ECMWF-ESA Workshop on Machine Learning
for Earth System Observation and Prediction

5 Oct 2020

On the Interpretation of Neural Networks Trained for Meteorological Applications

Imme Ebert-Uphoff (iebert@colostate.edu)

Affiliations:

- Cooperative Institute for Research in the Atmosphere (CIRA)
- Electrical and Computer Engineering, Colorado State University

CIRA
Connecting Models and Observations

COLORADO STATE
UNIVERSITY

This presentation is based on this recent paper:

Imme Ebert-Uphoff and Kyle Hilburn,

Evaluation, Tuning and Interpretation of Neural Networks for Working with Images in Meteorological Applications,

Bulletin of the American Meteorological Society (BAMS),

https://doi.org/10.1175/BAMS-D-20-0097.1,

Aug 31, 2020 (early online release).

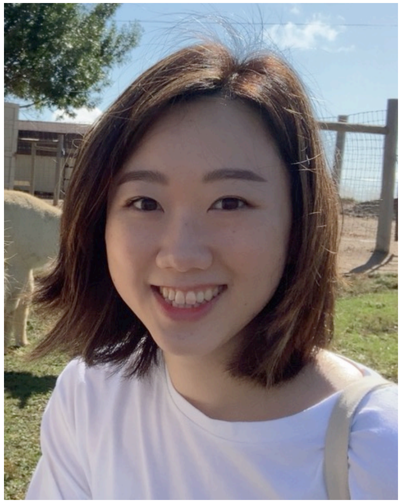
So work reported here is joint work by Kyle and myself.

Kyle Hilburn
CIRA, CSU

Wonderful Collaborators on related topics

LRP for science discovery

Kyle Hilburn
CIRA
Research Associate



Yoonjin Lee
ATS
Ph.D. student
(Kummerow group)

Ben Toms
ATS
Ph.D. student
(Barnes group)

Elizabeth Barnes
ATS
Associate Prof.

COLORADO STATE UNIVERSITY

All at Colorado State University

Motivation

ANNs

- Have emerged as promising tools in countless earth science related applications.
- **Perform amazingly well at many complex tasks.**
- ***If ANNs work fine, why do we care how they work?***

“Clever Hans” Strategies

Clever Hans: German horse in 1907 that was believed to know arithmetic.

- Horse would answer questions by tapping its hoof the right number of times.
- Even the owner thought it knew arithmetic.
- It even answered correctly if strangers asked the question!
- It took a team of scientists to figure out what was going on.
- Turns out: People tend to tense up until correct number of taps completed!
- **So Clever Hans gave the right answer – but for the wrong reason.**
- **Exploited a correlated behavior.**

“Clever Hans” Strategies in Machine Learning

Examples from the following paper
(also source of images on the following slides):

> Lapuschkin, Sebastian, et al. “**Unmasking Clever Hans Predictors and Assessing What Machines Really Learn**.” Nature Communications, vol. 10, no. 1, Mar. 2019, p. 1096, https://doi.org/10.1038/s41467-019-08987-4.

Considered Task:
- Object recognition.
- ML algorithm trained to detect many different objects in images.

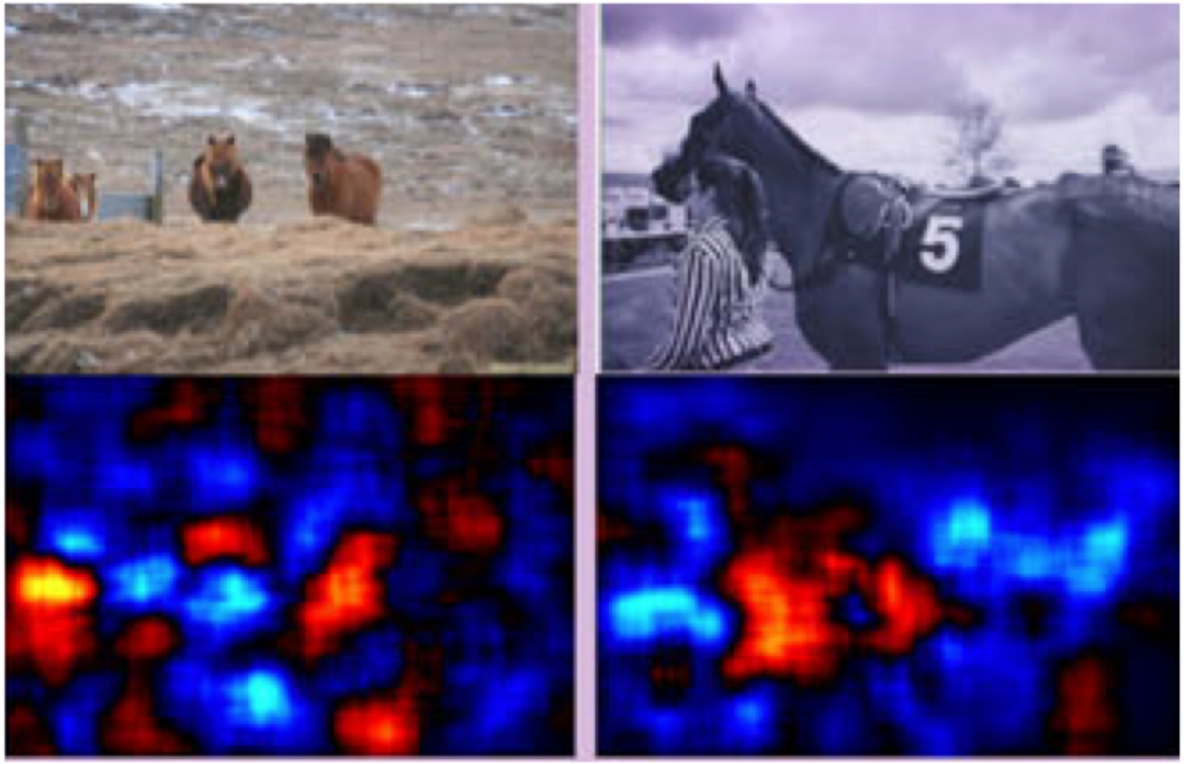
ML method used:
- Neural network (NN)

Specific task analyzed in paper:
- How does NN decide whether there is a *horse* in an image?
- Which strategies does it use to decide?

