Towards Interpretable Artificial Intelligence in the Atmospheric Sciences

or: how do we know what our models are doing?

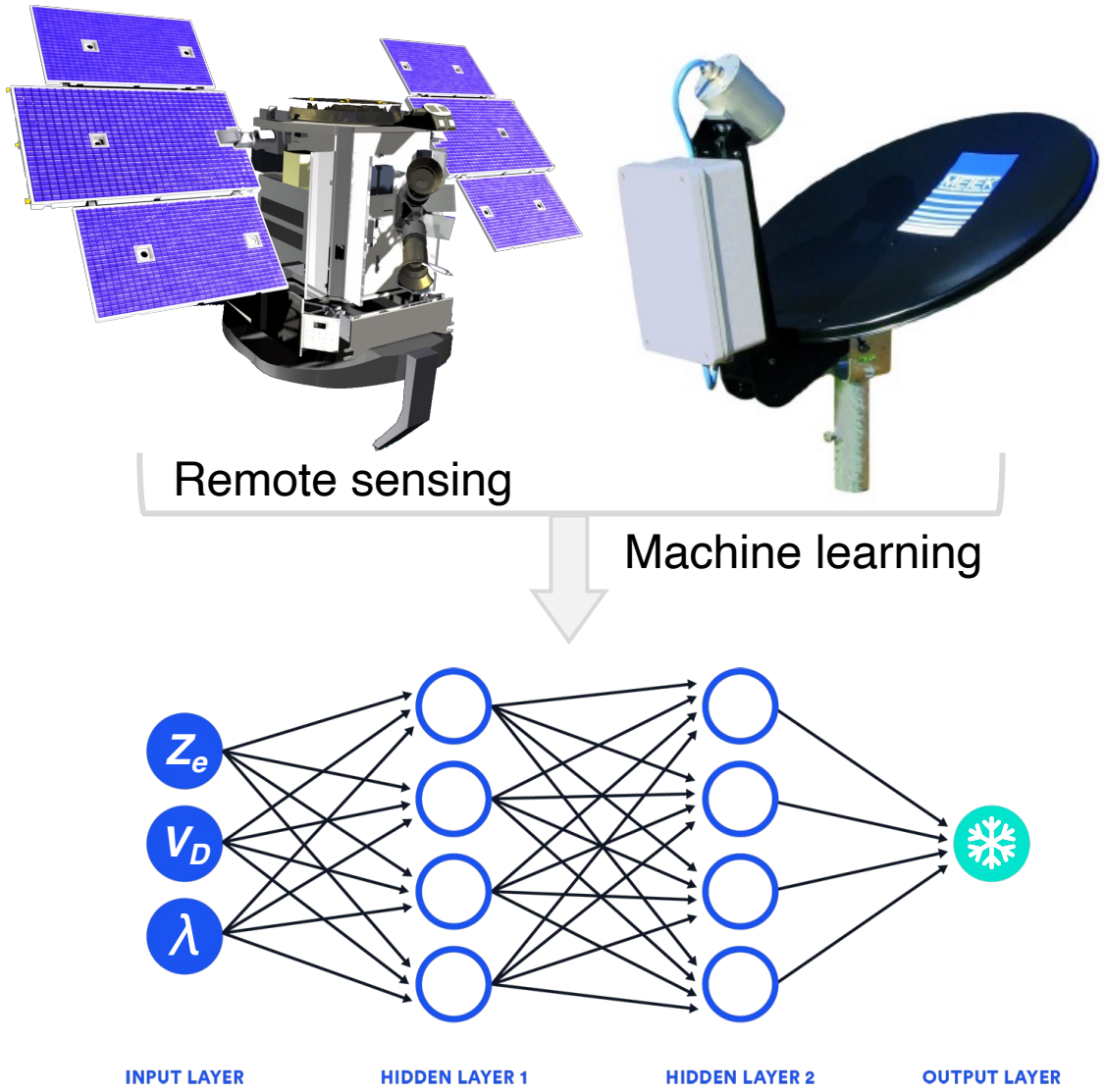
Fraser King[1], Claire Pettersen[1],
Christopher G. Fletcher[2], Derek Posselt[3]

[1]Climate and Space Sciences and Engineering, University of Michigan
[2]Dept. of Geography & Environmental Management, University of Waterloo
[3]Jet Propulsion Laboratory, California Institute of Technology

NASA NIP Grant # 80NSSC22K0789

Overview

1. ML in the Atmospheric Sciences
2. Random Forests for snowfall retrievals
3. Neural Networks for precipitation retrievals
4. Generative techniques for resolving radar blind zones
5. Towards model interpretability



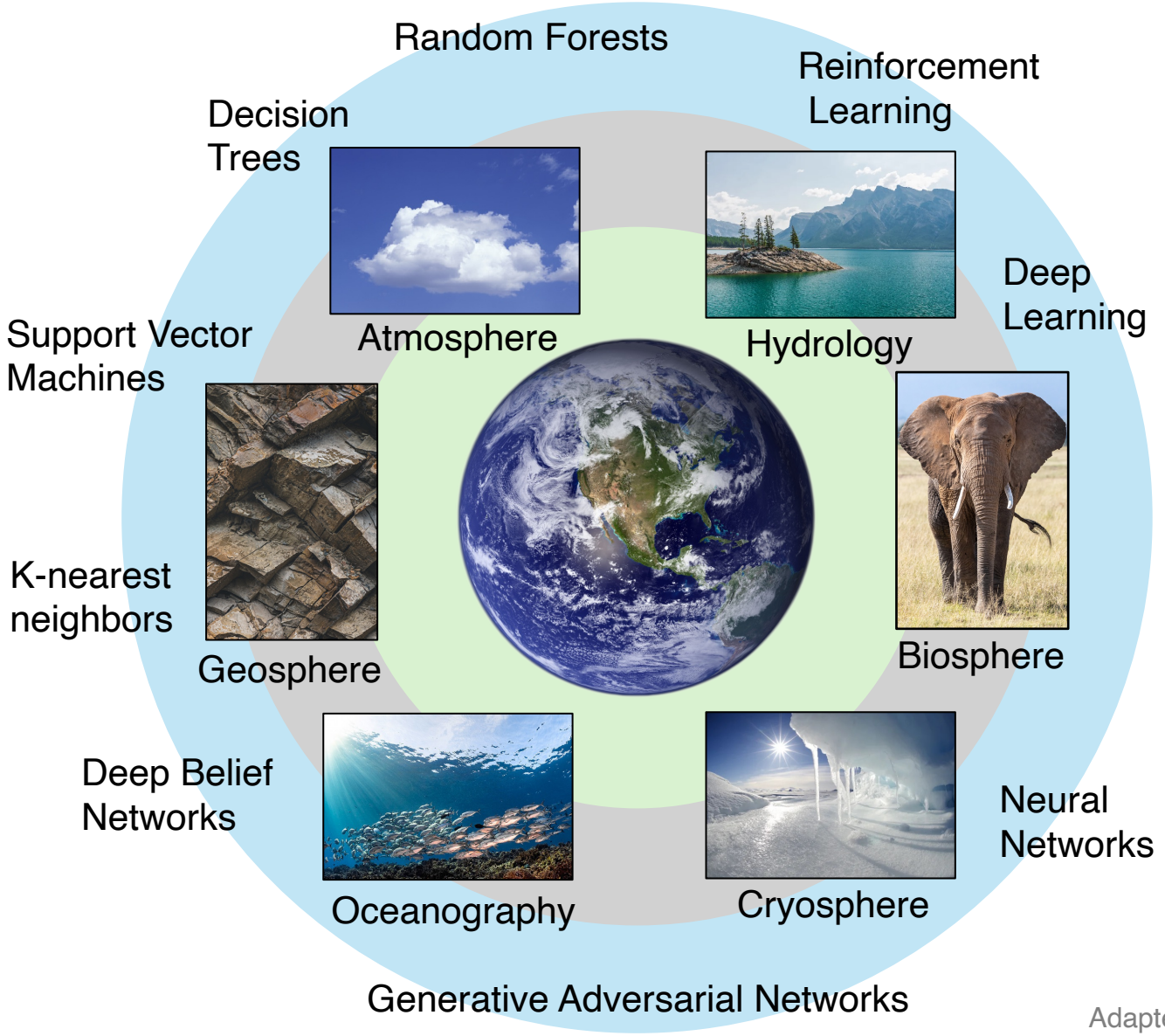
1

Machine Learning in the Atmospheric Sciences

How far have we come, what problems remain, and what can we do about them?

1. Earth ML

- Machine learning (ML) is an Artificial intelligence (AI) subset that allows machines to learn from data and make decisions
- ML has a deep history in the Geosciences. Remote sensing was an early adopter, along with applications in geomorphology, solid Earth geoscience, hydrogeophysics, seismology, and geochemistry
- For the purposes of this talk we will focus on one component: **the Atmosphere**

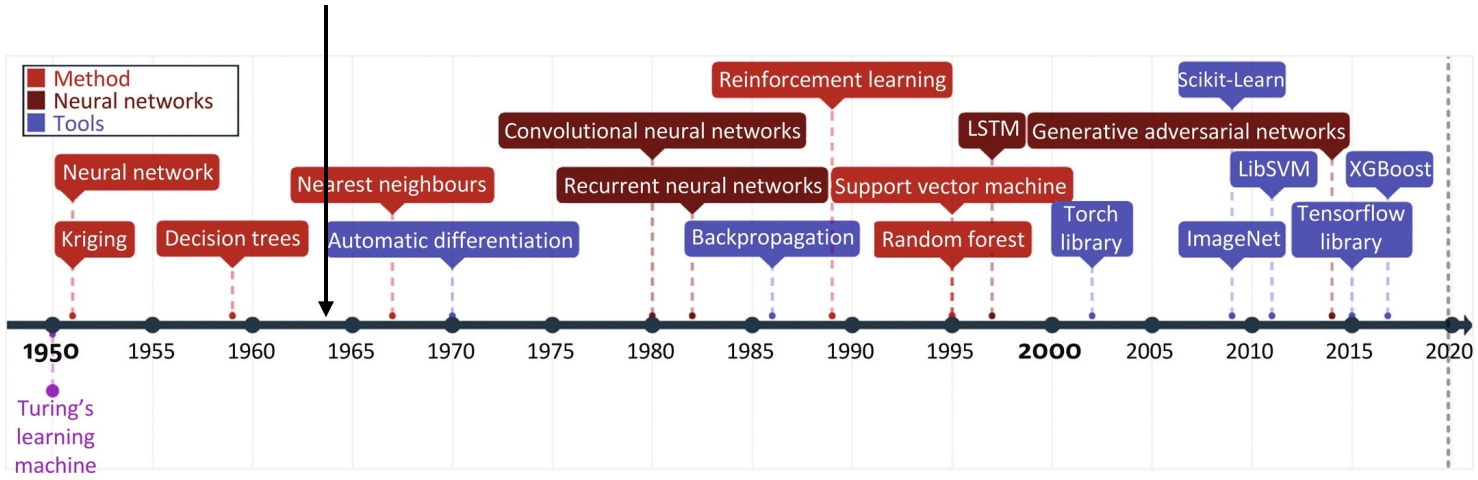


Adapted from (Sun et al., 2022)

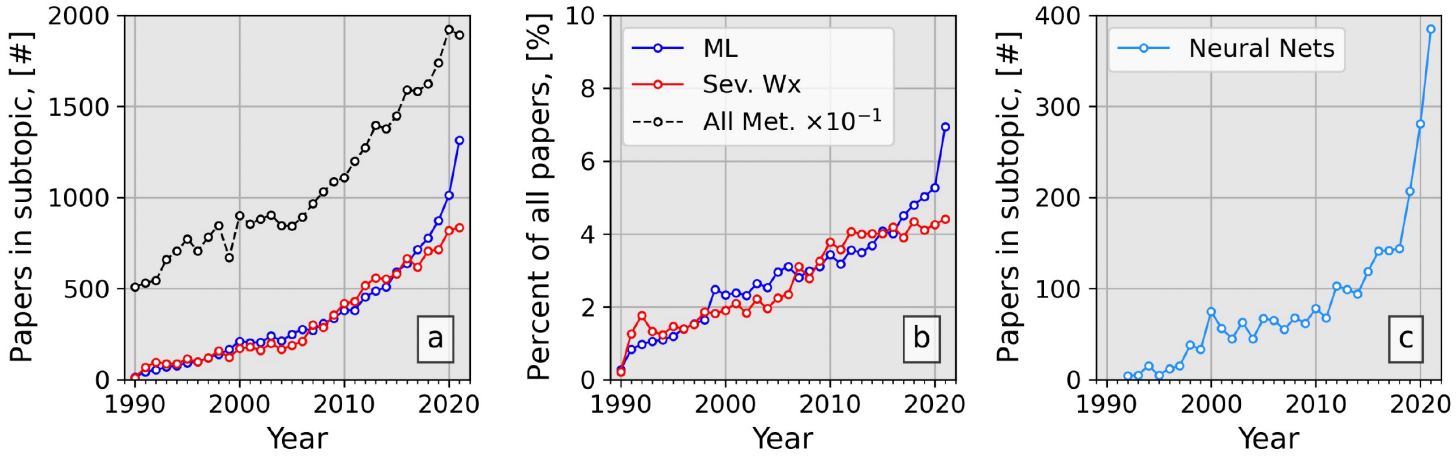
1. Origins

- The application of ML to problems in the Atmospheric Sciences is not novel
- The lack of computing resources, networking infrastructure and available global datasets were historically major limiting factors
- With improvements in both software and hardware capabilities, coupled with large observational datasets from reanalysis and remote sensing, ML has surged in popularity

Cyclicity of sediment deposits
(Preston and Henderson, 1964)

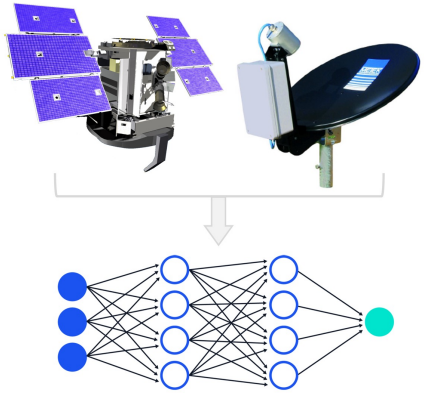


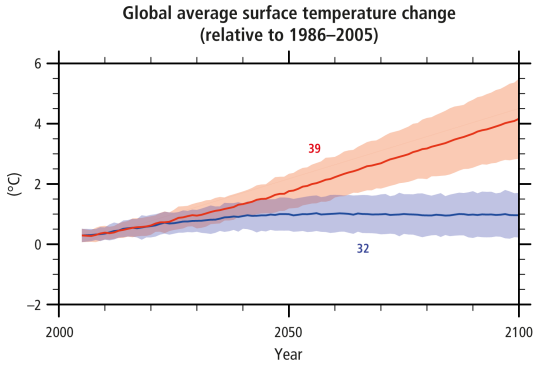
(Dramsch et al., 2020)



(Chase et al., 2022)

1. ML Challenges





Problem Identification → Model Development → Deployment & Maintenance → Integration & Decision Making

- What are the right questions to ask?
- Would a simpler method be sufficient?
- Are current AI methods mature enough for these problems?

