Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School

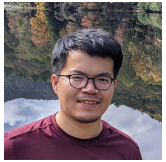
Day 4: Speakers

Amy McGovern (OU)

Imme Ebert-Uphoff (CSU)

Marie McGraw (CSU)

Ryan Lagerquist (CSU)



Douglas Rao (NOAA)

Ann Bostrom (UW)

Christopher Wirz (NCAR)

Not speaking today, but contributed slides and notebooks:

Katherine Haynes (CSU)

NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

LEAP

NOAA CENTER FOR ARTIFICIAL INTELLIGENCE

Radiant Earth Foundation EARTH IMAGERY FOR IMPACT

Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School

Day 4: Goals

- Learn about uncertainty lifecycle in environmental sciences and AI development
- Learn about common methods for uncertainty quantification and metrics to evaluate uncertainty
- Learn about different strategies for communicating uncertainty to different audiences

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH
LEAP
NOAA CENTER FOR ARTIFICIAL INTELLIGENCE
Radiant Earth Foundation
EARTH IMAGERY FOR IMPACT

Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School

Day 4: Agenda

- **9:00 Uncertainty quantification methods (Part 1)**
- 10:00 *Short brain & bio break*
- 10:10 Uncertainty quantification methods (Part 2)
- 10:45 *Short brain & bio break*
- 10:55 Communicating uncertainty (Part 3)
- 11:55 Lecture series wrap up!

Questions?

https://app.sli.do/event/1zumy91n

Or go to sli.do and use the code TAI4ES

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

LEAP

NOAA CENTER FOR ARTIFICIAL INTELLIGENCE

Radiant Earth Foundation
EARTH IMAGERY FOR IMPACT

Part 1: UQ in ML

Warm-up and refresher from yesterday

Let's do couple quick questions to get us back in the trustworthy AI mindset:

1. In your own words, tells us one thing you learned about selecting case studies yesterday

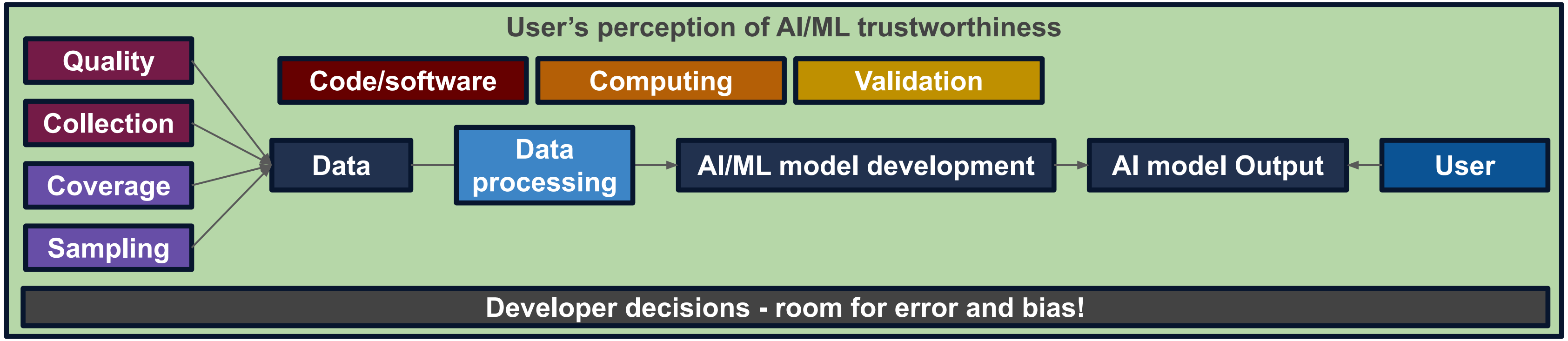
2. What was your favorite part of yesterday's lectures?

4.1. & 4.2. Go to sli.do and use the code TAI4ES

Opening discussion - building on yesterday's lecture

Where does uncertainty come into play? What types of uncertainty?

What are the potential implications of this uncertainty?



4.3. & 4.4. Go to sli.do and use the code TAI4ES

Overview of Part 1 & 2

1) Motivation and Examples

2) Classifications of Uncertainty

3) Evaluation criteria for uncertainty estimates.
   To answer the question: *What makes a good uncertainty estimate?*

BREAK

4) Selected methods for UQ estimation in ML algorithms.
   To answer the question: *How do you quantify uncertainty in ML?*

Motivation

Why do we want uncertainty estimates for our ML models?

1. **We want to know how much we can trust the ML method's answer for a specific sample.**
2. Side effect: An ML method that knows about its own uncertainty often gives better predictions, too!
3. We want to know whether we can improve the model (see types of uncertainty later - aleatory vs. epistemic) or whether we're dealing with internal variability that cannot be reduced.

Motivation

Simple example from

Chang, D.T. **Bayesian Neural Networks: Essentials.** arXiv preprint, v1, June 2021, https://arxiv.org/abs/2106.13594

Consider simple regression task with scalar output, i.e. predict scalar, y.

Traditional (deterministic) NN may yield as sample output for 10 samples:

```
Predicted: 5.8 - Actual: 6.0
Predicted: 5.7 - Actual: 5.0
Predicted: 5.9 - Actual: 6.0
Predicted: 6.3 - Actual: 6.0
Predicted: 6.3 - Actual: 8.0
Predicted: 5.8 - Actual: 5.0
Predicted: 4.9 - Actual: 6.0
Predicted: 5.1 - Actual: 5.0
Predicted: 6.4 - Actual: 6.0
Predicted: 5.8 - Actual: 5.0
```

1) Traditional (deterministic) NN

```
Predicted: 5.8 - Actual: 6.0
Predicted: 5.7 - Actual: 5.0
Predicted: 5.9 - Actual: 6.0
Predicted: 6.3 - Actual: 6.0
Predicted: 6.3 - Actual: 8.0
Predicted: 5.8 - Actual: 5.0
Predicted: 4.9 - Actual: 6.0
Predicted: 5.1 - Actual: 5.0
Predicted: 6.4 - Actual: 6.0
Predicted: 5.8 - Actual: 5.0
```

```
Prediction mean: 5.96, stddev: 0.69, 95% CI: [7.32 - 4.6]  - Actual: 6.0
Prediction mean: 5.83, stddev: 0.71, 95% CI: [7.24 - 4.43] - Actual: 5.0
Prediction mean: 5.81, stddev: 0.7,  95% CI: [7.17 - 4.44] - Actual: 6.0
Prediction mean: 6.14, stddev: 0.74, 95% CI: [7.59 - 4.69] - Actual: 6.0
Prediction mean: 6.81, stddev: 0.74, 95% CI: [8.26 - 5.35] - Actual: 8.0
Prediction mean: 5.46, stddev: 0.72, 95% CI: [6.86 - 4.05] - Actual: 5.0
Prediction mean: 5.4,  stddev: 0.72, 95% CI: [6.81 - 4.0]  - Actual: 6.0
Prediction mean: 5.12, stddev: 0.73, 95% CI: [6.56 - 3.69] - Actual: 5.0
Prediction mean: 6.75, stddev: 0.74, 95% CI: [8.19 - 5.3]  - Actual: 6.0
Prediction mean: 5.5,  stddev: 0.73, 95% CI: [6.93 - 4.07] - Actual: 5.0
```

2) Probabilistic NN here yields **mu** and **sigma** → can calculate 95% confidence interval.

- We get sigma value with each estimate. Tells us about confidence of NN prediction for that sample.
- One test for sigma: We can check how often actual value is within 95% CI.
  (But NN could cheat - just make **sigma** really large and actual value will always be in 95% CI.
  So this test is not enough to evaluate quality of estimates - see "sharpness criterion" later!)
- Also note: The actual estimates (prediction mean) have changed, too.
  Estimate itself often *better* in probabilistic ML than in deterministic ML. But not always.

Uncertainty Propagation – A Satellite CDR Example

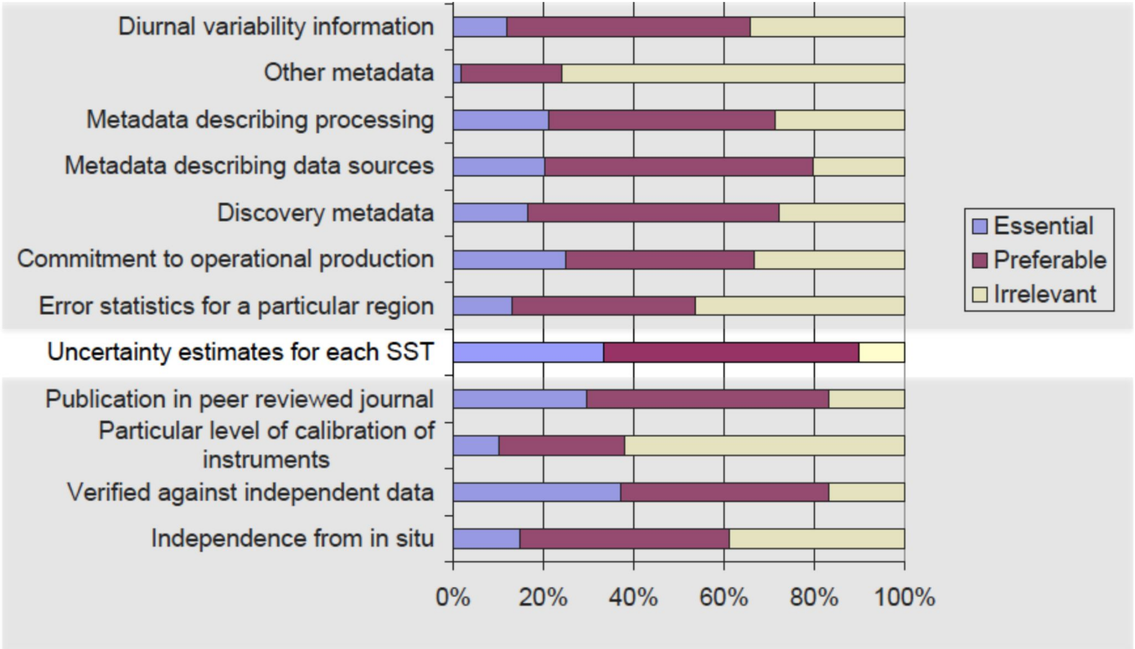
International requirements for different applications for sea surface temperature (SST) climate data records (CDR) to ensure reliable climate monitoring and prediction.

Source	Application	Uncertainty (K)			Horizontal res. (km)			Observing cycle (h)			Timeliness (h unless otherwise stated)		
		Goal	B/T	T/H	Goal	B/T	T/H	Goal	B/T	T/H	Goal	B/T	T/H
AOPC	Climate-AOPC	0.25	0.4	1	10	50	500	3	6	24	3	6	12
WCRP	CLIVAR	0.1	0.2	0.3	10	20	50	3	4	6	24	36	3 d
WCRP	Climate modelling research	0.5	1	2	50	100	250	1	3	12	30d	45d	60d
J Eyre	Global NWP	0.3	0.5	1	5	15	250	3	24	120	3	24	5 d
JF Mahfouf	High res. NWP	0.3	0.5	1	1	5	20	1	2	6	0.5	1	6
P Ambrosetti	Nowcasting / Very short range forecasting	0.5	0.8	2	5	10	50	3	6	24	3	6	24
JCOMM	Ocean applications	0.1	0.5	1	10	25	100	1	3	24	5m	1	6
JCOMM	Ocean applications	0.1	0.5	1	1	10	25	1	3	24	5m	1	6
JCOMM	Ocean applications	0.1	0.2	0.5	10	50	100	1	3	24	12	24	3 d
JCOMM	Ocean applications	0.1	0.2	0.5	5	10	25	6	24	72	1	2	3
JCOMM	Ocean applications	0.1	0.5	1	0.5	1	10	0.5	1	3	0.5	1	6
JCOMM	Ocean applications	0.1	0.2	0.5	1	5	10	3	12	24	1	2	3
OOPC	Climate-OOPC	0.1	0.126	0.2	1	8	500	1	3	24	3	5	12
L Ferranti	Seasonal and inter-annual forecasts	0.1	0.2	0.5	50	85.5	250	3	6	12	3	6	24
WCRP	CLIC	0.5	0.8	2	25	39.7	100	24	30	48	30 d	38 d	60 d

Uncertainty Propagation – A Satellite CDR Example

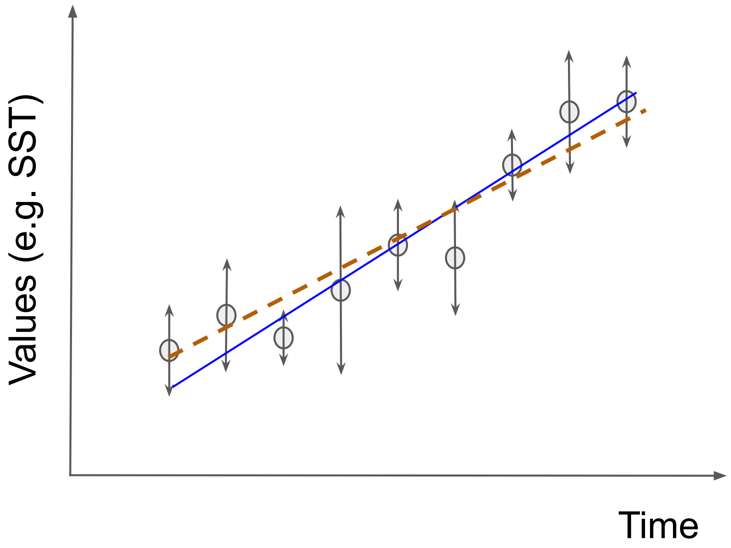
ESA's SST Climate Change Initiative User Requirements report:

More than 90% of surveyed users prefer to have uncertainty estimates with SST for their applications.



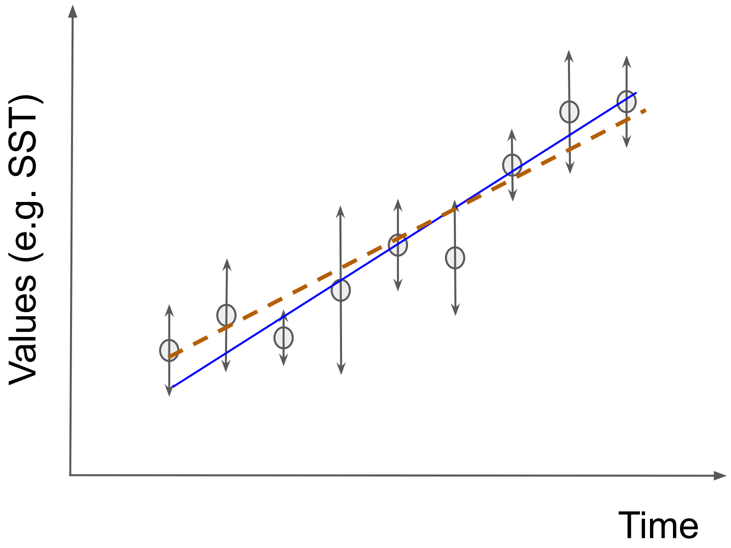
Credit: C. Merchant

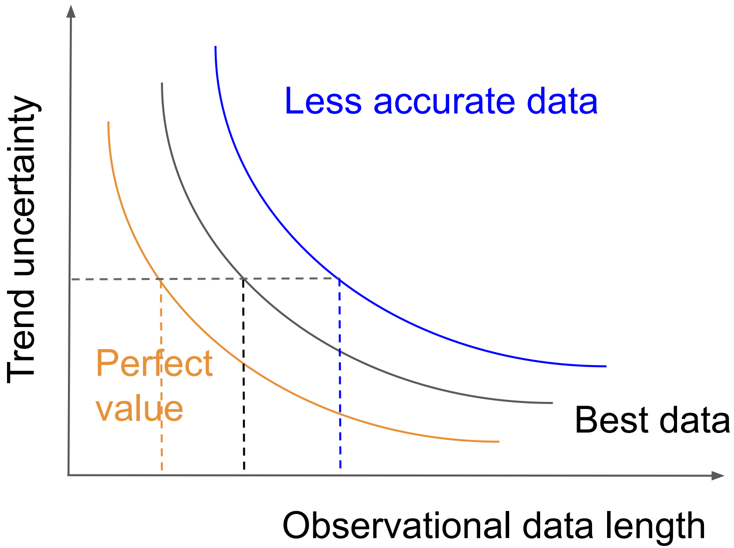
Impact of Data Uncertainty in Climate Monitoring



In climate analysis, we often rely on the time series for trend quantification, which is affected by the uncertainty & quality of the data (thinking about the noise/signal ratio).

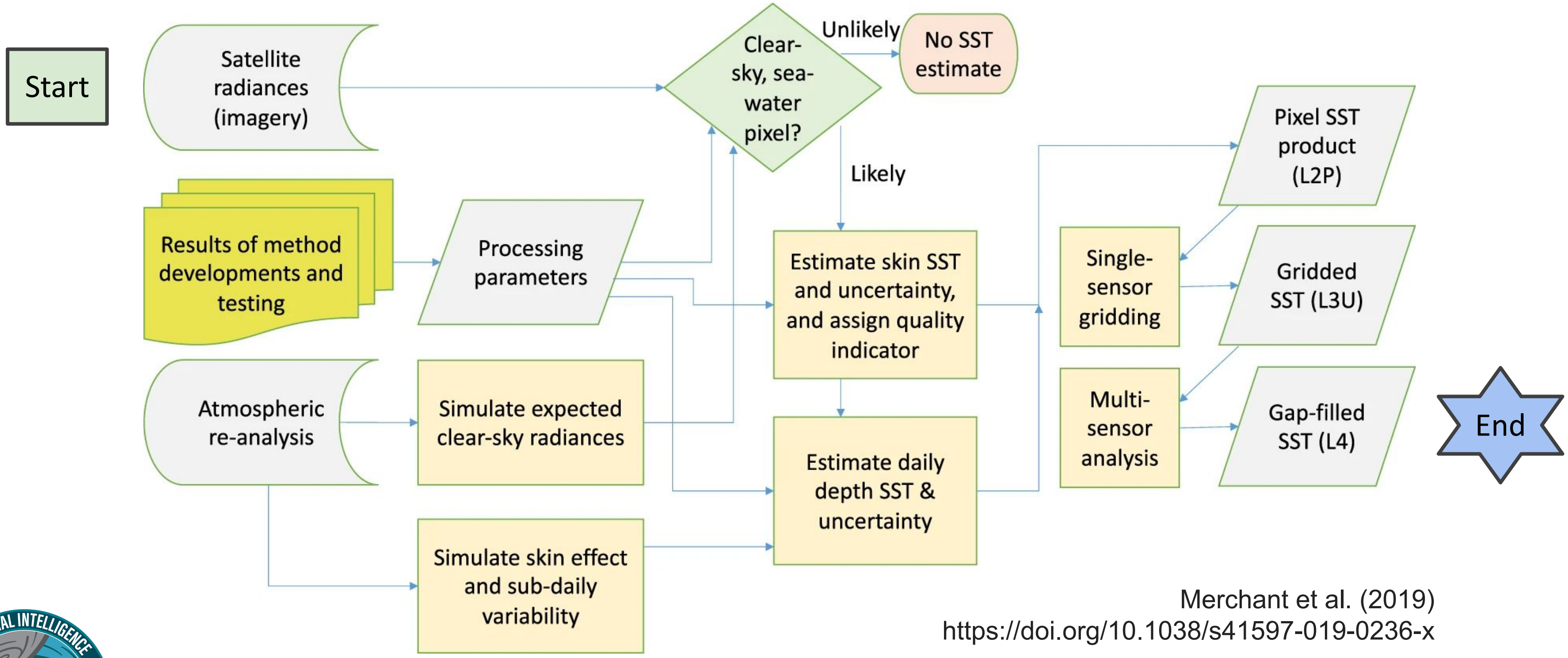
Impact of Data Uncertainty in Climate Monitoring





In climate analysis, we often rely on the time series for trend quantification, which is affected by the uncertainty & quality of the data (thinking about the noise/signal ratio).

A practical satellite product workflow



Merchant et al. (2019)
https://doi.org/10.1038/s41597-019-0236-x

Satellite data retrieval requires a series of transformation from raw signal to physical observations to usable products.

Uncertainty Propagation – A Satellite CDR Example

Level 0 (raw data)	Level 1b (radiance)	Level 2 (granular)	Level 3 (gridded)	Level 4+ (gap-filled)
Data digitisation; Sensor noise; Instrument failure; …	Sensor calibration; Geolocation error; …	Retrieval algorithm accuracy; Definition of the geophysical variables (e.g., skin temperature v.s. temperature at a depth); Dependency data …	Spatial-temporal sampling; Locally-correlated errors; …	Extrapolation / interpolation; Post-processing; …

Errors in the processing workflow will propagate through the workflow and sometimes being amplified / mitigated into the desired products/information. ML uncertainty typically corresponds to the uncertainty in Level 2 in this workflow.

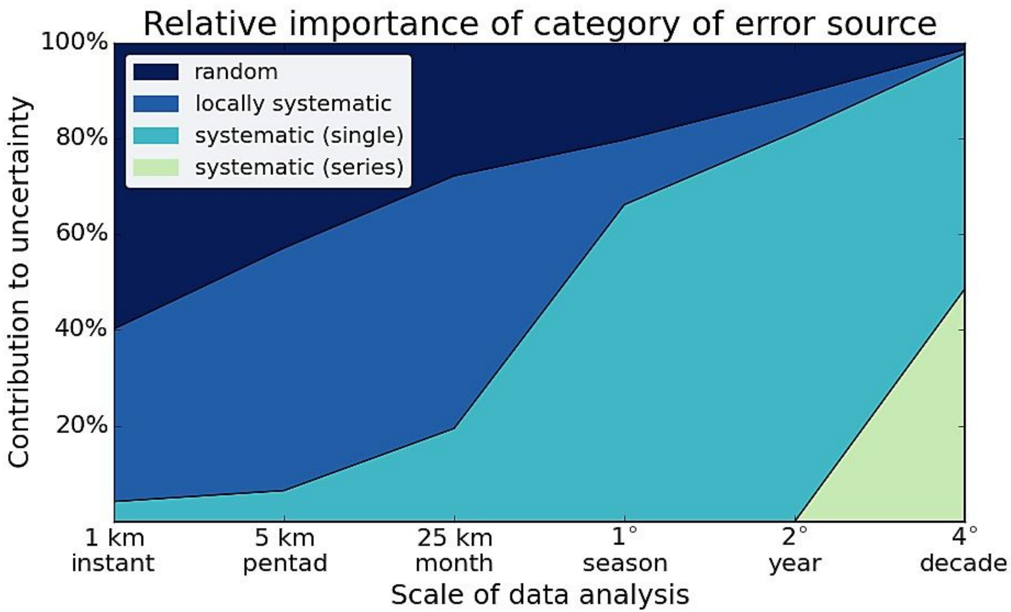
Uncertainty Propagation – A Satellite CDR Example

Level 0 (raw data)	Level 1b (radiance)	Level 2 (granular)	Level 3 (gridded)	Level 4+ (gap-filled)
Data digitisation; Sensor noise; Instrument failure; …	Sensor calibration; Geolocation error; …	Retrieval algorithm accuracy; Definition of the geophysical variables (e.g., skin temperature v.s. temperature at a depth); Dependency data …	Spatial-temporal sampling; Locally-correlated errors; …	Extrapolation / interpolation; Post-processing; …

Most common stages for ML use cases

Errors in the processing workflow will propagate through the workflow and sometimes being amplified / mitigated into the desired products/information. ML uncertainty typically corresponds to the uncertainty in Level 2 in this workflow.

Uncertainty Propagation – A Satellite CDR Example



Although there are different source for the uncertainty, the contribution from different sources depends on the applications at hand.

Application contexts are very important for uncertainty quantification.

Merchant et al. (2017) https://doi.org/10.5194/essd-9-511-2017

Uncertainty propagation

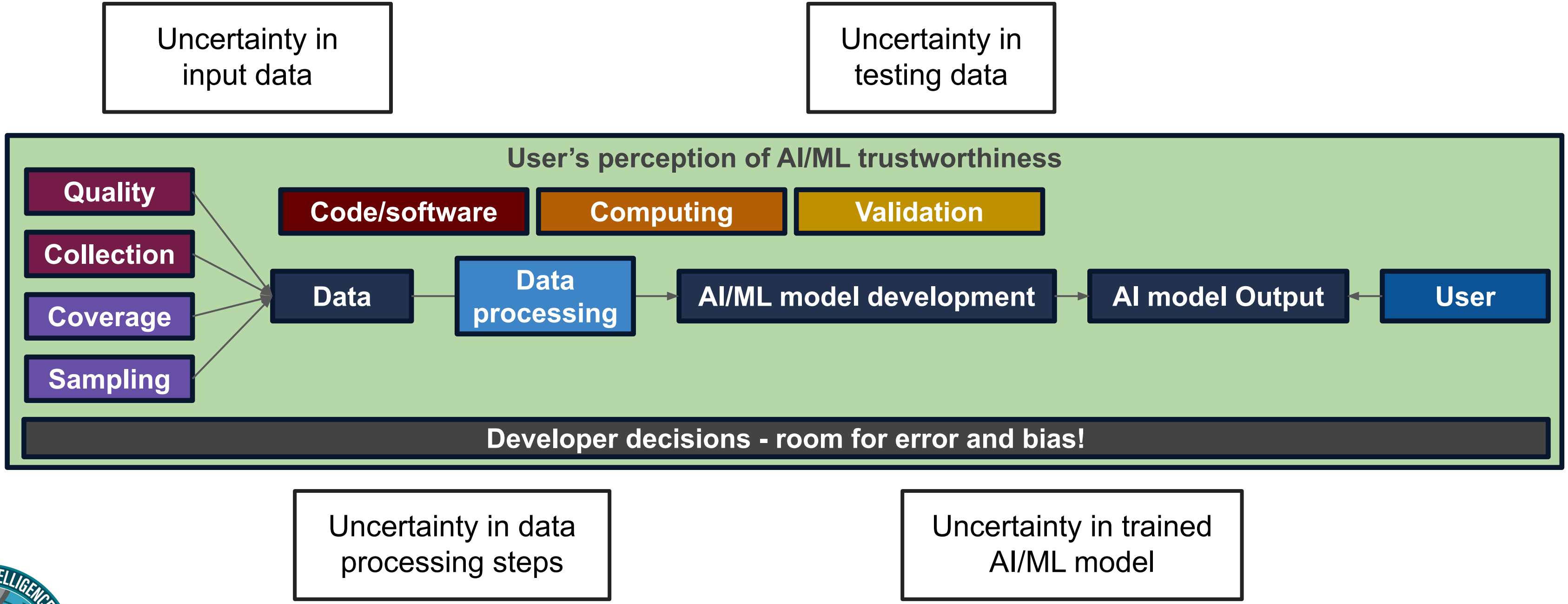
Assuming the target (y) is a function of multiple predictants (i.e., x_1, …, x_n)

$$y = f(x_1, .., .., x_n)$$

The overall uncertainty is the combination of the errors from each individual predictants while taking account of the correlated errors among different predictants (e.g., spatially correlated, temporally correlated, or physically correlated).

$$u^2 = \sum_{i}^{n} \left(\frac{\partial f}{\partial x_i}\right)^2 u_i^2(x_i) + 2\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\frac{\partial f}{\partial x_i}\right)\left(\frac{\partial f}{\partial x_j}\right) u(x_i, x_j)$$

Bring it back to AI/ML cases



Quick break to give you time to soak information in and ask questions!