Method used for analyzing strategies:
- NN visualization technique (LRP) → constructs attribution maps.

Detecting horses – Strategy 1 of algorithm

Input Images



Attribution maps:
In red is where the NN is looking to decide whether there is a horse.

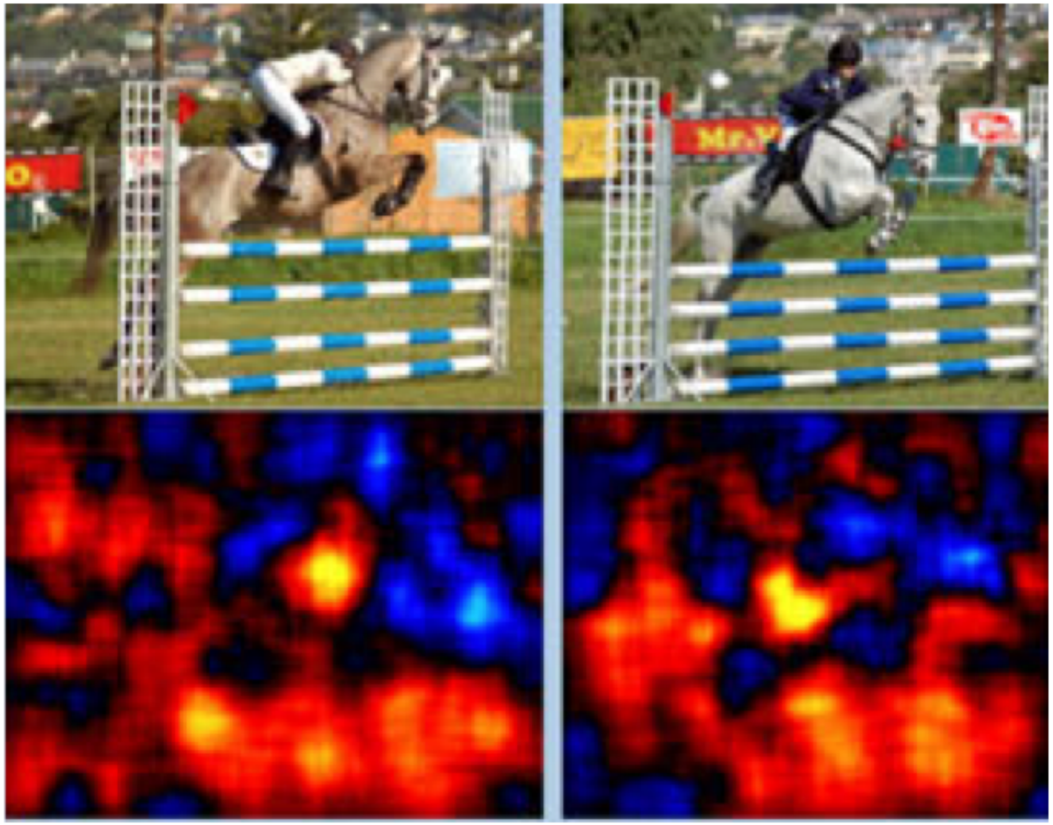
Red areas: increase confidence
Blue areas: decrease confidence
Black areas: not useful

Strategy 1: What does NN detect here?
NN detects mainly parts of horses.
Excellent strategy!

Detecting horses – Strategy 2 of algorithm

Input Images

This is where the NN is looking to decide.



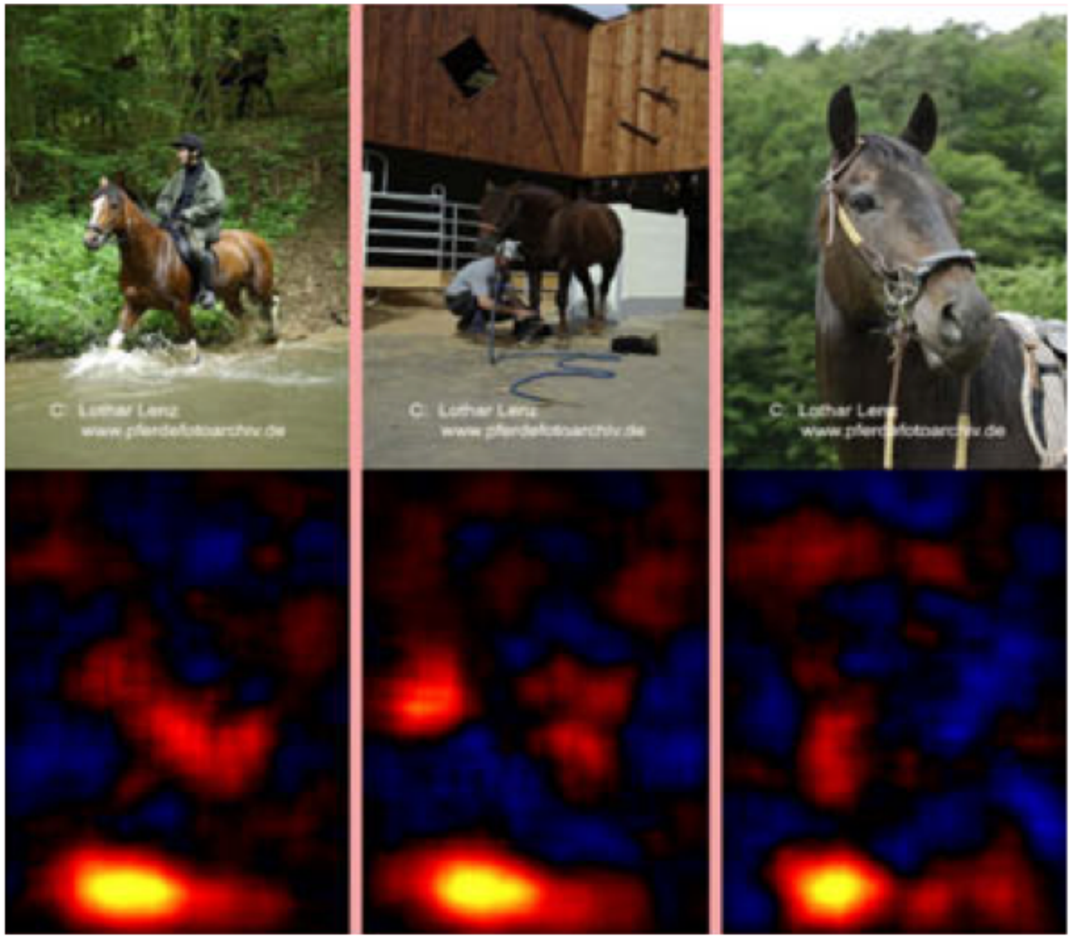
Strategy 2: What does NN detect here?