- Curating training data
- Model architecture
- Hyperparameter optimization
- Uncertainty quantification
- Robustness

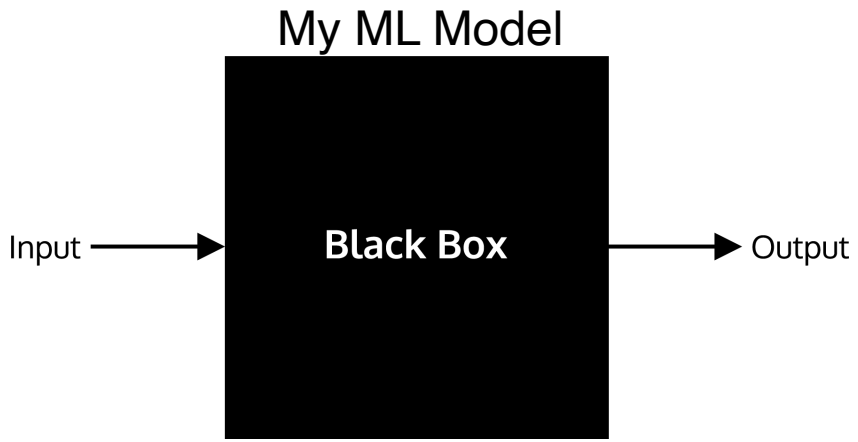
- Real-time processing and associated costs
- Operational management
- Permissions
- Open/closed source

- Integration with current techniques
- Ethical concerns
- AI Policies
- **Trust (Mechanistic Interpretability)**

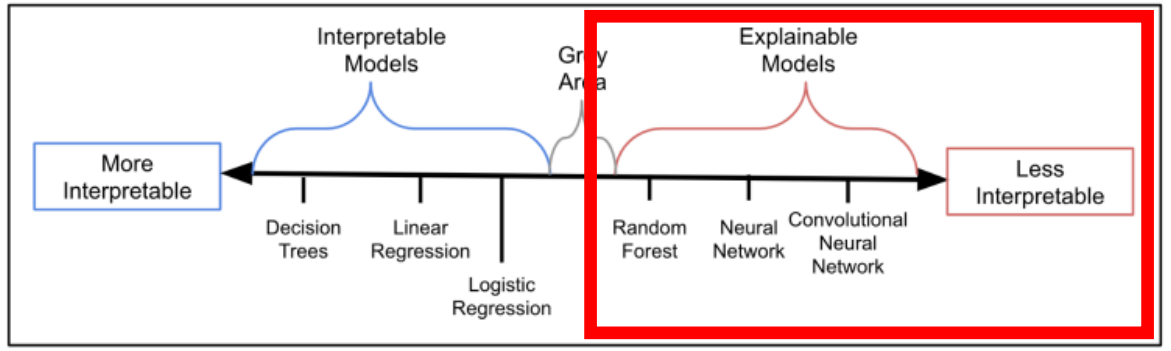
Adapted from
(Sun et al., 2022)

1. Towards Interpretability

- Many ML models are often considered *black boxes* as their inner workings are opaque to the observer
- This can lead to issues of **trust**, as biases or errors in the model decision making process may be difficult to identify
- This is especially relevant in Atmospheric Science, as models impact the daily lives of millions of people
- Explanatory techniques (e.g., LIME, SHAP) exist, and can help explain some NN behavior. But a method for definitive, comprehensive interpretable understanding remains to be seen



Let's focus here for this talk



(Conor O'Sullivan, 2020)

(McGill, 2010)

A Centimeter-Wavelength Snowfall Retrieval Algorithm Using Machine Learning

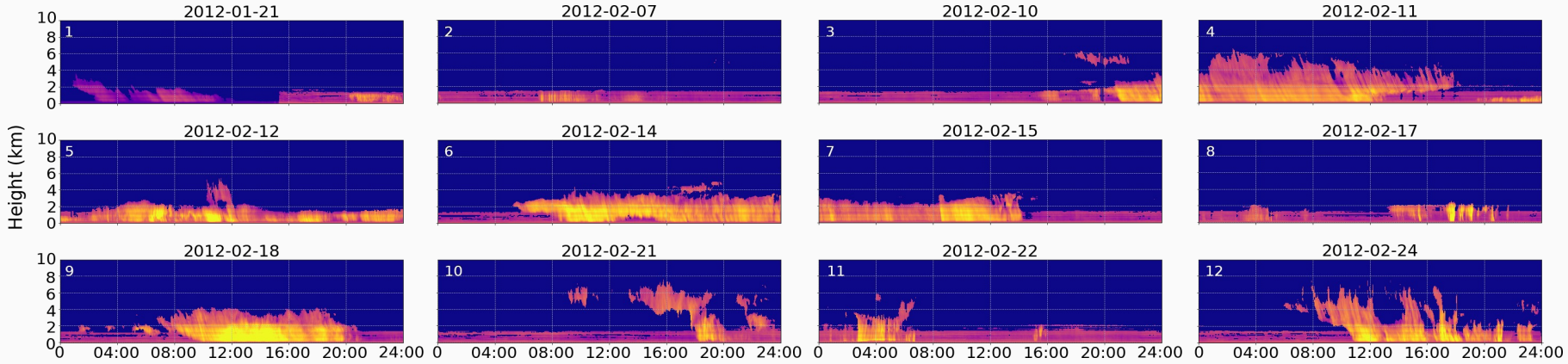
What retrieval accuracy can be achieved using a supervised machine learning algorithm (i.e., a random forest) when trained on surface radar observations?

https://doi.org/10.1175/JAMC-D-22-0036.1

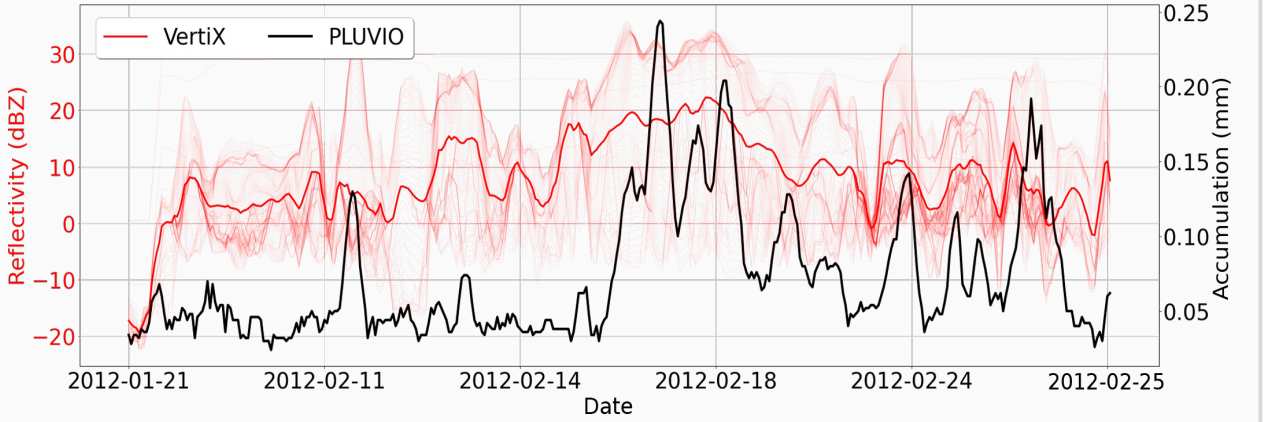
2. GCPEx

- Vertical radar data was collected from a GPM ground validation campaign
- This and collocated in situ snowfall data was used to train a random forest

Data



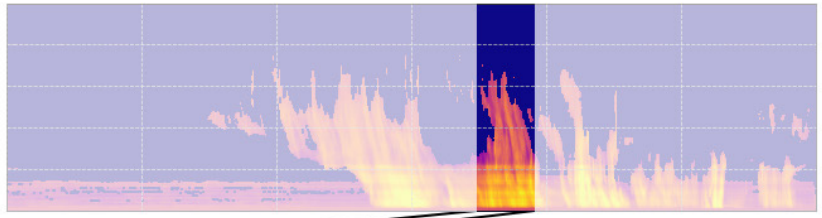
Idea

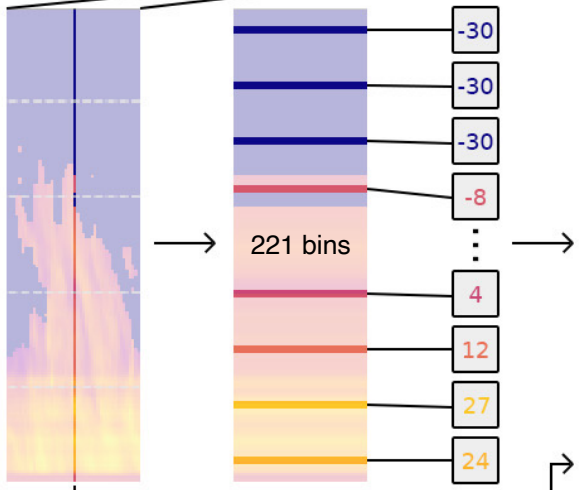


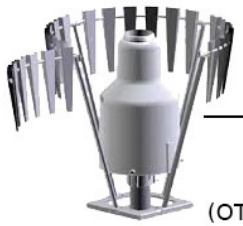
(King et al., 2022)

2. Random Forest Retrieval

GCPEx VertiX Radar – Feb. 24, 2012







0.05

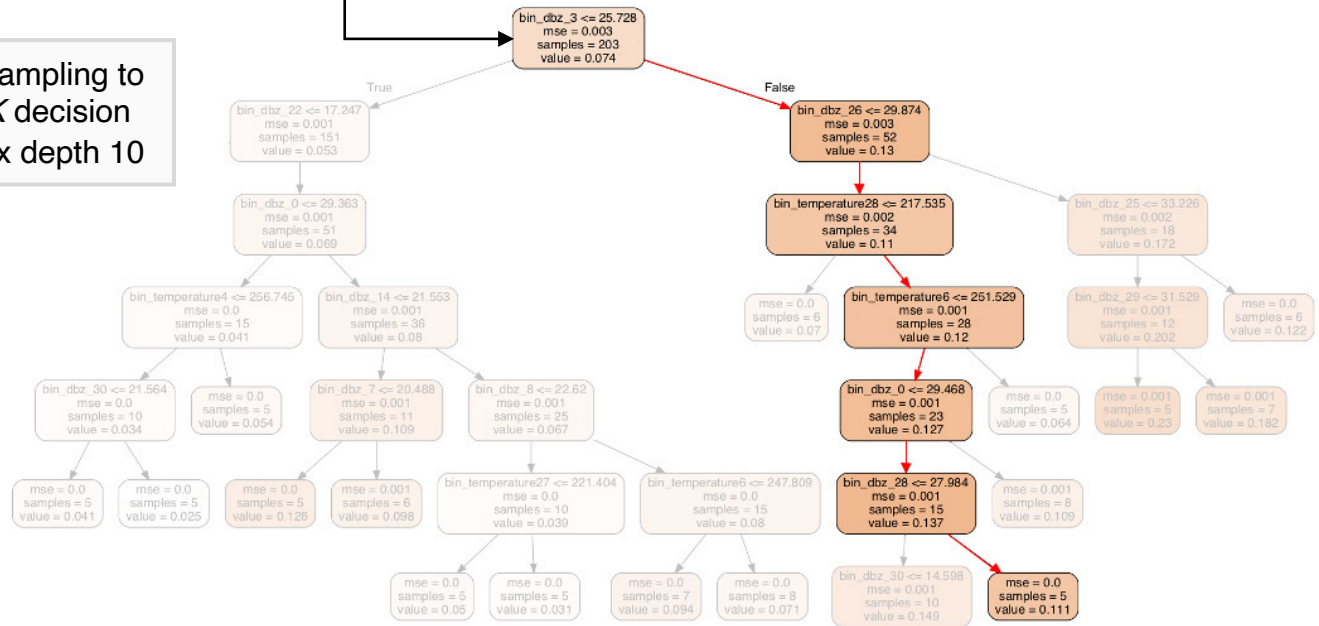
$$\begin{bmatrix} Z_{0_0} & Z_{1_0} & \dots & Z_{m_0} \\ Z_{0_1} & Z_{1_1} & \dots & Z_{m_1} \\ Z_{0_2} & Z_{1_2} & \dots & Z_{m_2} \\ \vdots & \vdots & \vdots & \vdots \\ Z_{0_n} & Z_{1_n} & \dots & Z_{m_n} \\ T_{0_0} & T_{1_0} & \dots & T_{m_0} \\ T_{0_1} & T_{1_1} & \dots & T_{m_1} \\ T_{0_2} & T_{1_1} & \dots & T_{m_2} \\ T_{0_n} & T_{1_2} & \dots & T_{m_n} \\ \overline{y_0} & \overline{y_1} & \dots & \overline{y_m} \end{bmatrix}$$

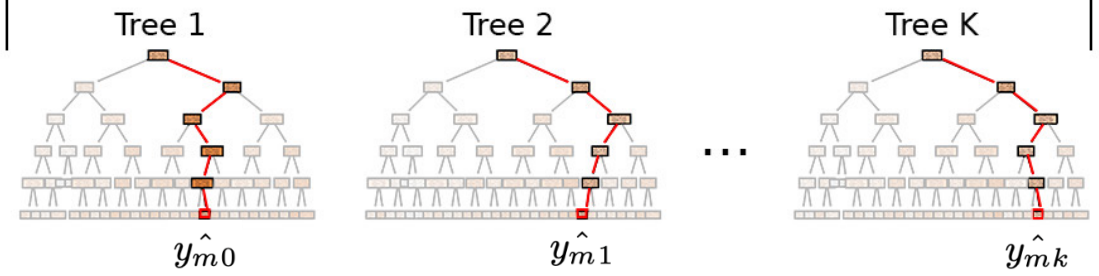
(2n+1, m)

$$\hat{y_m} = \frac{1}{K} \sum_{k=1}^{K} \hat{y_{mk}}$$

Surface snow accumulation (mm)

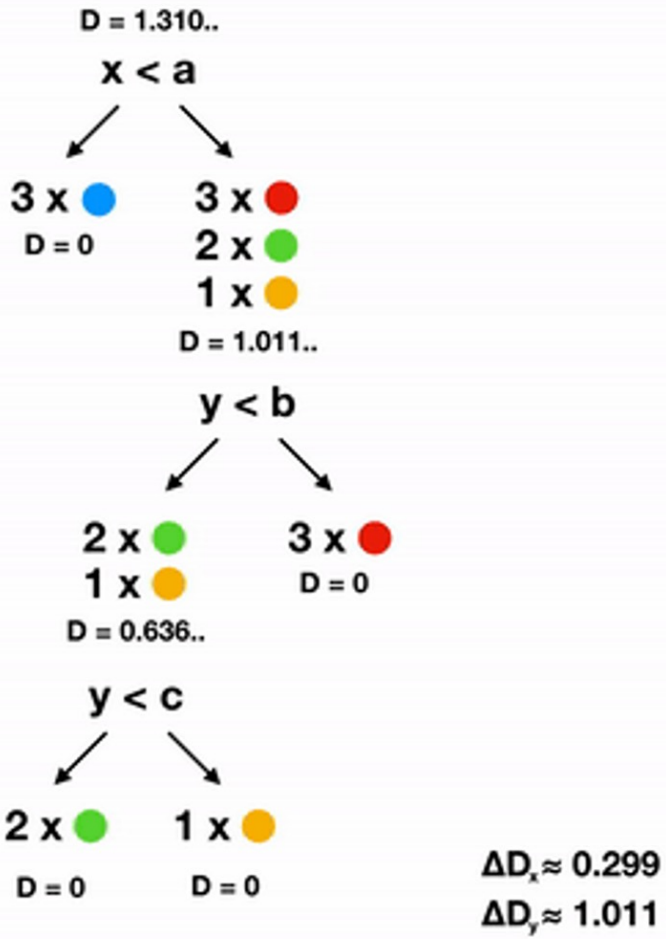
Bootstrap sampling to generate *K* decision trees of max depth 10

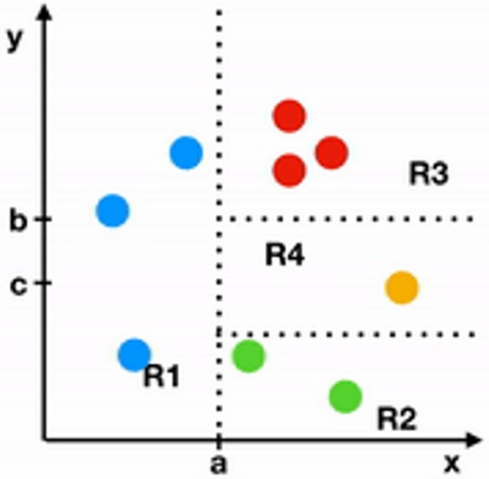




2. RF Explainability

- The RF calculates feature importance (in regression models) by averaging the decrease in mean squared error (MSE) across all trees when a feature is used for splitting
- It measures how much each feature contributes to the predictive power of the model by comparing the performance of trees with and without the feature
- Issue of "*Can't see the forest for the trees*"