Go to sli.do and use the code TAI4ES

Motivating examples

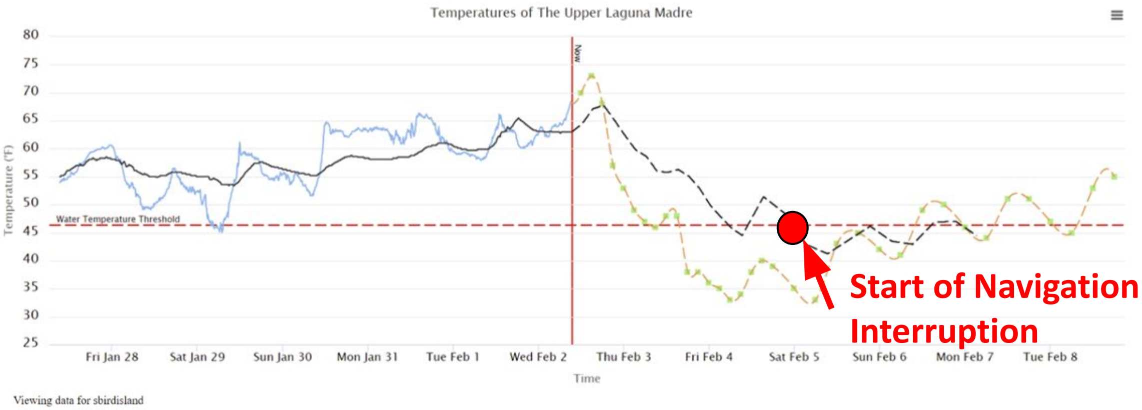
Two examples of practical use of UQ:

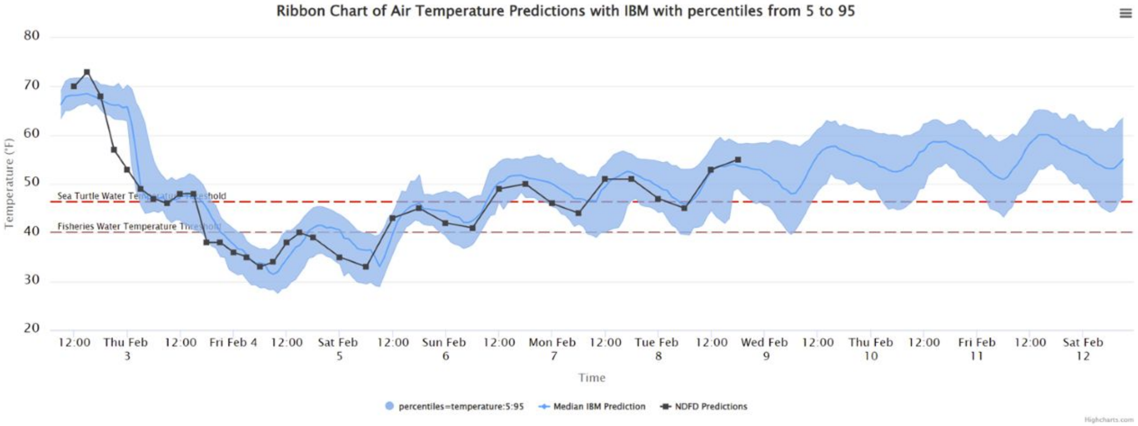
1. **Cold Stunning predictions**

2. **Estimating precipitation from satellite imagery**

Uncertainty: Cold Stunning Predictions

- Water temperature below 8C for ~24 hrs leads to sea turtle cold stunnings
- AI (shallow neural nets) used since 2008 to predict onset and duration of cold stunnings (black dash line)
- AI Predictions allow for interruption of navigation, staging of resources, …
- Here, example for Feb 2022 cold stunning predictions (400+ sea turtles)





Research: IBM/AI2ES providing ensemble air temperature predictions (right)
(1) Create ensemble ANN predictions
(2) Quantify uncertainty in AI temperature and threshold crossings predictions

How to best quantify, visualize, communicate uncertainty?

Co-production of models with stakeholders?

Satellite Application of UQ estimate

Orescanin, M., Petković, V., Powell, S.W., Marsh, B.R. and Heslin, S.C., 2021. Bayesian Deep Learning for Passive Microwave Precipitation Type Detection. IEEE Geoscience and Remote Sensing Letters.

Task:

Classify precipitation type (stratiform or convective) based on passive MW imagery.

Goal:

Provide two outputs:

1. Map of precipitation type: indicates stratiform/convective per pixel.
2. Map of uncertainty: indicates how much to trust classification per pixel.

Method used:

Bayesian neural network. More details later.

Satellite Application of UQ estimate

Orescanin et al., Bayesian Deep Learning for Passive Microwave Precipitation Type Detection, 2021.

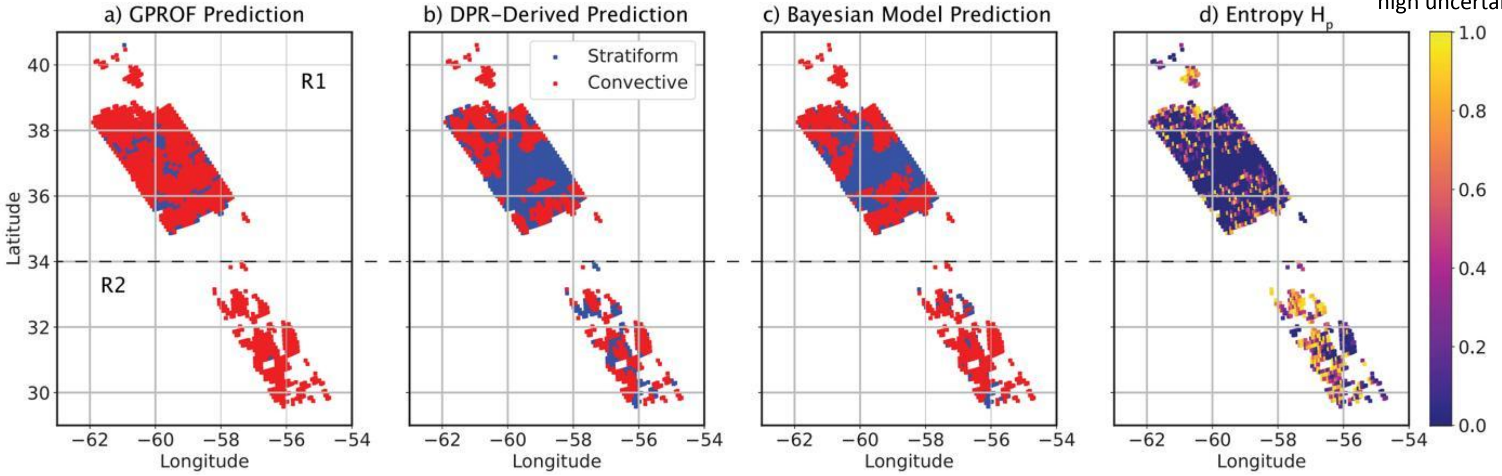
Classification from baseline operational algorithm

Ground truth (label from DPR)

Classification from Bayesian NN

Uncertainty estimate from Bayesian NN

high entropy = high uncertainty



Next topic:
2) Classifications of Uncertainty

Classifications of Uncertainty

There are many different ways to classify uncertainty.
Different classifications arise from

1. Needs of the specific application and end user;
2. Information available and approaches used to develop uncertainty estimates.

It can be very confusing to encounter all these different classifications in the literature.
Do not be surprised to see those. It’s not you - it’s the nature of the field.

Example: two approaches that use very different classification

1. **Component-based approach**, i.e. modeling uncertainty of each component separately (based on expert knowledge), then propagating contributions.
2. **Typical AI approach**, i.e. being given only data set and AI model, and no information about components.

Classification of Uncertainty in AI

Aleatory uncertainty: the <u>natural randomness in the underlying process</u>.

- Also known as: statistical, stochastic or irreducible uncertainty.
- Classic Example:
  - Tossing a perfect coin (50-50 probability)
  - Even the best model of this system cannot predict outcome of tossing a coin, because of its stochastic properties.
- **Irreducible:** This uncertainty can be estimated, but not eliminated.

Epistemic uncertainty: the <u>scientific uncertainty due to limited data and knowledge</u>.

- Also known as: systemic, model, or reducible uncertainty.
- Uncertainty based on our ignorance - we just do not know the system and its state well enough.
- Classic example:
  - Training a machine learning model with few data samples.
  - Uncertainty can be reduced by feeding more appropriate data and/or adding more physical knowledge.
- **Reducible:** This uncertainty can be reduced with better models & data, e.g., by collecting and feeding in more data or by choosing a better ML method.

Total Uncertainty = Aleatory Uncertainty + Epistemic Uncertainty

Aleatory vs. Epistemic uncertainty

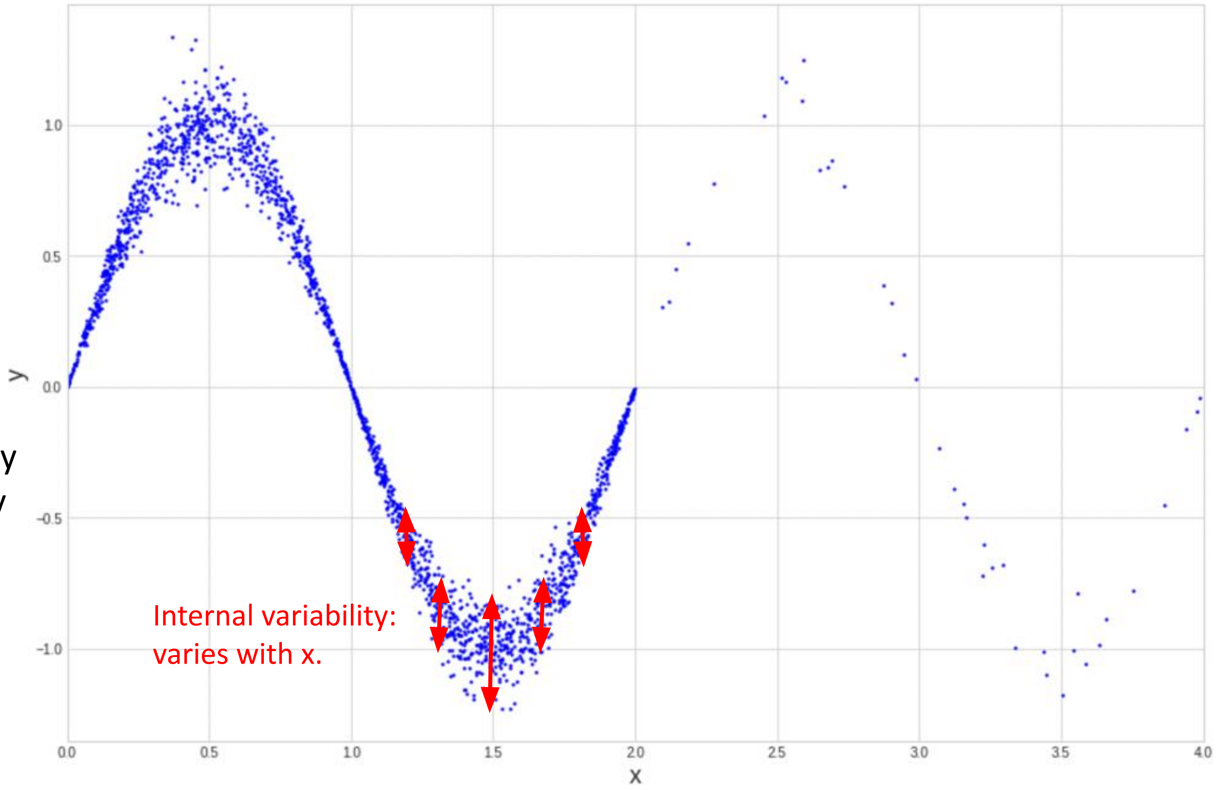
Distribution of training data:

A) y(x) is normal distributed for all x ∈ [0,4] with
- **$\mu = \sin(x \cdot \pi)$**
- **$\sigma = 0.1 \cdot \sin(x \cdot \pi)$**

Represents internal variability.

B) Sampling in x varies:
- For x ∈ [0,2]: high density
- For x ∈ [2,4]: low density

Represents sampling rate in x.



Aleatory vs. Epistemic uncertainty

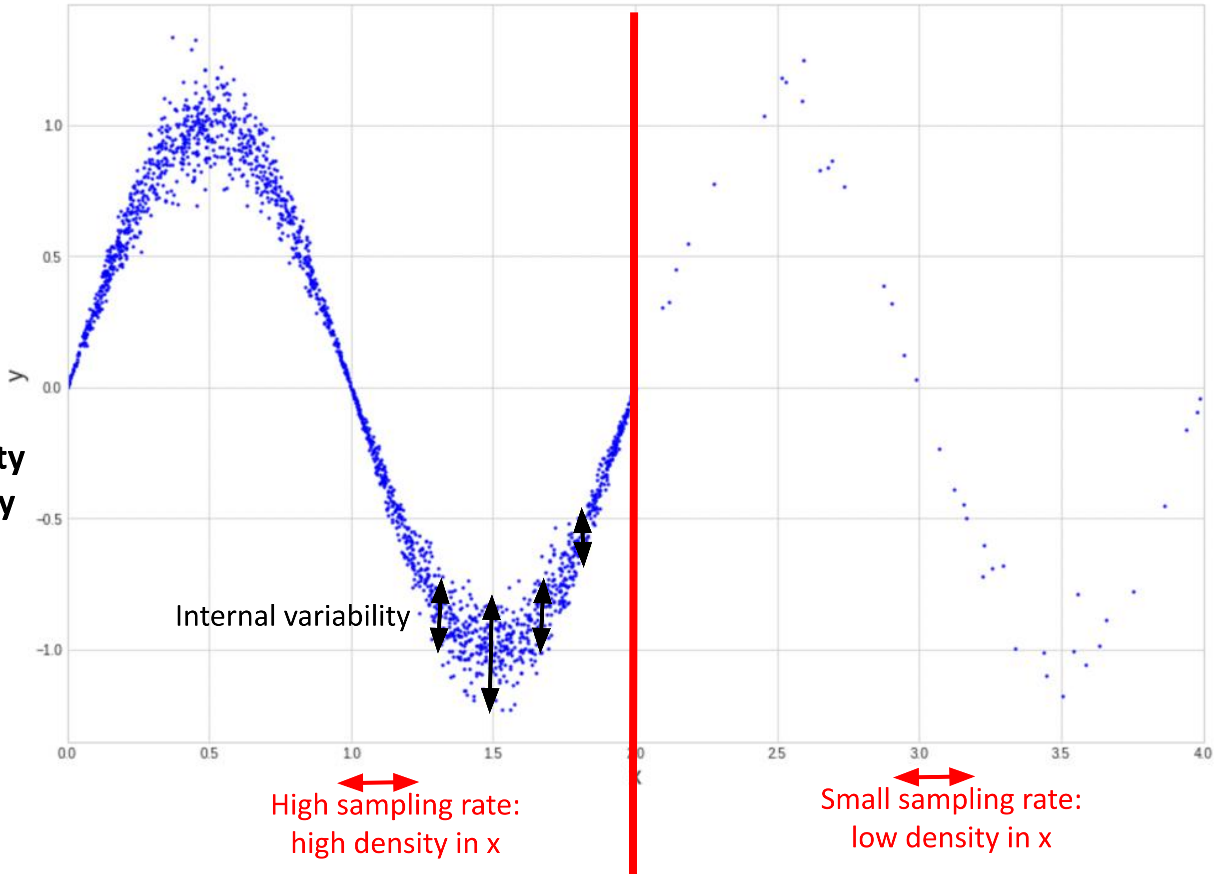
Distribution of training data:

A) y(x) is normal distributed
for all x ∈ [0,4] with
- μ = sin(x · π)
- σ = 0.1 · sin(x · π)

Represents internal variability.

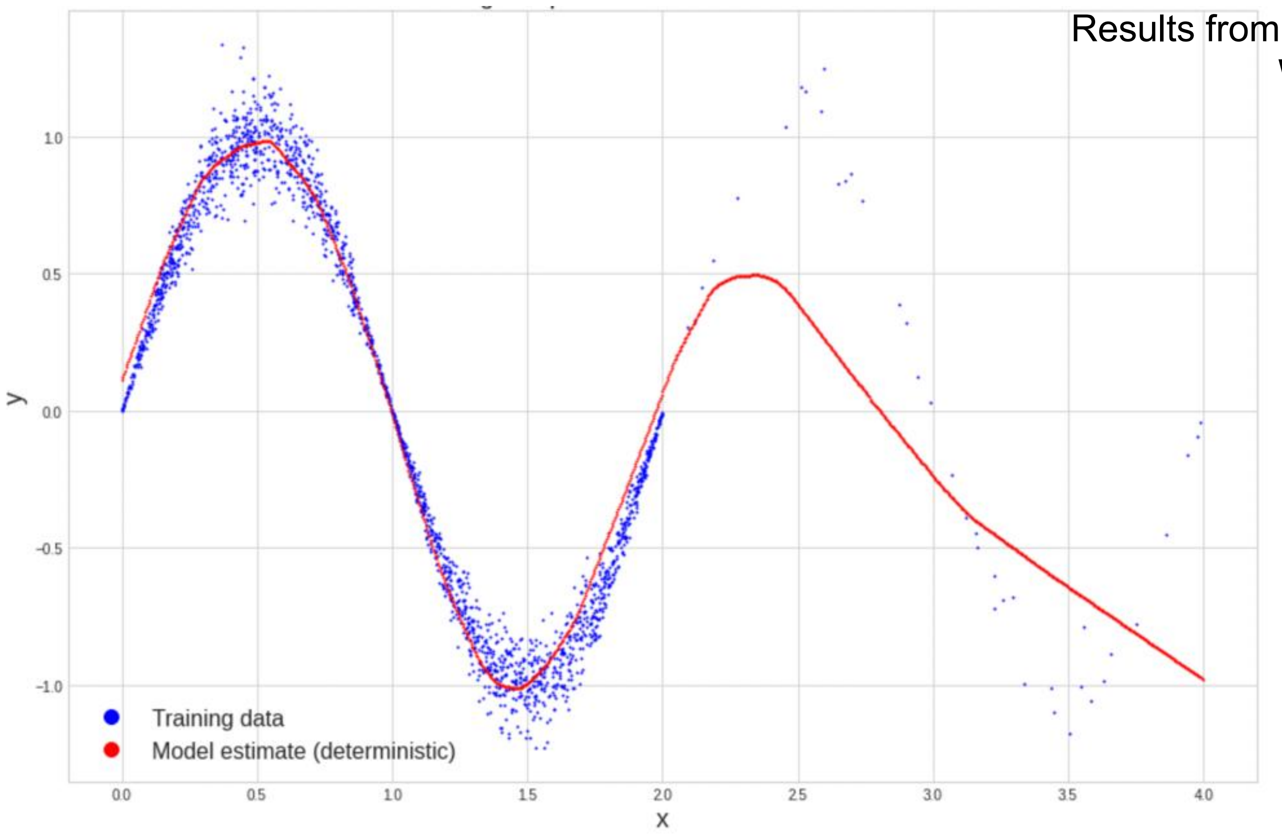
B) Sampling in x varies:
- **For x ∈ [0,2]: high density**
- **For x ∈ [2,4]: low density**

Represents sampling rate in x.

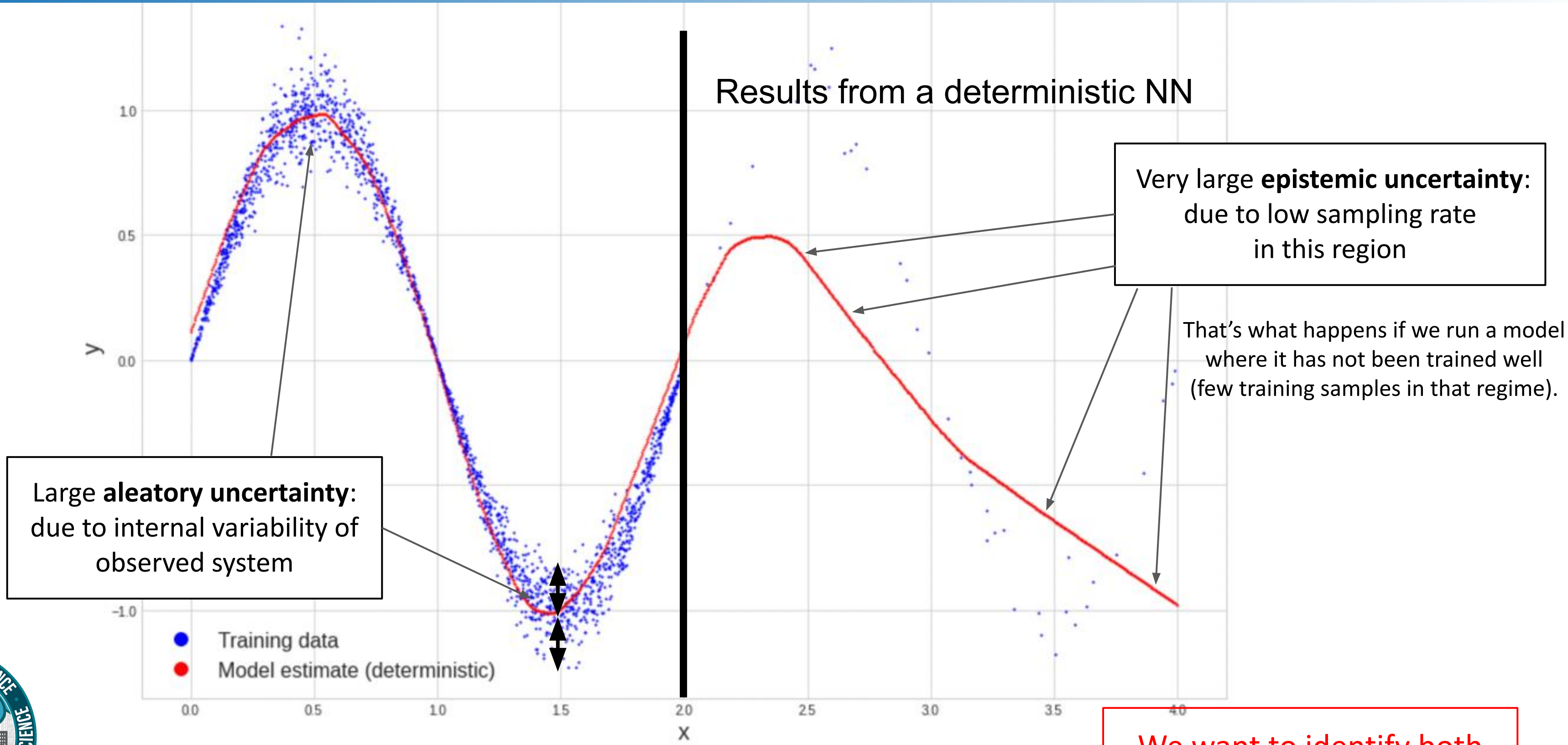


Aleatory vs. Epistemic uncertainty

Results from a deterministic NN:
What do we notice?



Aleatory vs. Epistemic uncertainty



We want to identify both types of uncertainty!

Do UQ methods capture both?

Different UQ methods capture different types of uncertainty.

In theory:

1. Non-Bayesian methods capture only aleatory (irreducible) uncertainty.
2. Bayesian methods can capture epistemic uncertainty.
3. Some combinations of methods can capture both - in theory!

In practice:

- What the various methods capture **in practice** is yet another question altogether!
- Estimates vary greatly between methods.

→ That's why evaluation criteria for uncertainty estimates are so important!
→ Next big topic.

But first - a little exercise.

Sample classification in environmental science

Uncertainty categories adopted from:

Beucler et al., Machine Learning for Clouds and Climate, book chapter in "Clouds and Climate", AGU Geophysical Monograph Series, https://www.essoar.org/doi/abs/10.1002/essoar.10506925.1

Uncertainty categories:

1. **Stochastic**: due to internal climate variability or the chaotic nature of flow, etc.
2. **Observational**: due to measurement and representation errors
3. **Structural**: due to incorrect model structure
4. **Parametric**: due to incorrect model parameters

Question: Can we map these four categories to aleatory vs. epistemic?

Reminder:

- Aleatory uncertainty: natural randomness in the underlying process.
- Epistemic uncertainty: scientific uncertainty due to limited data or knowledge.

4.5. Go to sli.do and use the code TAI4ES

Classification of Uncertainty - ES example

1. **Stochastic**: due to internal climate variability or the chaotic nature of flow, etc.
2. **Observational**: due to measurement and representation errors (e.g., satellite retrieval error)
3. **Structural**: due to incorrect model structure (e.g., ML model type)
4. **Parametric**: due to incorrect model parameters (e.g., ML training)

Which ones are aleatory vs. epistemic?

- **Clearly aleatory (irreducible)**: stochastic
- **Clearly epistemic (lack of knowledge)**: structural, parametric
- **But what about observational?**
  Seems to have both aleatory and epistemic components.
  Some of it could be reduced by better knowledge of sensor system (e.g., satellite),
  but some of it is inherent internal variability of sensor system - so both?

Key lessons: 1) Different classifications do not easily map to each other.
2) Many classifications are valid and make sense in their own way.

Quick break to give you time to soak information in and ask questions!

Go to sli.do and use the code TAI4ES

Next topic:
3) Evaluating uncertainty estimates

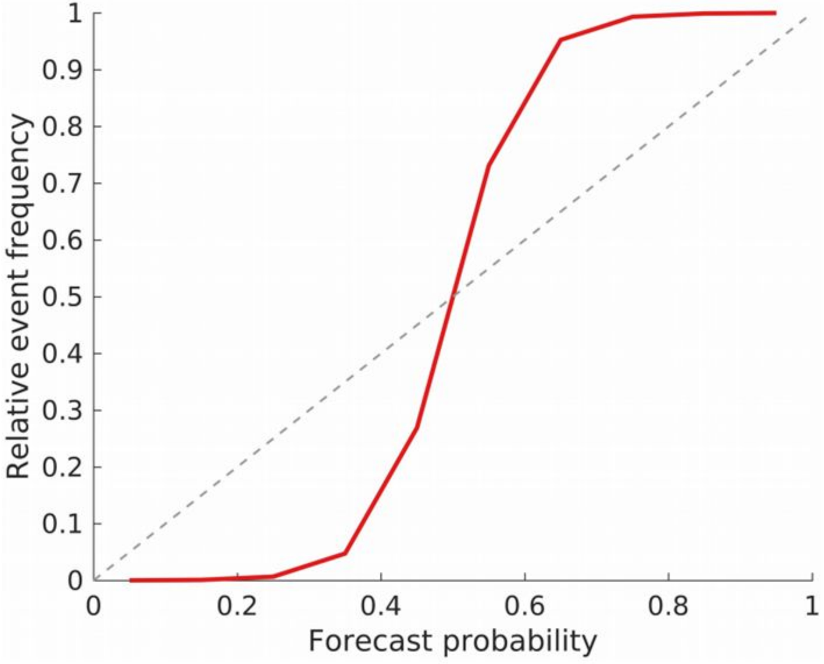
Evaluating uncertainty estimates

This section will discuss five evaluation tools:

a) Reliability curve and attributes diagram
b) Spread-skill plot
c) Probability integral transform (PIT) and PIT histogram
d) Discard test
e) The continuous ranked probability score (CRPS)

3a) Reliability curve and attributes diagram

- The reliability curve is used to evaluate probabilistic predictions.
- Typically used for binary classification (predicting a yes-or-no event).
- Reliability curves can also be modified for regression (shown later).



- Reliability curves evaluate only the central prediction (not uncertainty).
- However, in studies involving UQ, reliability curves are commonly used to evaluate the central prediction.

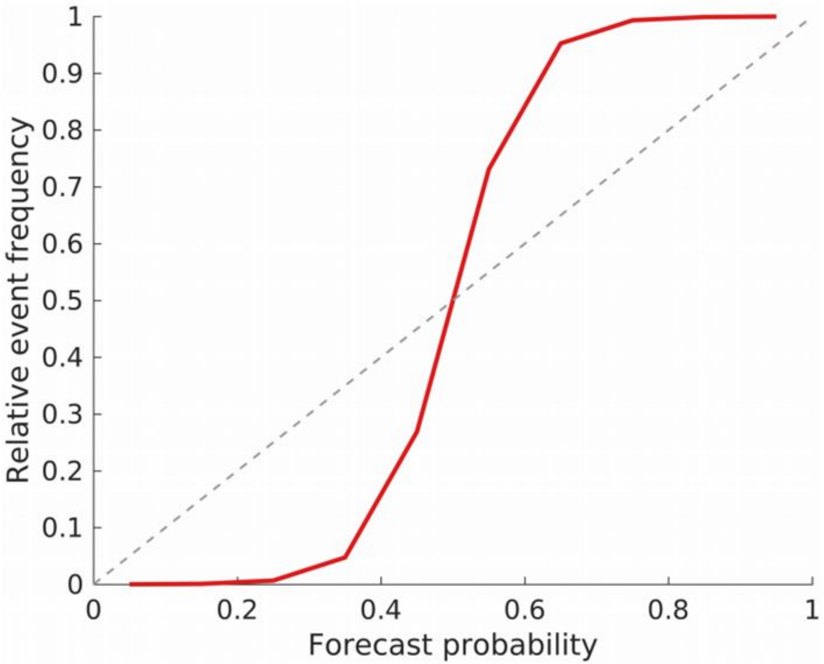
3a) Reliability curve and attributes diagram

- Reliability curves are used even if the model is not truly probabilistic.
- Most ML models for classification output confidence scores (pseudo-probabilities), which are not true probabilities.
- However, most people just call these “probabilities” and use the reliability curve to evaluate how calibrated these “probabilities” are.
- There is nothing wrong with using the reliability curve for this purpose, as long as you recognize the difference between pseudo-probabilities and true probabilities.

Take-home point: you can use the reliability curve to evaluate true probabilities or pseudo-probabilities, but be careful with terminology.

