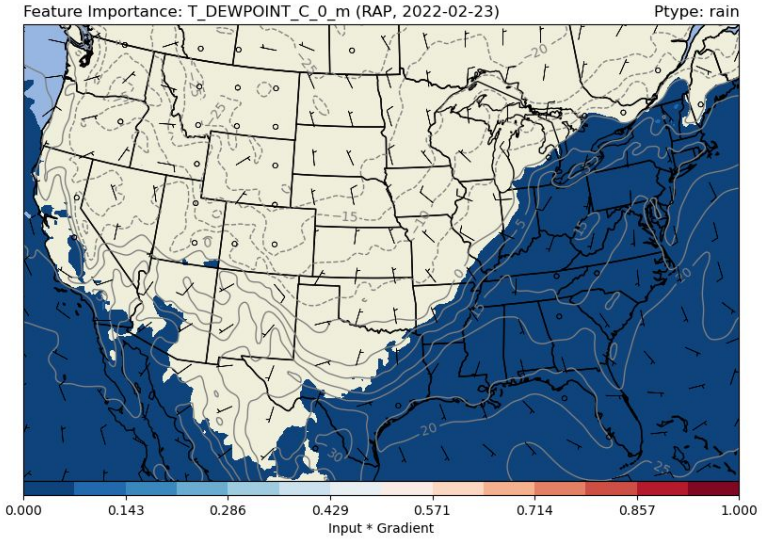
Feature Importance by precipitation type: Dew Point