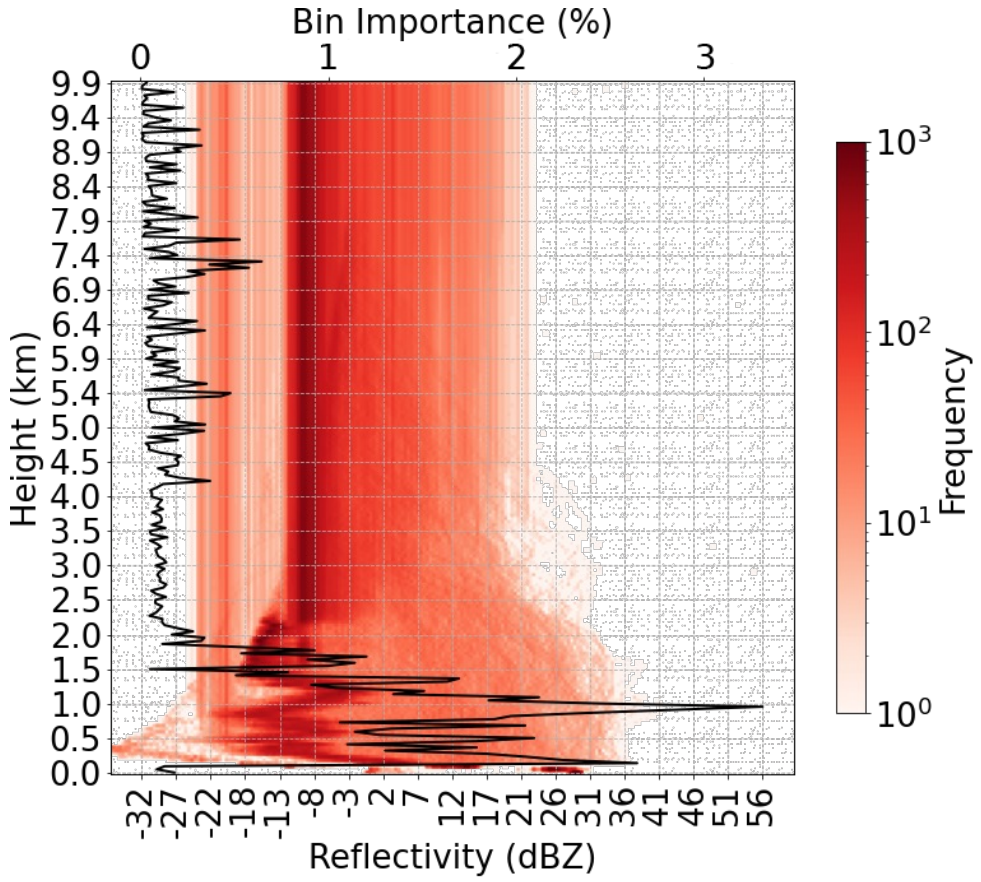




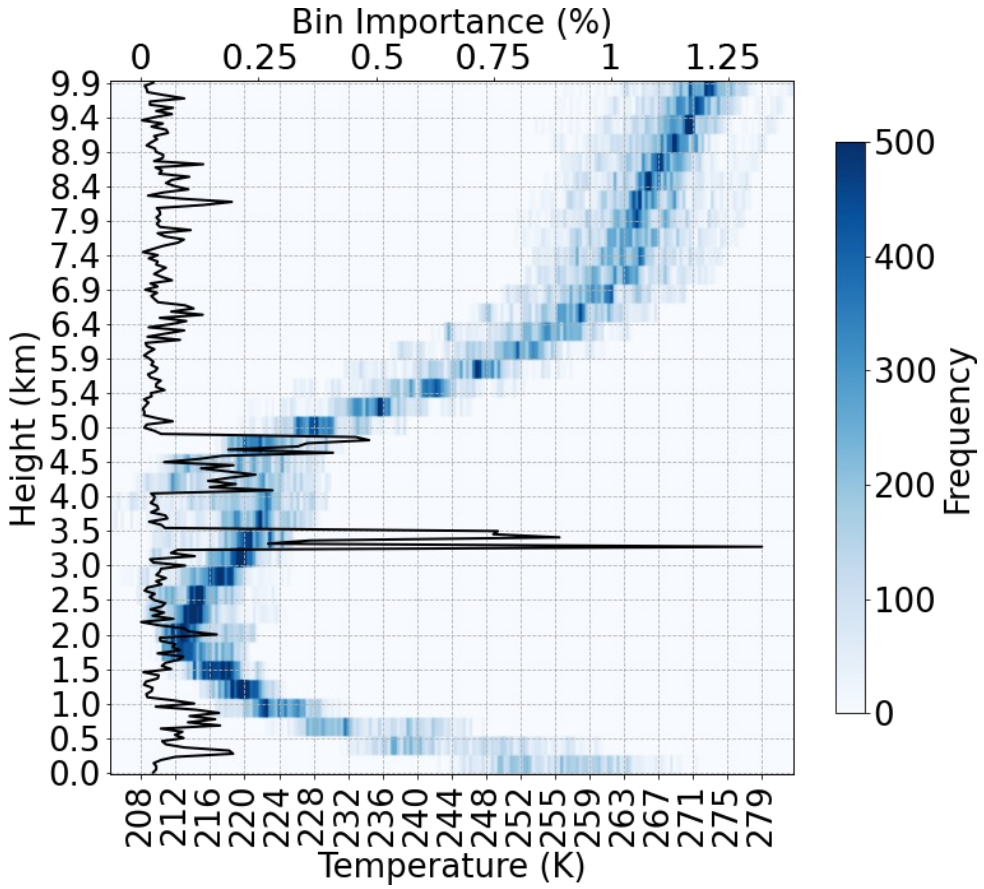
(Tom Grigg, 2019)

2. Feature Importance

Reflectivity



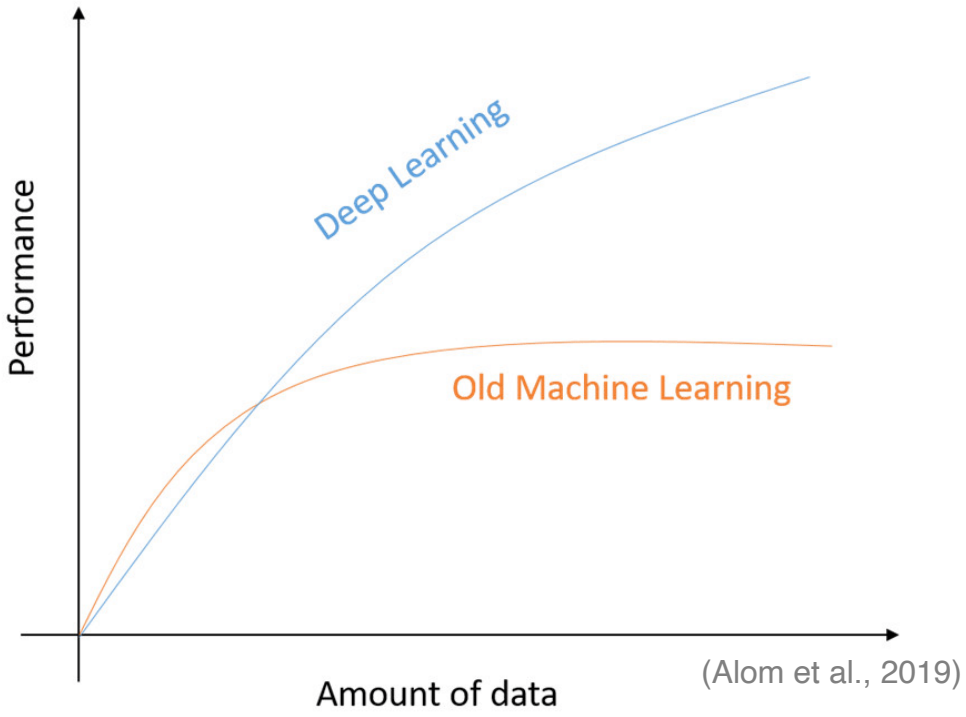
Temperature



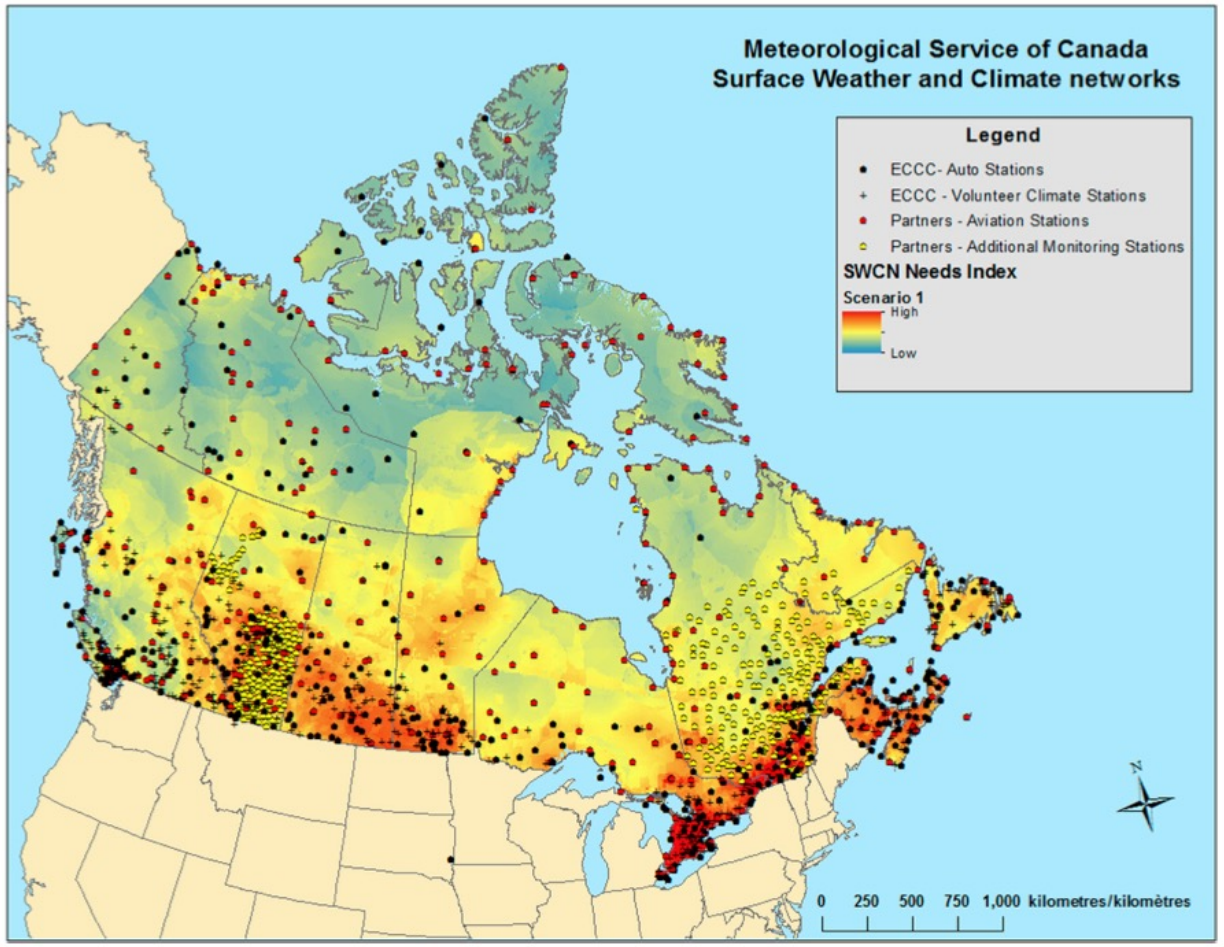
- We note a spike in reflectivity importance in the lowest 2 km of the atmosphere (this will be important later!)

2. Deep Learning

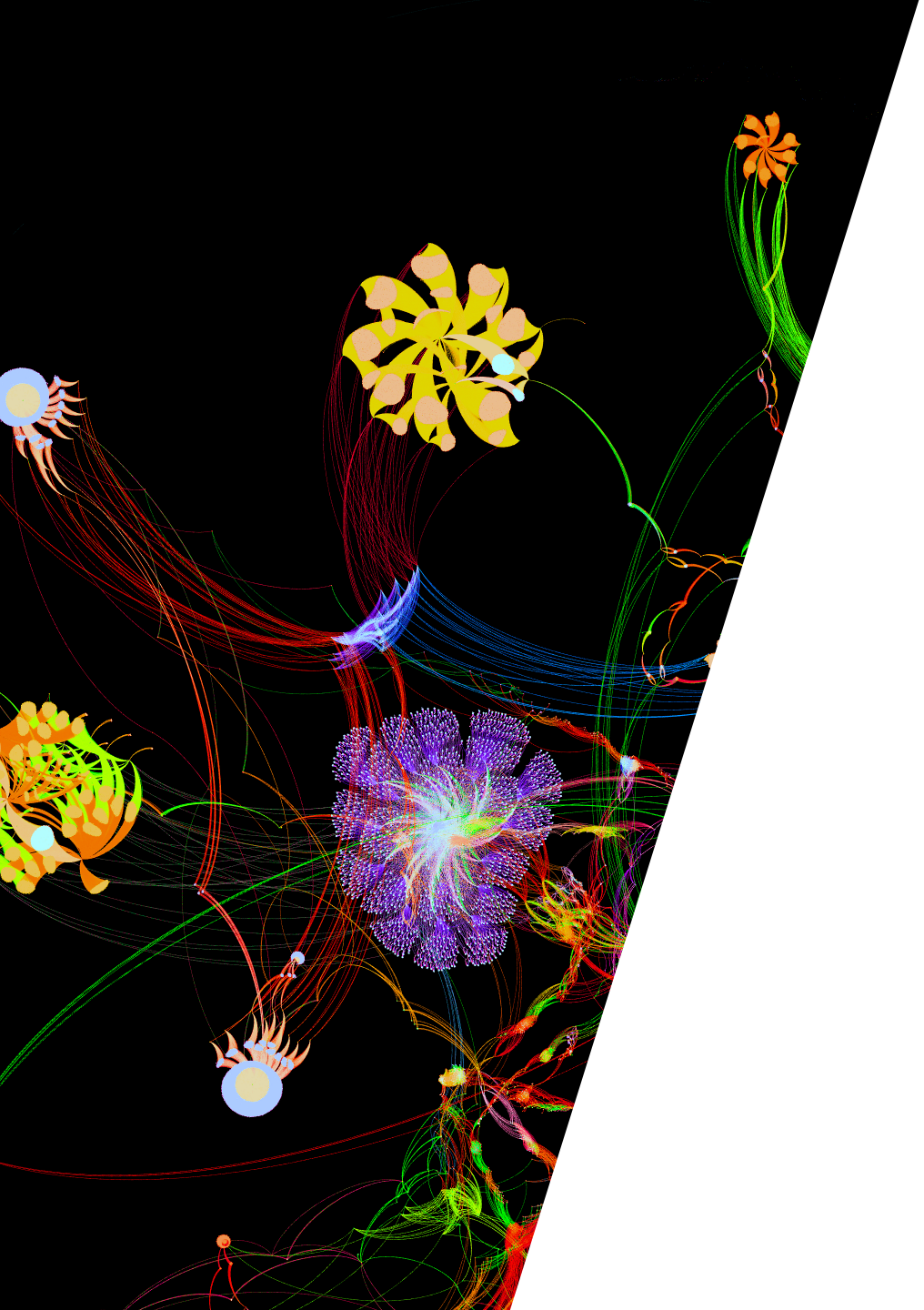
- With access to a wider observational network, could we train a deep neural network to derive a novel, high accuracy, precipitation retrieval?



(Alom et al., 2019)



(Mekis et al., 2018)



3

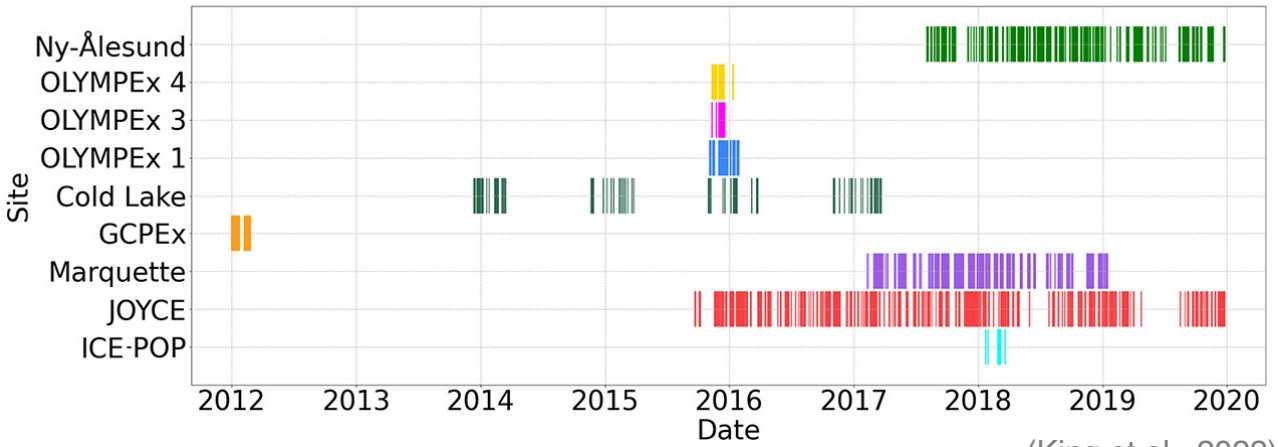
DeepPrecip: A deep neural network for precipitation retrievals

Can we generalize the previous model to new regional climates? Does the different architecture provide new insights into retrieval behavior?