3a) Reliability curve and attributes diagram

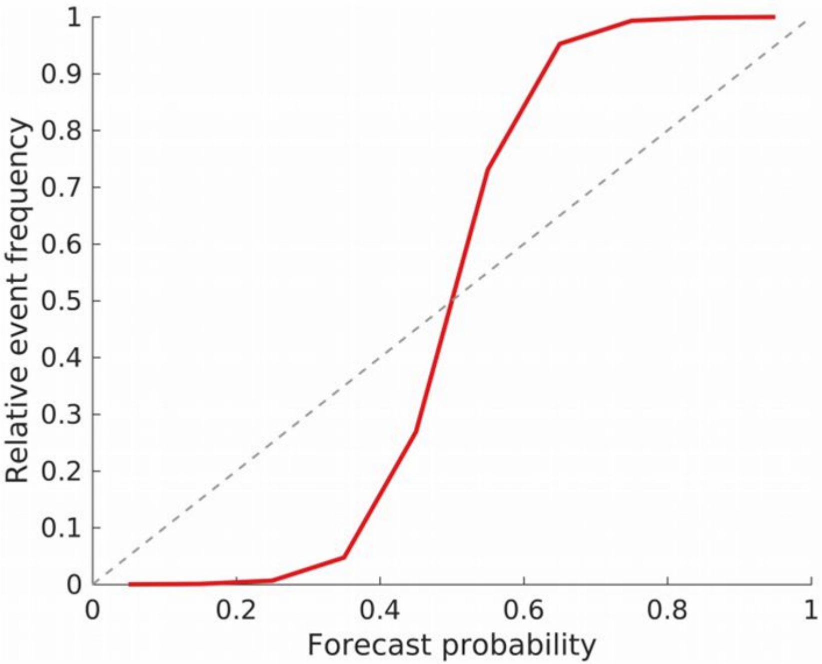
- The reliability curve plots predicted event probability vs. conditional event frequency.
- The reliability curve is binned by predicted probability, often into 10 bins:
    - 0-10%
    - 10-20%
    - …
    - 90-100%
- Thus, each point is a mean over all examples in one bin:
    - *x*-coordinate: mean predicted probability in bin
    - *y*-coordinate: observed event frequency in bin
- Thus, the reliability curve answers the question:



"Given predicted probability p, how likely is the event to actually occur?"

3a) Reliability curve and attributes diagram

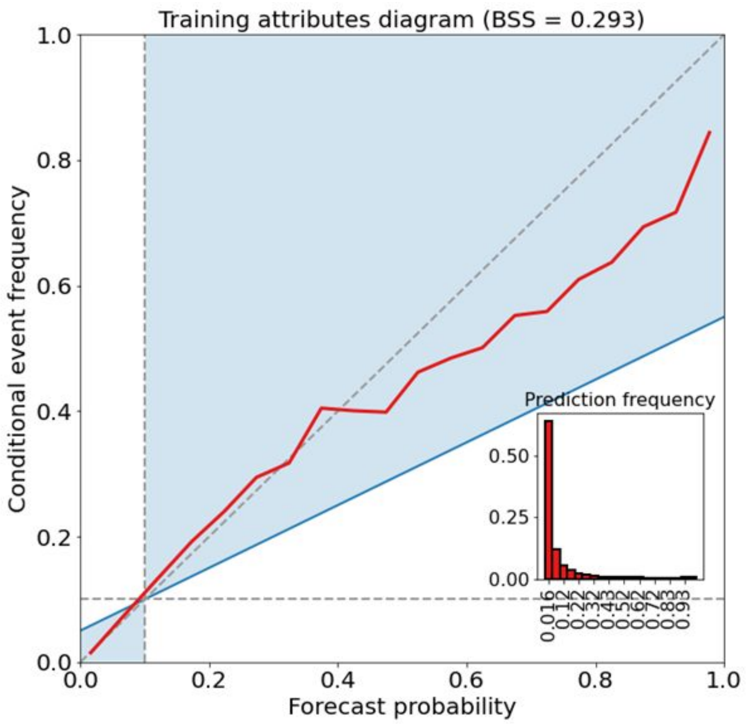
- Ideally, conditional event frequency should always = forecast probability.
- In other words, in cases where the model says probability = *p,* the true event frequency (*f*) should be *p*.
- Dashed grey line: perfect reliability, where *f* = *p* for all bins.
- Points below grey line: predicted probability is too high, or model is “overconfident”.
- Points above grey line: predicted probability is too low, or model is “underconfident”.



3a) Reliability curve and attributes diagram

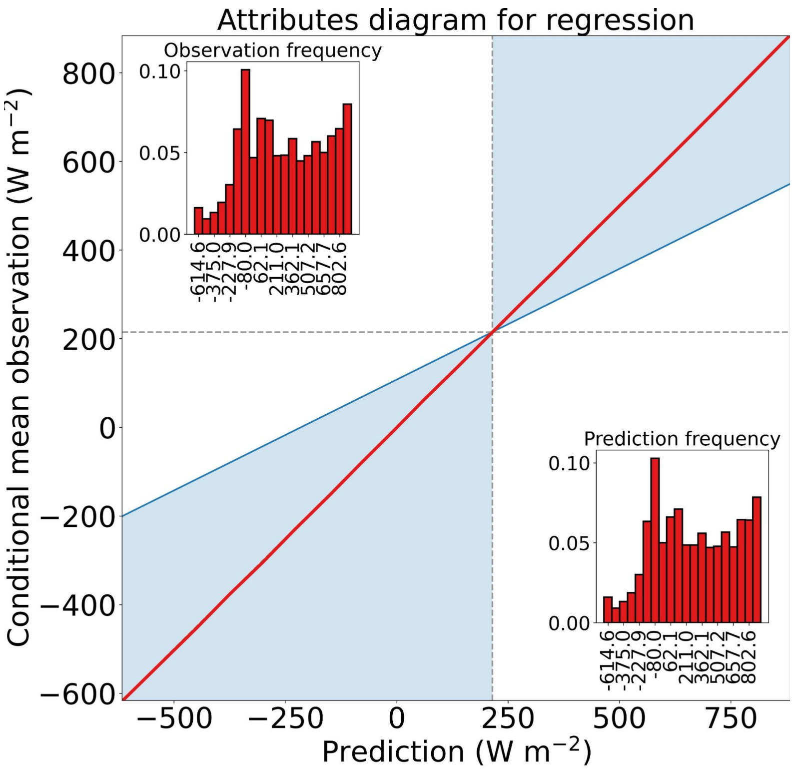
- The attributes diagram (Hsu and Murphy 1986) is a reliability curve with extra reference lines:
    - Diagonal grey line = perfect reliability, as before
    - Vertical grey line = climatology (event frequency over all data)
    - Horizontal grey line = no resolution
    - Blue shading = positive-skill area
- A model with no resolution follows the no-resolution line.
- A climatological model (one that always predicts p = event frequency over all data) has a reliability curve with one point, at the intersection of the climo and no-resolution lines.

"Positive skill" means Brier skill score > 0.



3a) Reliability curve and attributes diagram

- **The attributes diagram can also be adapted for regression problems.**
- **Differences are summarized below**, letting the target variable be *z*.
  - The *x*-axis is the model-predicted *z*-value – a real number that in general can range from (-∞, +∞) – instead of a probability.
  - The *y*-axis is the conditional mean observed *z*-value – a real number that in general can range from (-∞, +∞) – instead of an event frequency.
  - The perfect-reliability line is still the 1-to-1 line.
  - The no-resolution line is at $y = z_{climo}$, and the climatology line is at $x = z_{climo}$, where z_{climo} is the average *z*-value over the full dataset.
  - The interpretation of the perfect-reliability, climatology, and no-resolution lines is the same.
  - The positive-skill area shows where the mean squared error (MSE) skill score (MSESS), rather than the BSS, is positive.



3b) The spread-skill plot

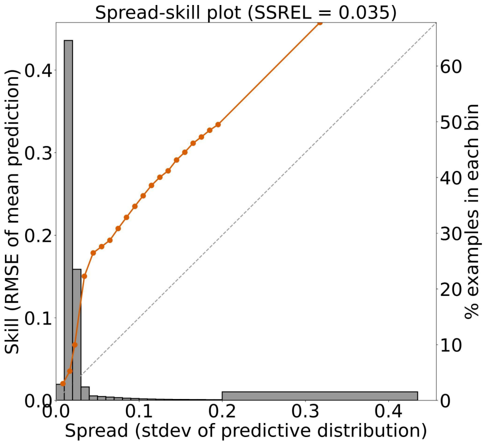
- Similar to reliability curve but may be used *only* for models that include uncertainty.
- Can be used for classification or regression.

- *x* = predicted model spread
  - Mean standard deviation of model's predictive distribution

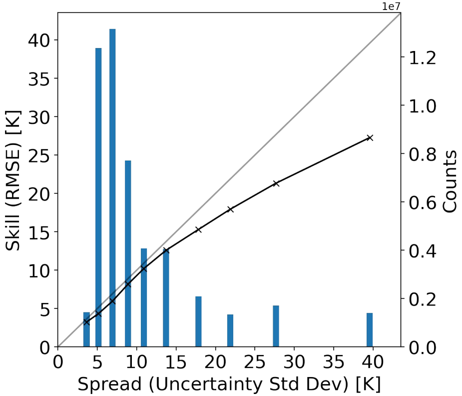
- *y* = RMSE of model's mean prediction

- Each point corresponds to one bin of spread values.
  - Just like, in reliability curve, each point corresponds to one bin of forecast probs.
- The spread-skill plot answers the following question:

"Given predicted model spread, what is model error?"



Spread-skill plot for classification task



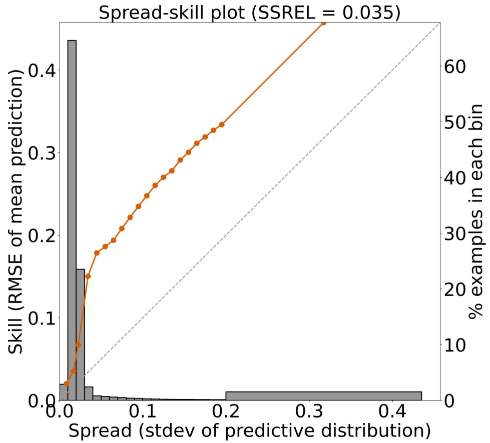
Spread-skill plot for regression task

3b) The spread-skill plot

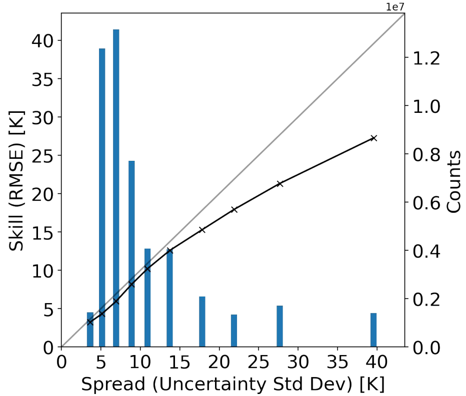
- For a model with perfectly calibrated uncertainty estimates, the spread-skill plot follows the 1-to-1 line.
    - At points below the 1-to-1 line (bins where spread > error), the model is overspread or “underconfident”.
    - At points above the 1-to-1 line, the model is underspread or “overconfident”.

- We also include a histogram to show the number of cases in each spread bin.

- Overall quality of spread-skill plot can be summarized by mean distance from the 1-to-1 line, which we call the spread-skill reliability (SSREL).



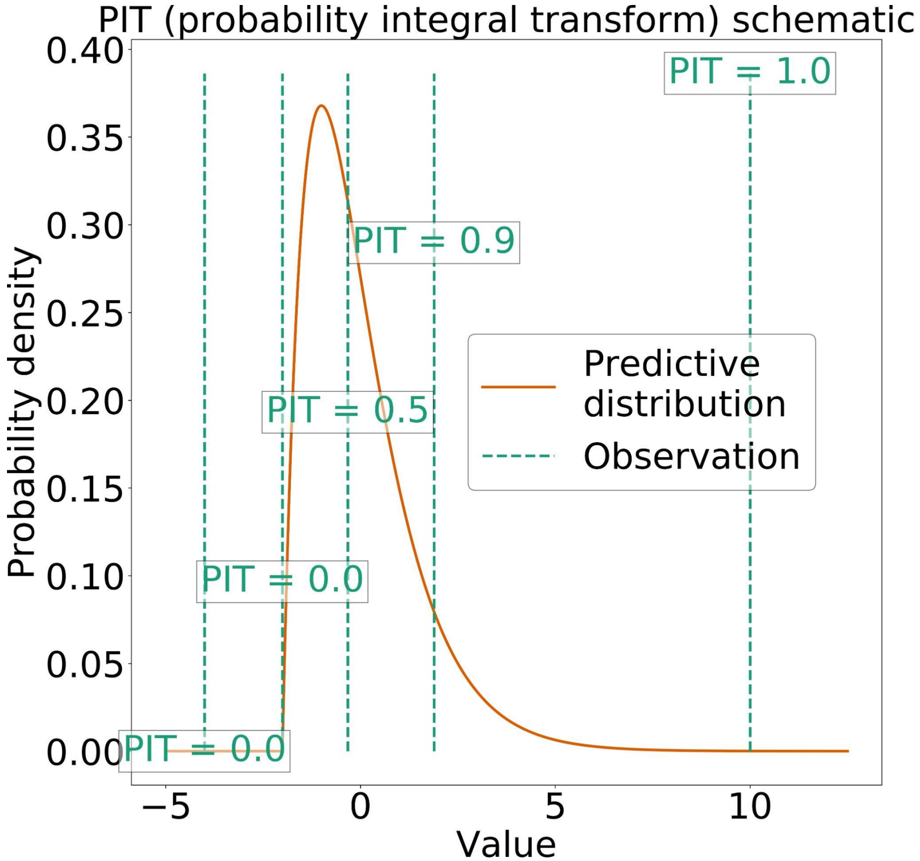
Spread-skill plot for classification task



Spread-skill plot for regression task

3c) The PIT histogram

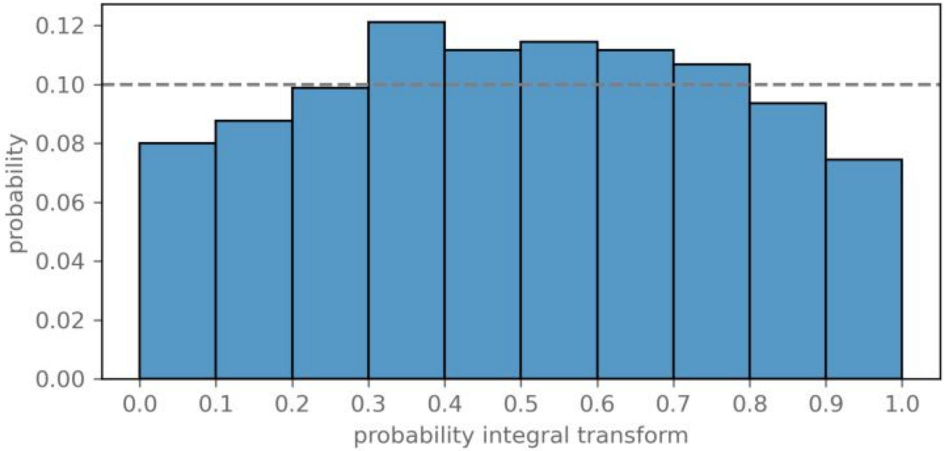
- PIT = probability integral transform
- Definition: value that predictive CDF (cumulative density function) attains at observed value
- Alternate definition: quantile of observed value in distribution of predictions



- Examples:
    - Observed value = median of predictive distribution ⇒ PIT = 0.5
    - Observed value = max of predictive distribution ⇒ PIT = 1.0
    - Observed value = min of predictive distribution ⇒ PIT = 0.0

3c) The PIT histogram

- The PIT histogram is a histogram of all PIT values (one for each example).
- For a perfectly calibrated model, the PIT histogram is uniform.
  - In other words, all PIT values occur with the same frequency.
- If the PIT histogram has a hump in the middle, the model has too much spread or is "underconfident" (below; Figure 15 of Barnes *et al.* 2021).



- If the PIT histogram has humps on the sides (a lot of values near 0 or 1), the model has too little spread or is "overconfident".

3c) The PIT histogram

- Atmospheric scientists are typically more familiar with the rank histogram, invented by Talagrand and discussed in Hamill (2001).

- The PIT histogram is a generalization of the rank histogram.

- The rank histogram is used for ensembles, where the ensemble contains a finite number of models and thus generates a finite number of predictions.

- The PIT histogram (and all other evaluation tools in Section 3) can be used for any method that generates a predictive distribution, whether the distribution is created by:

  a) collecting deterministic predictions from each member of an ensemble;
  b) predicting the quantiles of a distribution;
  c) predicting the parameters (*e.g.*, mean and standard deviation for Gaussian) of a distribution;
  d) anything else.

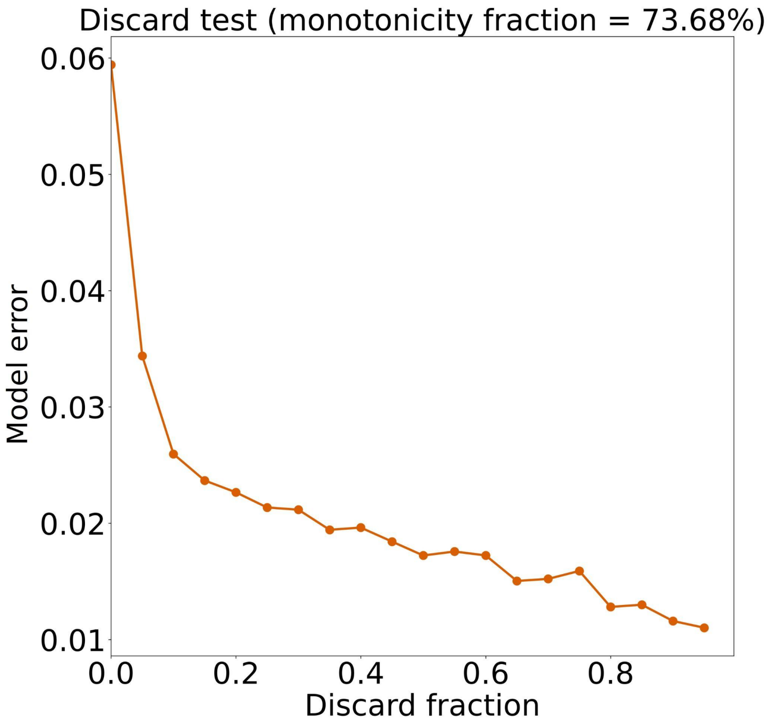
- Happily, the rank histogram and PIT histogram can be interpreted in the same way (uniform = perfectly calibrated; bunched in middle = underconfident; bunched at sides = overconfident).

3d) The discard test

- **The discard test compares model error versus the fraction of highest-uncertainty cases discarded.**

- **Procedure:**
    - Select a fraction *f* (example: 0.1 or 10%).
    - Discard the *f* * 100% of cases with highest uncertainty.
    - Compute the model error before and after discarding.
    - Did the model error decrease after discarding? If so, good.

- For a model with well calibrated uncertainty estimates, error should decrease monotonically as the discard fraction increases.

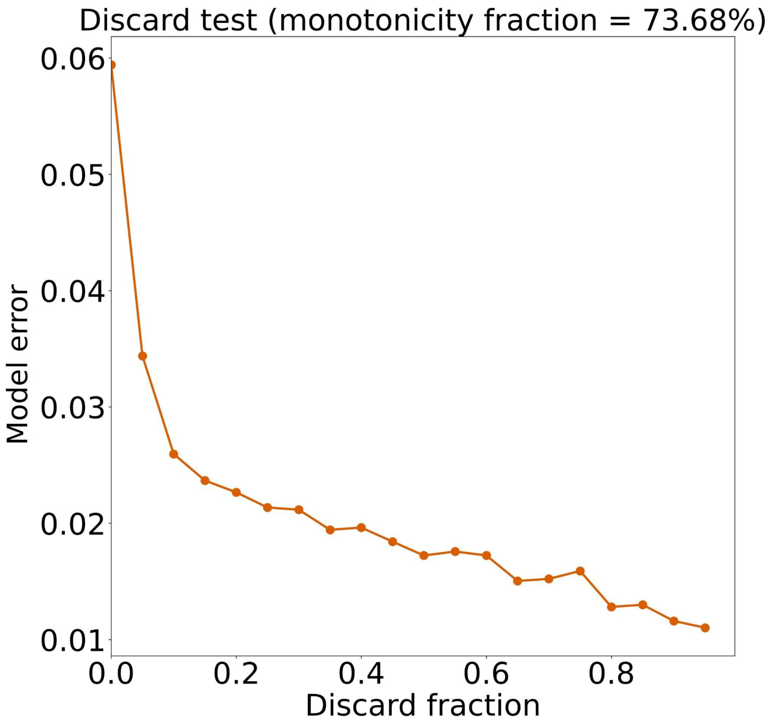


3d) The discard test

- **The overall quality of the discard test can be summarized by the monotonicity fraction (MF):**

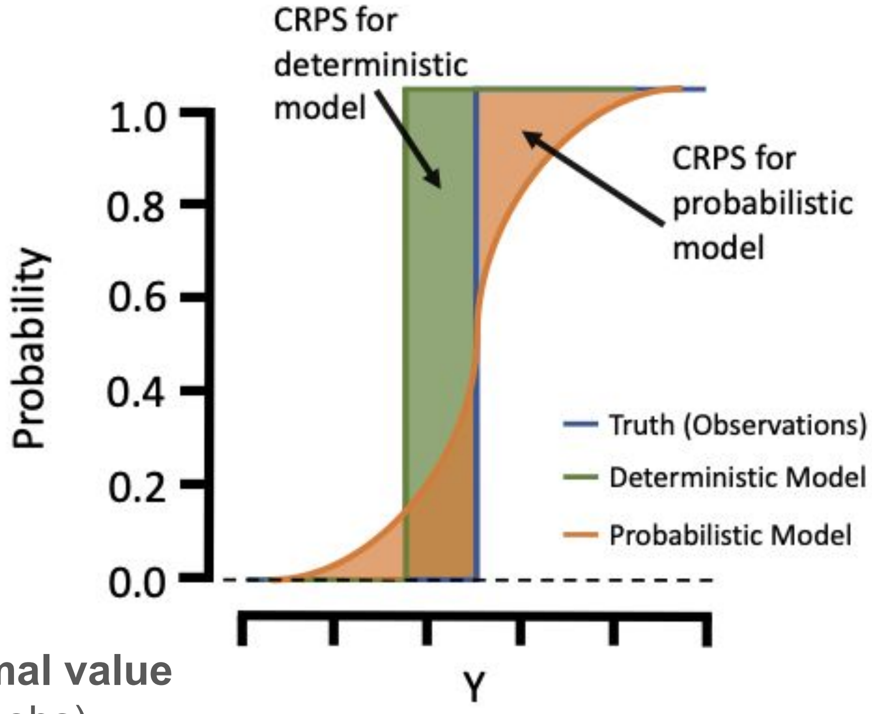
$$\mathrm{MF} = \frac{1}{N_f - 1} \sum_{i=1}^{N_f - 1} \mathcal{I}(\epsilon_i \geq \epsilon_{i+1})$$

- N_f is the number of discard fractions used
- ε_i is the model error with the i^{th} discard fraction
- $\mathcal{I}()$ is the indicator function, which evaluates to 1 if the condition is true and 0 if the condition is false



3e) The continuous ranked probability score (CRPS)

CRPS: comparison between **probabilistic models** and **deterministic models**



CRPS varies from [0, ∞), and the **optimal value is 0** (no difference between model and obs)

3e) The continuous ranked probability score (CRPS)

CRPS: comparison between **probabilistic models** and **deterministic models**

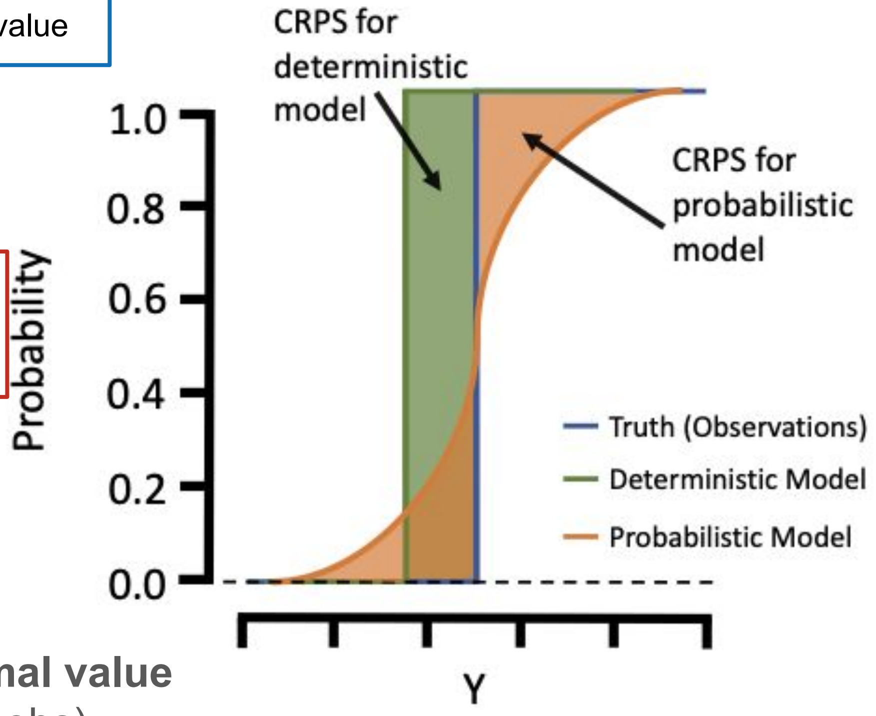
Predicted value

Observed value

$$CRPS(F, y_{true}) = \int_{-\infty}^{\infty} \left(\boxed{F(y_{pred})} - \boxed{\mathcal{H}}(y_{pred} - y_{true}) \right)^2 dy$$

Predicted CDF

Heaviside step function (1 if y_{pred} > y_{true}; 0 otherwise)



CRPS varies from [0, ∞), and the **optimal value is 0** (no difference between model and obs)

3e) The continuous ranked probability score (CRPS)

- Big **advantage** of CRPS:
  - Considers the full predictive PDF, not just the mean or standard deviation or certain quantiles

- CRPS can be used as a loss function to train neural networks (more in methods)
  - Creates an ensemble of predictions that can be used to quantify uncertainty
  - Ensemble members are trained to represent the true PDF and do not require any *a priori* distribution information

- Original work using CRPS and providing excellent derivations
  - Hersbach (2000): Decomposition of the CRPS for ensemble prediction systems
  - Gneiting et al (2005): Calibrated probabilistic forecasting using minimum CRPS estimation
  - Gneiting and Raftery (2007): Strictly proper scoring rules, prediction, and estimation
  - Székely and Rizzo (2005): A new test for multivariate normality

Overview of UQ-evaluation methods

Method	Classification	Regression	What it tells us
Attributes diagram	✔	✔	Evaluates only central prediction, not uncertainty estimates. Class: observed event frequency as a function of predicted event probability, Brier score, Brier skill score Reg: mean observed target value as a function of predicted target value, mean squared error (MSE), MSE skill score
Spread-skill plot	✔	✔	Model error as a function of predicted model spread. If uncertainty is perfectly calibrated, this plot follows the 1-to-1 line.
PIT histogram		✔	Distribution of PIT values. If uncertainty is perfectly calibrated, this distribution is uniform, so the PIT histogram is flat.
Discard test	✔	✔	Model error vs. discard fraction. If uncertainty is well calibrated, error decreases monotonically as discard fraction increases, *i.e.,* as more high-uncertainty samples are dropped.

Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School

Day 4: Agenda

- 9:00 Uncertainty quantification methods (Part 1)
- **10:00 *Short brain & bio break***
- 10:10 Uncertainty quantification methods (Part 2)
- 10:45 *Short brain & bio break*
- 10:55 Communicating uncertainty (Part 3)
- 11:55 Lecture series wrap up!

Questions?

https://app.sli.do/event/1zumy91n

Or go to sli.do and use the code TAI4ES

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

LEAP

NOAA CENTER FOR ARTIFICIAL INTELLIGENCE

Radiant Earth Foundation
EARTH IMAGERY FOR IMPACT

Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School

Day 4: Agenda

- 9:00 Uncertainty quantification methods (Part 1)
- 10:00 *Short brain & bio break*
- **10:10 Uncertainty quantification methods (Part 2)**
- 10:45 *Short brain & bio break*
- 10:55 Communicating uncertainty (Part 3)
- 11:55 Lecture series wrap up!