NN detects the poles – indicative of horses in provided samples.

Faulty reasoning: What if there is pole, but no horse?

Can lead to false positives (false alarms)!

Detecting horses – Strategy 3 of algorithm

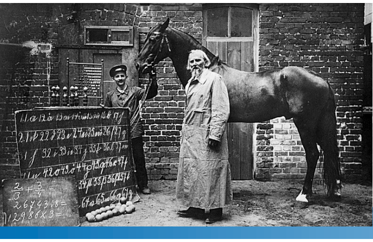


Strategy 3: What does the NN detect in these images?

Html tags on the bottom of the images.

Bad strategy – would not be there in real world.

Likely to lead to false negatives (= misses).



NN learned Clever Hans strategies!

ML algorithm might *also* give the right result, but for the wrong reason!

- Don't blame the algorithm.
- Algorithm did exactly what it was supposed to do, namely, to ***discover and use most helpful correlations/patterns in the data to perform its task.***
- ***Just like Clever Hans – the algorithm exploited correlations in the "training data".***
- It worked for all examples given!

Problem:
- Many correlations are present in data – but not representative of real world.
- If we use those: does not generalize. Faulty reasoning.

Example in meteorology:
- Large hail mainly reported in highly populated areas.
- Should we conclude that large hail only occurs in high population areas?
- Of course not!

Conclusions:
- Using NNs as black box is not a good idea.
- Need to better understand what NN is doing.

NN interpretation

1. **Sample application and corresponding NN**:
   The GREMLIN model

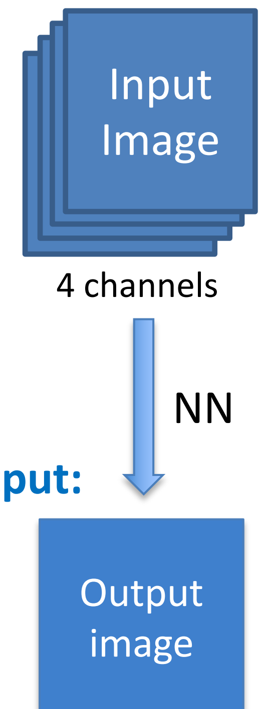
2. **NN interpretation – The tools** – illustrated for GREMLIN

3. **NN interpretation – Sample workflow**

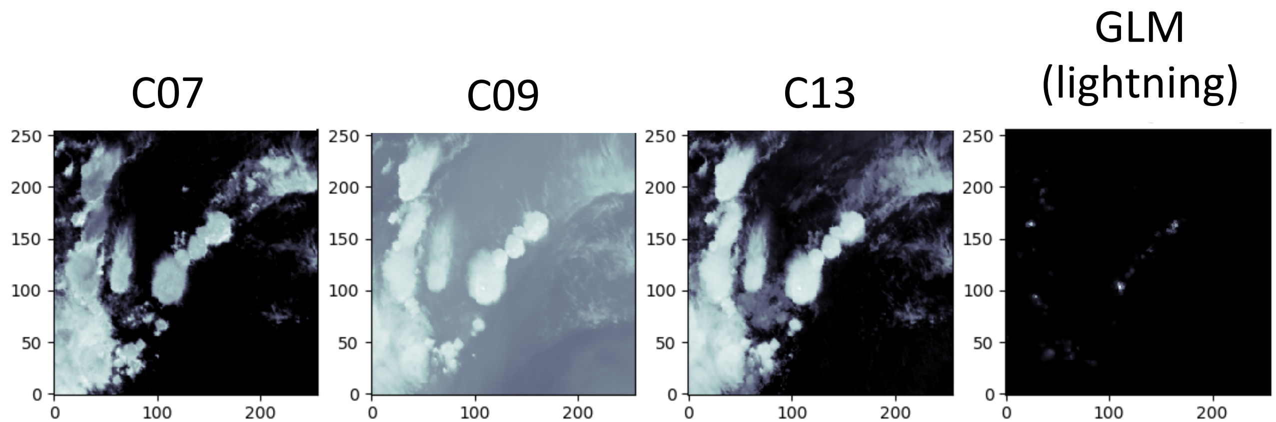
Sample Application:
Generating synthetic radar images from GOES imagery

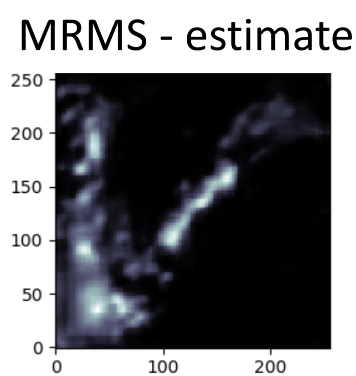
Input: GOES Channels C07, C09, C13, GLM. Output: MRMS (radar).

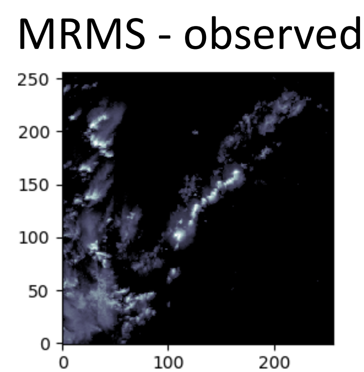
Input:



Output:







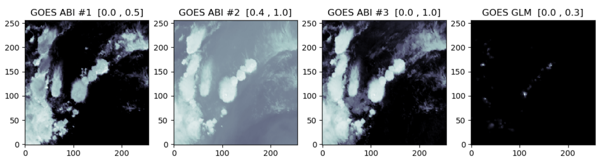
Kyle Hilburn

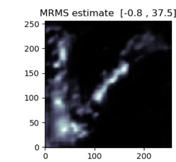
Motivation: GOES imagery is available in all of CONUS, but MRMS is not.