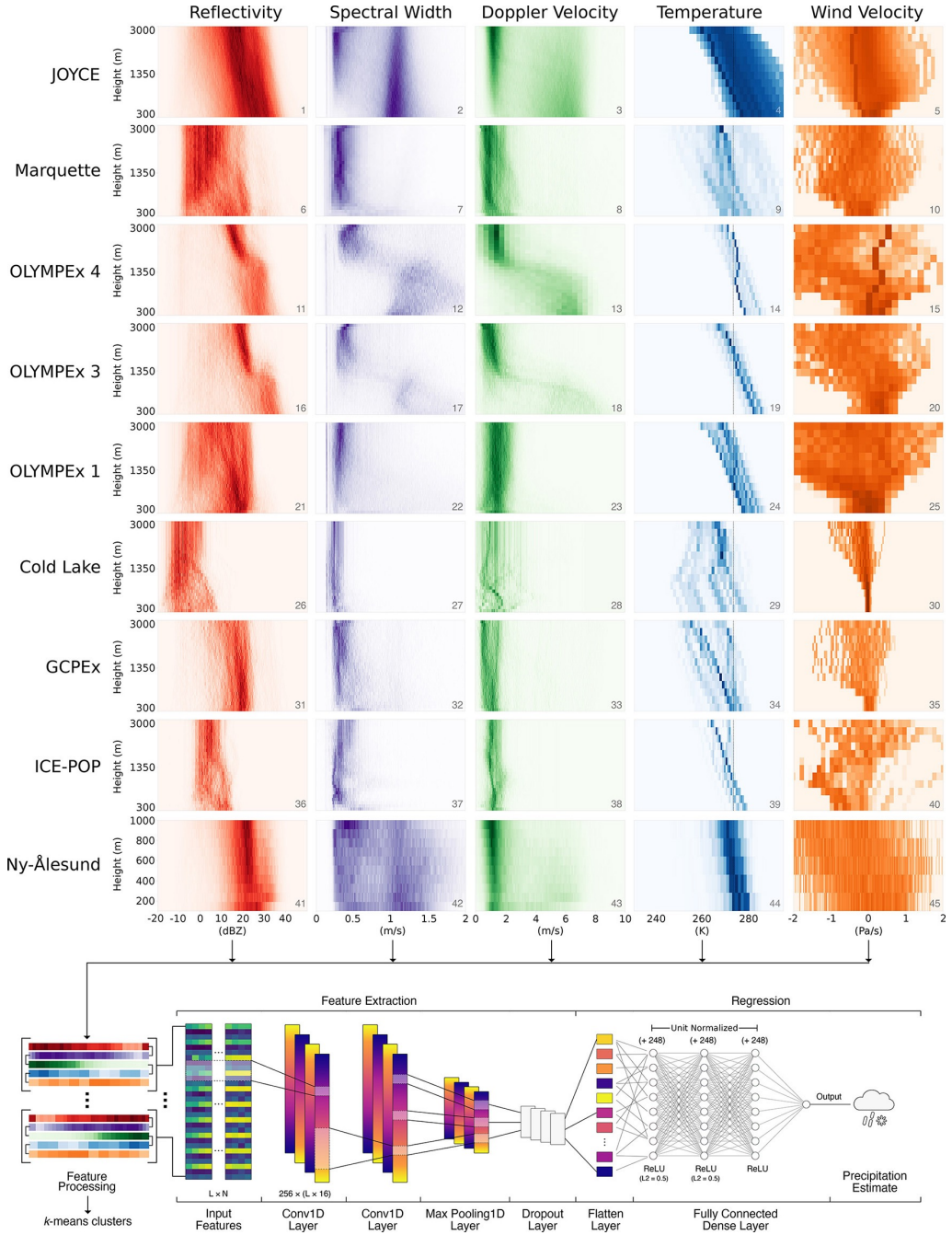
https://doi.org/10.5194/amt-15-6035-2022

3. DeepPrecip

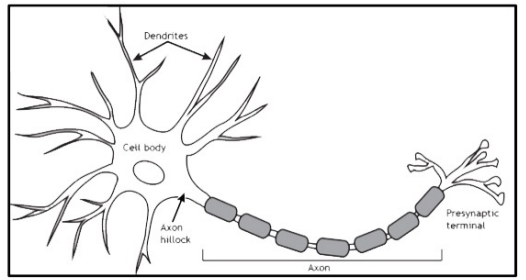
- Data collected from 9 sites spread across the northern hemisphere
- Each site has a collocated MRR and Pluvio gauge
- **Key:** observations from multiple regional climates and periods



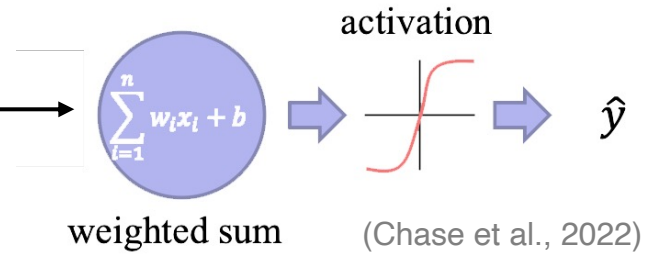
(King et al., 2022)



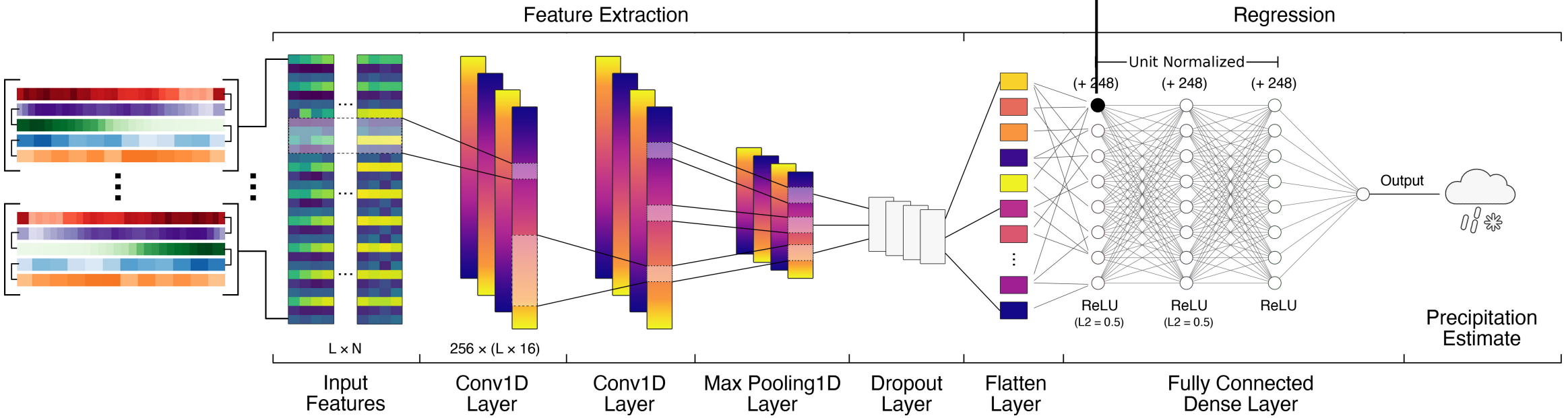
3. Model Architecture



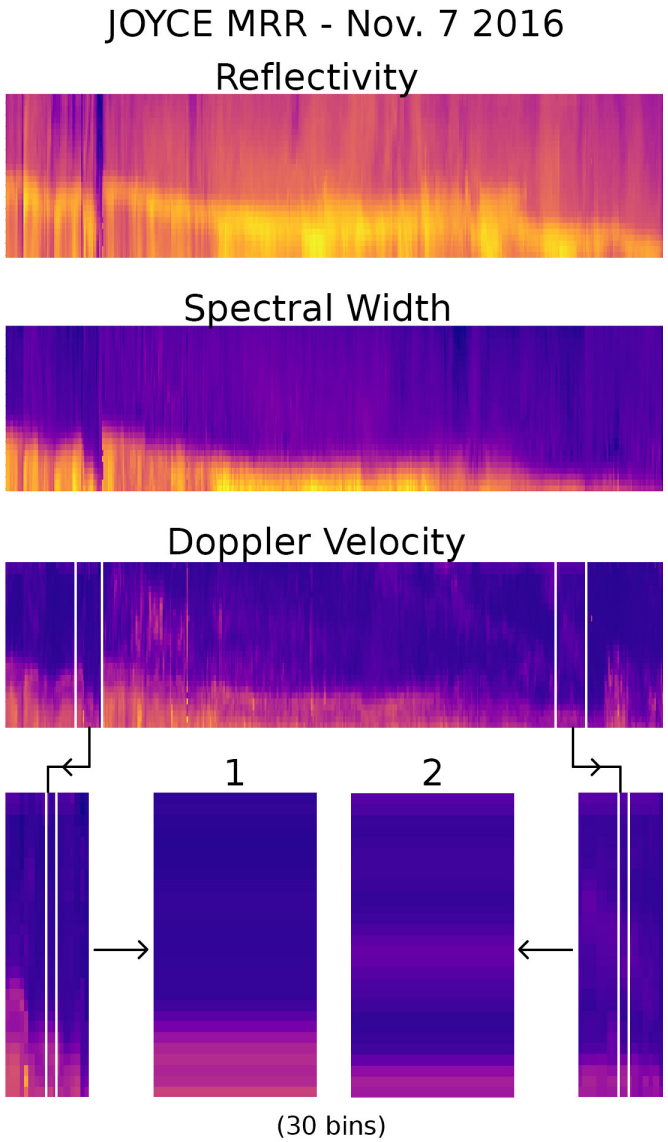
- What makes DeepPrecip different from the previous RF?
- DeepPrecip is a 1D convolutional neural network with two primary system components responsible for:
  1. Feature Extraction
  2. Snowfall Regression

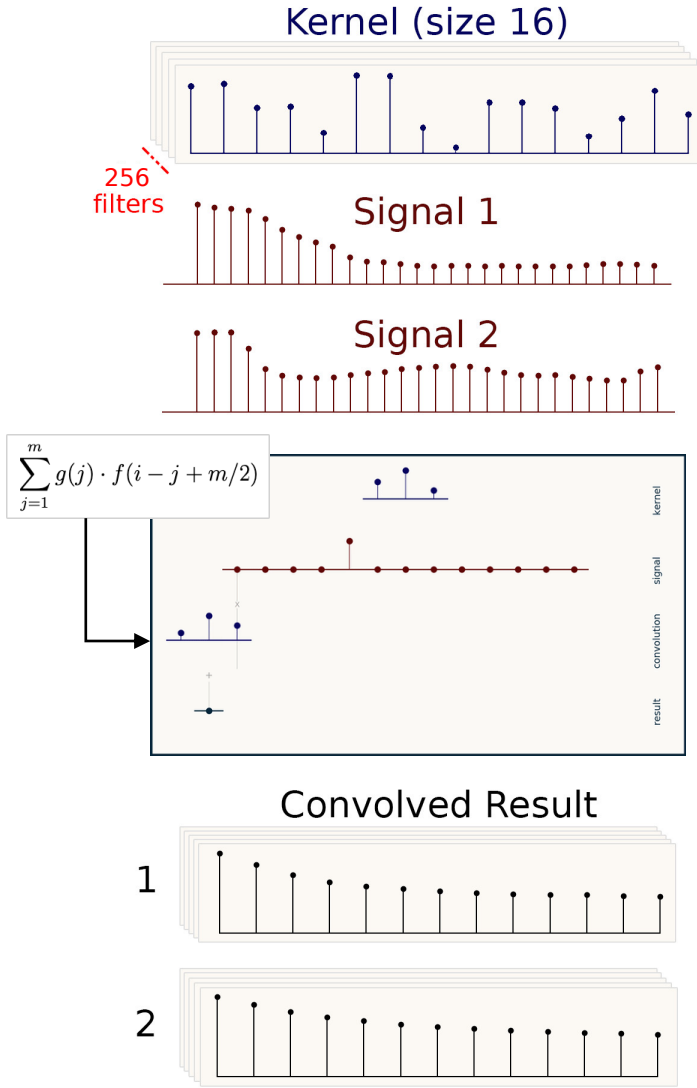


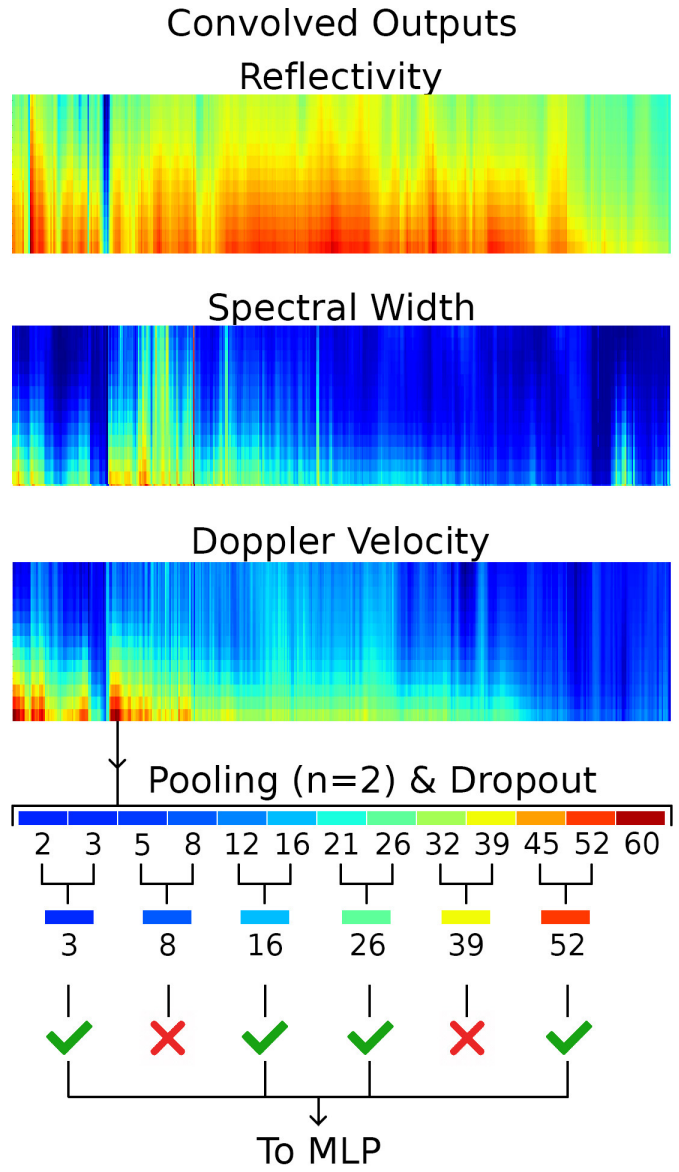
(Chase et al., 2022)



3. Feature Extraction

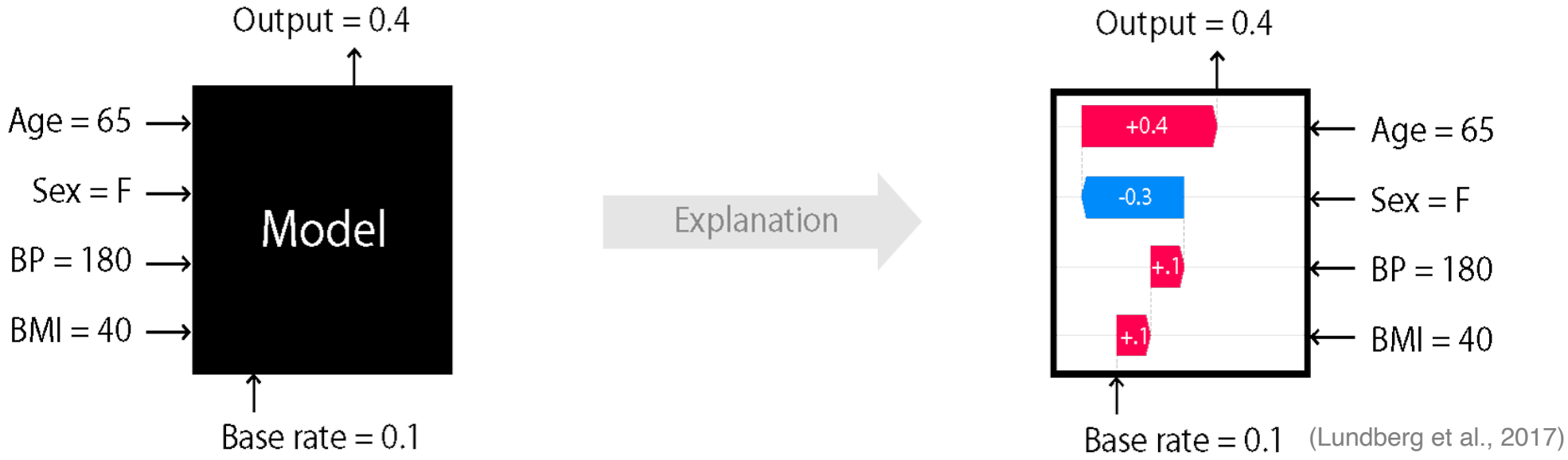






3. SHapley Additive exPlanations (SHAP)

- Feature attribution is a group of explainability techniques used to explain how machine learning models make decisions (e.g., SHAP or LIME)