Questions?

https://app.sli.do/event/1zumy91n

Or go to sli.do and use the code TAI4ES

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

NOAA CENTER FOR ARTIFICIAL INTELLIGENCE

Radiant Earth Foundation
EARTH IMAGERY FOR IMPACT

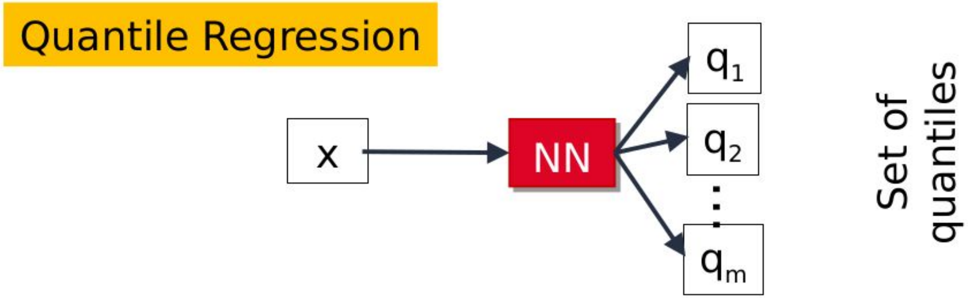
Methods

- We chose 6 UQ methods for this presentation; we think these 6 methods represent the most common/promising approaches.

1. Quantile regression (also works for ML models other than NNs)
2. CRPS loss function
3. Parametric prediction
4. Deep ensembles
5. Monte Carlo dropout - time permitting
6. Bayesian neural network - time permitting

Method 1: Quantile regression

- **Quantile regression (QR) involves directly predicting the quantiles of a probability distribution.**
- This means that for each data sample, instead of predicting a single number (the mean or “expected value” or “maximum-likelihood estimate”), we predict several numbers (quantile-based estimates).
- In early work, median regression (predicting the 50th percentile) was seen as an alternative to least-squares linear regression, which predicts the mean.
- **QR works for many types of ML, not just neural nets.**



Method 1: Quantile regression

- **The "trick" is to train the model with the quantile loss function:**

$$\mathcal{L} = \begin{cases} q \quad \left|y_{\text{true}} - y^q_{\text{pred}}\right| & , \quad \text{if } \; y_{\text{true}} > y^q_{\text{pred}}; \\ (1-q)\left|y_{\text{true}} - y^q_{\text{pred}}\right| & , \quad \text{if } \; y_{\text{true}} \leq y^q_{\text{pred}}. \end{cases}$$

- q is the desired quantile level, ranging from [0, 1]
- y_{true} is the correct value
- y_{pred}^q is the estimated value at quantile level q

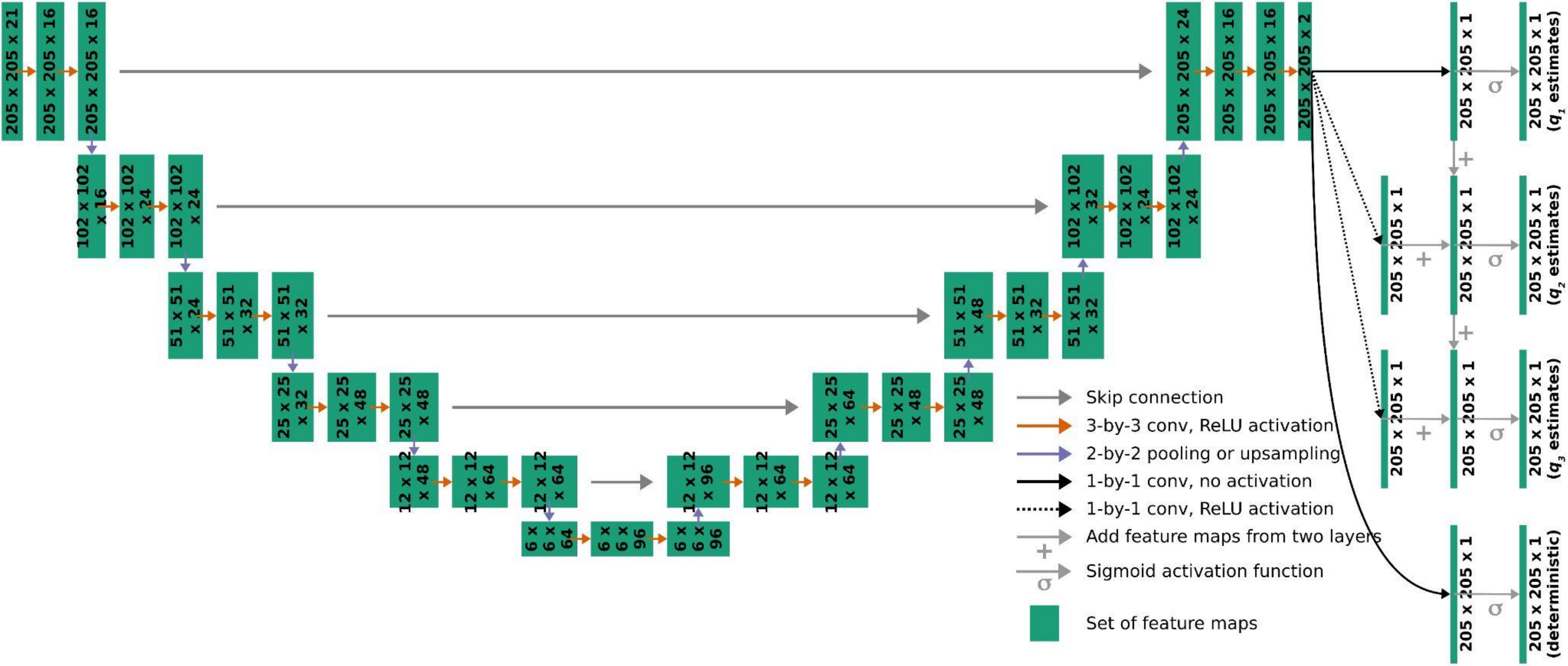
- Large values of q penalize underprediction ($y_{\text{pred}}^q < y_{\text{true}}$) more than overprediction ($y_{\text{pred}}^q > y_{\text{true}}$), encouraging the model to output large y_{pred}^q.
- Conversely, small values of q encourage the model to output small y_{pred}^q.

Method 1: Quantile regression

- To estimate multiple quantiles with NNs, a common approach is to train a separate NN for each quantile.
- Because the different NNs are trained independently, this approach does not prevent the problem of quantile-crossing, where the estimated value y_{pred}^{q} decreases as the quantile level q increases.
    - Example: the 25th-percentile rainfall prediction is 30 mm but the 75th-percentile prediction is 20 mm.

- Thus, **we have developed a novel NN architecture that completely prevents quantile-crossing** (see [notebook]).
- For any consecutive pair of quantile levels, q_{i-1} and q_i, the estimate $y_{\text{pred}}^{q_i}$ must be $\geq$ $y_{\text{pred}}^{q_{i-1}}$.
- To satisfy this condition, we express $y_{\text{pred}}^{q_i}$ as the sum of $y_{\text{pred}}^{q_{i-1}}$ and a positive term.
- We implement this with Add() layers and the ReLU activation function.

- **Below: U-net architecture with quantile regression.**
- Task is binary classification: take 205-by-205 satellite images and predict convection at 1-hour lead.



Method 2: Using CRPS loss function

This method is for any machine learning model with a loss function

- Implemented by Gneiting and Raftery (2007)
- Can be evaluated for any distribution using Monte Carlo techniques to generate ensemble members representative of the distribution

$$\mathrm{CRPS}^*(F, y_{true}) = E_F \mid \dot{Y} - y_{true} \mid \; - \; \tfrac{1}{2} E_F \mid \dot{Y} - \dot{Y}' \mid$$

MAE between NN predictions and y_{true}

Half the predicted spread (MAE of pairwise differences between ensemble members)

$\dot{Y}$ = Predictions from all ensemble members (i.e., randomly-drawn sample with the distribution of y_{pred})

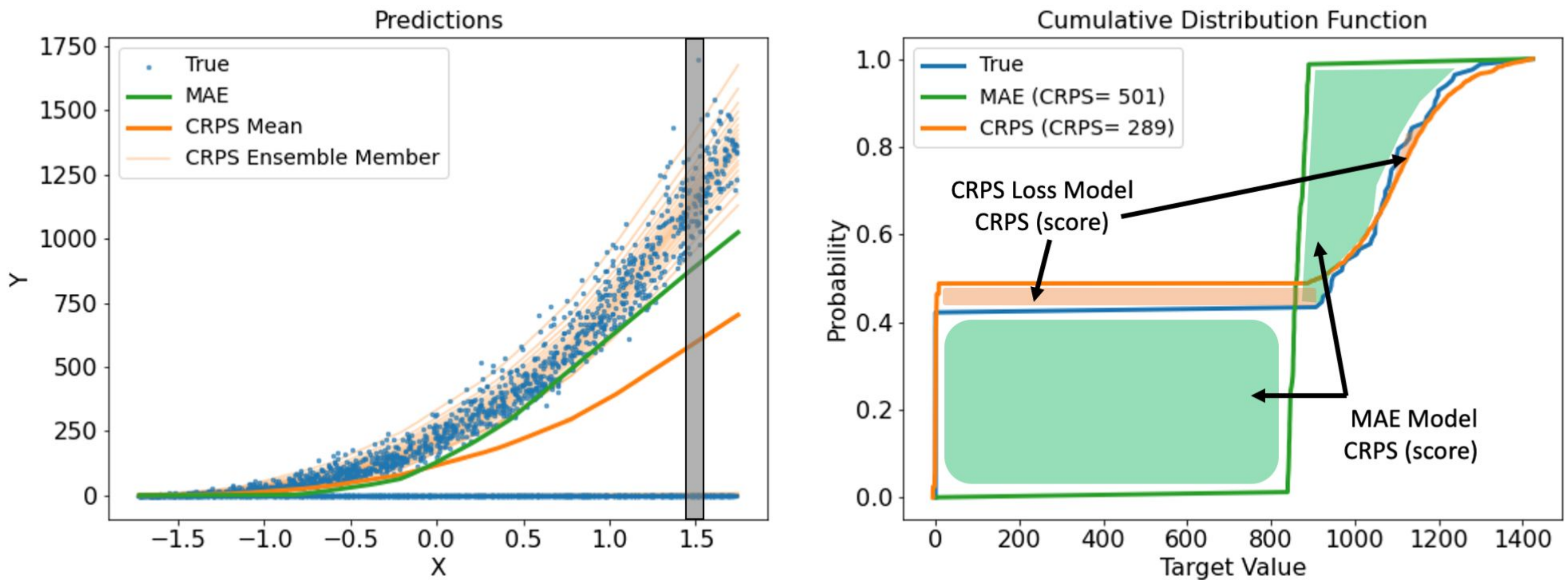
$\dot{Y}'$ = Transposed copy of the predictions

E_F = Evaluation function (reduces dimensionality to single value, often the mean)

y_{true} = Observations; y_{pred} = Predictions

- CRPS* = Negative orientation of CRPS
  - Can be reported in same units as observations
  - Reduces to Mean Absolute Error (MAE) for a single ensemble member

Method 2: CRPS loss function



Probabilistic CRPS does well at capturing **different regimes**
Do not need to know data distribution *a priori*
Evaluate carefully–looking only at **mean** (for example) would make it look like MAE is better than CRPS

Based on Brey (2021)

Method 2: CRPS loss function

- Recent work in probabilistic ML using CRPS as the loss function:
  - Chapman et al. (2022): Probabilistic prediction from deterministic atmospheric river forecasts with deep learning
  - Ghazvinian *et al.* (2021): A novel hybrid artificial neural network parametric scheme for postprocessing medium-range precipitation forecasts
  - Grönquist et al. (2021): Deep learning for post-processing ensemble weather forecasts
  - Brey (2021): CRPS-Net, A package for making and working with probabilistic predictions
  - Scher and Messori (2020): Ensemble methods for neural network-based weather forecasts
  - Rasp and Lerch (2018): Neural Networks for postprocessing ensemble weather forecasts

- Notebook demonstrating the CRPS loss function
- Notebook demonstrating different UQ methods and evaluation metrics
  - Regression task with six sample datasets
  - CRPS loss function, Monte Carlo dropout, parameters of probability distribution
  - Attributes diagram, spread-skill plot, PIT histogram, discard test
- Notebooks demoing MC dropout and quantile regression, including evaluation methods

Method 3: Parametric Prediction

Premise: Instead of having NN output a single prediction, output a **probability distribution**

Pros: Does not require any modification to NN architecture (just change output layer and loss function); does not require assumptions of linearity, normality, etc

Cons: Must specify distribution *a priori*

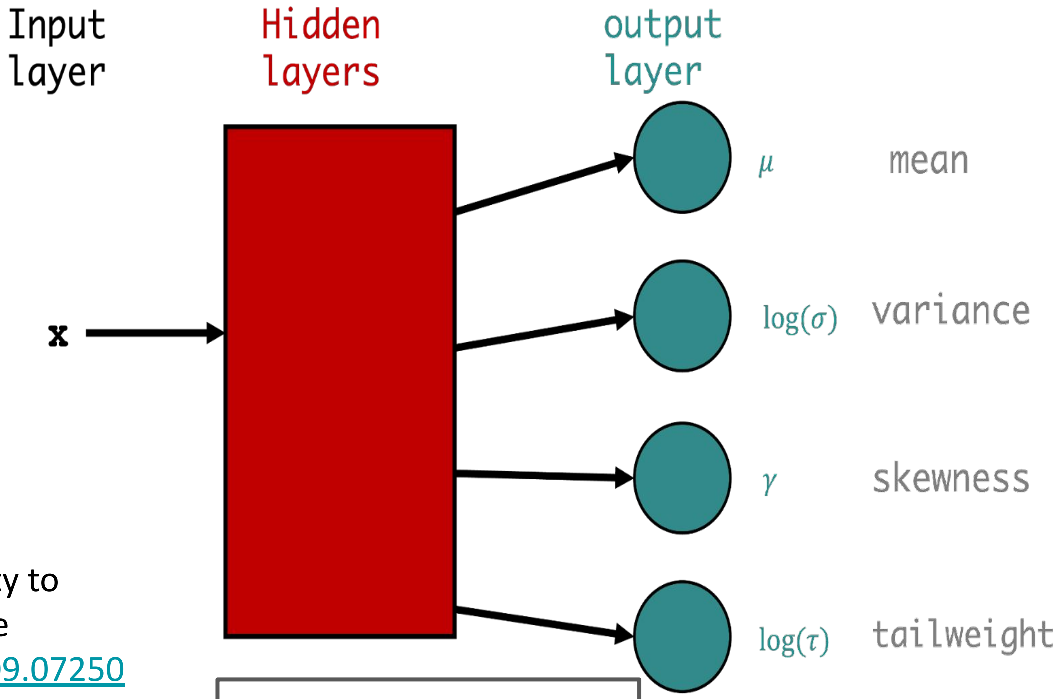


Figure 8 from Barnes et al. (2021)

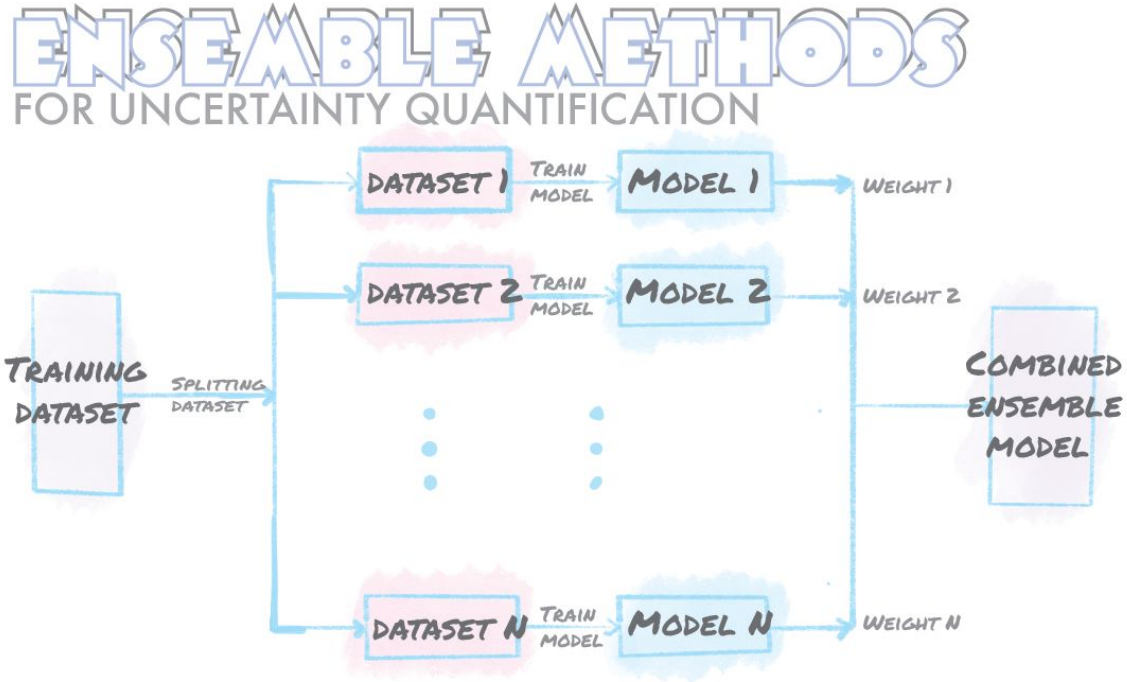
References and Resources
Barnes et al. (2021): Adding Uncertainty to Neural Network Regression Tasks in the Geosciences, https://arxiv.org/abs/2109.07250

Notebook with implementation of this method

Method 4: Deep Ensemble

Ensemble techniques use the diversity of various model trained on slightly different data / features / initialization to estimate the predictive uncertainty.

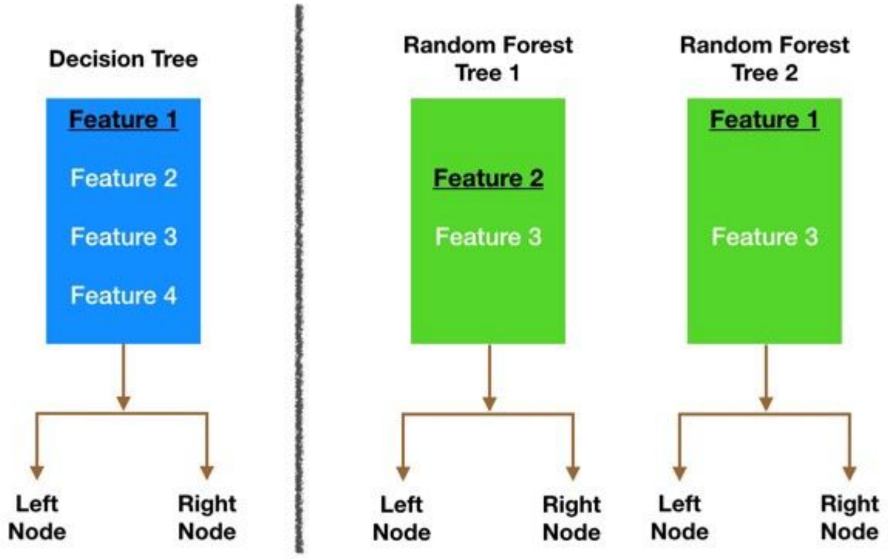
Bootstrapping/bagging are the most commonly used ensemble technique.



Credit: https://www.rossidata.com/UncertaintyQuantificationandEnsembleLearning

Ensemble explained via random forest

Decision tree



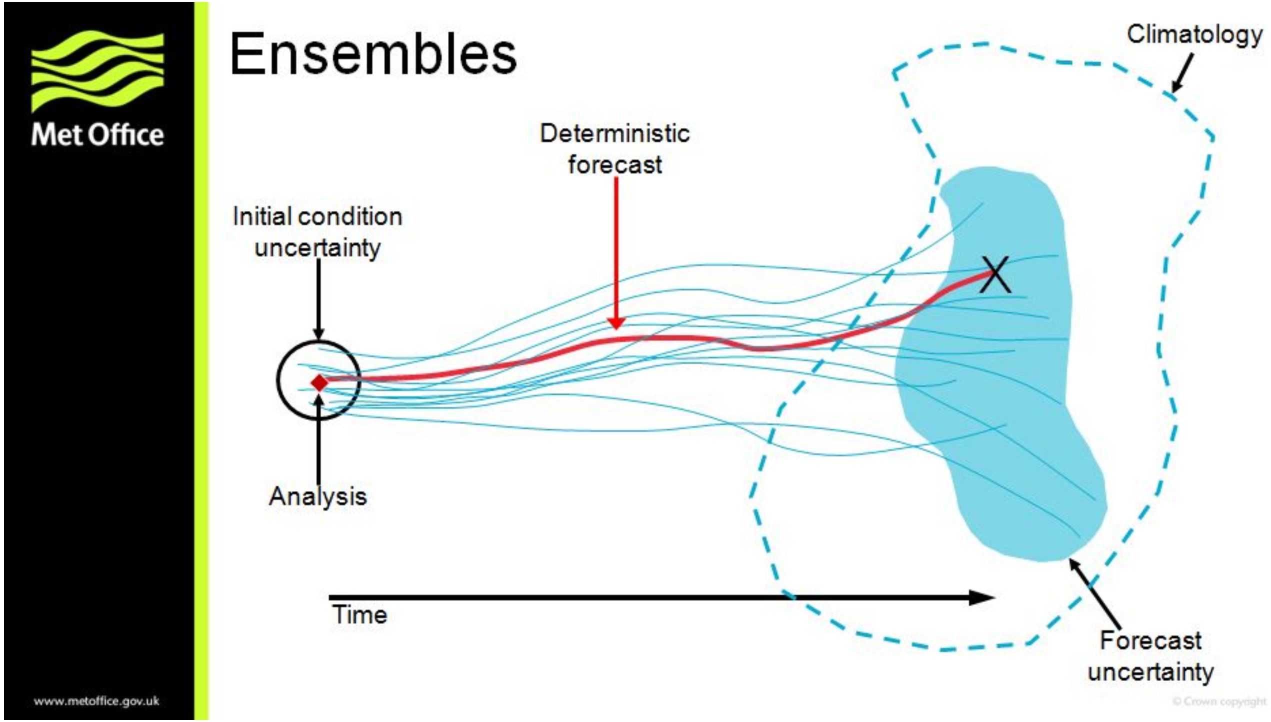
Random forest

From a single learner/model to an ensemble of learners/models

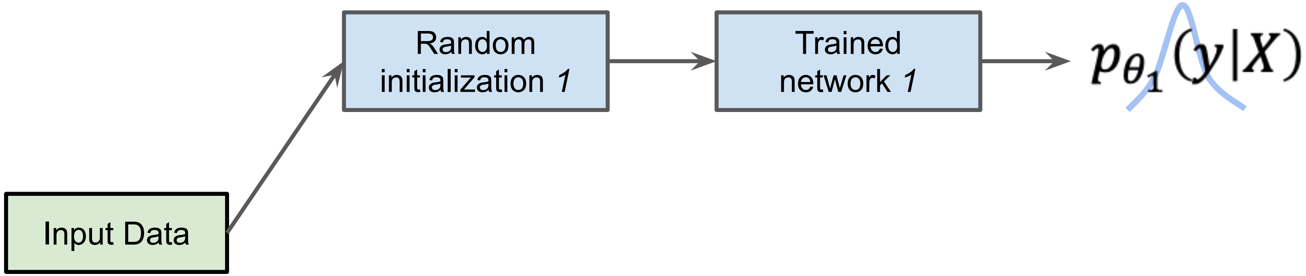
Credit: https://towardsdatascience.com/understanding-random-forest-58381e0602d2

A common example of ensemble for model uncertainty

We often use forecast / prediction from an ensemble of model runs to quantify the uncertainty in weather/climate modeling.

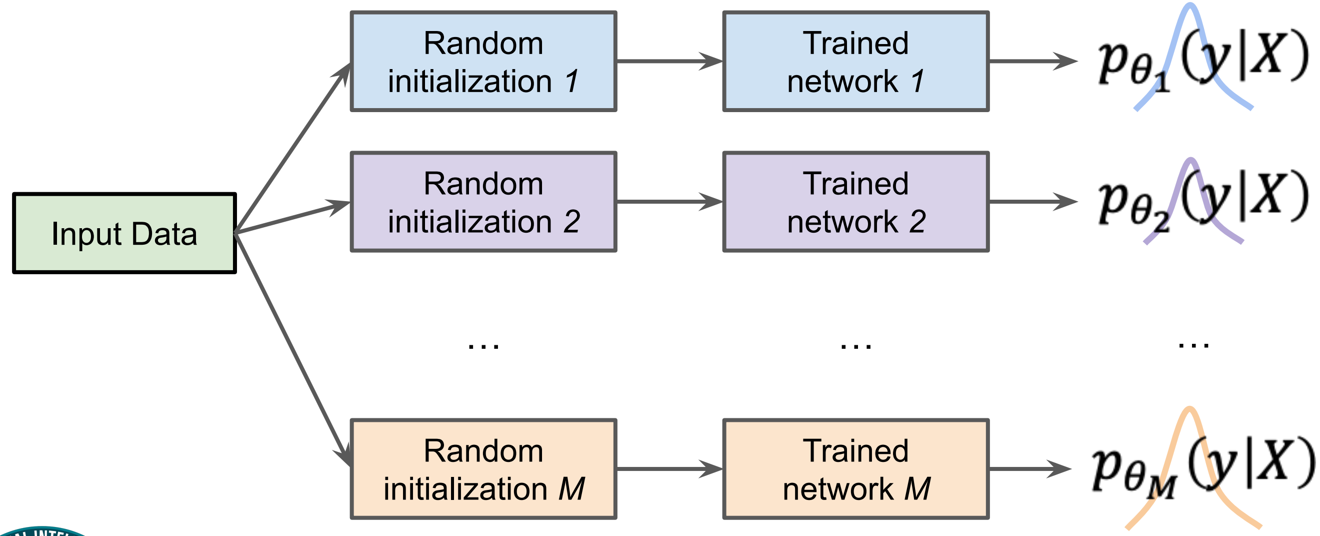


Method 4: Deep Ensemble



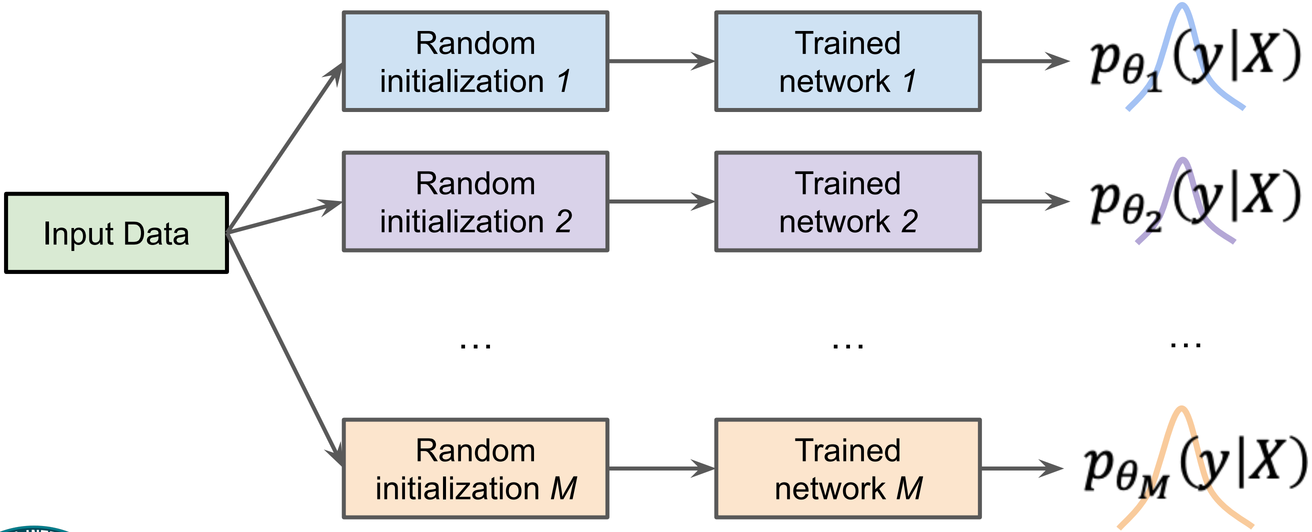
Lakshminarayanan et al. (2017) "Simple and scalable predictive uncertainty estimation using deep ensembles." Advances in neural information processing systems 30.

Method 4: Deep Ensemble



Lakshminarayanan et al. (2017) "Simple and scalable predictive uncertainty estimation using deep ensembles." Advances in neural information processing systems 30.