CNN architecture

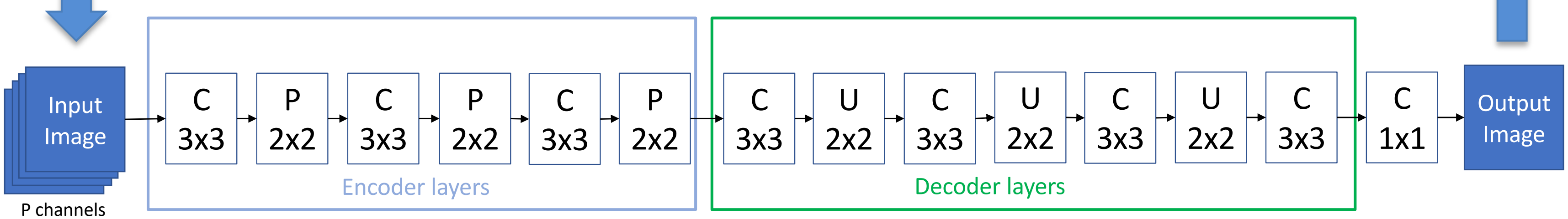
Input: GOES channels

Output: MRMS estimate





Standard encoder-decoder network



C = convolution layer

P = pooling layer (downsampling)

U = upsampling

Numbers: size of filters/masks

Final model is called:

GREMLIN = “GOES Radar Estimation via Machine Learning to Inform NWP”

Typical Situation

- We trained a neural network.
- Reasonably happy with its performance.
- But now we would like to know:

How does the NN do its task?
Which strategies does it use?
Are those strategies reasonable?
Clever Hans strategies or trustworthy ones?

NN interpretation – The Tools

Tools discussed here:

1. Ablation studies
2. Layer-wise relevance propagation (LRP)
3. Using Synthetic inputs

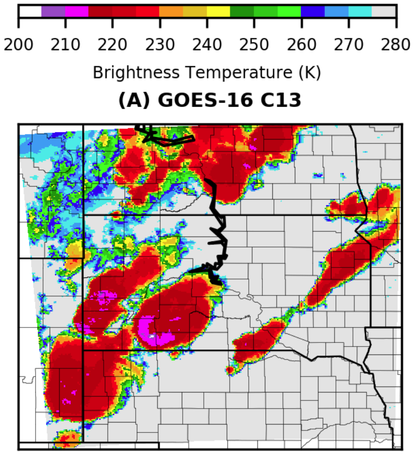
There are many other tools. See for example:

McGovern A, Lagerquist R, Gagne DJ, Jergensen GE, Elmore KL, Homeyer CR, Smith T., Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*. Aug 22, 2019.

Tool 1: Ablation Study

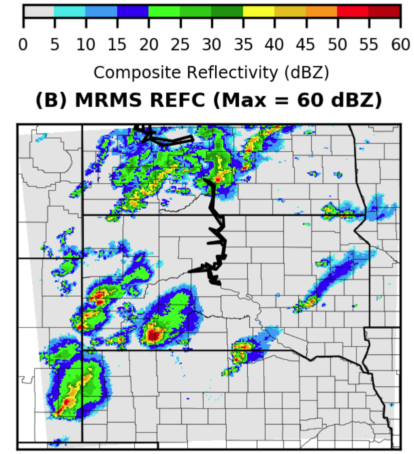
1. Reduce capabilities of NN
2. Retrain simplified NN
3. Analyze:

- What performance do we lose?
- What's still the same?
- What's different?
- So which NN feature is needed for which capability?



Input: GOES Ch. 13

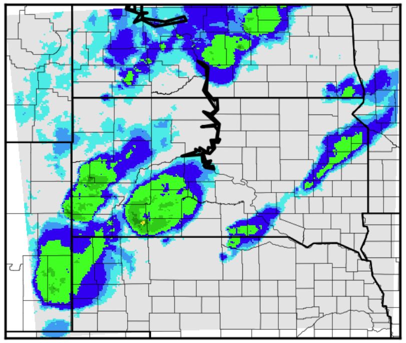
NN



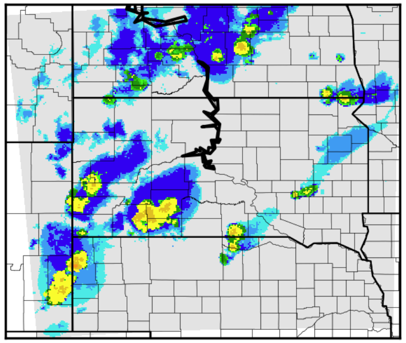
Desired Output (MRMS)

Ablation experiment for our NN (GREMLIN): estimate MRMS using simplified NNs

Convolution mask: cut down to (1x1)
→ No spatial patterns used by NN

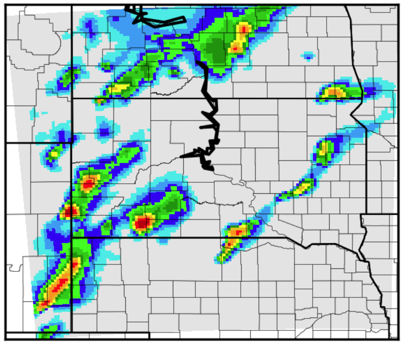


Input: only Ch. 13

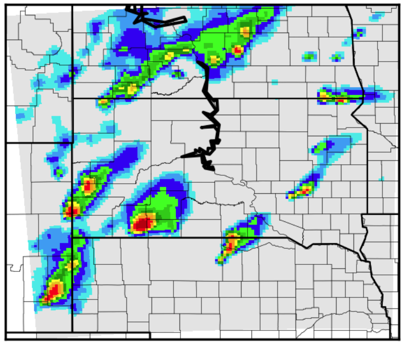


Input: all 4 channels

Convolution mask: regular (3x3)
→ Spatial patterns are used by NN



Input: only Ch. 13



Input: all 4 channels

Simplest model → Increasing model complexity → Full model

Tool 2: Layer-Wise Relevance Propagation (LRP)

1. Pick a sample
2. Use LRP to find out:
   Where in the input is the NN focusing to come up with its answer?

LRP provides:

Attribution maps – just like for the horses earlier.