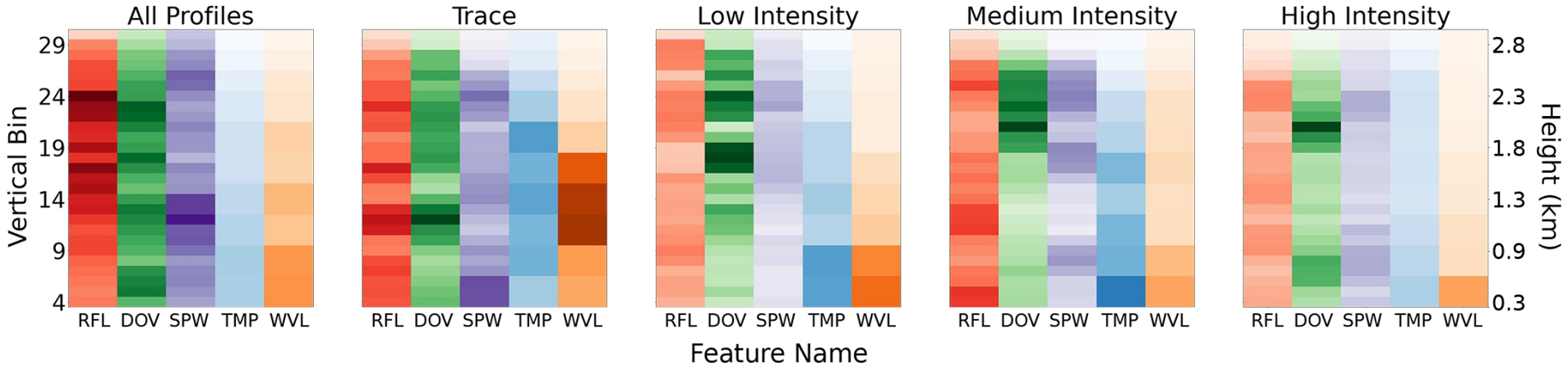
(Lundberg et al., 2017)

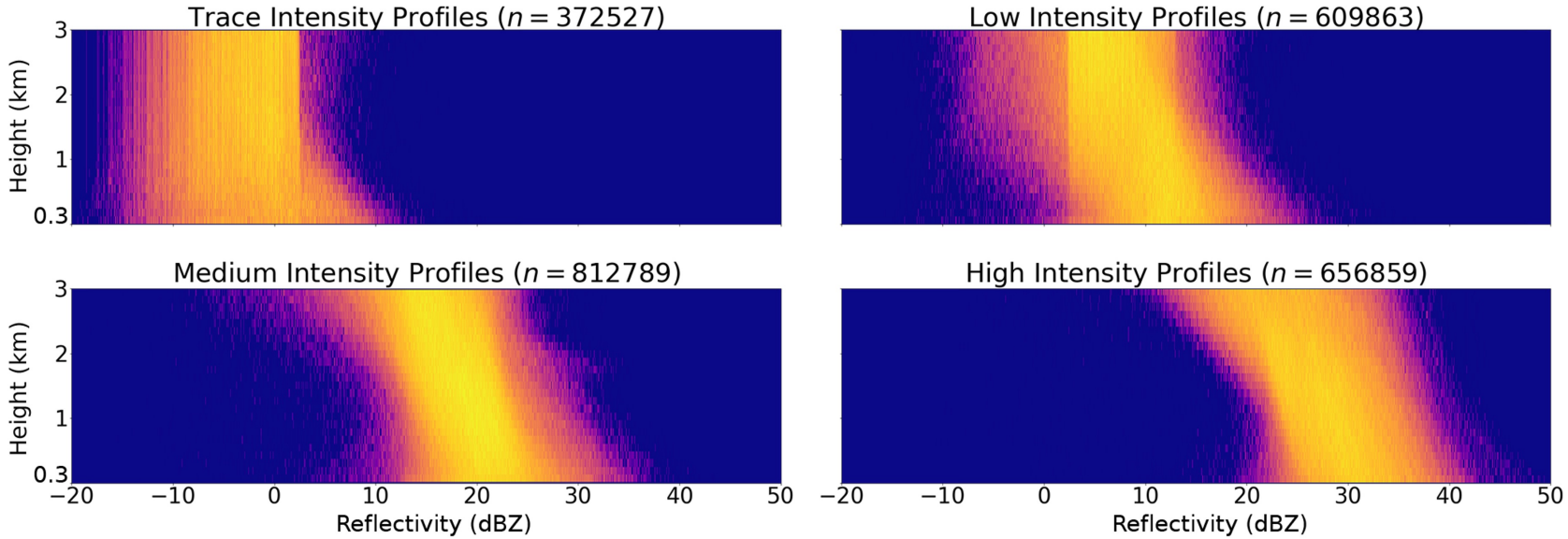
- At a high level, the Shapley value is computed by carefully perturbing input features and seeing how changes to the input features correspond to the final model prediction

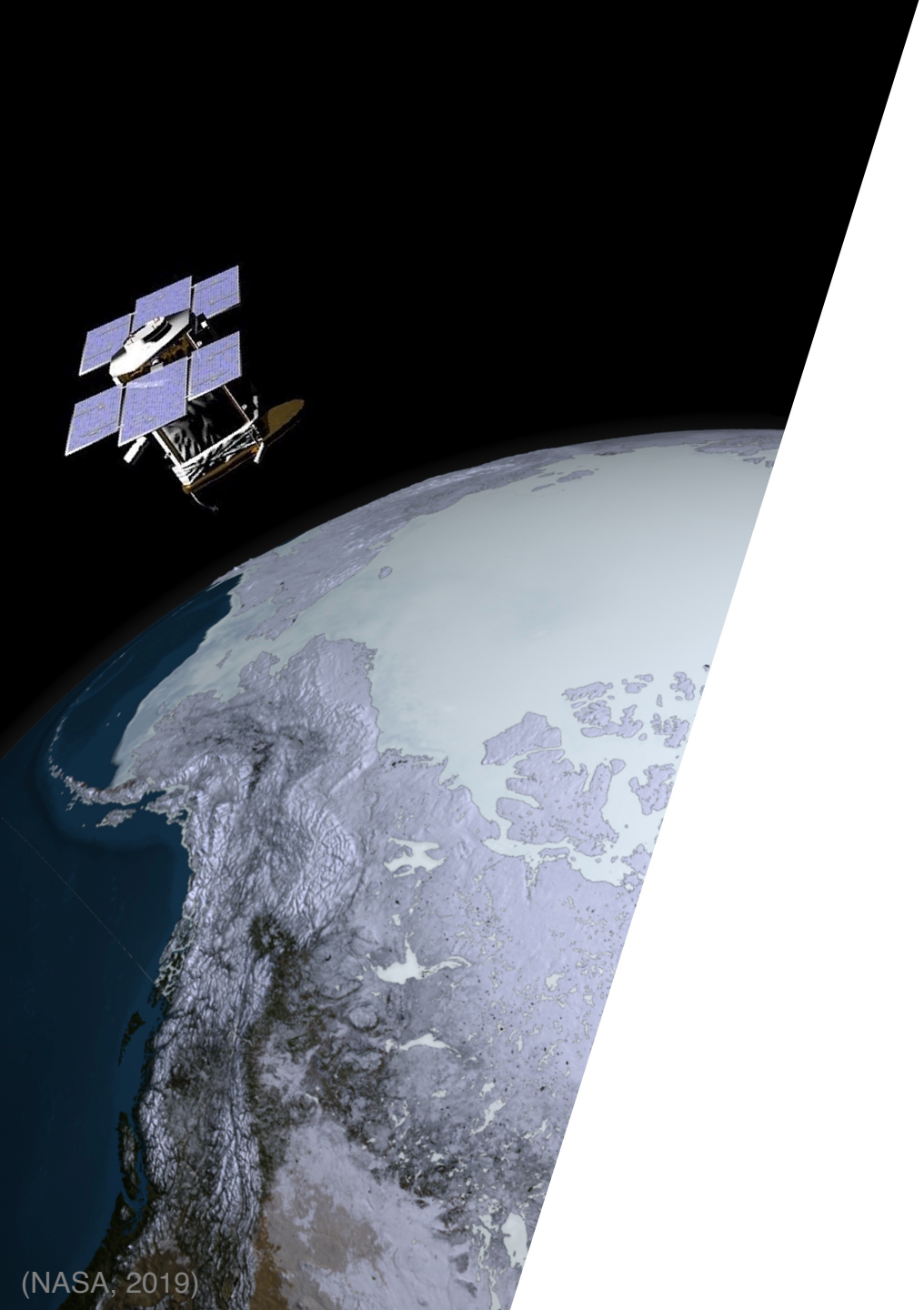
- The Shapley value of a given feature is then calculated as the average marginal contribution to the overall model score

3. SHAP Feature Importance

- We broke the data up into groups using a standard k-means approach
- Note that darker colors represent a higher ”importance score”
- We once again see that the region below 2 km is very important in retrieval accuracy!







(NASA, 2019)

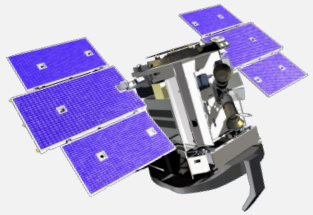
4

Development of a full-scale connected U-Net for reflectivity inpainting in spaceborne radar blind zones

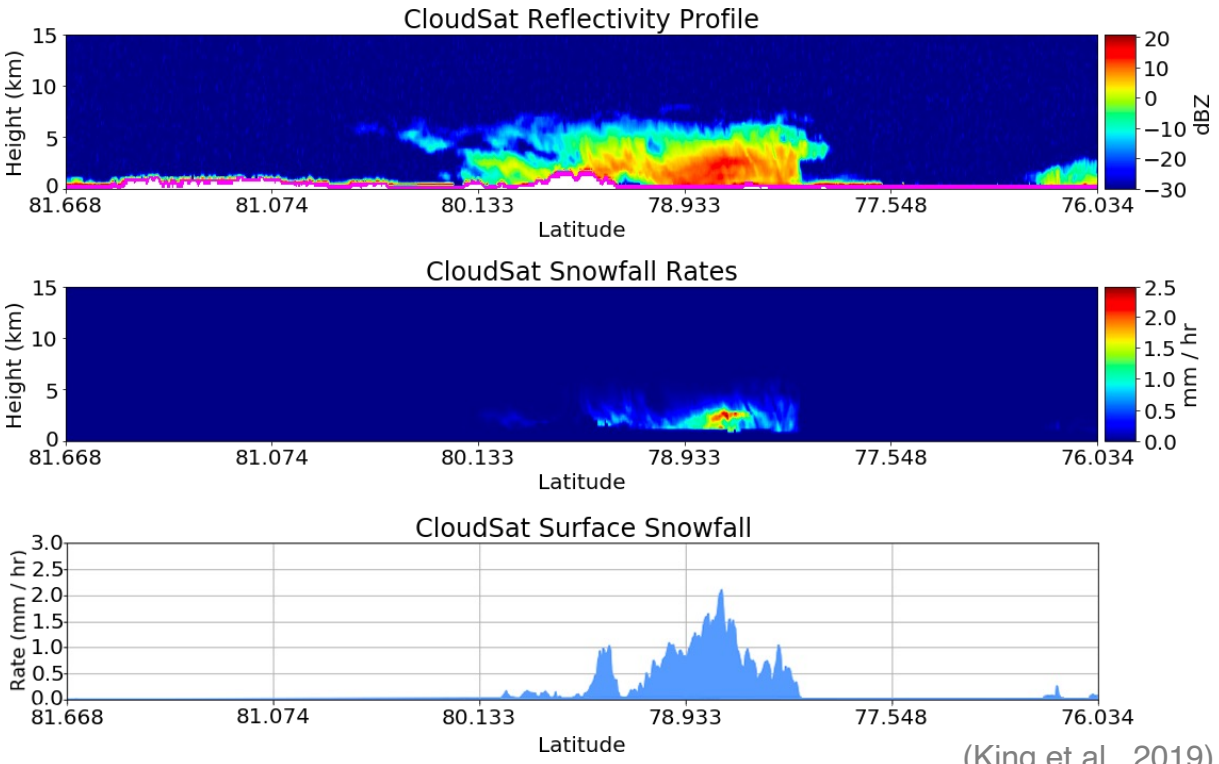
What role do generative models have in improving known issues from satellite-derived remote sensing precipitation estimates in the lowest 2 km of the atmosphere?

https://doi.org/10.1175/AIES-D-23-0063.1

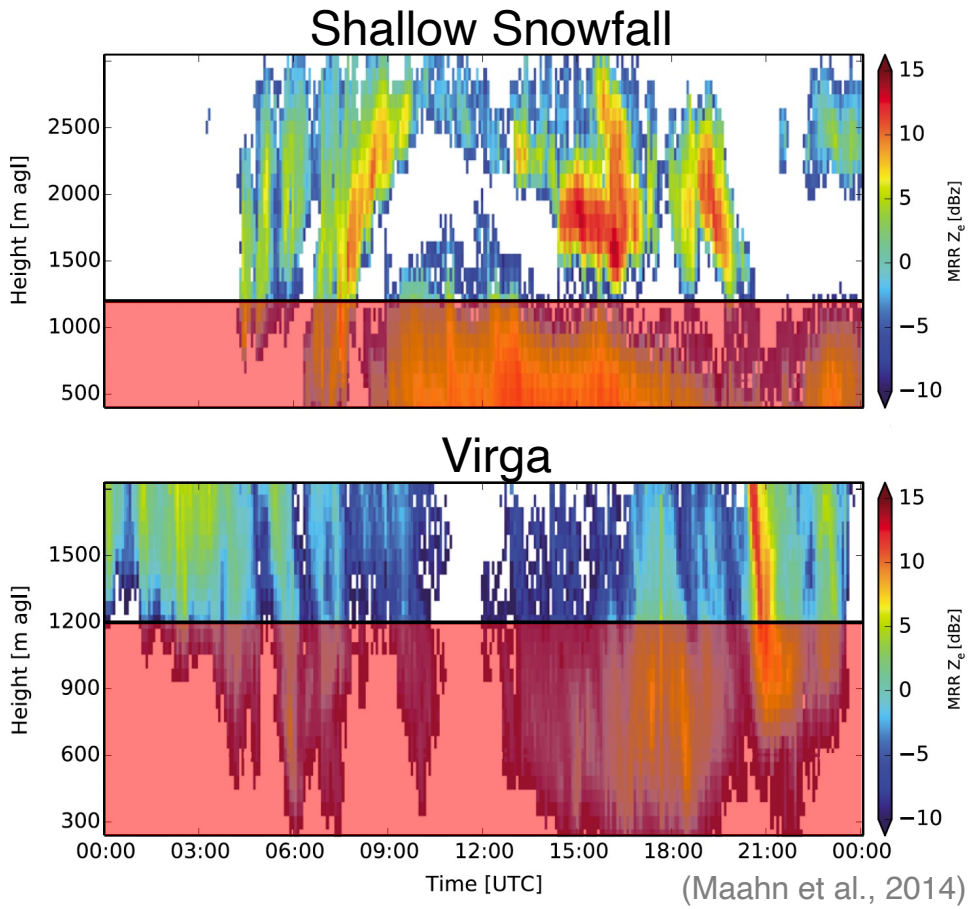
4. The Radar Blind Zone



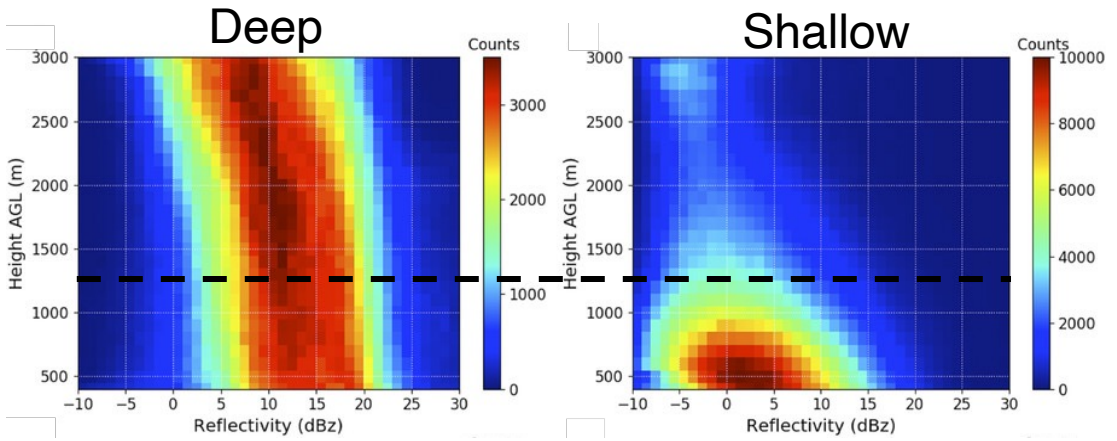
The CloudSat Cloud Profiling Radar (CPR) allows us to look inside of clouds to view hydrometeor activity