Method 4: Deep Ensemble



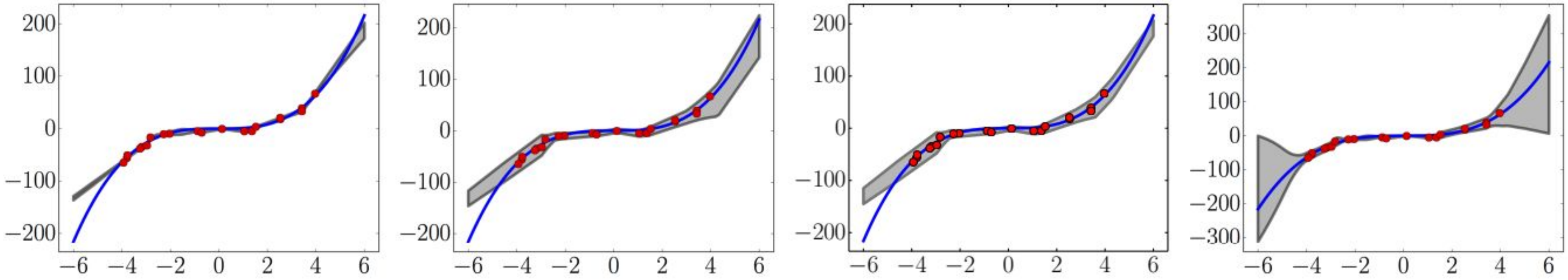
$$p(y|X) = \frac{1}{M}\sum_{i=1}^{M} p_{\theta_i}(y|X)$$

Assumption of uniformly-weighted mixture

Lakshminarayanan et al. (2017) "Simple and scalable predictive uncertainty estimation using deep ensembles." Advances in neural information processing systems 30.

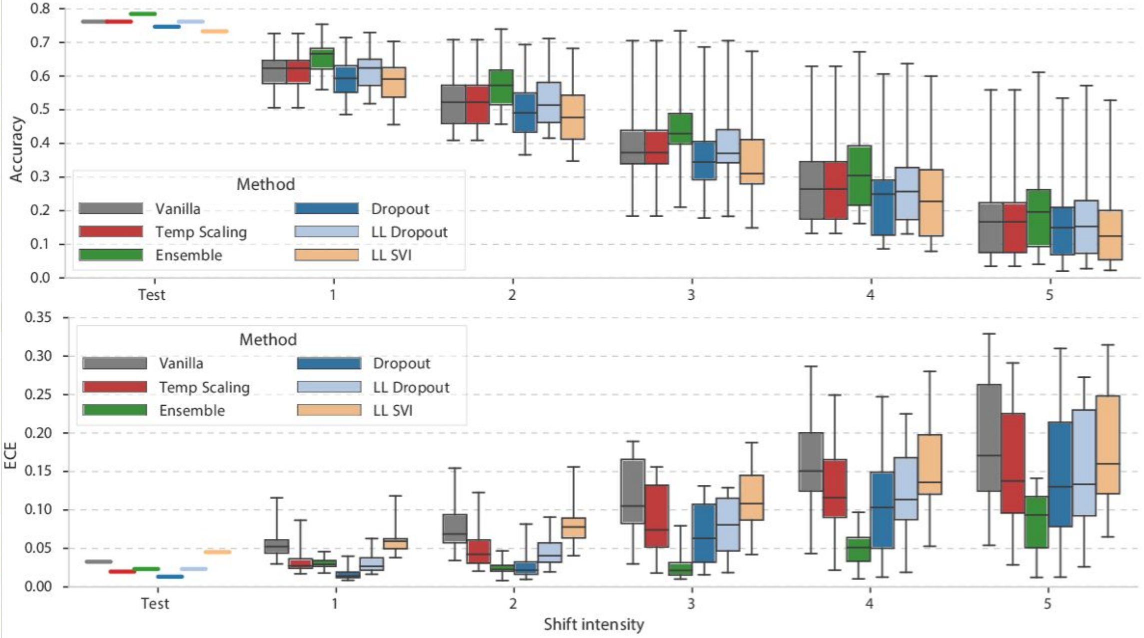
Method 4: Deep Ensemble

Deep ensemble is able to capture the uncertainty of the machine learning model that include both aleatoric and epistemic.



Lakshminarayanan et al. (2017) "Simple and scalable predictive uncertainty estimation using deep ensembles." Advances in neural information processing systems 30.

Method 4: Deep Ensemble

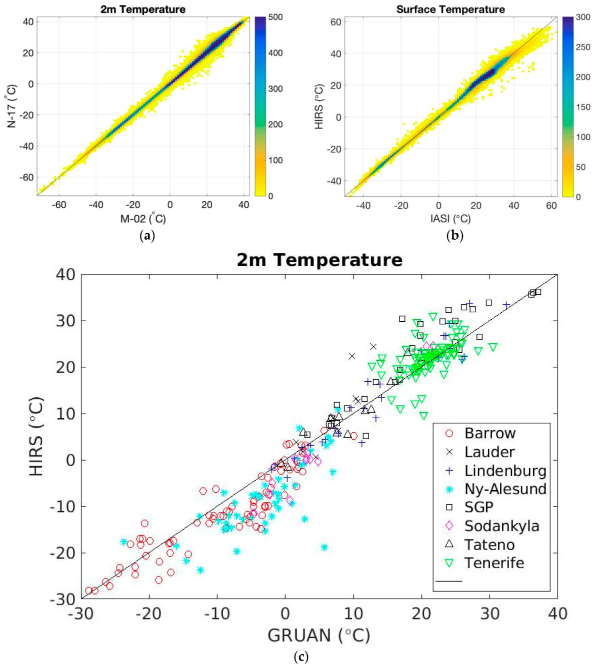


Deep ensemble is able to capture uncertainty even under “data shift”.

Ovadia, et al. "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift." *Advances in neural information processing systems* 32 (2019).

Deep ensemble example in satellite retrieval

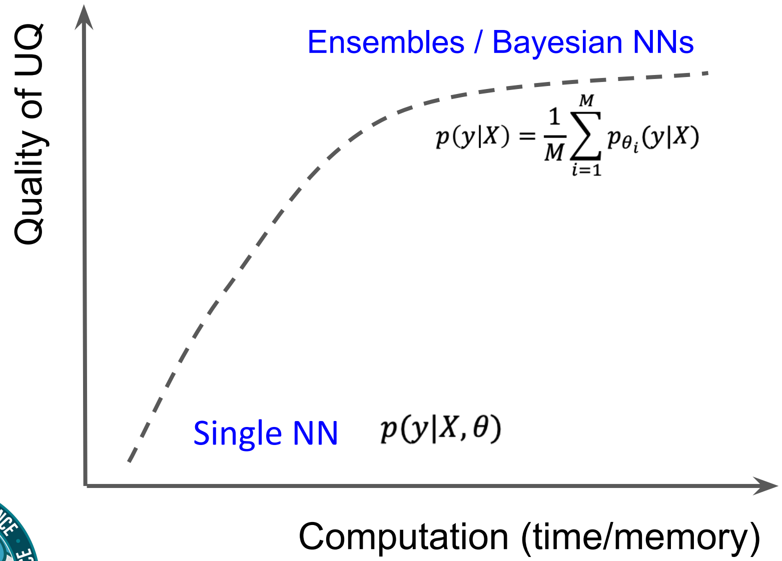
Using long term satellite records (HIRS) to estimate the temperature and humidity at different pressure levels (10 levels), 3-layer NN (Matthews et al., 2019).



Mean target	Mean output	Mean 95% PI (low)	Mean 95% PI (high)	PI % coverage
280.5637	280.5626	276.8789	284.2451	94.36%
275.8684	275.8694	272.3181	279.4166	94.66%
268.2491	268.2485	266.0995	270.3981	94.75%
261.2067	261.2067	259.2162	263.1963	94.81%
252.9137	252.9134	250.8242	255.0034	94.64%
242.7405	242.7421	240.3515	245.1342	94.99%
230.4737	230.4741	228.2630	232.6822	94.75%
218.7811	218.7802	215.6937	221.8664	94.37%
207.8232	207.8239	204.8009	210.8469	94.78%
212.1718	212.1729	208.0538	216.2923	94.54%

Comments on Deep Ensemble

Deep ensemble is an useful tool for estimating uncertainty but it comes at a price – high demand for computing and memory (similar to BNN you will see later).



The trade-off between the computational cost and the quality of uncertainty estimation is hard to be generalized and should be addressed for your own use cases.

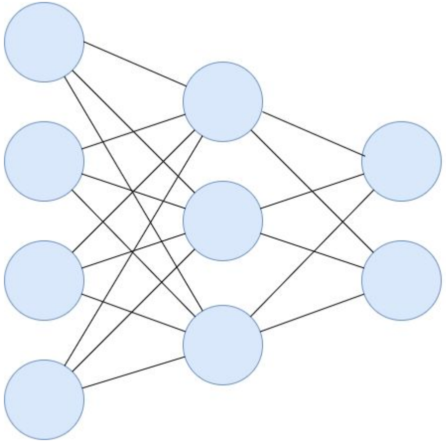
There are active ongoing developments of practical and general UQ methods in AI/ML.

Credit: Lakshminarayanan (2022)

Method 5: Monte Carlo Dropout

What is dropout? Randomly drop some neurons in network - typically during training.

Regular Neural Network



Same network with two neurons "dropped out" (eliminated from network)

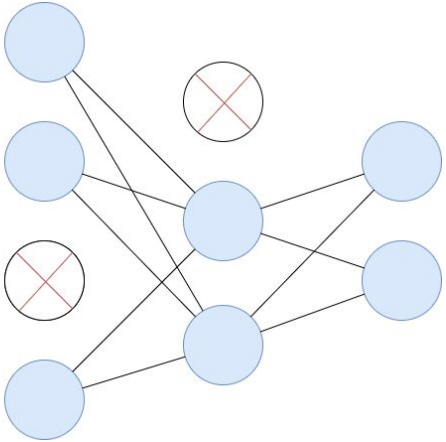


Image source: Michal Oleszak, **Monte Carlo Dropout**, *Towards data science*, Sep 20, 2020.
https://towardsdatascience.com/monte-carlo-dropout-7fd52f8b6571

Method 5: Monte Carlo Dropout

Standard use of dropout: Dropout used during training - to avoid overfitting

- Dropout is *usually* used only during NN training to avoid overfitting:
- Idea: during training randomly ignore neurons according to specified drop out probability:
    - neurons with drop-out rate=0 → never dropped
    - neurons with drop-out rate=0.5 → dropped about 50% of the time.
    - **Dropout rate** becomes an additional hyperparameter
- For each batch: decide which neurons are dropped. Different neurons active for each batch!
- NN is forced to learn to distribute signals across many neurons (redundancy), because it cannot rely on any neuron to be connected.
- Since different neurons are dropped in each batch, we effectively create an **ensemble**

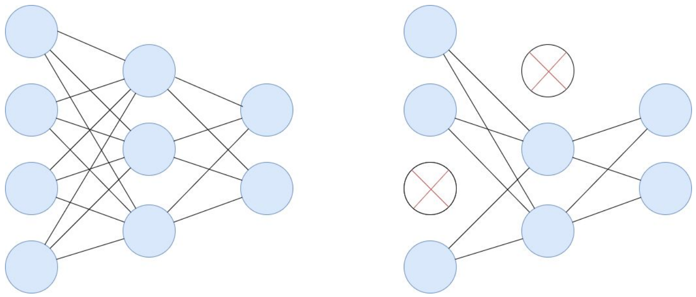


Image source: Michal Oleszak, **Monte Carlo Dropout**, *Towards data science*, Sep 20, 2020.
https://towardsdatascience.com/monte-carlo-dropout-7fd52f8b6571

Method 5: Monte Carlo Dropout

Different use - to obtain uncertainties: Drop-out during prediction (after having trained with dropout)

- Dropout can be interpreted as Bayesian approximation of Gaussian process.
- Idea: Use dropout when running the model to generate predictions
  ⟶ each dropout version provides a different NN model
  ⟶ provides ensemble of NN models
  ⟶ ensemble of predictions ⟶ can get uncertainty estimate from that ensemble.
- Loss function: Best coupled with using special loss function during training, see next slide.
- *It was shown that MC Dropout can be interpreted as a special case of Bayesian inference - although originally it was not derived as such. See:*
  - Gal, Yarin, and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning." International conference on machine learning. PMLR, 2016.

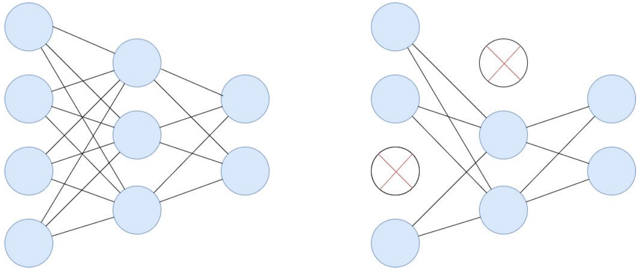
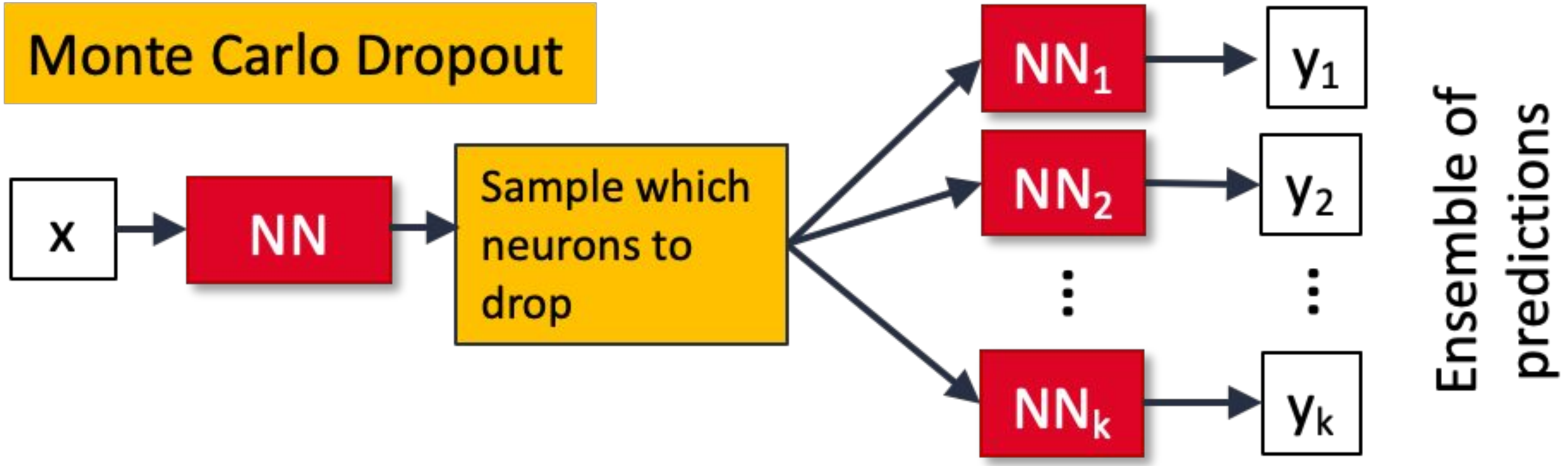


Image source: Michal Oleszak, **Monte Carlo Dropout**, *Towards data science*, Sep 20, 2020.
https://towardsdatascience.com/monte-carlo-dropout-7fd52f8b6571

Method 5: Monte Carlo Dropout



Pros:

- Extremely easy to implement:
  just add dropout layers to NN and make sure they stay on during inference.

Cons:

- Slow at inference time.
- We have not found them to give great results for our applications.

Method 6: Bayesian Neural Network

Standard (deterministic) NN:

- Weights & biases are parameters to be learned;
- Activation functions are fixed (pre-selected).

Bayesian Neural Networks = Neural Networks where

- either weights & biases,
- or activation functions

in the layers are probabilistic.

Most common BNN type - we will only focus on this type here:

- **Activation functions are fixed;**
- **Weights and biases are modeled as probabilistic.**

Method 6: Bayesian Neural Network

Standard (Deterministic) Neural Network:
all weights and biases are scalars to be learned

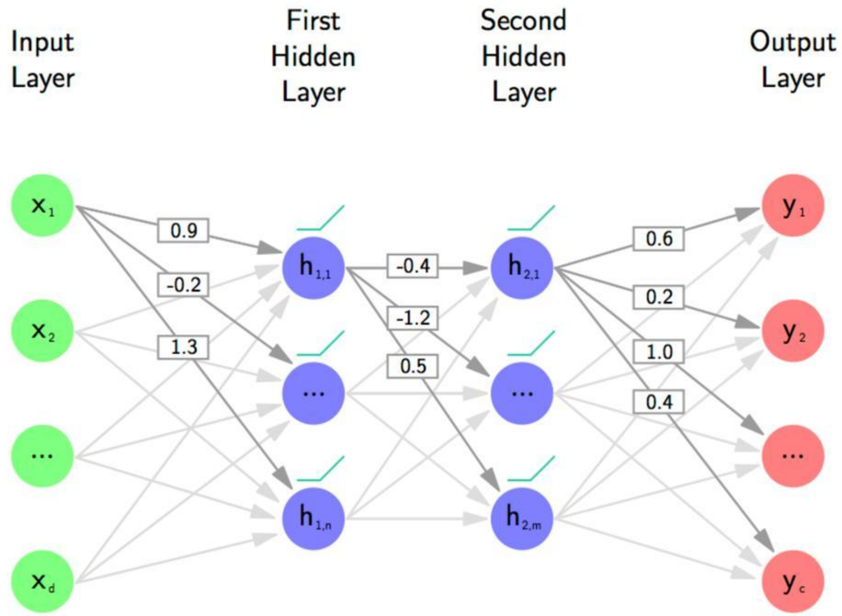


Image credit: Gluon Educational Resources, Chapter 18 on Variational methods and uncertainty.
https://gluon.mxnet.io/chapter18_variational-methods-and-uncertainty/bayes-by-backprop.html

Method 6: Bayesian Neural Network

Bayesian Neural Network:

- **Probabilistic layers:**
  all weights and biases are probability distributions to be learned.

- Shown here:
  - Normal distribution (mu, sigma) assumed for each weight.
  - But does not have to be Gaussian.

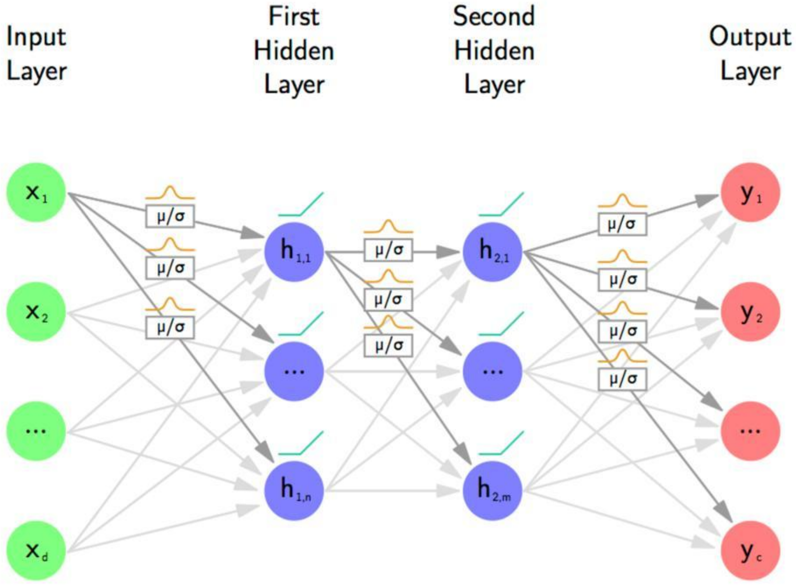


Image credit: Gluon Educational Resources, Chapter 18 on Variational methods and uncertainty.
https://gluon.mxnet.io/chapter18_variational-methods-and-uncertainty/bayes-by-backprop.html

Method 6: BNNs - implementation

- TensorFlow's probability (TFP) library provides probabilistic layers to implement BNNs.

- **In theory, we can just replace deterministic layers by probabilistic layers using TFP. But in practice it's not that easy:**
    - Large memory requirements;
    - Large computational requirements for training BNNs.

- Experience by other groups working in environmental science:
    - Many research groups report that they can only implement 2-3 probabilistic layers before running out of memory.
    - **Success story:** One group successfully **converted all layers of a deep, complex NN** into probabilistic layers, using their own implementation.

  Which group and application could it be?
  Hint: You've already seen it mentioned in this presentation.

BNNs for predicting precipitation from satellite imagery

Orescanin, M., Petković, V., Powell, S.W., Marsh, B.R. and Heslin, S.C., 2021. Bayesian Deep Learning for Passive Microwave Precipitation Type Detection. IEEE Geoscience and Remote Sensing Letters.

Method used:

- Bayesian neural network - making all NN weights probabilistic.

Application:

- **Classify precipitation type (stratiform or convective) based on passive MW imagery**
- Namely, map from raw GMI data to precipitation type (stratiform/convective)
- Goal: provide two outputs:
    i. Map of precipitation type: indicates stratiform/convective per pixel.
    ii. Map of uncertainty: indicates how much to trust classification per pixel.

Set-up:

- 14 million samples available for training and testing.
- Ground truth (labels): obtained from dual-frequency precipitation radar (DPR)
- Input: passive Microwave Imagery (GMI)

BNNs for predicting precipitation from satellite imagery

Approach:

- NN architecture: CNN of type ResNet
- **Baseline model: <u>deterministic</u> ResNet**
  Comment: ResNet is a deep network with lots of parameters.
- **New model: <u>probabilistic</u> ResNet**
  - Turn *all* layers of ResNet into Bayesian layers:
    → all weights of all layers are modeled as Gaussian distribution.
    → That doubles # of parameters:
      each weight, *w,* is replaced by two parameters: (*mu, sigma*)
- **Key comments:**
  - Implementation: Implemented in TensorFlow probability.
  - Impressive implementation: they manage to turn *all* layers of ResNet into Bayesian layers, without running out of memory!
  - But still needs lots of data (they have 14M samples!)

BNNs for predicting precipitation from satellite imagery

<u>Deterministic</u> ResNet: 86% accuracy, no uncertainty estimate.

<u>Probabilistic</u> ResNet: 90% accuracy, and yields uncertainty estimate.

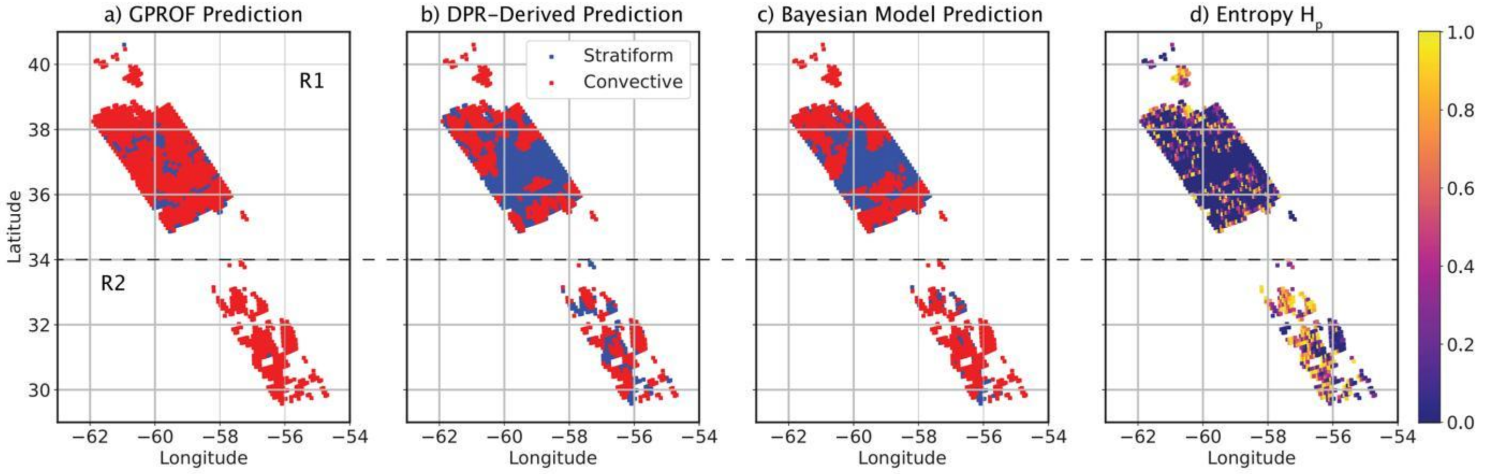
high entropy =
high uncertainty

Classification from baseline operational algorithm

Ground truth (label from DPR)

Classification from Bayesian NN

Uncertainty estimate from Bayesian NN



Orescanin et al., Bayesian Deep Learning for Passive Microwave Precipitation Type Detection, 2021.

Other success stories in environmental science

- Vandal *et al.* (2018): "**Quantifying uncertainty in discrete-continuous and skewed data with Bayesian deep learning**"
    - Successfully use Bayesian deep learning for precipitation
- Clare *et al.* (2021): "**Combining distribution-based neural networks to predict weather forecast probabilities**"
    - Turn regression problem into multi-class classification, by splitting continuous var into bins
    - Then use neural net with softmax activation
- Scher and Messori (2021): "**Ensemble methods for neural-network-based weather forecasts**"
    - Test 4 ensemble methods: random initial perturbations, PCA-based perturbations, retraining many times, Monte Carlo dropout
- Foster *et al.* (2021): "**Probabilistic machine learning estimation of ocean mixed layer depth from dense satellite and sparse in-situ observations**"
    - Test 4 ML methods with UQ – including Monte Carlo dropout, predicting mean and variance, deep ensemble
- Ortiz *et al.* (2022): "**Decomposing satellite-based classification uncertainties in large earth-science datasets**"
    - Decompose uncertainty into aleatoric and epistemic components
    - This decomp helps users make informed decisions about high-uncertainty cases (*e.g.,* need to collect more data vs. augment existing data)
- Chapman *et al.* (2022): "**Probabilistic predictions from deterministic atmospheric river forecasts with deep learning**"
    - Compare dynamical ensemble vs. neural networks vs. analogue ensemble for UQ. Find that NNs have many advantages.

Draft paper and Jupyter notebooks from our team for your use

- Draft paper (49 pages) - includes most of the UQ methods and eval tools discussed here.
- Notebook for CRPS loss function
- Notebook demonstrating wide variety of UQ methods and evaluation tools
  - Application: regression task with 6 synthetic datasets
  - UQ methods: CRPS loss function, Monte Carlo dropout, parametric prediction
  - Evaluation tools: attributes diagram, spread-skill plot, PIT histogram, discard test
- Monte Carlo dropout
  - Application: classifying hand-written digits
  - Includes spread-skill plot, discard test, and case studies
- Same as above but for quantile regression

Ryan
Lagerquist
(CSU)

Katherine
Haynes
(CSU)

Big thanks to **Ryan Lagerquist** and **Katherine Haynes** for creating these notebooks.

Other Suggested Reading

Resources that we found particularly helpful as entry point:

1. Dürr, O., Sick, B. and Murina, E., 2020. **Probabilistic deep learning**: With python, keras and tensorflow probability. Manning Publications. (book)
2. Blog comparing different UQ methods including MC Dropout, Deep Ensemble, GPR, Qunatile Regression - https://www.inovex.de/de/blog/uncertainty-quantification-deep-learning/ (blog post)
3. Dr. Steven Brey's detailed explanation of a probabilistic implementation of the CRPS - https://github.com/TheClimateCorporation/ensemble/blob/main/notebooks/intro_to_probabilistic_predictions.ipynb (Github)
4. Review of UQ in Deep Learning - https://doi.org/10.1016/j.inffus.2021.05.008 (article)
5. Valentin Jospin, L., Buntine, W., Boussaid, F., Laga, H. and Bennamoun, M. **Hands-on Bayesian Neural Networks - a Tutorial for Deep Learning Users**, arXiv preprint, v2, Sept 2021, https://arxiv.org/abs/2007.06823 (article)
6. Chang, D.T. , **Bayesian Neural Networks: Essentials.** arXiv preprint, v1, June 2021, https://arxiv.org/abs/2106.13594 (article)
7. Ortiz, P., Orescanin, M., Petković, V., Powell, S.W. and Marsh, B., 2022. **Decomposing Satellite-Based Classification Uncertainties in Large Earth Science Datasets**. IEEE Transactions on Geoscience and Remote Sensing, 60, pp.1-11. (article)

Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School

Day 4: Agenda

- 9:00 Uncertainty quantification methods (Part 1)
- 10:00 *Short brain & bio break*
- 10:10 Uncertainty quantification methods (Part 2)
- **10:45 *Short brain & bio break***
- 10:55 Communicating uncertainty (Part 3)
- 11:55 Lecture series wrap up!

Questions?

https://app.sli.do/event/1zumy91n

Or go to sli.do and use the code TAI4ES

Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School

Day 4: Agenda

- 9:00 Uncertainty quantification methods (Part 1)
- 10:00 *Short brain & bio break*
- 10:10 Uncertainty quantification methods (Part 2)
- 10:45 *Short brain & bio break*
- **11:00 Communicating uncertainty (Part 3)**
- 11:55 Lecture series wrap up!

Questions?