Color code here:

Red = where the NN focuses.

White = NN thinks there's no relevant information

LRP experiment for our NN (GREMLIN)

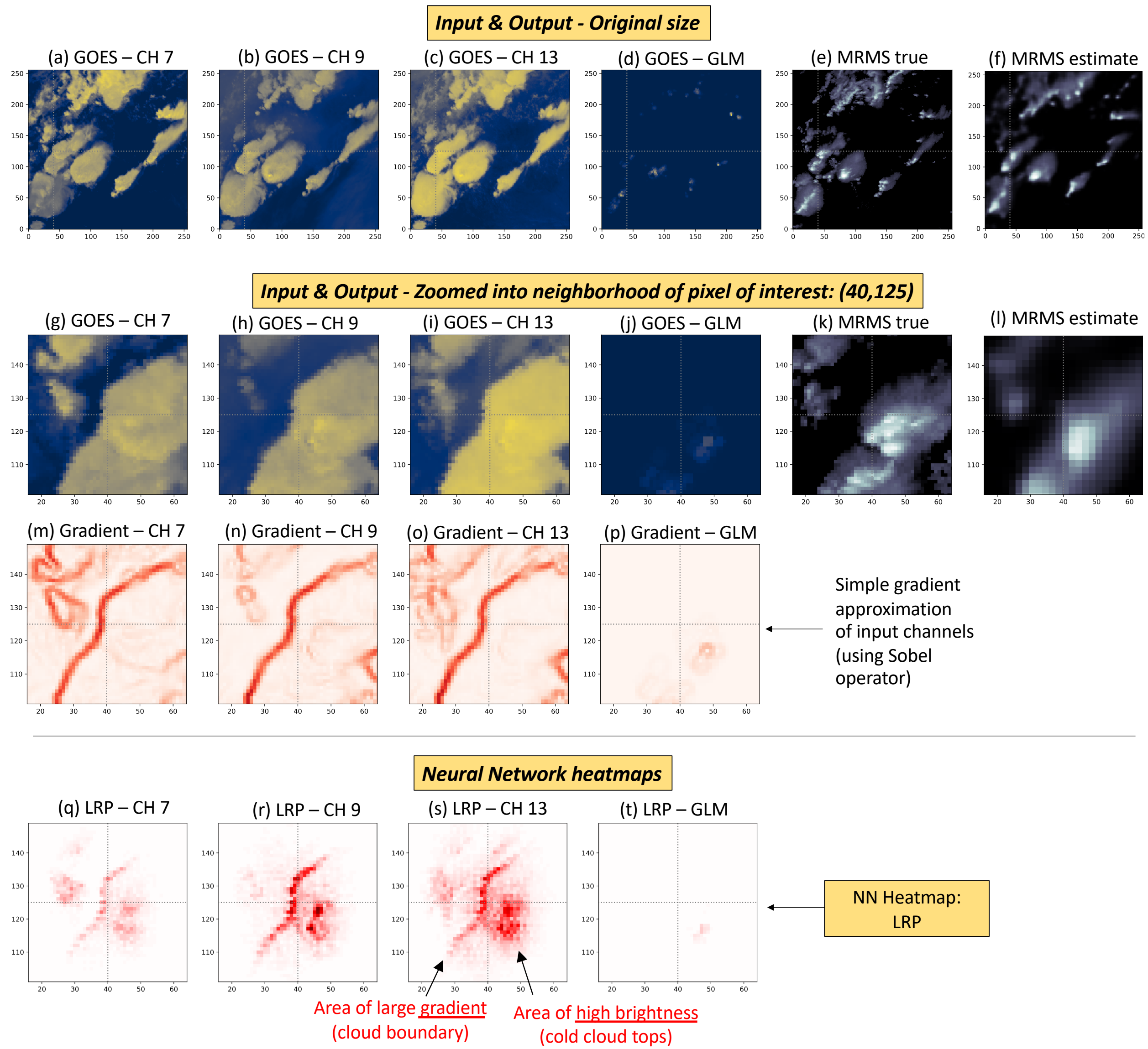
Question:

How does NN know when to create **large** MRMS estimates?

Method:

1. Select samples where MRMS estimate is high.

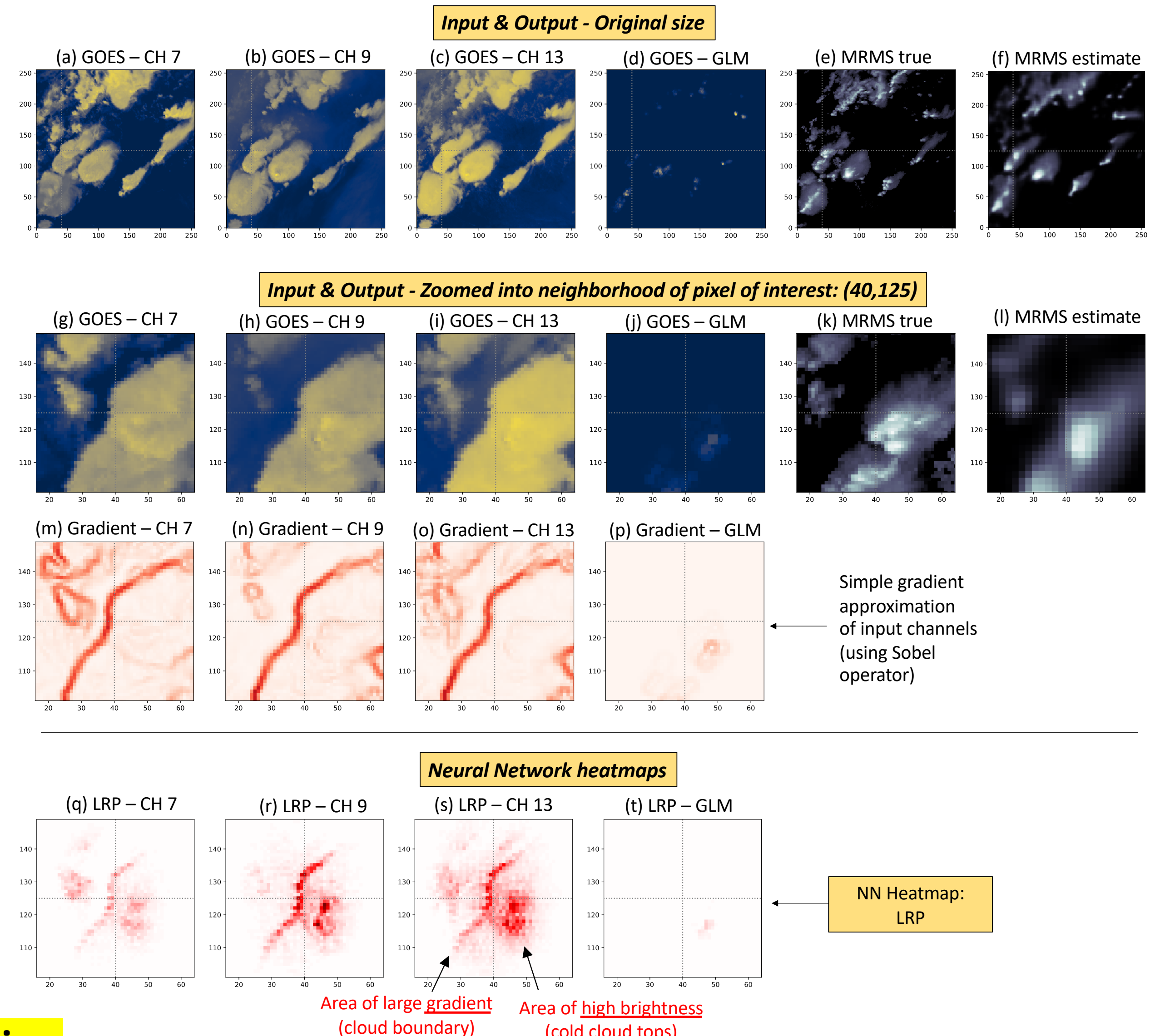
2. Where is NN looking in input?



LRP experiment for our NN (GREMLIN)

Question:
How does NN know when to create **large** MRMS estimates?

Method:
1. Select samples where MRMS estimate is high.
2. Where is NN looking in input?



LRP found three strategies:

NN creates large MRMS values only when it encounters:

1) Strong lightning (biggest trigger), 2) Cold cloud tops, or 3) Cloud boundaries.

Side note -
For those of you who know **saliency maps**:

Do saliency maps give similar results to LRP here?

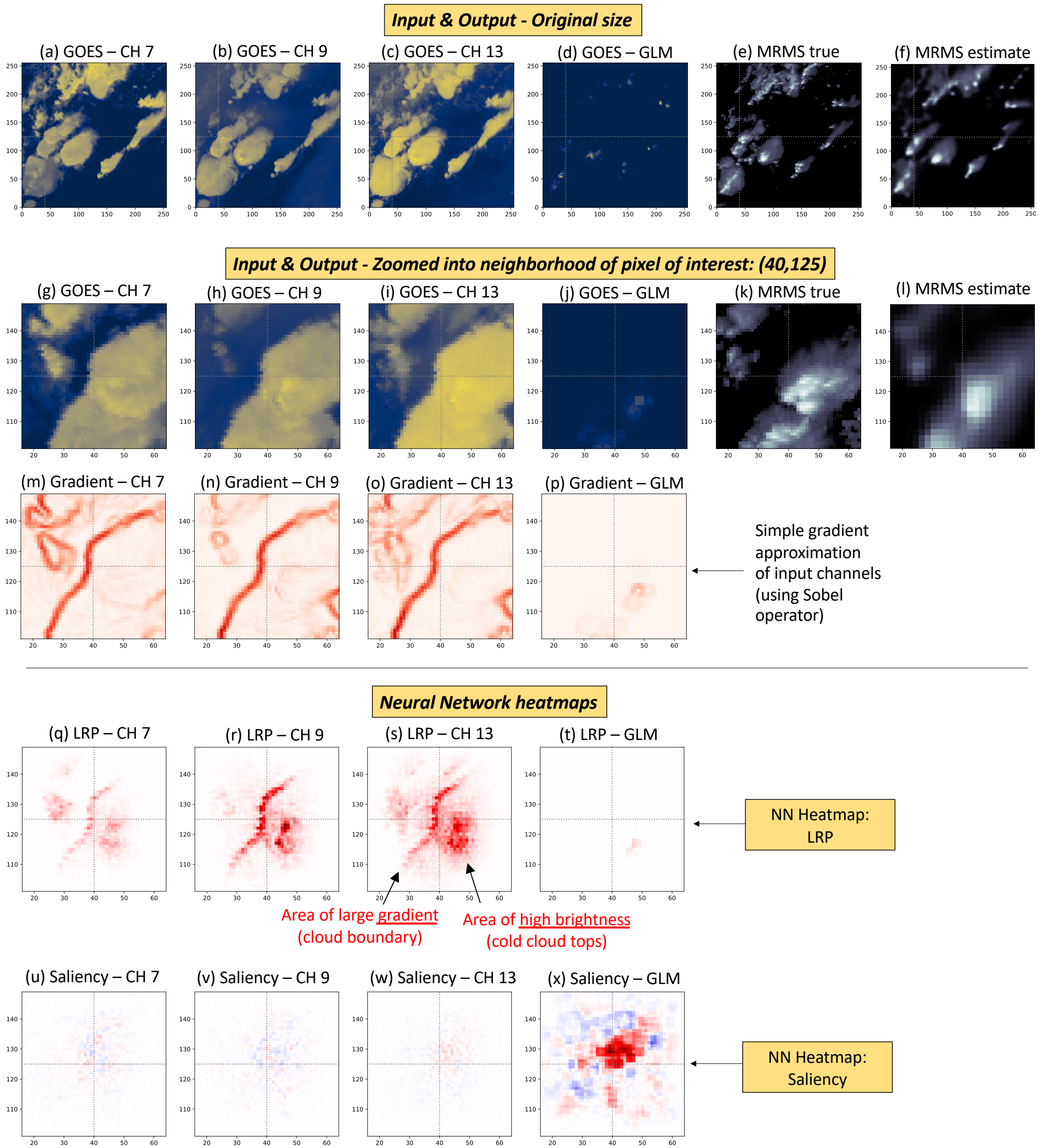
Answer: No!