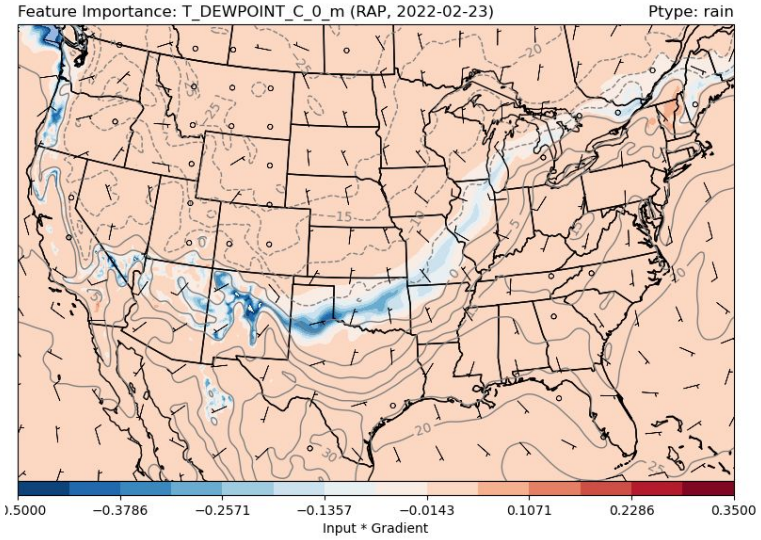


Feature Importance by precipitation type: Dew Point

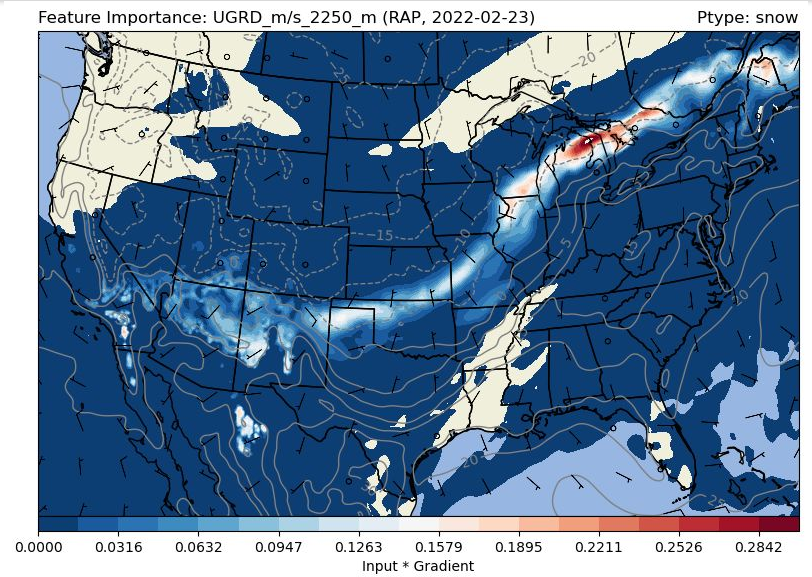


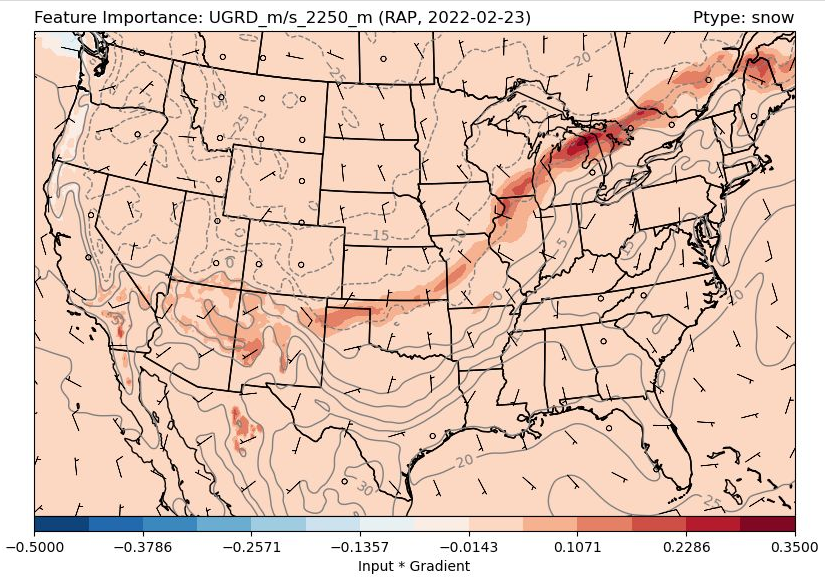


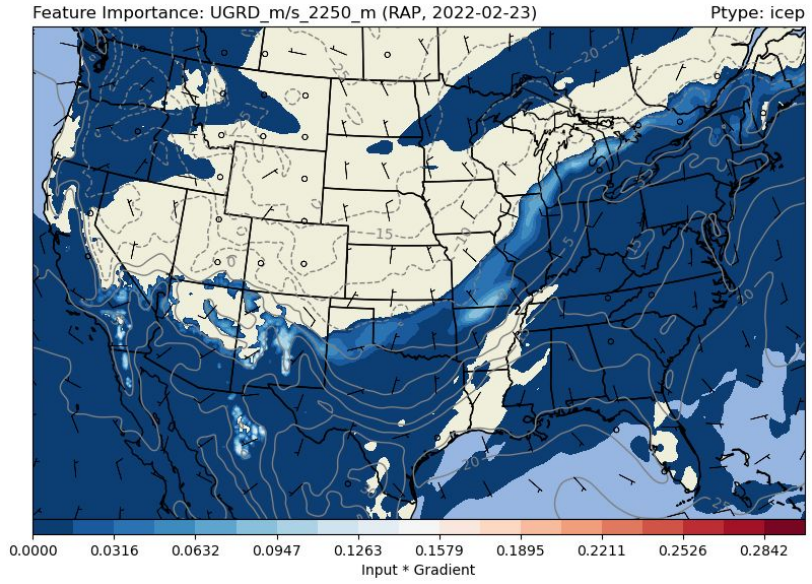