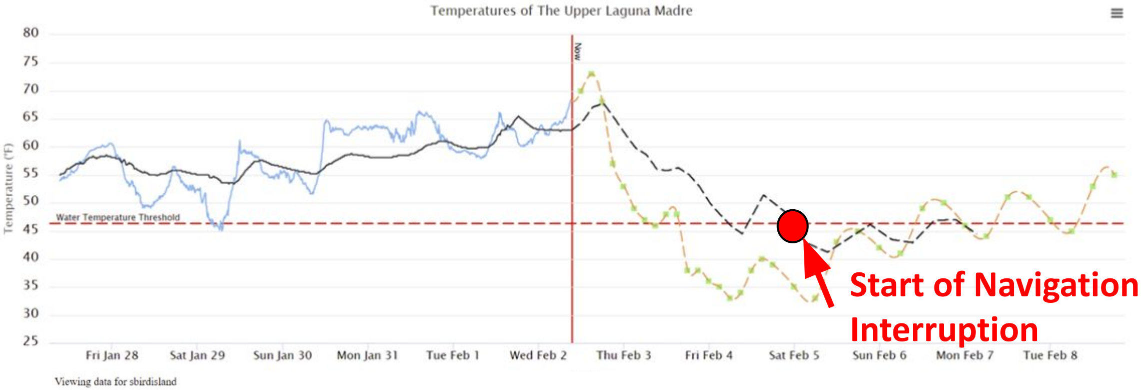
https://app.sli.do/event/1zumy91n

Or go to sli.do and use the code TAI4ES

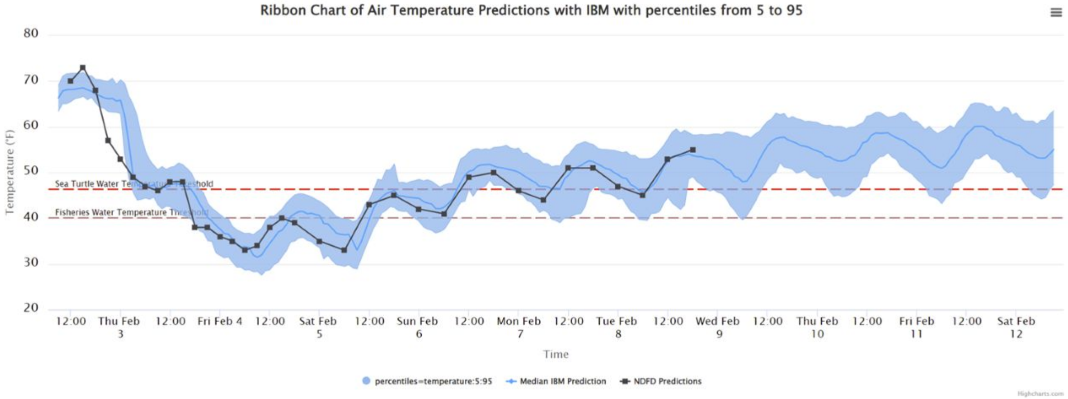
Part 3: Risk communication and Uncertainty

Uncertainty: Cold Stunning Predictions

- Water temperature below 8C for ~24 hrs leads to sea turtle cold stunnings
- AI (shallow neural nets) used since 2008 to predict onset and duration of cold stunnings (black dash line)
- AI Predictions allow for interruption of navigation, staging of resources, …
- Here, example for Feb 2022 cold stunning predictions (400+ sea turtles)



Research: IBM/AI2ES providing ensemble air temperature predictions (right)
(1) Create ensemble ANN predictions
(2) Quantify uncertainty in AI temperature and threshold crossings predictions



ARTIFICIAL INTELLIGENCE · RISK COMMUNICATION · ENVIRONMENTAL SCIENCE

How can we best quantify, visualize, communicate uncertainty?

4.6. Go to sli.do and use the code TAI4ES

Definition of Trust (Reminder from Monday)

- Trust is the **willingness of a party to be vulnerable to the actions of another** party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party. (e.g., Mayer et al 1995)
- Trust: In the presence of uncertainty, **the degree to which someone does or does not rely on**, or put faith in, someone or something (Wirz et al.)
  - Definition is purposefully broad, so as to capture the many different definitions and related dimensions of trust. Our definition of trust is designed to capture trust in all forms.
- Trust is the **relationship between a trustor and a trustee**: the trustor trusts the trustee. Trust is dynamic, evolves with interactions, and is easier to lose than gain.

AI2ES Definition: Trust is the willingness to assume risk by relying on or believing in the actions of another party.

Characterizing and communicating risk and uncertainty

“**Risk** is a situation or event where something of **human value** (including humans themselves) is at stake and where the **outcome is uncertain**.”

“**Risk** is an **uncertain consequence** of an event or an activity with respect to **something that humans value**.”

– *Aven and Renn, 2009 Journal of Risk Research*

Quick survey to help us think about “risk” in the context of AI

4.8. Go to sli.do and use the code TAI4ES

Risk perceptions: Psychometrics

What makes risk acceptable?

The risk factor space

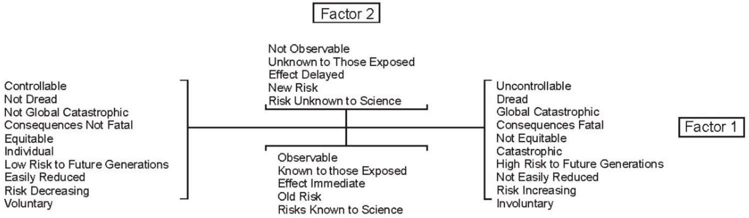
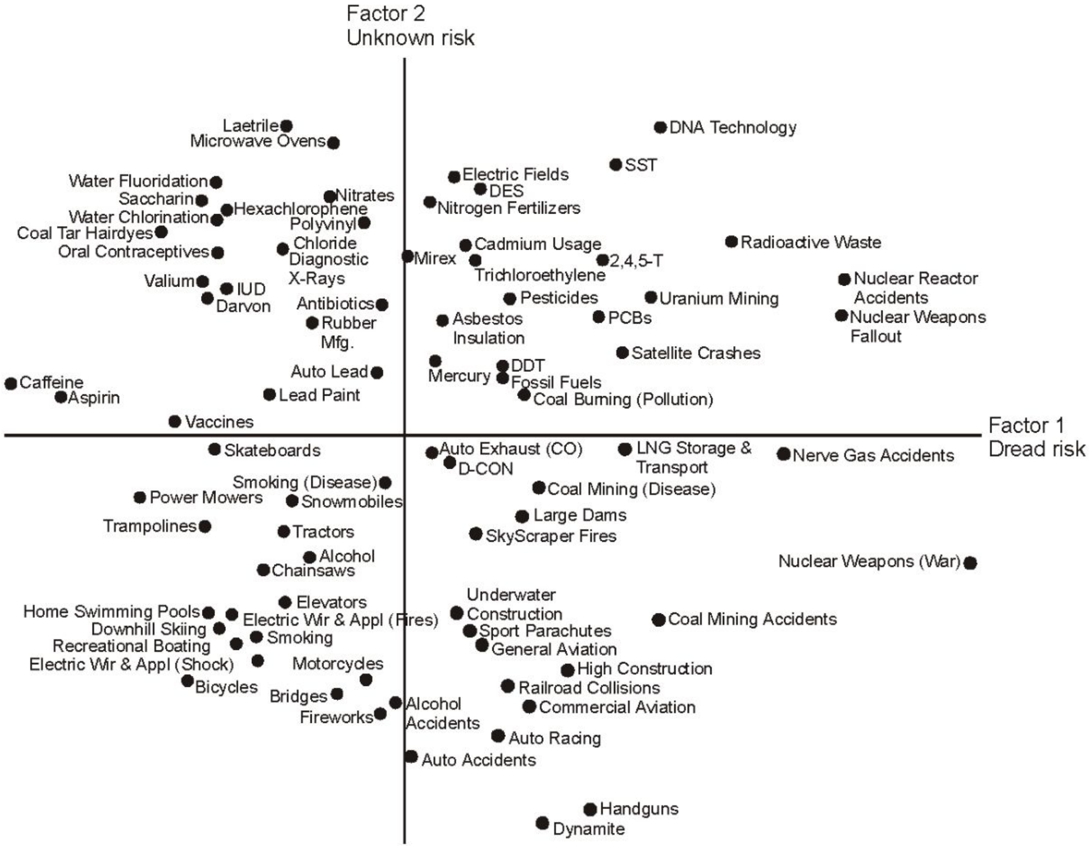
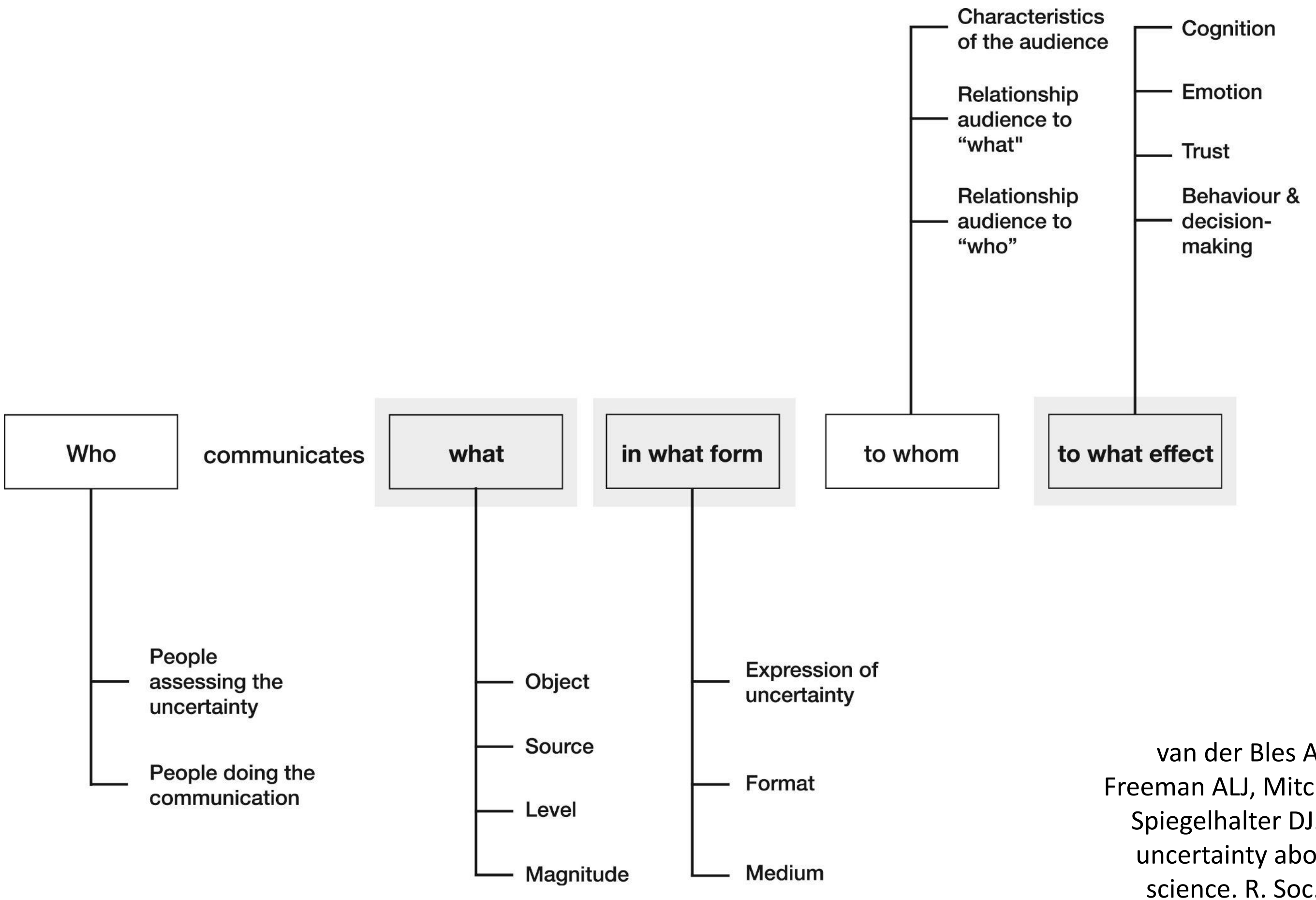


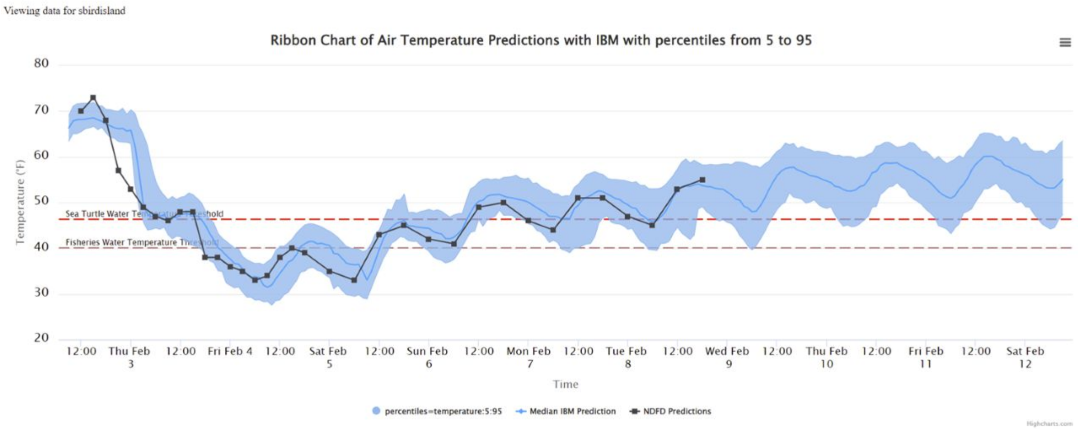
Figure 3. Location of 81 hazards on Factors 1 and 2 derived from the interrelationships among 15 risk characteristics. Each factor is made up of a combination of characteristics, as indicated by the lower diagram. Source: Slovic (1987).

Paul Slovic, Science, 1987





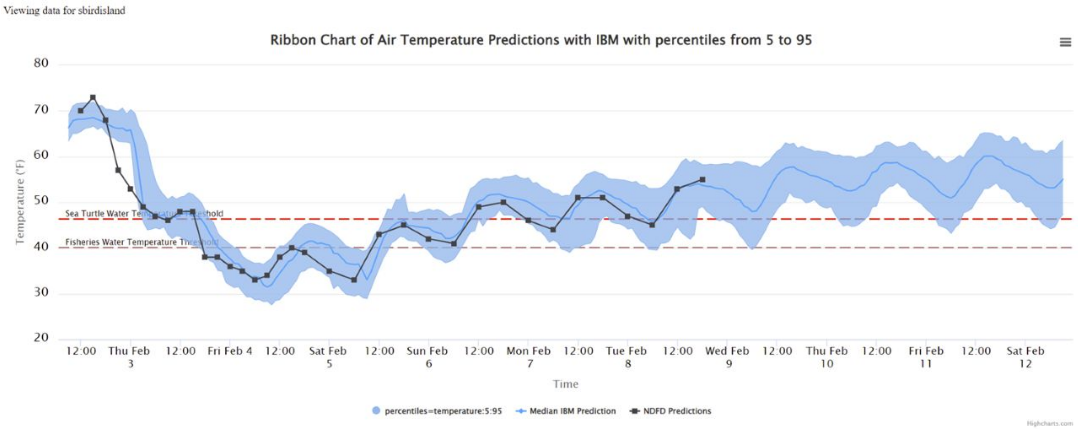
van der Bles AM, van der Linden S, Freeman ALJ, Mitchell J, Galvao AB, Zaval L, Spiegelhalter DJ. 2019 Communicating uncertainty about facts, numbers and science. R. Soc. open sci. 6: 181870.



Research: IBM/AI2ES providing ensemble air temperature predictions (left)
(1) Create ensemble ANN predictions
(2) Quantify uncertainty in AI temperature and threshold crossings predictions

Who communicates

- People assessing the uncertainty
- People doing the communication



Research: IBM/AI2ES providing ensemble air temperature predictions (left)
(1) Create ensemble ANN predictions
(2) Quantify uncertainty in AI temperature and threshold crossings predictions

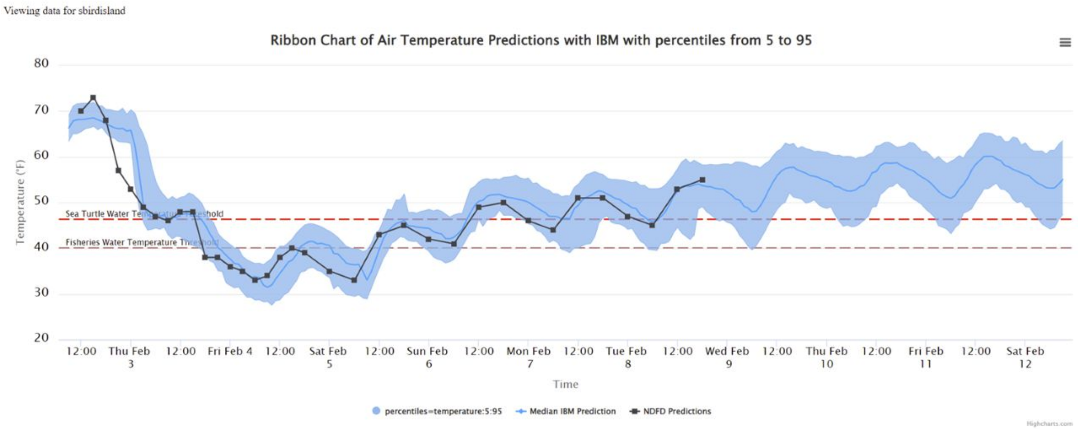
Who communicates

- People assessing the uncertainty
- People doing the communication

- Are their intentions good, and aligned with your best interests? (**value similarity**)
- Do they have the right expertise? (**competence**)

Top: Marshall Shepherd, University of Georgia Photographic Services
Bottom: Jeff Masters, Wunderground

(Earle, 2010; Oreskes & Conway, 2011; Freudenburg et al., 2008; Bloodhart et al., 2015)



Research: IBM/AI2ES providing ensemble air temperature predictions (left)
(1) Create ensemble ANN predictions
(2) Quantify uncertainty in AI temperature and threshold crossings predictions

Who communicates

- People assessing the uncertainty
- People doing the communication

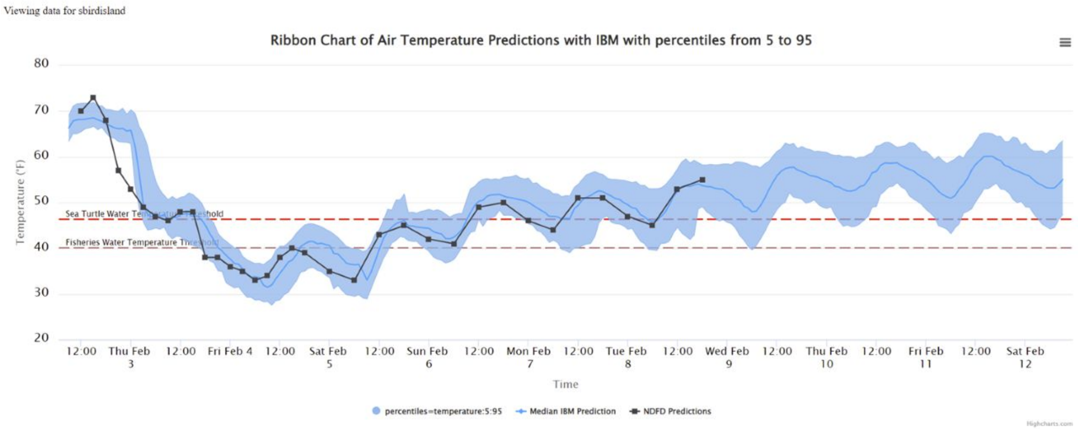
- Are their intentions good, and aligned with your best interests? (**value similarity**)
- Do they have the right expertise? (**competence**)

- Local weathercasters tend to be trusted for information about weather and climate.

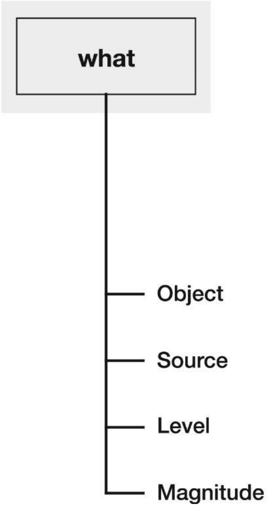
- Others (e.g., politicians) with unaligned interests may communicate uncertainty strategically: “merchants of doubt,” Scientific Certainty Argumentation Methods (SCAMs)

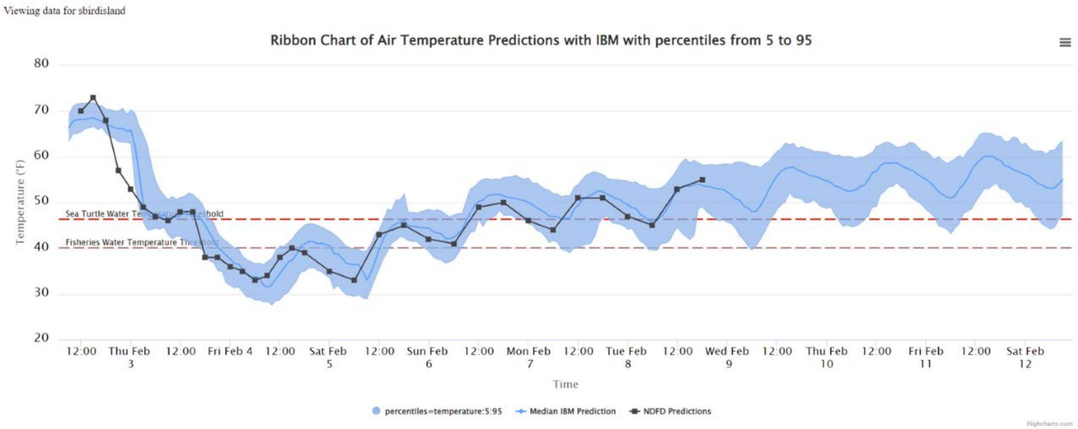
Top: Marshall Shepherd, University of Georgia Photographic Services
Bottom: Jeff Masters, Wunderground

(Earle, 2010; Oreskes & Conway, 2011; Freudenburg et al., 2008; Bloodhart et al., 2015)



Research: IBM/AI2ES providing ensemble air temperature predictions (left)
(1) Create ensemble ANN predictions
(2) Quantify uncertainty in AI temperature and threshold crossings predictions

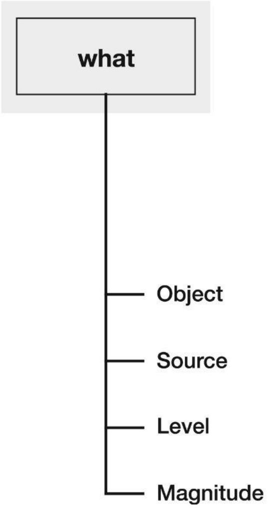




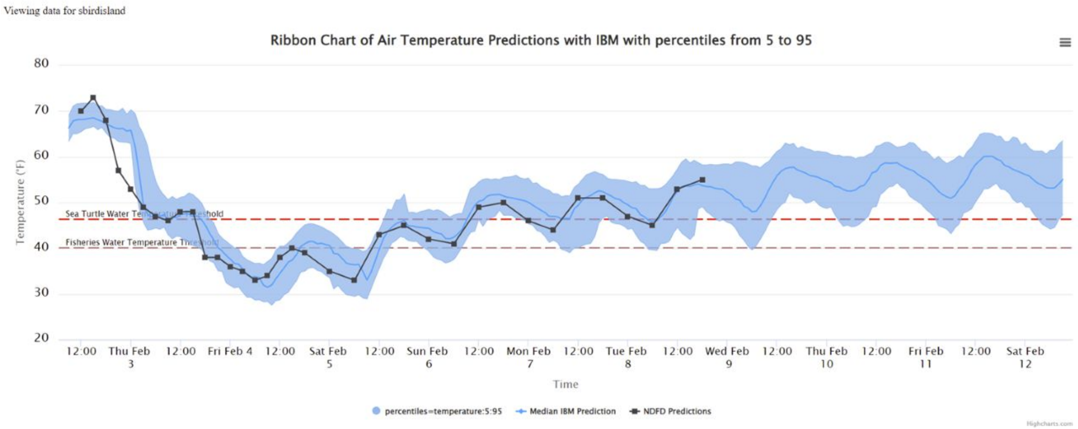
Research: IBM/AI2ES providing ensemble air temperature predictions (left)

(1) Create ensemble ANN predictions

(2) Quantify uncertainty in AI temperature and threshold crossings predictions

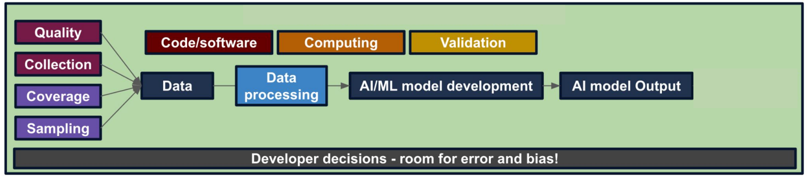


Facts, numbers, scientific models and hypotheses



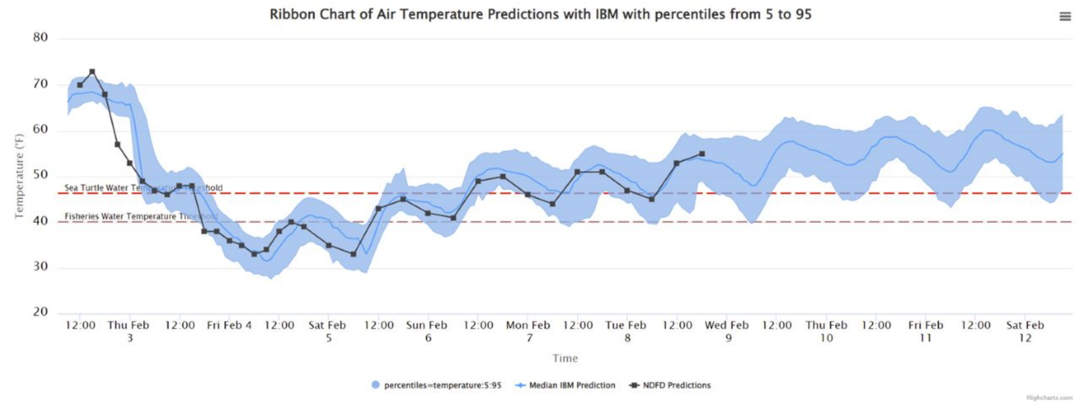
Research: IBM/AI2ES providing ensemble air temperature predictions (left)
(1) Create ensemble ANN predictions
(2) Quantify uncertainty in AI temperature and threshold crossings predictions

what



- **Object** — Facts, numbers, scientific models, and hypotheses
- **Source** — Data or model variability, biases, or other shortcomings
- **Level** — Direct uncertainties, as shown above
- **Magnitude**

Viewing data for sbirdisland



Research: IBM/AI2ES providing ensemble air temperature predictions (left)
(1) Create ensemble ANN predictions
(2) Quantify uncertainty in AI temperature and threshold crossings predictions

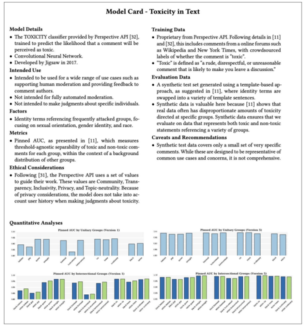
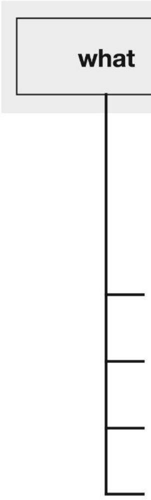


Figure 3: Example Model Card for two versions of Perspective API's toxicity detector.