(King et al., 2019)

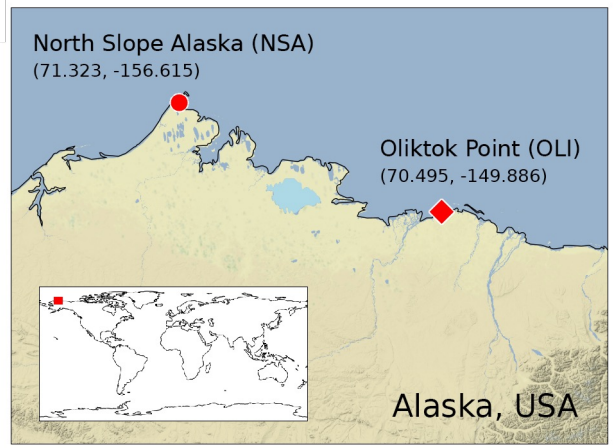


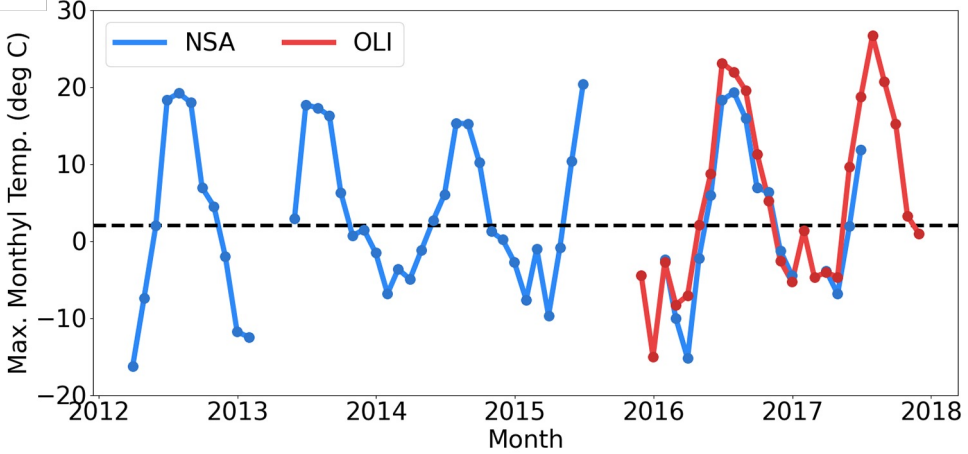
(Maahn et al., 2014)



(Pettersen et al., 2020)

4. Training Data

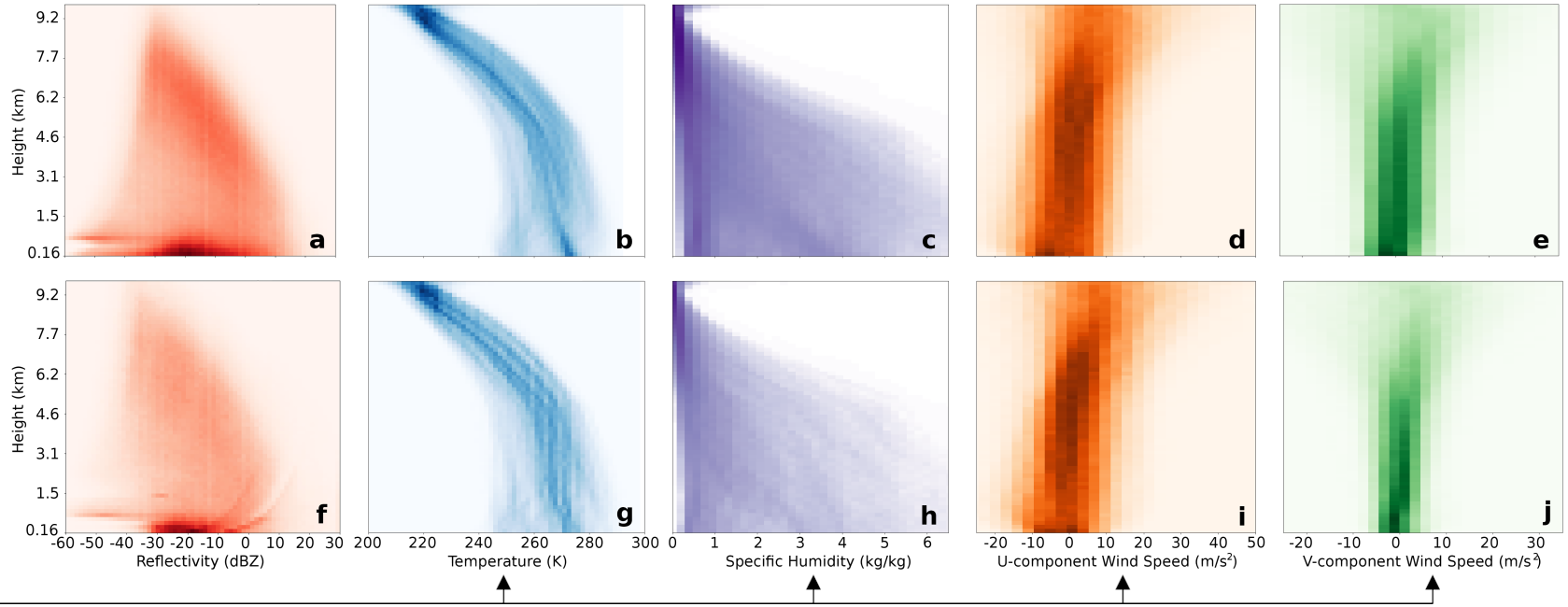




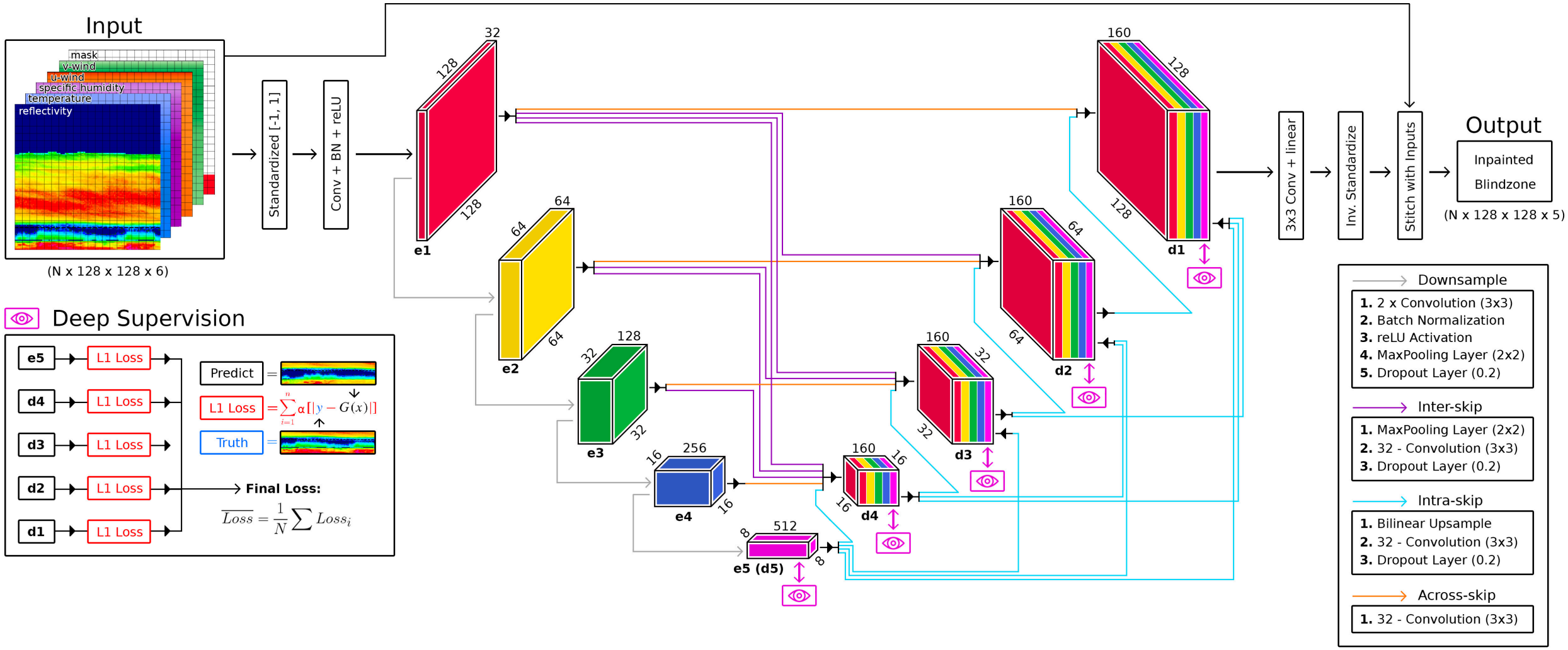
- Data comes from two Arctic locations along the northern Alaskan coast (NSA & OLI)
- We focused on using cold season observations when temperatures were below 2°C

KAZR-CLOUDSAT
This VAP applies reflectivity offsets to surface KaZR observations to align the reflectivities more closely with those observed by CloudSat (Kollias et al., 2019)

ERA5
Collocated atmospheric data from ERA5 is also aligned to provide the models with additional context



4. U-Net Architecture

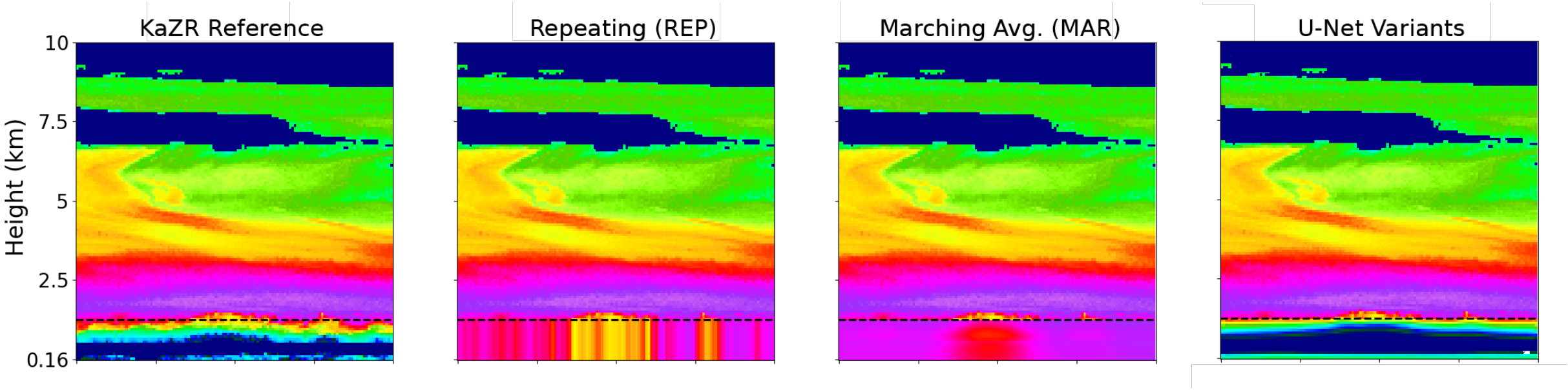


(King et al., 2024)

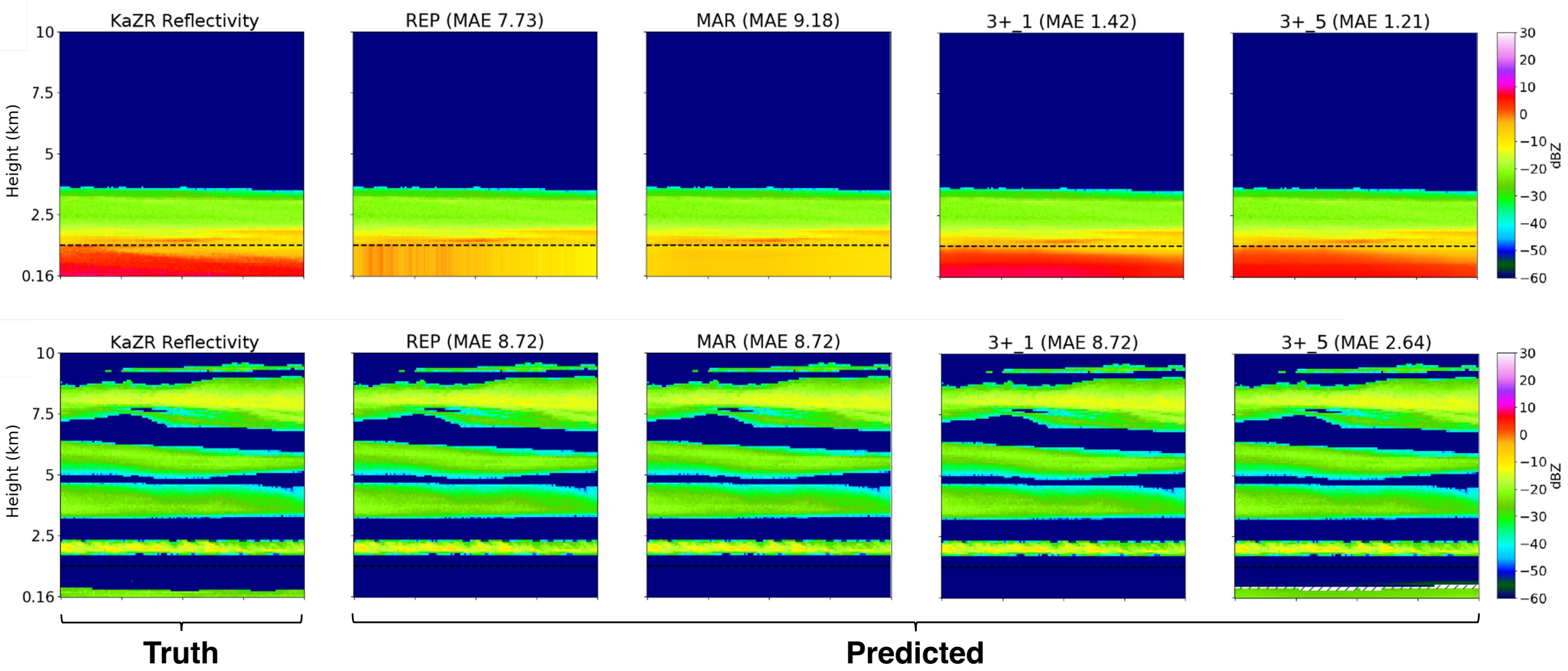
4. Models

- Five models are compared in this analysis including 2 traditional linear techniques and 3 CNN-based approaches
- This allows us to assess performance using both “simple” and “complex” inpainting schemes

Name	ID	Scheme	N_{params}	Inputs	Input Shape	Output Shape
Repeating Extrapolation	REP	Structure-based	–	r	1×128	16×128
Marching Average	MAR	Smooth structure	–	r	4×128	16×128
UNet++	++	CNN	7,000,804	r,t,q,u,v	$128 \times 128 \times 6$	$128 \times 128 \times 5$
3Net+ Single Channel	3+_1	CNN	6,789,620	r	$128 \times 128 \times 2$	128×128
3Net+ Multi Channel	3+_5	CNN	6,789,620	r,t,q,u,v	$128 \times 128 \times 6$	$128 \times 128 \times 5$