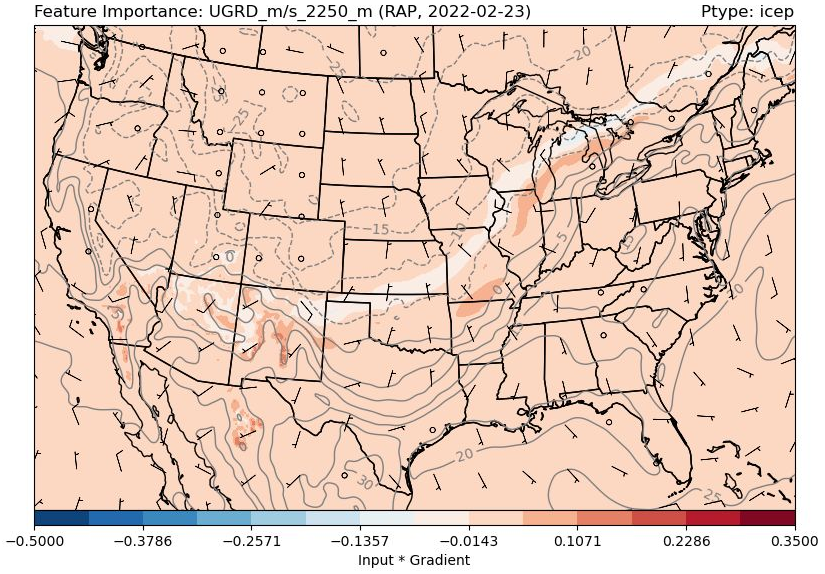


Feature Importance by precipitation type: UGRD

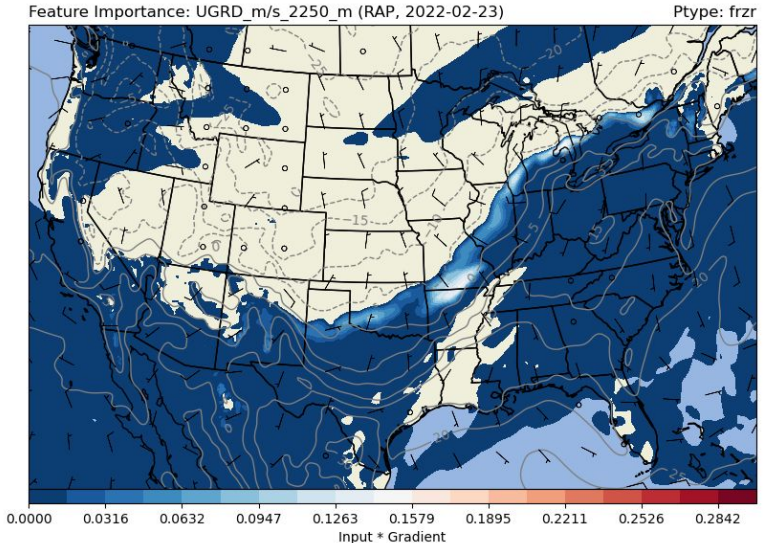


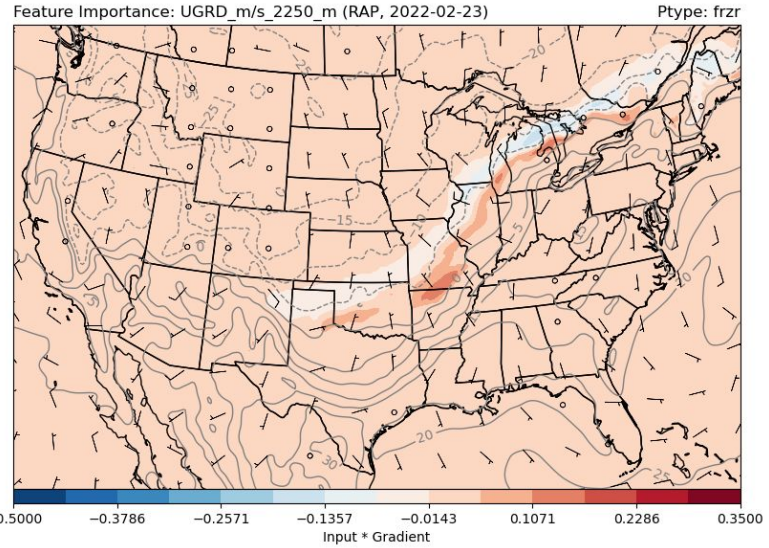


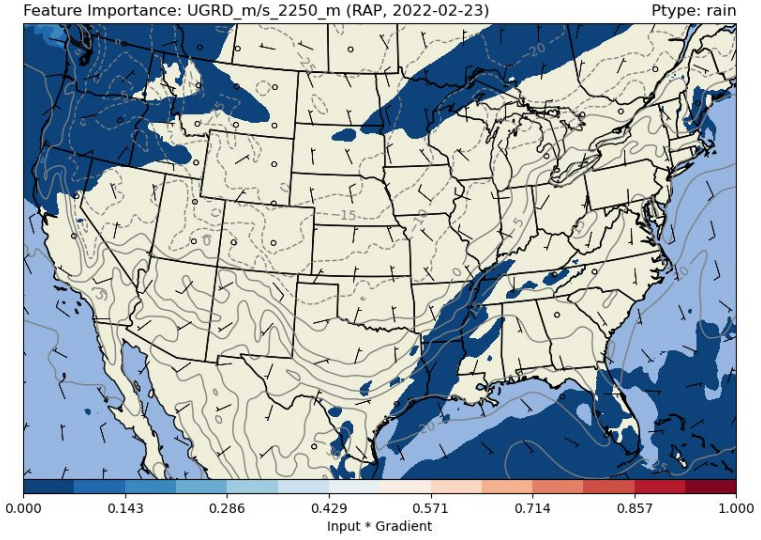