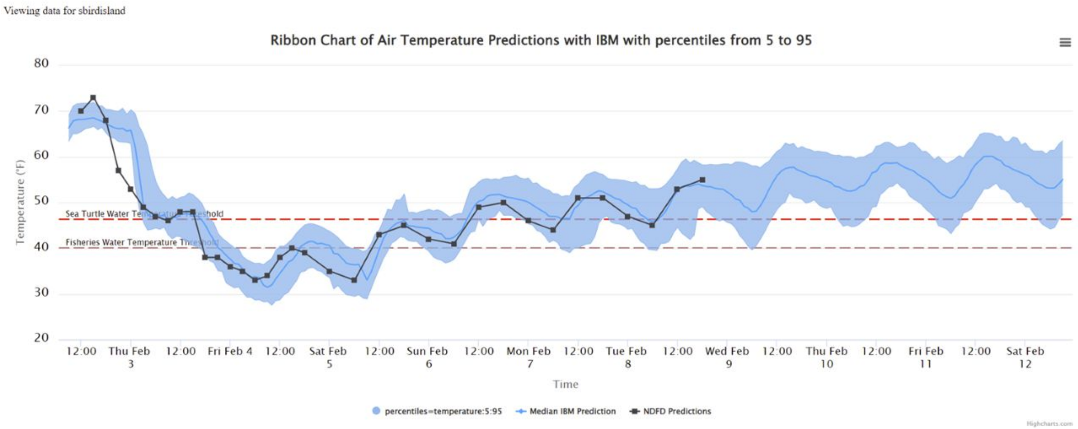


Object — Facts, numbers, scientific models, and hypotheses

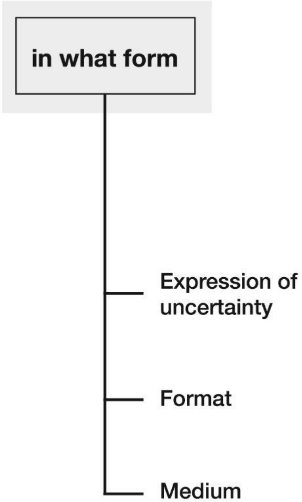
Source — Data or model variability, biases, or other shortcomings

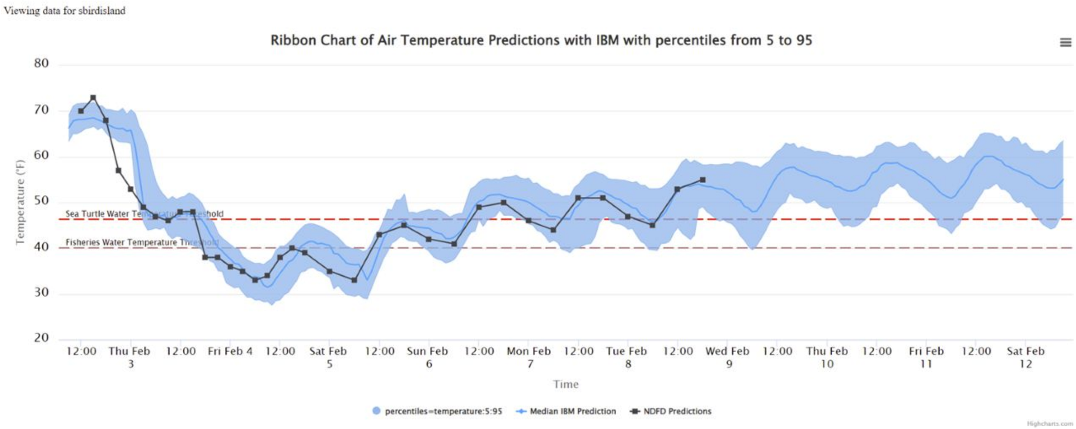
Level — Uncertainty can be direct, or indirect (e.g., quality of evidence)

Magnitude — How big the uncertainties are matter in decision making!

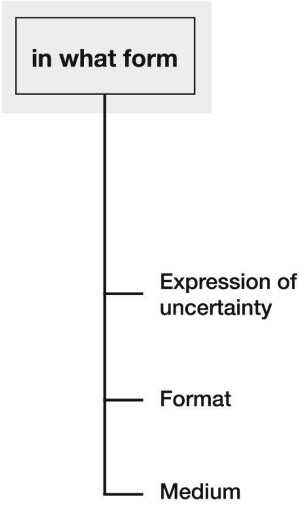


Research: IBM/AI2ES providing ensemble air temperature predictions (left)
(1) Create ensemble ANN predictions
(2) Quantify uncertainty in AI temperature and threshold crossings predictions





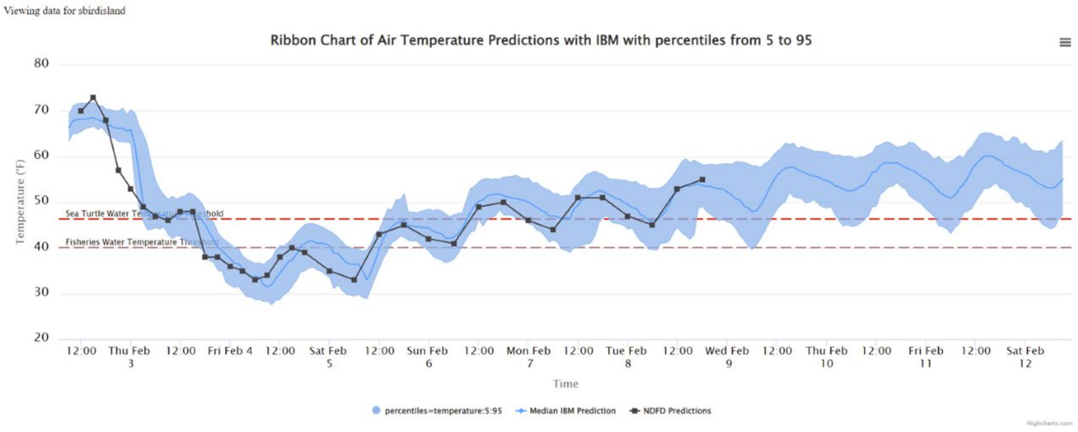
Research: IBM/AI2ES providing ensemble air temperature predictions (left)
(1) Create ensemble ANN predictions
(2) Quantify uncertainty in AI temperature and threshold crossings predictions



Communicating numerical risks:

- Use absolute risks (but also provide relative risks when dealing with potential catastrophic events).
- For single unique events, use percent chance if possible, or if necessary, "1 in X."
- When appropriate, express chance as a proportion, a frequency, or a percentage—it is crucial to be clear about the reference class.

(Spiegelhalter, 2017)



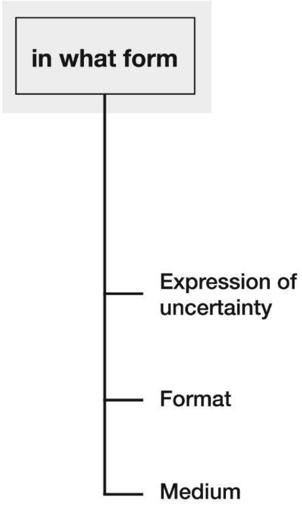
Research: IBM/AI2ES providing ensemble air temperature predictions (left)
(1) Create ensemble ANN predictions
(2) Quantify uncertainty in AI temperature and threshold crossings predictions

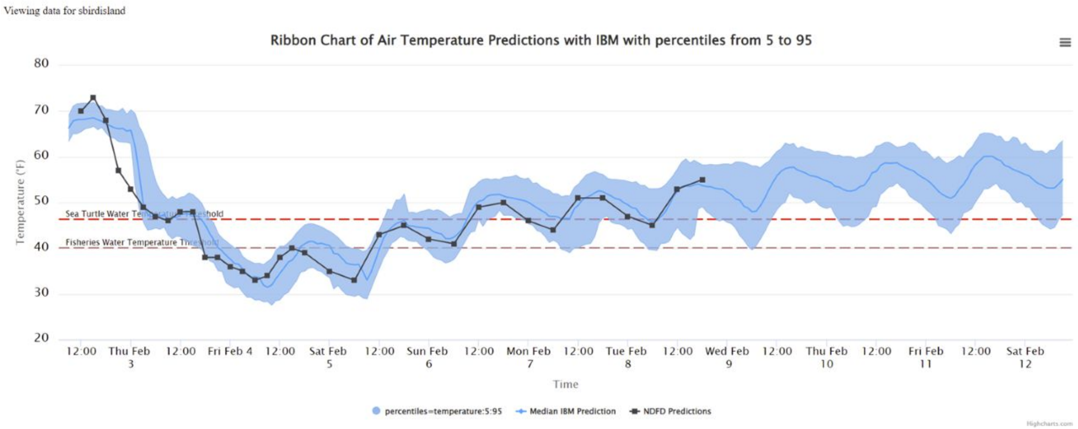
Communicating numerical risks:

- Use absolute risks (but also provide relative risks when dealing with potential catastrophic events).
- For single unique events, use percent chance if possible, or if necessary, “1 in X.”
- When appropriate, express chance as a proportion, a frequency, or a percentage—it is crucial to be clear about the reference class.

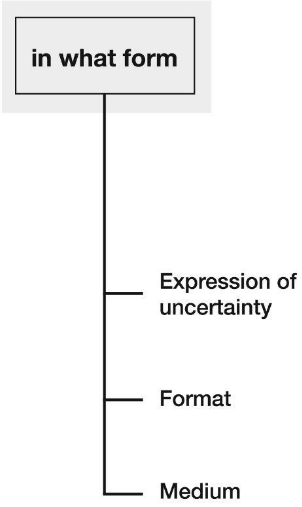
(Spiegelhalter, 2017)

“More important than the choice of format is **being absolutely clear as to what the probability actually means** (Morgan et al. 2009), which requires careful specification of the reference class (Gigerenzer & Galesic 2012).”
- Spiegelhalter, 2017 p 38





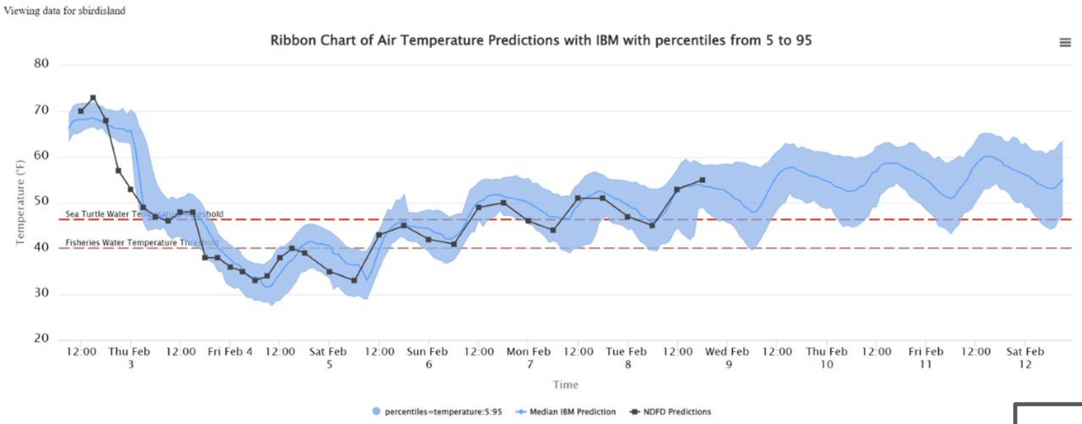
Research: IBM/AI2ES providing ensemble air temperature predictions (left)
(1) Create ensemble ANN predictions
(2) Quantify uncertainty in AI temperature and threshold crossings predictions



Communicating numerical risks (continued):

- To avoid framing bias, provide percentages or frequencies both with and without the outcome.
- Keep the denominator fixed when making comparisons with frequencies, and use an incremental risk format.
- Be explicit about the time interval.
- Be aware that comparators can create an emotional response.
- For more knowledgeable audiences, consider providing quantitative epistemic uncertainty about the numbers and qualitative assessment of confidence in the analysis.
- More sophisticated metrics can be made for technical audiences, but this only serves to exclude others.

(Spiegelhalter, 2017)



Research: IBM/AI2ES providing ensemble air temperature predictions (left)
(1) Create ensemble ANN predictions
(2) Quantify uncertainty in AI temperature and threshold crossings predictions

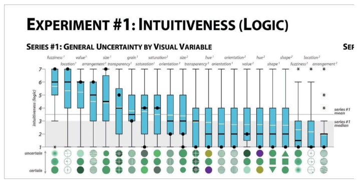
Decreasing precision

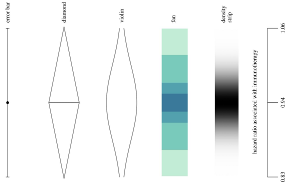
i. A full explicit probability distribution
ii. A summary of a distribution
iii. A rounded number, range or an order-of-magnitude assessment
iv. A predefined categorisation of uncertainty
v. A qualifying verbal statement
vi. A list of possibilities or scenarios
vii. Informally mentioning the existence of uncertainty
viii. No mention of uncertainty
ix. Explicit denial that uncertainty exists

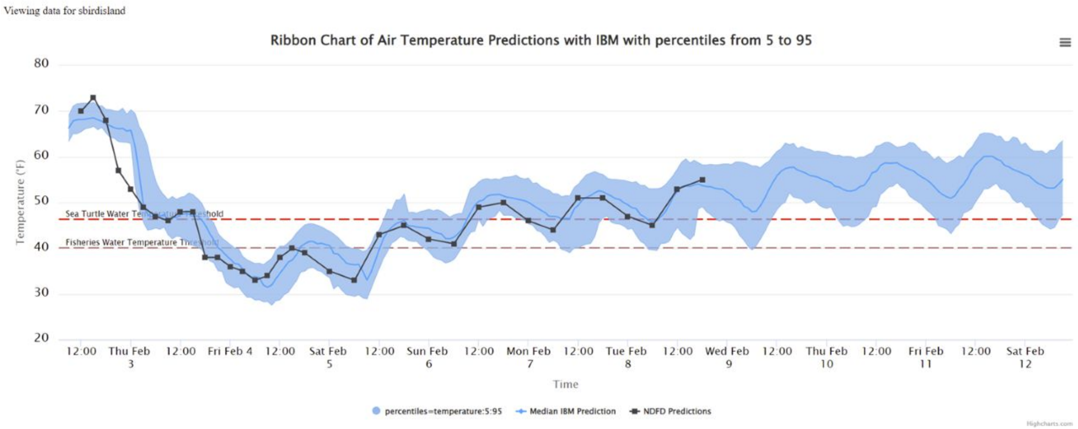
Alternative expressions for communicating direct uncertainty about a fact, number or scientific hypothesis.

in what form

- Expression of uncertainty
- Format
- Medium







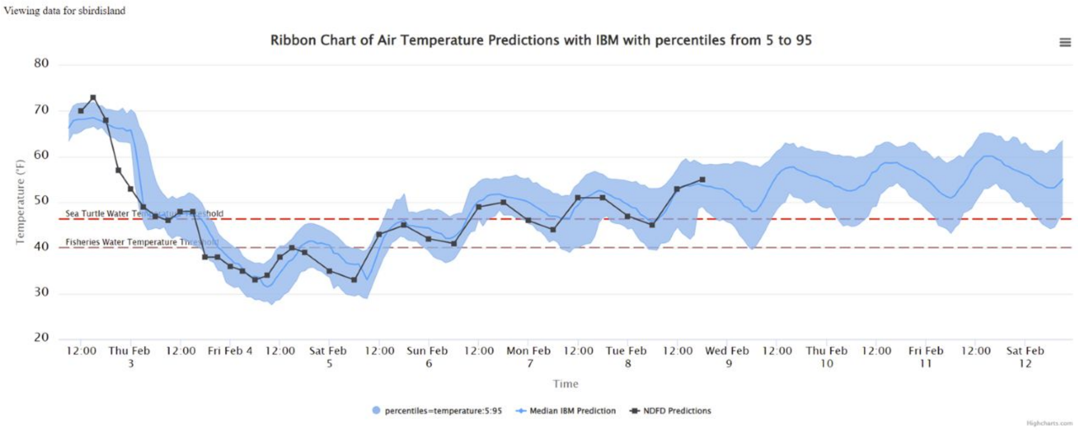
Research: IBM/AI2ES providing ensemble air temperature predictions (above)
(1) Create ensemble ANN predictions
(2) Quantify uncertainty in AI temperature and threshold crossings predictions

Characteristics of the audience

Relationship audience to "what"

Relationship audience to "who"

to whom



Research: IBM/AI2ES providing ensemble air temperature predictions (above)
(1) Create ensemble ANN predictions
(2) Quantify uncertainty in AI temperature and threshold crossings predictions

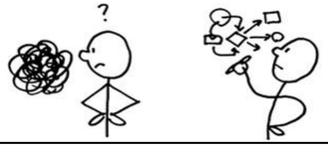
- Numeracy

Characteristics of the audience

Relationship audience to "what"

Relationship audience to "who"

to whom

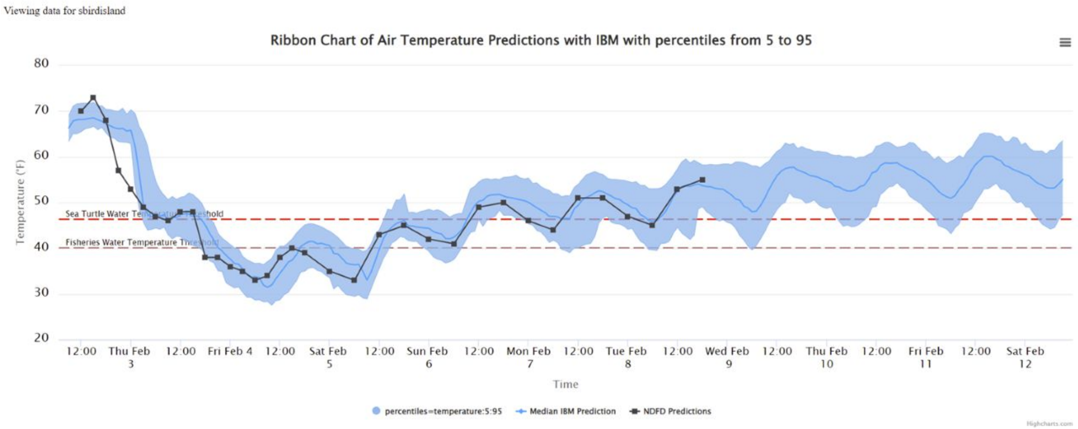


Which is larger?
1/100,000
1/10,000

Peters et al. (2007) found that ~10% of respondents had trouble with this kind of comparison.

Innumeracy increases use of heuristics (rules of thumb, mental shortcuts)

(Peters et al., 2007; Peters & Levin, 2008; Reyna et al., 2009)



- Numeracy
- Graphicacy

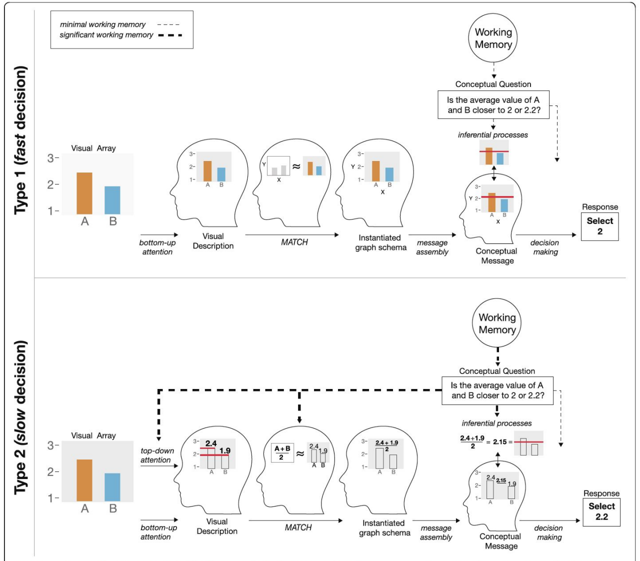
Characteristics of the audience

Relationship audience to “what"

Relationship audience to “who”

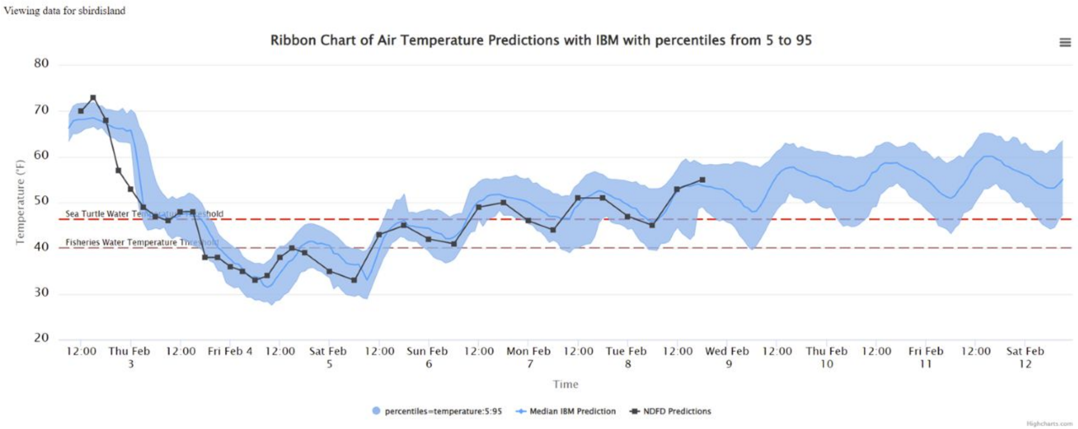
to whom

Research: IBM/AI2ES providing ensemble air temperature predictions (above)
(1) Create ensemble ANN predictions
(2) Quantify uncertainty in AI temperature and threshold crossings predictions



Padilla et al., 2018

Fig. 5 Examples of a fast Type 1 (top) and slow Type 2 (bottom) decision outlined in our proposed model of decision making with visualizations. In these examples, the viewer's task is to decide if the average value of bars A and B are closer to 2 or 2.2. The thick dotted line denotes significant working memory and the thin dotted line negligible working memory



Research: IBM/AI2ES providing ensemble air temperature predictions (above)
(1) Create ensemble ANN predictions
(2) Quantify uncertainty in AI temperature and threshold crossings predictions

to whom

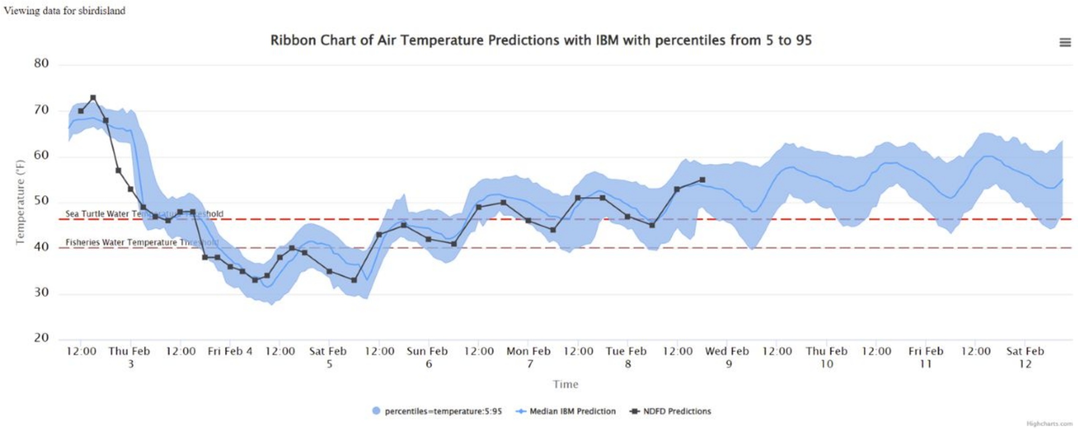
- Characteristics of the audience
  - Numeracy
  - Graphicacy
  - Mental models
- Relationship audience to "what"
- Relationship audience to "who"

WFO forecaster in 2009 on hurricanes impacting Miami/Dade:

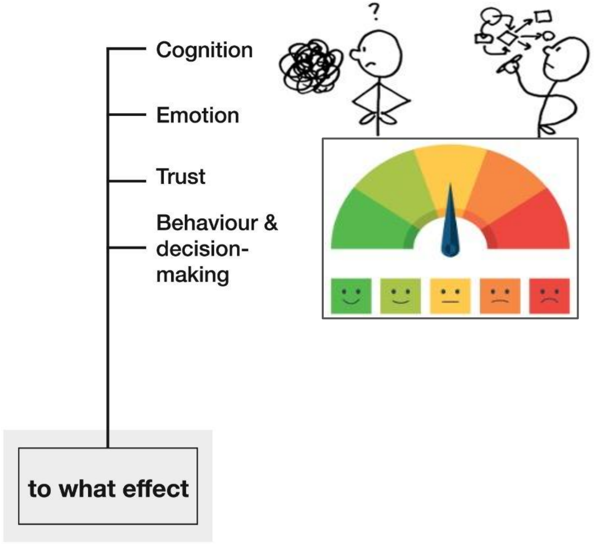
"...at least the way that I envision it, if I was going out on the street and told somebody that **a hurricane was coming**, the first thing they'd want to know is **how strong it is**. And **if I told them it was going to produce 15- to 20-inches of rain, that probably wouldn't answer their question.** They want to know **how strong it is, based on the winds**. If I say it has winds of 80 miles per hour, that may not be as bad. If I say it has winds of 180 miles an hour, that would probably really scare a lot of people. I think that, especially in this area, **people use Hurricane Andrew as a benchmark, being a Category 5 hurricane**. And if the winds are forecasted to be less than that, people may not be as concerned as if the winds were supposed to be as strong as a comparable Category 5 hurricane. "

(unpublished quote, see Bostrom et al. 2016)

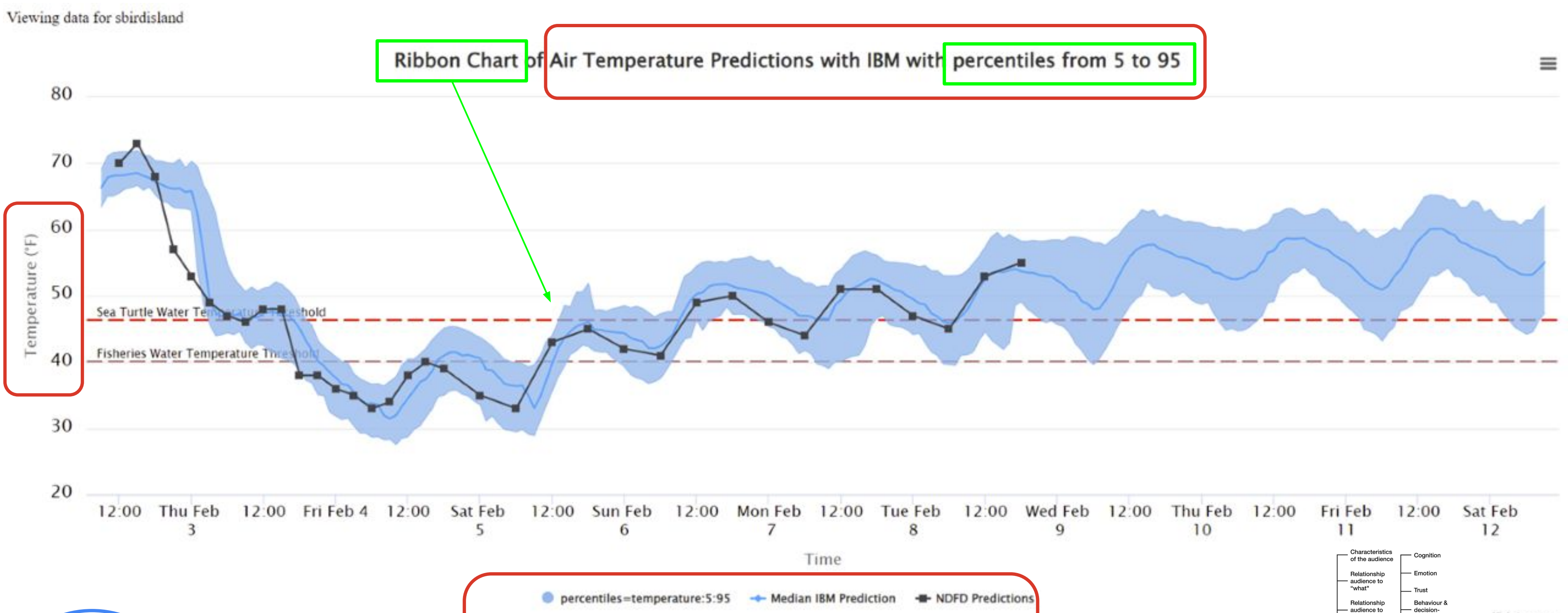
ARTIFICIAL INTELLIGENCE · RISK COMMUNICATION · ENVIRONMENTAL SCIENCE



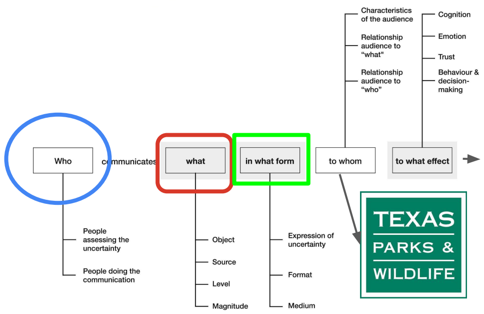
Research: IBM/AI2ES providing ensemble air temperature predictions (above)
(1) Create ensemble ANN predictions
(2) Quantify uncertainty in AI temperature and threshold crossings predictions



- Communicating uncertainty can increase trust in the information, affect attitudes toward the messenger, and may sometimes delay decision making - more research needed!
- Align the format with the decision, for example:
  5% of models predict a temperature below the sea turtle temperature threshold
- And test your communications!