LRP found three strategies for high MRMS values:

1) Strong lightning.
2) Cold cloud tops.
3) Cloud boundaries.

Saliency only found strongest strategy:

1) Strong lightning.



But: saliency maps can be applied to any NN, while LRP only implemented so far for simpler ones.

Tool 3: Synthetic Inputs

1. Create synthetic inputs that represent different meteorological scenarios – use parameters to be able to specify different scenarios.
2. Conduct experiments with NN:
    a) Feed in inputs for specific meteorologic conditions – how does NN respond?
    b) Find inputs that generate minimal / maximal response.

Why create synthetic inputs?

1. **Controlled experiments:**
   Only desired meteorological scenario present in image - isolated.
2. **Can generate and test an unlimited number and type of scenarios:**
   Even scenarios not included in training samples.

Synthetic Inputs for GREMLIN

Kyle Hilburn

Synthetic experiments designed by Kyle Hilburn

- **Sum of Generalized Elliptical Gaussians (GEG). See equations below.**
- **Use as synthetic data for GOES Ch. 13.**
- Can choose parameters to generate different meteorological scenarios.
- Feed into GREMLIN.
- Observe behavior.

We are using a sum of Generalized Elliptical Gaussians (GEG) model with an outer Gaussian G_o that represents the thunderstorm anvil and an inner Gaussian G_i that represents the overshooting top. The synthetic brightness temperature T is a function of (x,y) with the parameters: location x_0 and y_0, amplitude A, size S, aspect α, orientation θ, and sharpness (exponent) p for the outer and inner Gaussians, denoted with subscripts o and i:

$$\hat{x}_{o,i} = \left(x - x_{0,o,i}\right)\cos\theta_{o,i} - \left(y - y_{0,o,i}\right)\sin\theta_{o,i} \tag{3a}$$

$$\hat{y}_{o,i} = \left(x - x_{0,o,i}\right)\sin\theta_{o,i} + \left(y - y_{0,o,i}\right)\cos\theta_{o,i} \tag{3b}$$

$$T_{o,i} = exp\left(-1\left(\frac{\hat{x}_{o,i}^2}{2S_{o,i}^2} + \frac{\hat{y}_{o,i}^2}{2\left(S_{o,i}\alpha_{o,i}\right)^2}\right)^{p_{o,i}}\right) \tag{3c}$$

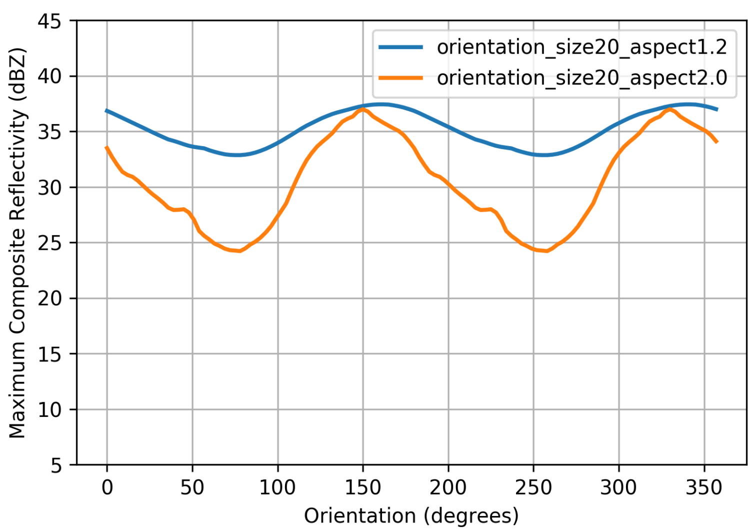
$$T = A_o T_o + A_i T_i \tag{3d}$$

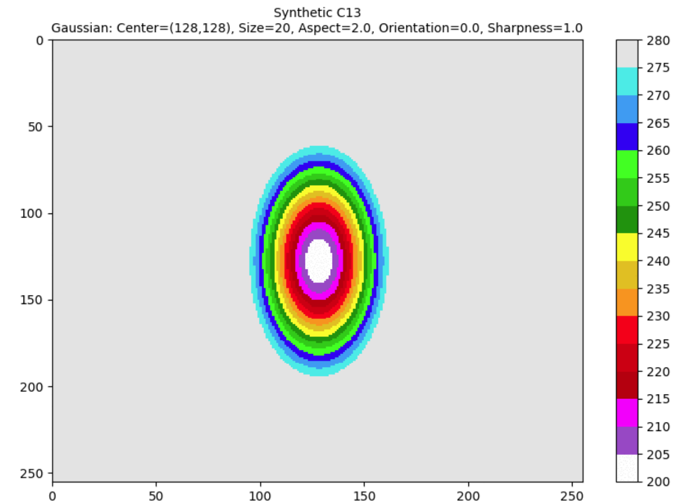
Kyle Hilburn

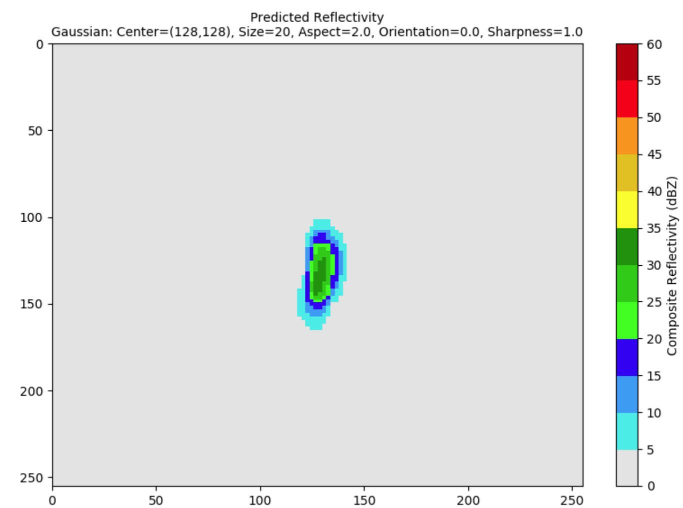
PowerPoint version of this slide
includes animation here:
varying orientation & NN output