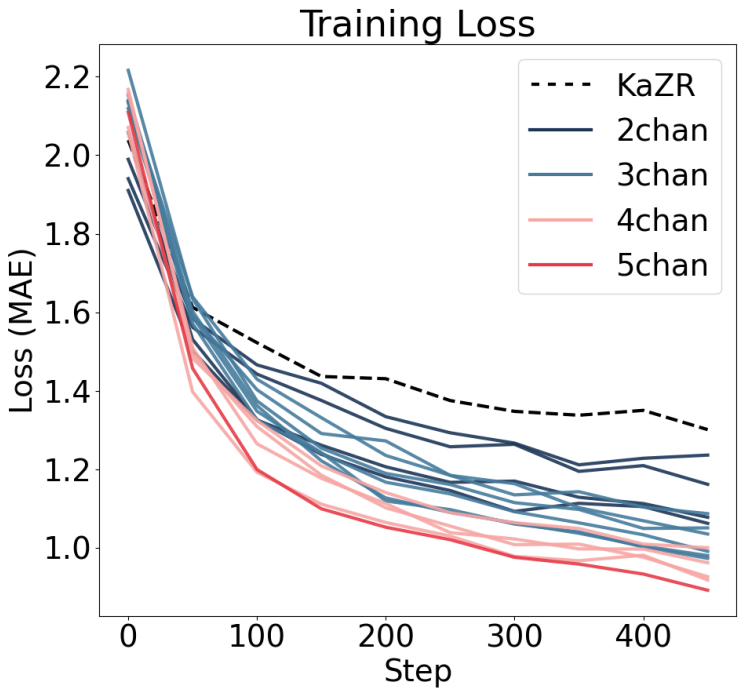


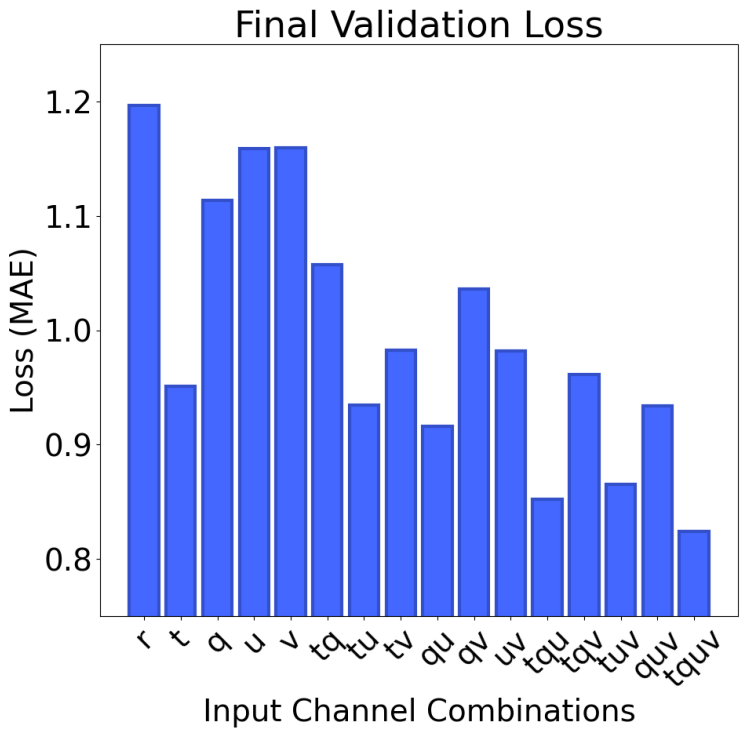
4. Inpainting Results

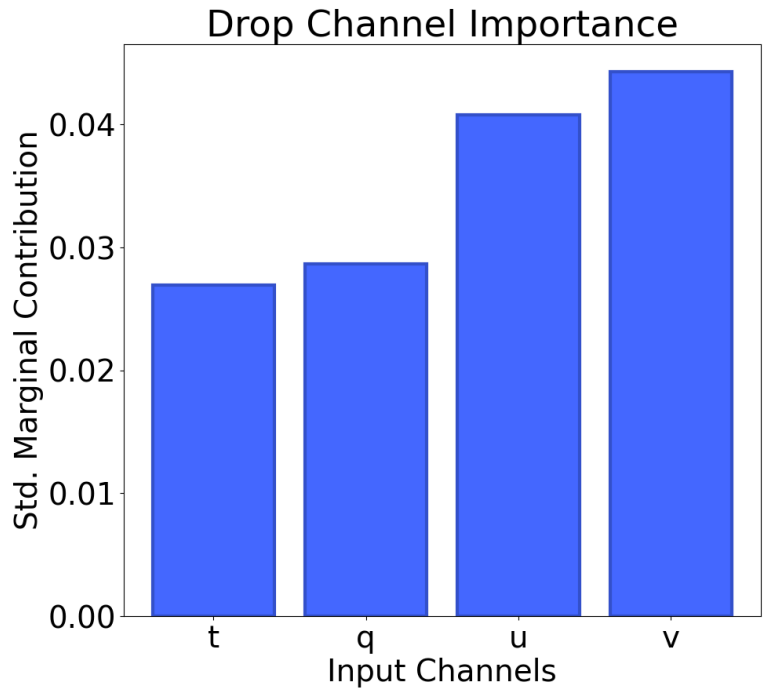


4. Drop Channel Importance

- How is my model making these choices? Can we learn something about the internal decision making process like we did with the RF and DeepPrecip?
- How important are the ERA5 variables? Do they add value over just using the KaZR?

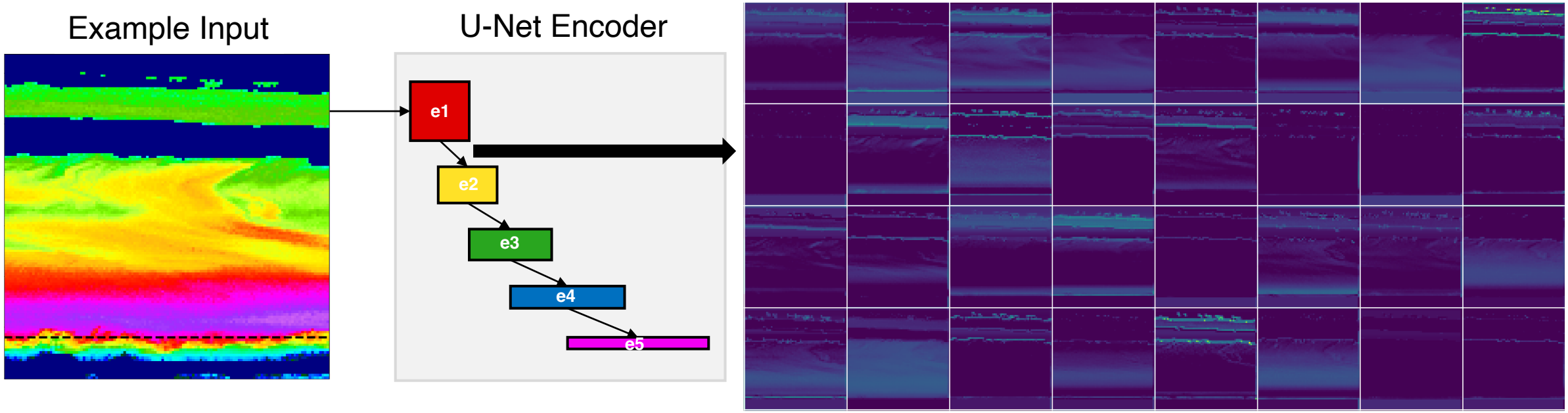






4. Feature Maps

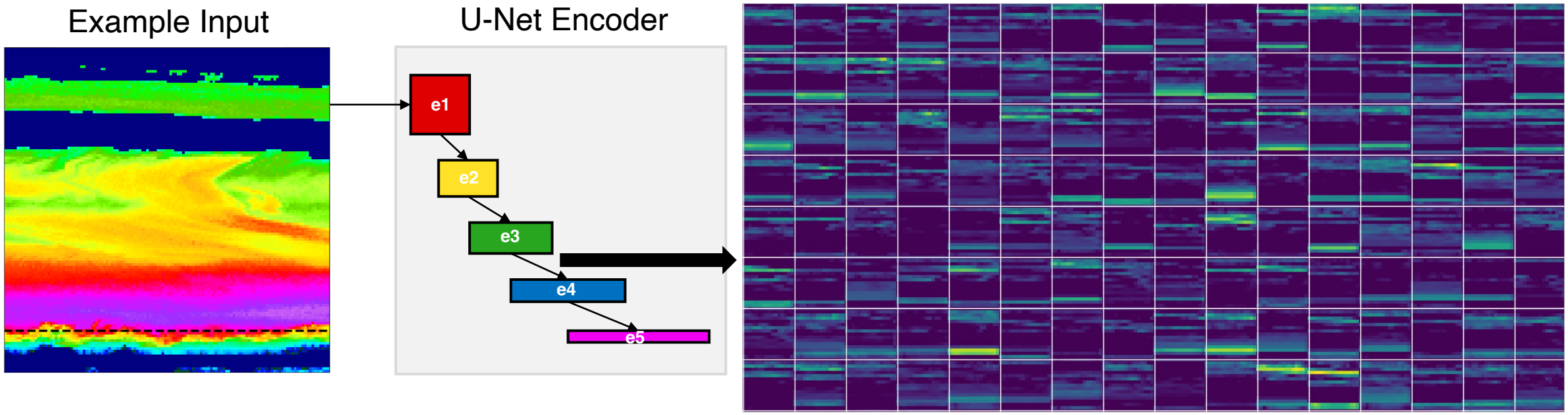
- Using this U-Net architecture, the model appears to learn to relate features in cloud aloft to blind zone reflectivity structures



- Cloud edges, blind zone threshold reflectivity, cloud gaps and reflectivity gradients are common structures identified as being important contributors to inpainting accuracy

4. Feature Maps

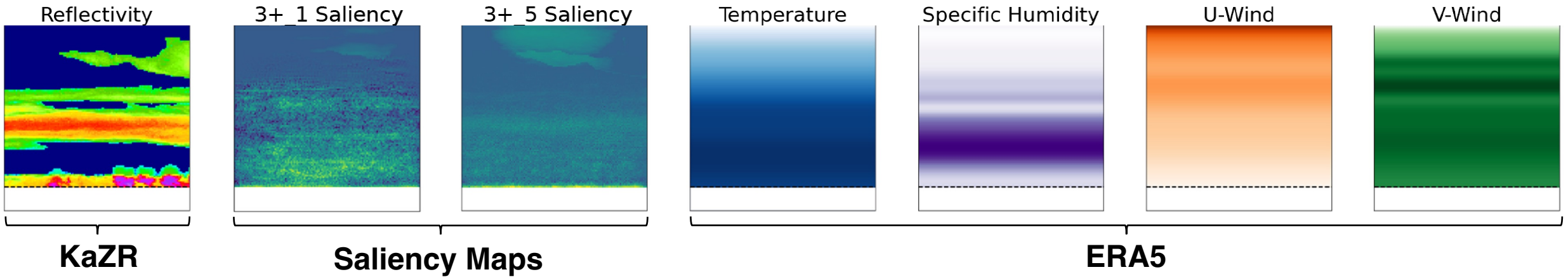
- Using this U-Net architecture, the model appears to learn to relate features in cloud aloft to blind zone reflectivity structures



- Cloud edges, blind zone threshold reflectivity, cloud gaps and reflectivity gradients are common structures identified as being important contributors to inpainting accuracy

4. Saliency Maps

- By examining a handful of saliency maps, we gain some insight into regions of importance at inference which we can then attempt to connect back to physical processes



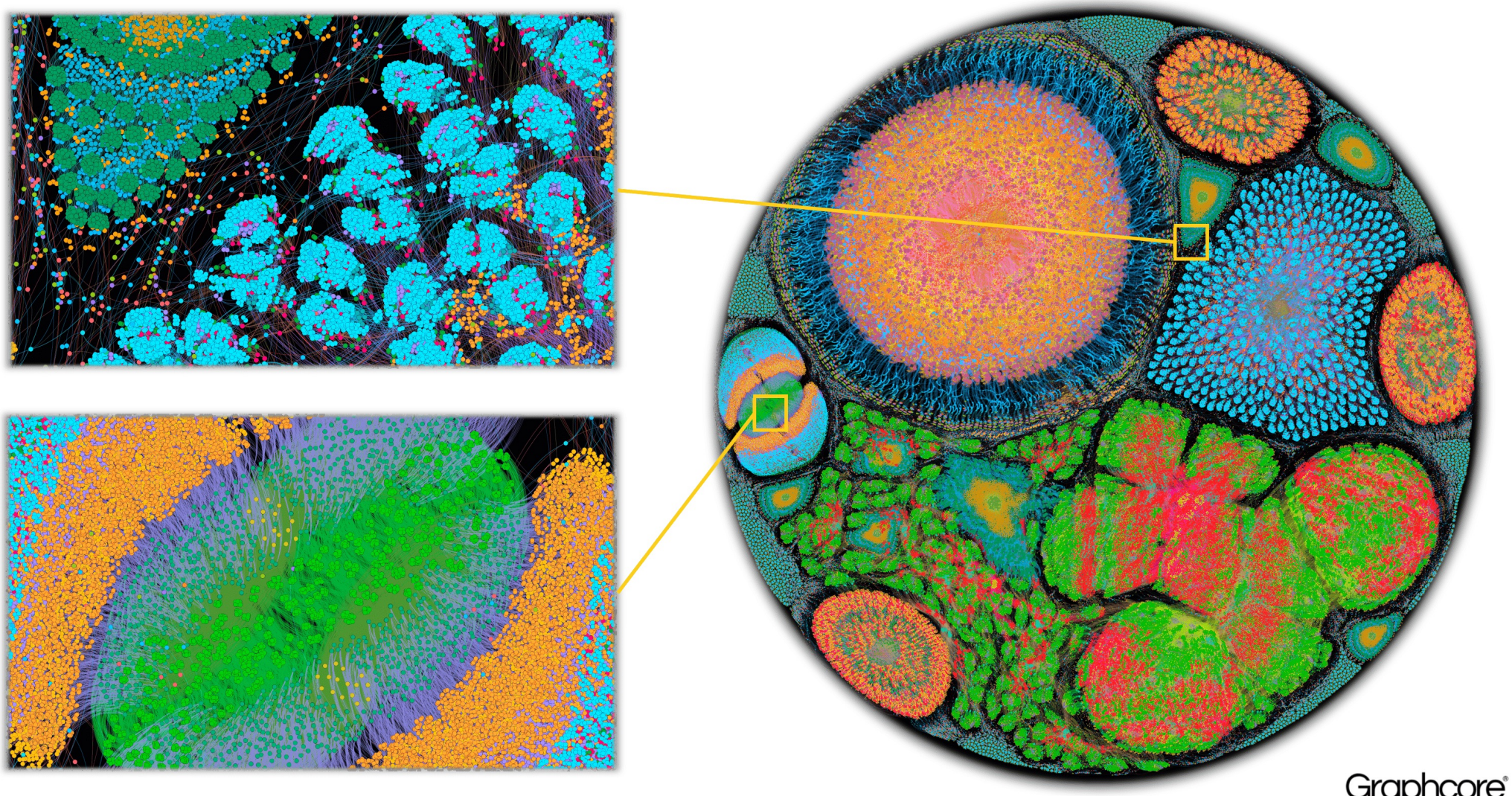
- The most important regions tend to be near the blind zone threshold or in reflectivity gradients
- Areas **without** reflectivity information also appear as important in data sparse cases
- ERA5 information near the tropopause appears as a halo of significance in the 3+_5 model

(Olah et al., 2017)

Future Concepts for Interpretable Machine Learning

Are we able to distinguish highly interpretable features from a suite of simple toy precipitation models?