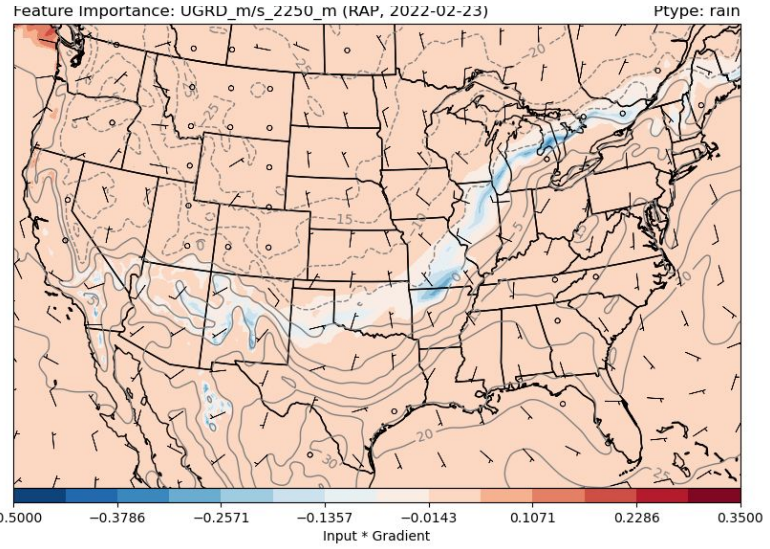


Feature Importance by precipitation type: UGRD

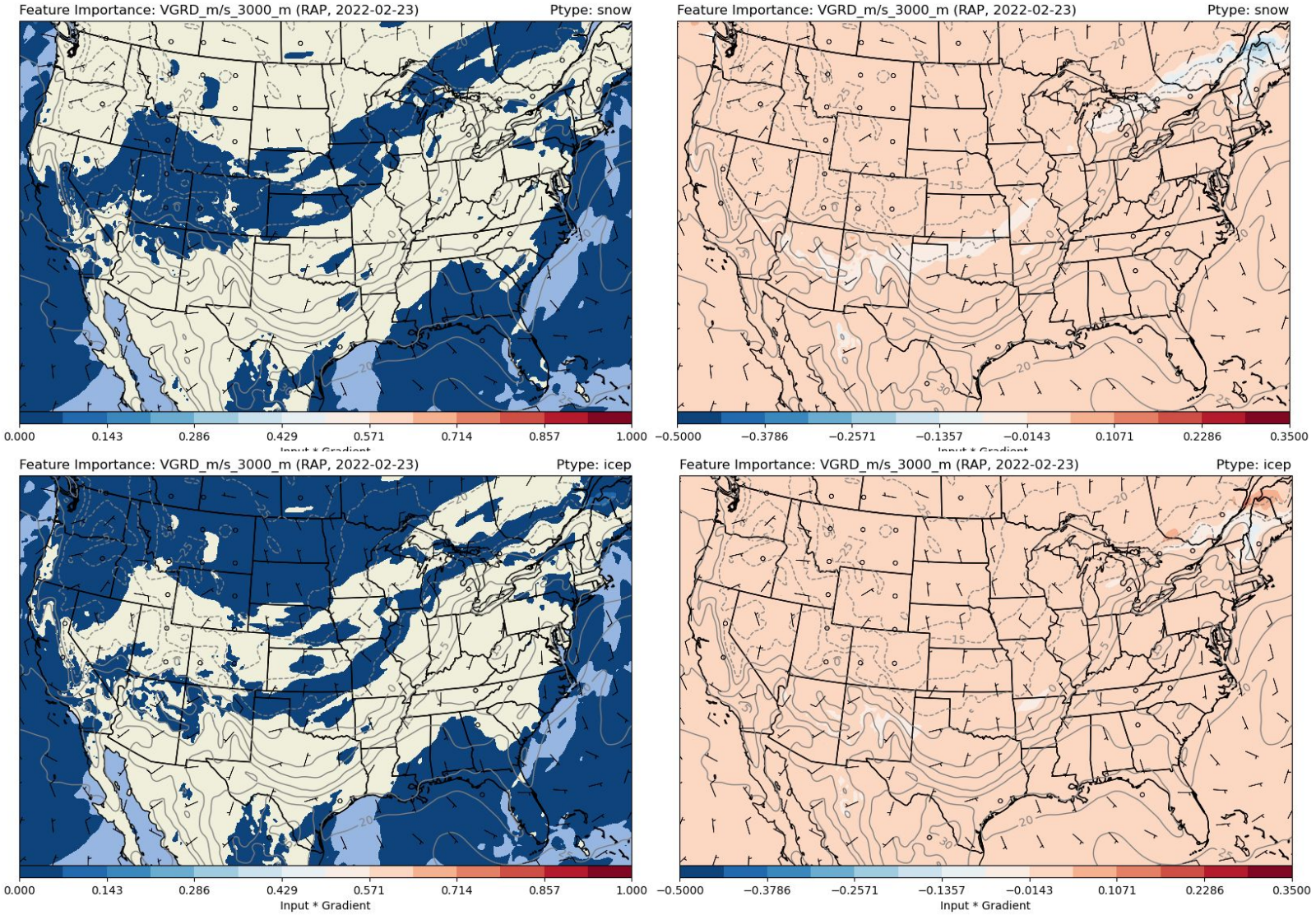








Feature Importance by precipitation type: VGRD



Feature Importance by precipitation type: VGRD