Gaussian: Varying Orientation

- Loop for size=20, aspect=2.0
- Maximum response near 150 or 330 deg
- Sinusoidal behavior makes sense, and maximum response angles must be statistically related to wind shear in training dataset?







NN interpretation – The Workflow

- Typical workflow: next slide

- Key point:
  Workflow only possible with close collaboration of ML expert and environmental scientist.

Sample Workflow - Interpretation using Subsets of Input Samples

Interpretation method here: Choose and analyze a set of inputs.
Example: Synthetic inputs (as seen before).

Sample strategies to select a set of input samples:

a. Biggest successes and failures
b. Grouping by true class/value
c. Grouping by single meteorological property
d. Clustering of Input Samples
e. Modifying Input Samples
f. Creating synthetic input samples

Sample Analysis Step

1) Meteorologist selects a set of input samples to focus on.

Steps 2-3: NN evaluation for set of samples (incl. visualization)

2) Run set of samples through NN.

3) For set of samples look at
- Corresponding outputs of NN
- Results from NN visualization tools (e.g., LRP)
- Performance measures – for entire set and outliers.

Steps 4-6: Hypothesis generation and testing, experimental design

4) Meteorological expert interprets results from Step 3.

5) Based on observations in Step 4:
Generate hypothesis of NN strategy.

6) If needed: Design follow-up experiment to test hypothesis.

Experimental design

Sample Workflow – Key Observations

Key observations:

1. **NN interpretation = process of hypothesis generation and testing.**

2. **NN interpretation tools are *just useful tools* in that process.**

3. **Environmental scientist is crucial in every step!**

4. More generally:
   We need **close collaboration** between ML expert and environmental scientist for every step of ML
   a) algorithm development,
   b) evaluation,
   c) tuning,
   d) interpretation.

New NSF-sponsored AI Institute: AI2ES

Two names:

1) NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography.

2) AI for Environmental Science (AI2ES)
   See https://www.ai2es.org/

- NSF award: $20M. Award period: 2020-2025.
- Lead: Amy McGovern @ Univ. of Oklahoma.

Collaborating Institutions and Partners (founding members):

Academic partners:
- Univ. of Oklahoma
- Texas A&M – Corpus Christi
- Colorado State Univ.
- North Carolina State Univ.
- Univ. at Albany
- Univ. of Washington
- Del Mar College

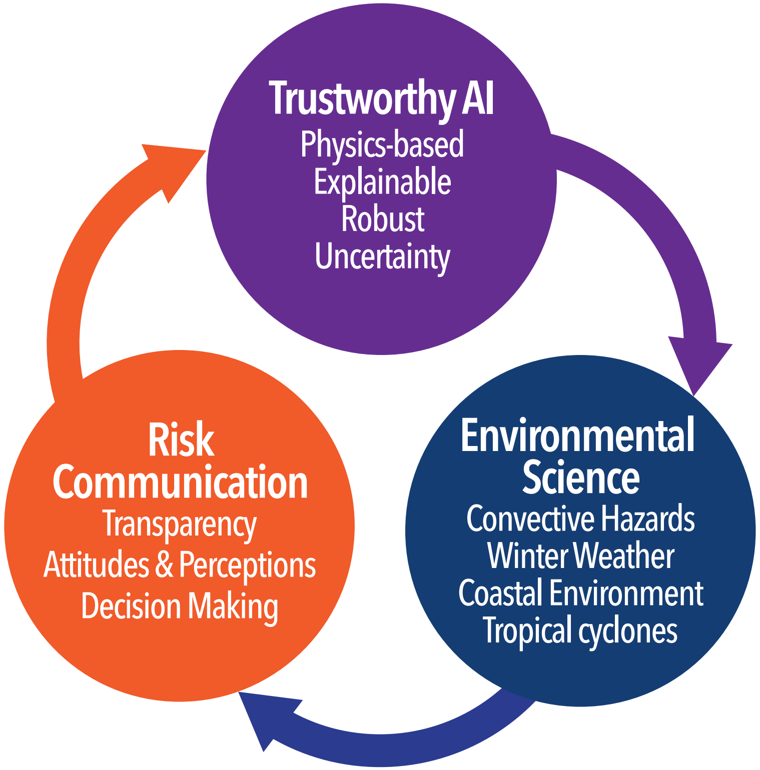
Federally funded research lab:
- NCAR

Private industry partners:
- Google
- IBM Weather
- NVIDIA
- Disaster Tech

Federal partner:
- NOAA

New NSF-sponsored AI Institute: AI2ES

- We are part of AI2ES (www.ai2es.org)
- **NN interpretation tools will be integrated into the AI institute.**
- **We will work with social scientists to identify what types of explanations are meaningful to end users (such as forecasters, public, etc.)**



We will be looking for post-docs and graduate students for the AI institute soon – at CSU and elsewhere!

More details on our NN interpretation work can be found here:

Imme Ebert-Uphoff and Kyle Hilburn,
Evaluation, Tuning and Interpretation of Neural Networks for Working with Images in Meteorological Applications,
Bulletin of the American Meteorological Society (BAMS),
Aug 31, 2020 (early online release).

Thank you!

Questions or Suggestions?

CIRA
Connecting Models and Observations