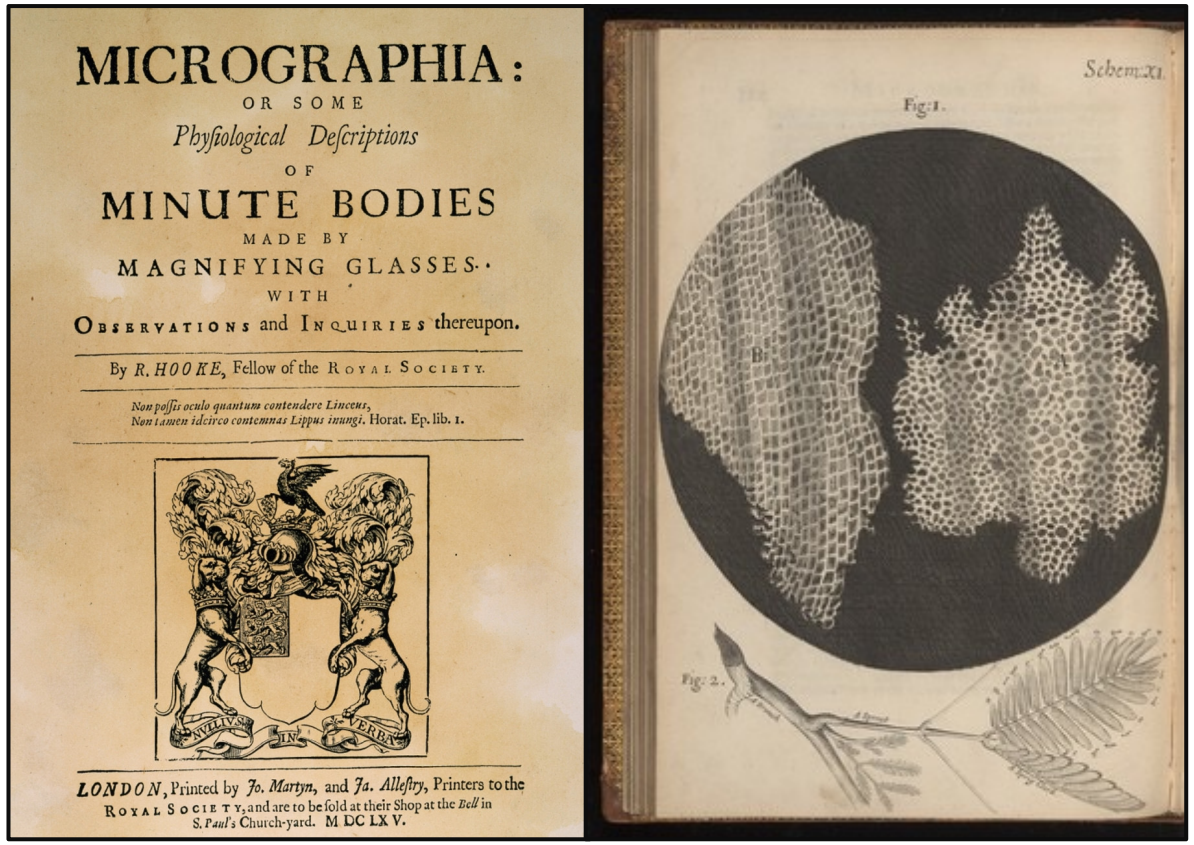
5. Neural Networks are Complex!



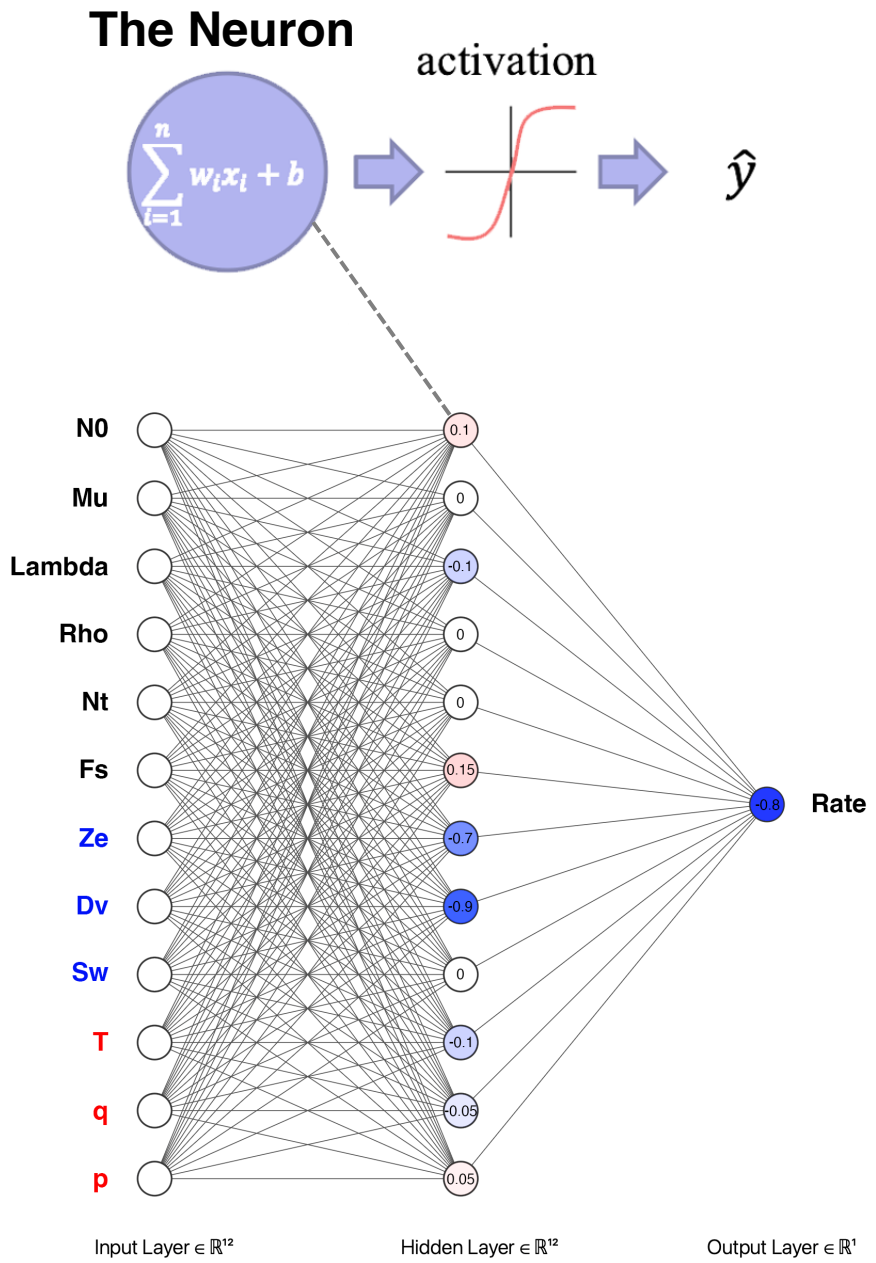
Graphcore®

5. Let's Zoom In!

- Many pivotal moments in the history of Science have been instances where Science "*zoomed in*"

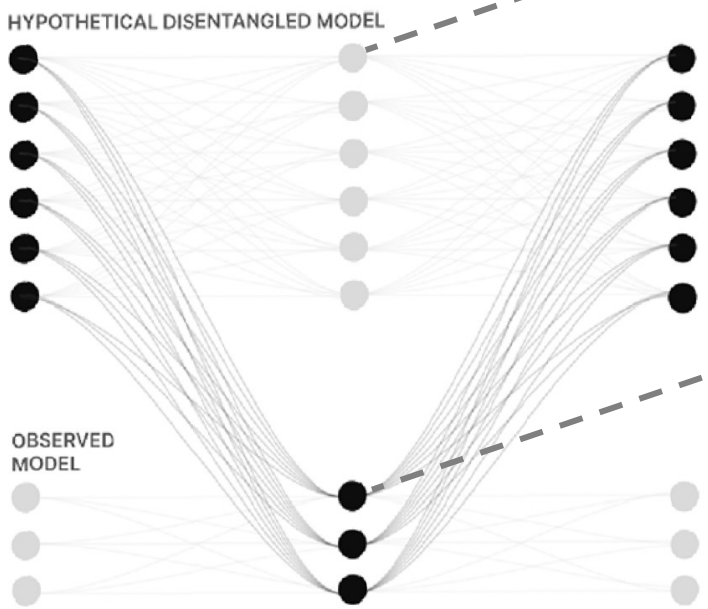


Hooke's Micrographia (1666)



5. Polysemanticity and Superposition

- Many neurons are *polysemantic* in nature and respond to mixtures of seemingly unrelated inputs
- This leads to network *superposition* where a neural network represents more independent "features" of the data than it has neurons



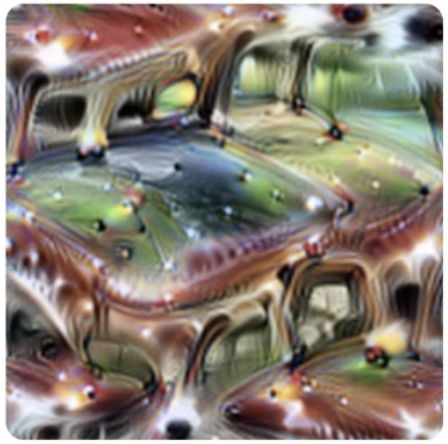
Monosemantic Neuron

Neuron 4b:409

Dataset examples for neuron 4b:409

(Olah et al., 2017)

Polysemantic Neuron

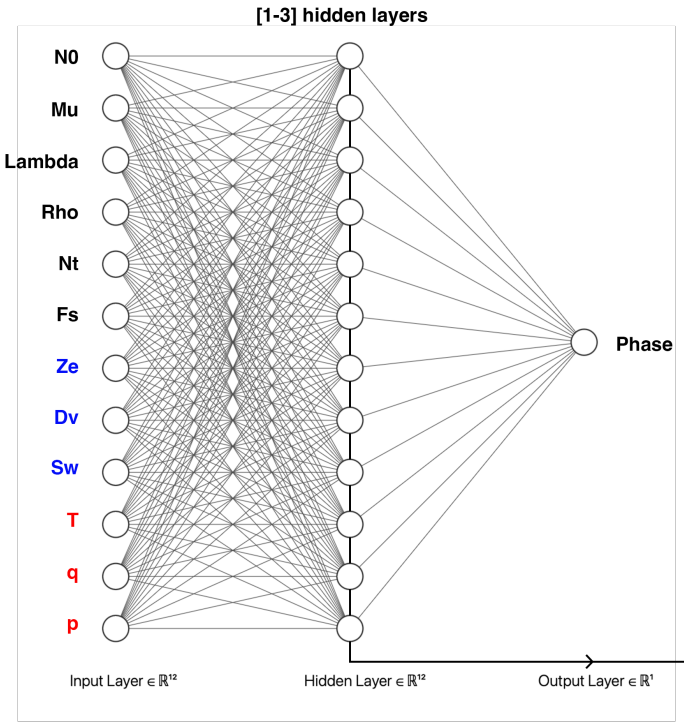


4e:55 is a polysemantic neuron which responds to cat faces, fronts of cars, and cat legs. It was discussed in more depth in Feature Visualization [4].

(Olah et al., 2017)

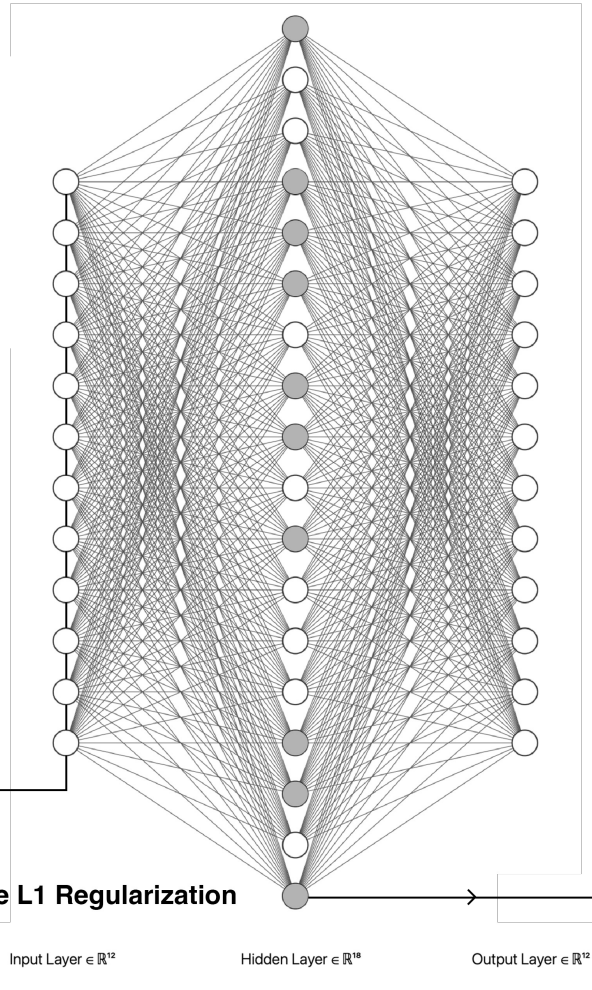
5. Future: Finding Physical Circuits

1. Train MLP classifier



Create database of layer activations

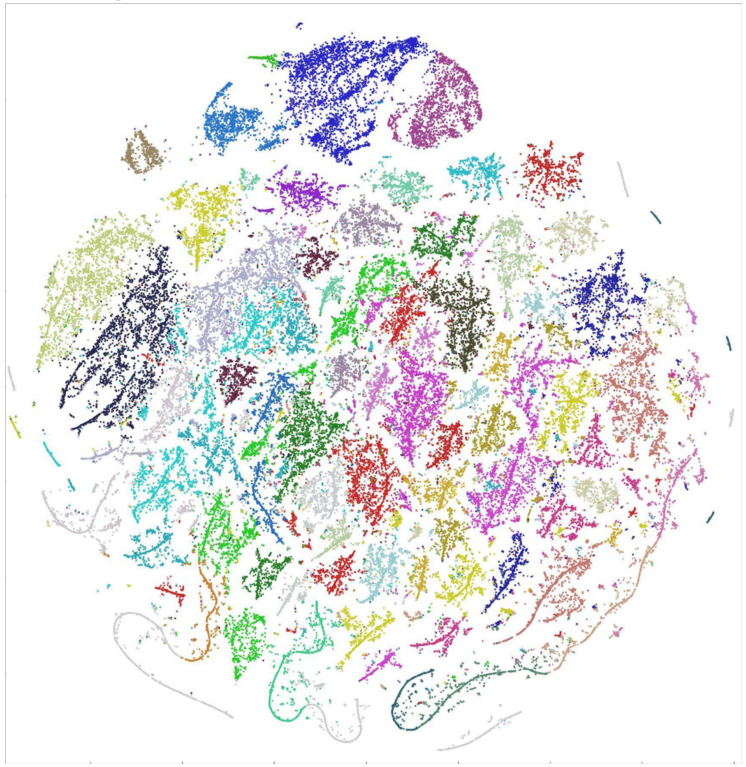
2. Train SAE on MLP activations



Use L1 Regularization

3. Interpret Sparse Dictionary Features

- Case-by-case analysis
- t-SNE/PCA
- Physical circuit identification



4. Enhanced NN interpretability and trust in the Atmospheric Sciences?

5. More Information

https://distill.pub

Distill

Sept. 2, 2021

PEER-REVIEWED

Understanding Convolutions on Graphs

Ameya Daigavane, Balaraman Ravindran, and Gaurav Aggarwal

Understanding the building blocks and design choices of graph neural networks.

Sept. 2, 2021

PEER-REVIEWED

A Gentle Introduction to Graph Neural Networks

Benjamin Sanchez-Lengeling, Emily Reif, Adam Pearce, and Alexander B. Wiltschko

What components are needed for building learning algorithms that leverage the structure and properties of graphs?

https://transformer-circuits.pub

Transformer Circuits Thread

Transformer Circuits Thread

Articles

FEBRUARY 2024

Circuits Updates — February 2024

A collection of small updates from the Anthropic Interpretability Team.

JANUARY 2024

Circuits Updates — January 2024

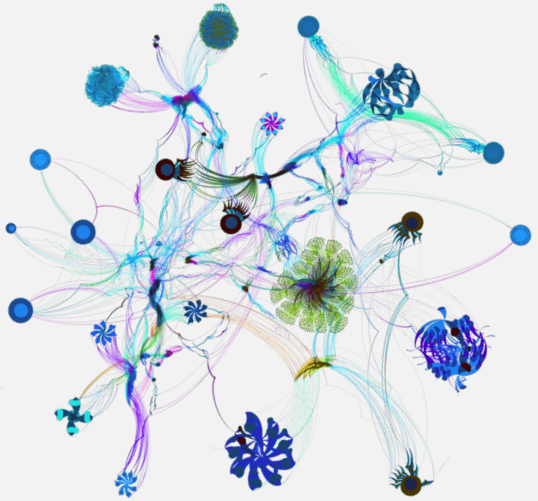
A collection of small updates from the Anthropic Interpretability Team.

Summary

Challenges Remain

Model trust continues to be a big issue in the application of ML in the Atmospheric Sciences

Model Development



We have examined the behavior of multiple ML models related to clouds and precipitation

Interpretability

There are new, exciting methods of mechanistic interpretability being developed and refined

Thank You

Questions?

NASA NIP Grant # 80NSSC22K0789