Research:
IBM/AI2ES providing ensemble air temperature predictions (above)
(1) Create ensemble ANN predictions
(2) Quantify uncertainty in AI temperature and threshold crossings predictions



Expressing risk and uncertainty in words

- "verbal phrases are *vague*, in the sense that the same term (e.g., a chance, not certain) can be used to characterize a whole range of numerical probabilities" *-Budescu & Wallsten, 1995*

Table 1. A Taxonomy of Risk Concepts

Risk Concept	Sample Cognition	Distinguishable From	Level of Precision	Evaluability	Illustrative Gist Statements	Illustrative Emotional Meanings
Possibility	Might happen, might not	Will/will not	Minimal	Very high	"It could happen to me."	"I am at risk." *(Implies negative feelings if for a bad outcome)*
Relative/Comparative Possibility	More likely	Less likely/equally likely	Vague	High	"It is more likely to happen to me than to others." "I am more likely to have this happen to me than to have that happen to me."	"I have a worse risk than others." "I have a worse risk of this than that."
Categorical Possibility	High chance	Normal/average	Defined by categories	Depends on categories	"I am a person who has a high chance of this happening."	"I have a bad risk."
Relative Probability	50% more likely	Other ratios, e.g., 40% more likely	Ratio only	High for ratio, low for meaning	"I have a risk that is higher to this degree."	"I have a worse risk than others."
Absolute Probability	12%	Other probabilities, e.g., 13%	Level	Low	"My risk is this."	*Unclear without background knowledge*
Comparative Probability	12% vs. 8%	Other combinations, e.g., 15% vs. 10%, 12% vs. 11%	Level, with ratio by calculation	High	"My (group's) risk is this, which is higher than another's (group's) risk." "My risk is this if I do X, which is higher than my risk if I do Y which is that."	"My risk is worse than their risk is." "My risk is bad and worse if I do X."
Incremental Probability	4% more likely	Other increments, e.g., 5% more likely	Change in level	High for difference	"My risk will change that much if I do this."	"My risk will change a lot (or a little)." *(Affect depends on comparison to baseline)*

- Zikmund-Fisher, 2013

Expressing risk and uncertainty in words

- "verbal phrases are *vague*, in the sense that the same term (e.g., a chance, not certain) can be used to characterize a whole range of numerical probabilities" *-Budescu & Wallsten, 1995*
- Semantic and pragmatic implications of these expressions include: **hedges** (maybe), **outcome valence** (risk, hope), and **directionality** (occurrence or non-occurrence of a target event).

Table 1. A Taxonomy of Risk Concepts

Risk Concept	Sample Cognition	Distinguishable From	Level of Precision	Evaluability	Illustrative Gist Statements	Illustrative Emotional Meanings
Possibility	Might happen, might not	Will/will not	Minimal	Very high	"It could happen to me."	"I am at risk." *(Implies negative feelings if for a bad outcome)*
Relative/Comparative Possibility	More likely	Less likely/equally likely	Vague	High	"It is more likely to happen to me than to others." "I am more likely to have this happen to me than to have that happen to me."	"I have a worse risk than others." "I have a worse risk of this than that."
Categorical Possibility	High chance	Normal/average	Defined by categories	Depends on categories	"I am a person who has a high chance of this happening."	"I have a bad risk."
Relative Probability	50% more likely	Other ratios, e.g., 40% more likely	Ratio only	High for ratio, low for meaning	"I have a risk that is higher to this degree."	"I have a worse risk than others."
Absolute Probability	12%	Other probabilities, e.g., 13%	Level	Low	"My risk is this."	*Unclear without background knowledge*
Comparative Probability	12% vs. 8%	Other combinations, e.g., 15% vs. 10%, 12% vs. 11%	Level, with ratio by calculation	High	"My (group's) risk is this, which is higher than another's (group's) risk." "My risk is this if I do X, which is higher than my risk if I do Y which is that."	"My risk is worse than their risk is." "My risk is bad and worse if I do X."
Incremental Probability	4% more likely	Other increments, e.g., 5% more likely	Change in level	High for difference	"My risk will change that much if I do this."	"My risk will change a lot (or a little)." *(Affect depends on comparison to baseline)*

- Zikmund-Fisher, 2013

Communicating uncertainty

vs.

confidence

(*a*)

term*	**likelihood of the outcome**
virtually certain	99–100% probability
very likely	99–100% probability
likely	66–100% probability
about as likely as not	33–66% probability
unlikely	0–33% probability
very unlikely	0–10% probability
exceptionally unlikely	0–1% probability

*additional terms (*extremely likely*: 95–100% probability, *more likely than not*: >50–100% probability, and *extremely unlikely*: 0–5% probability) may also be used when appropriate.

(*b*)

agreement ↑

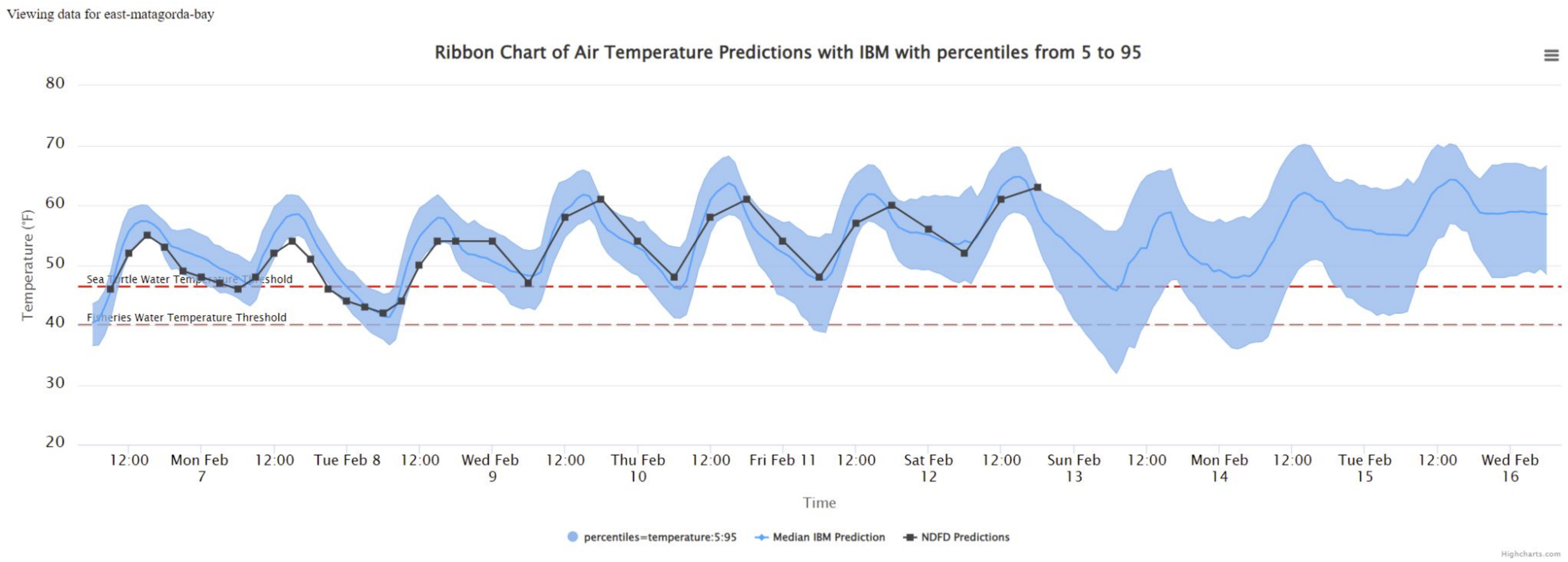
high agreement limited evidence	*high agreement medium evidence*	*high agreement robust evidence*
medium agreement limited evidence	*medium agreement medium evidence*	*medium agreement robust evidence*
low agreement limited evidence	*low agreement medium evidence*	*low agreement robust evidence*

evidence (type, amount, quality, consistency) →

confidence scale

Van der Bles et al 2018 (IPCC WG1 AR5, 2013)

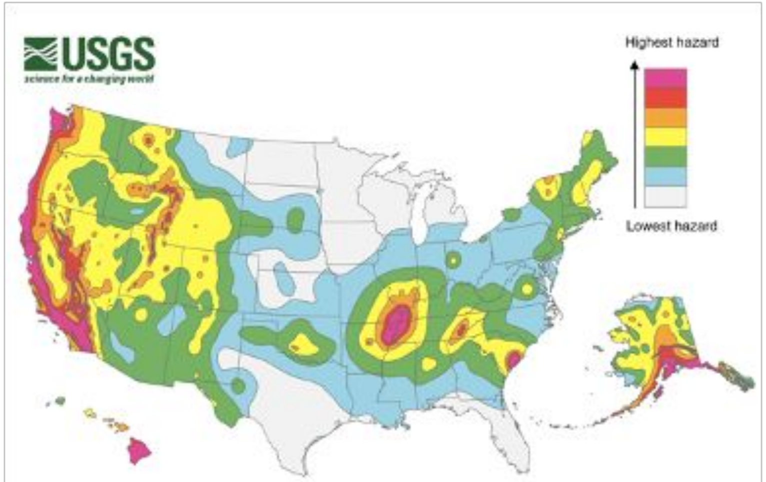
4.9. Go to sli.do and use the code TAI4ES



Pretend this is your model and it's Saturday February 12th and you the Port authority is asking you "will the temperature go below our sea turtle water temp threshold (the top red line) and how sure are you?"

How would you communicate certainty and your confidence?

Communicating uncertainty visually: Data classification in maps

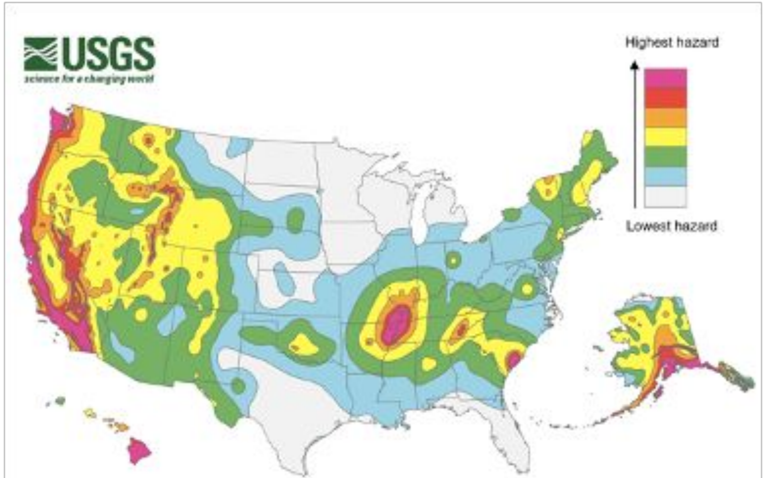


If and how data are binned:

“Our operating assumption is that everything west of Interstate 5 will be toast.” - The Really Big One, New Yorker

“An unfocused unclassed map is a more accurate representation of the risk data than a focused classed map.” - Severtson et al 2013, p 813

Communicating uncertainty visually: Data classification in maps

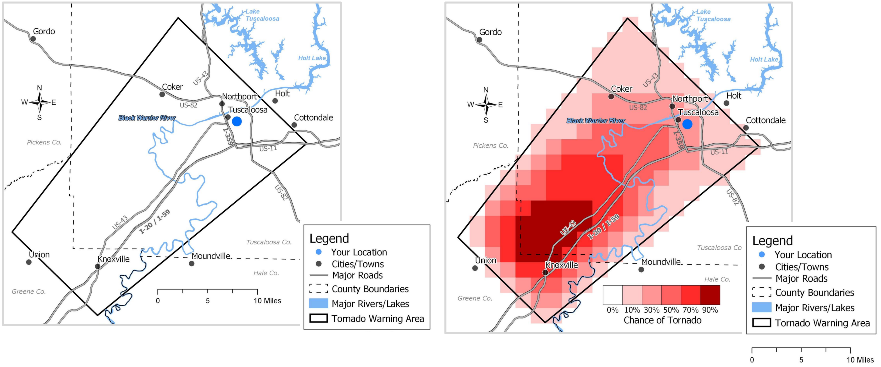


If and how data are binned:

"Our operating assumption is that everything west of Interstate 5 will be toast." - The Really Big One, New Yorker

"An unfocused unclassed map is a more accurate representation of the risk data than a focused classed map." - Severtson et al 2013, p 813

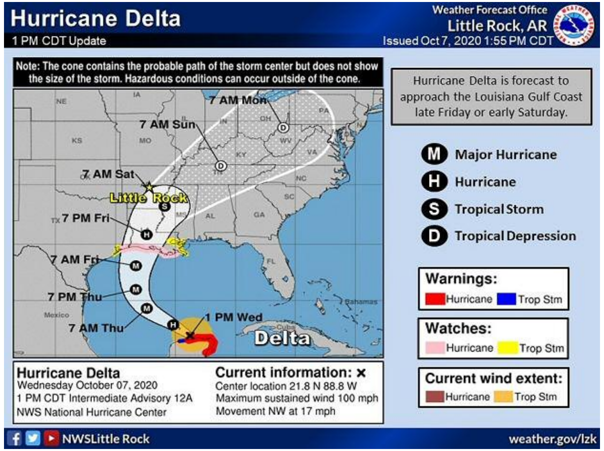
"Except, when you put a boundary on it, then people probably think if they're on one side of the boundary or the other there's a huge difference in probability when there isn't." - Scientist 3 Thompson et al 2015



Qin, C., S. Joslyn, S. Savelli, J. Demuth, R. Morss, and K. Ash, The Impact of Probabilistic Tornado Warnings on Risk Perceptions and Responses. *J. of Experimental Psychology-Applied. (under review)*

Key components of risk information processing:
Understanding, evaluative reaction, behavioral tendencies

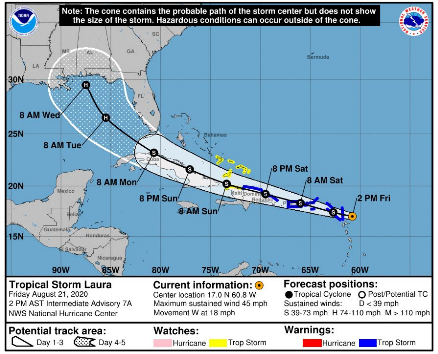
Evaluability is a function of understanding and context

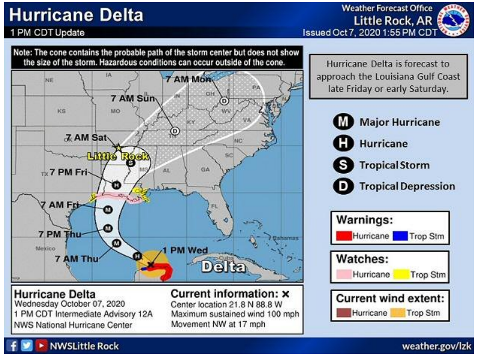


- Familiarity with a visualization drives preferences, also graphicacy, visualization format, and hurricane characteristics in combination influence hurricane forecast track interpretations (Millett et al. 2021)

Key components of risk information processing:
Understanding, evaluative reaction, behavioral tendencies

Evaluability is a function of understanding and context

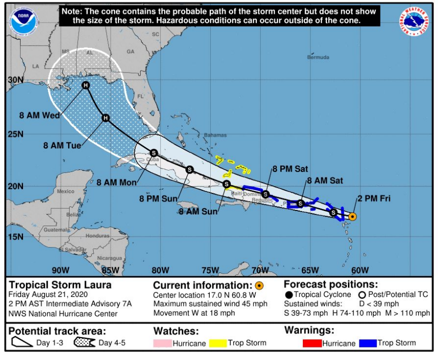


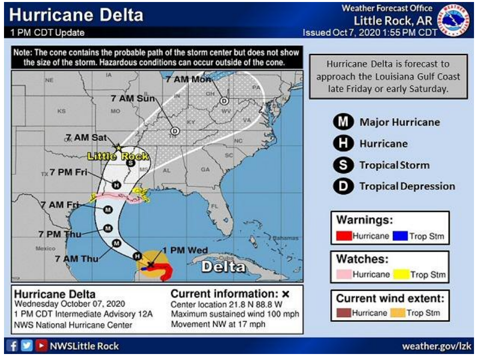


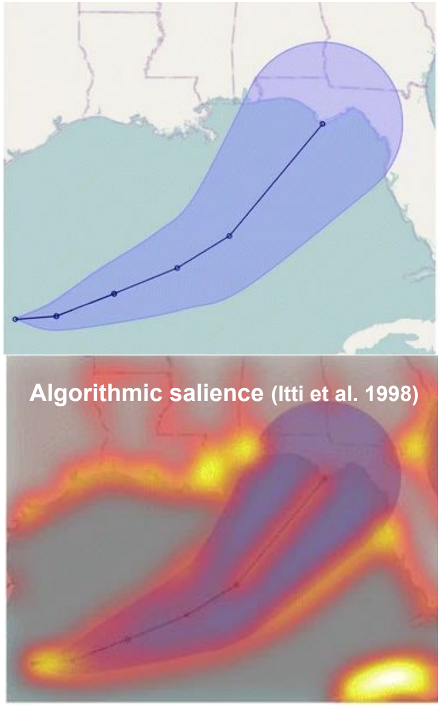
- When Hurricane Delta hit in October 2020, people in Louisiana were still recovering from Laura.

Key components of risk information processing:
Understanding, evaluative reaction, behavioral tendencies

- **Evaluability** is a function of understanding, and context







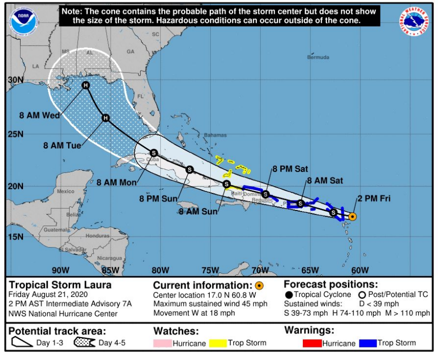
a

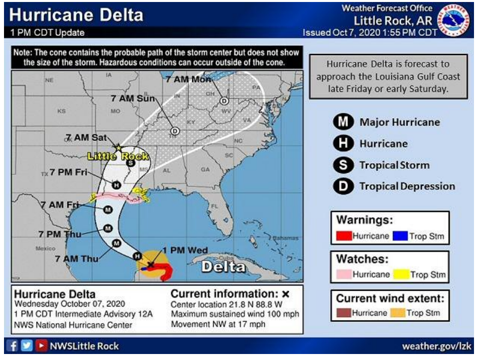
From Padilla et al.(2017)

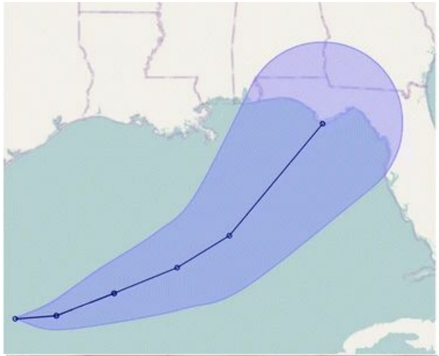
- Boundaries lead people to conceptualize the data as categorical (Tversky, 2005)
- “Deterministic construal error” (Joslyn & Savelli, 2021)

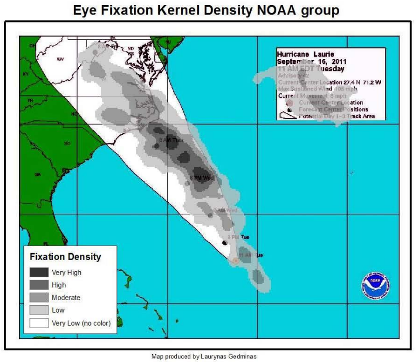
Key components of risk information processing:
Understanding, evaluative reaction, behavioral tendencies

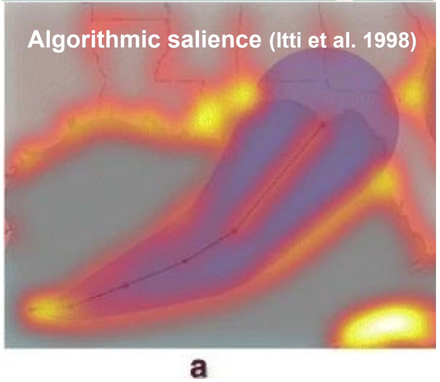
- **Evaluability** is a function of understanding, but also of context











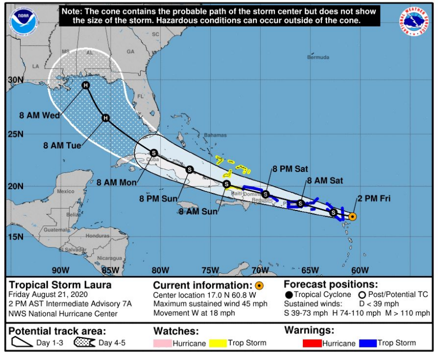
- Fixation on the center part of the cone of uncertainty and the legend (Gedminas, 2011)

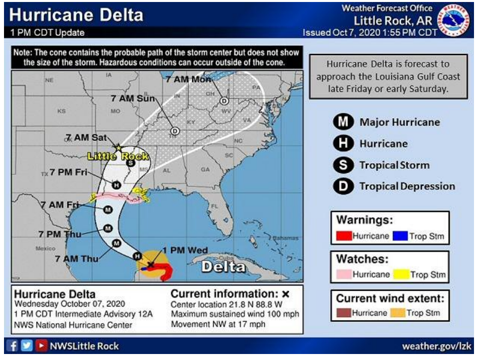
- Boundaries lead people to conceptualize the data as categorical (Tversky, 2005)
- “Deterministic construal error” (Joslyn & Savelli, 2021)

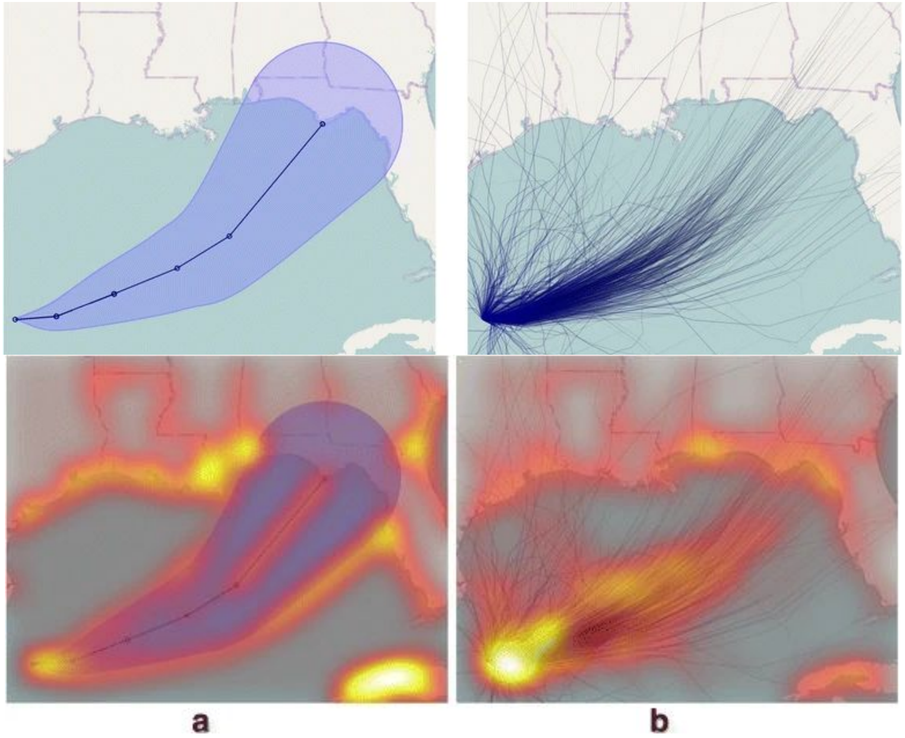
ARTIFICIAL INTELLIGENCE · RISK COMMUNICATION · ENVIRONMENTAL SCIENCE

Key components of risk information processing:
Understanding, evaluative reaction, behavioral tendencies

- **Evaluability** is a function of understanding, but also of context





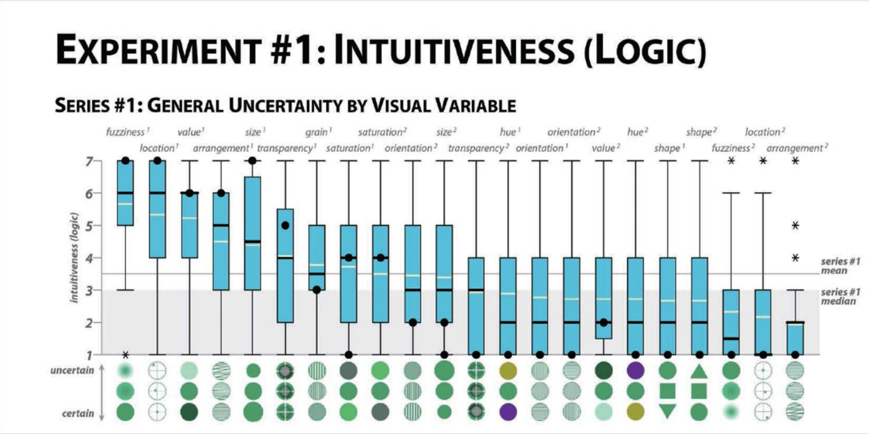


Participants viewing the ensemble display (b) were more likely to report that the display indicated the forecasters were less certain about the path of the hurricane over time compared to the cone (a), in an experiment by Padilla et al. (2017)

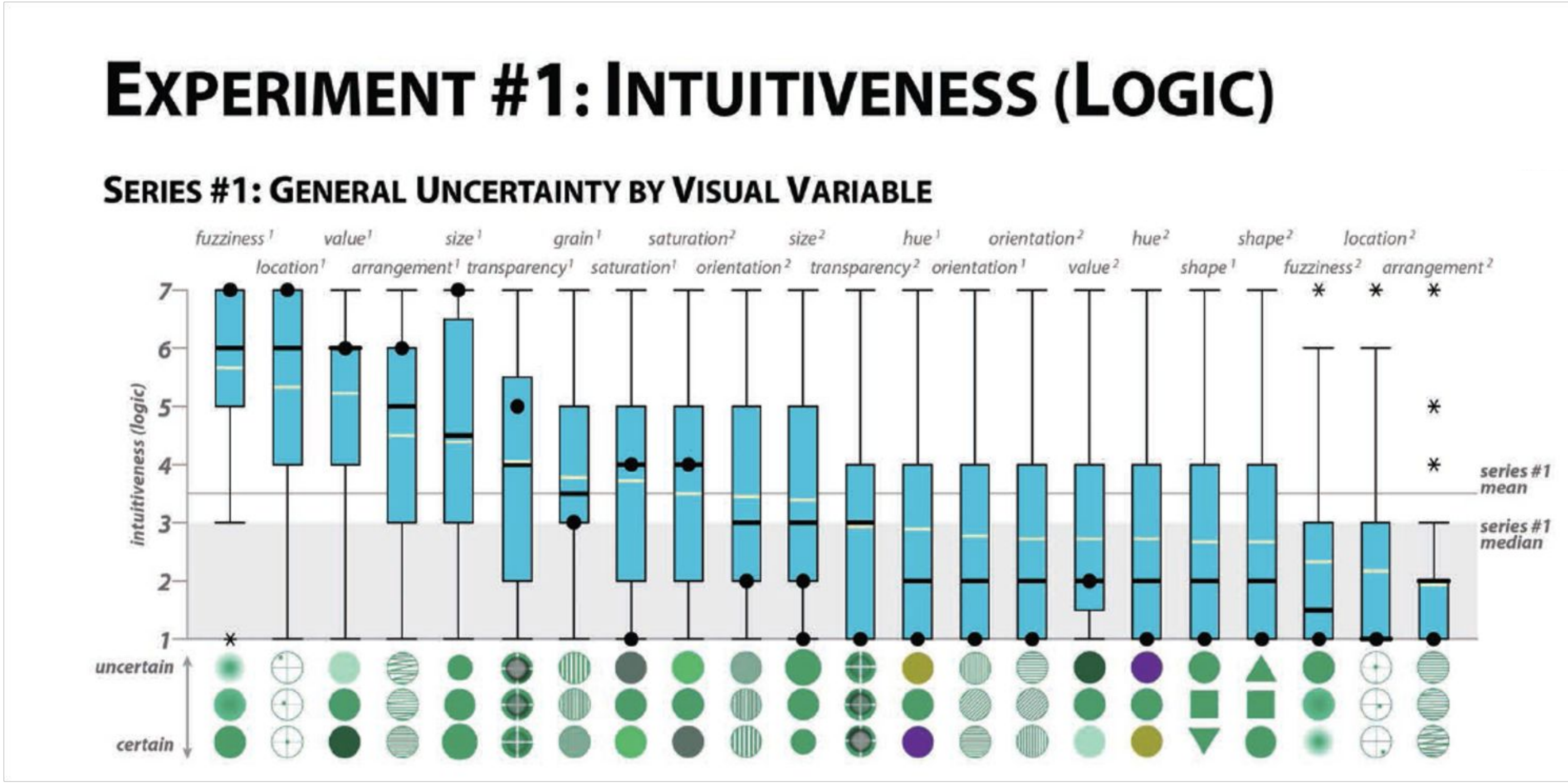
Communicating uncertainty visually

MacEachren et al. experimentally test visualizations of nine types of uncertainty, to examine effects of numerous visual attributes:

- Location
- Size
- Color hue, value and saturation,
- Grain
- Orientation
- Shape
- Arrangement
- Clarity/fuzziness
- Resolution (of boundaries and images)
- Transparency



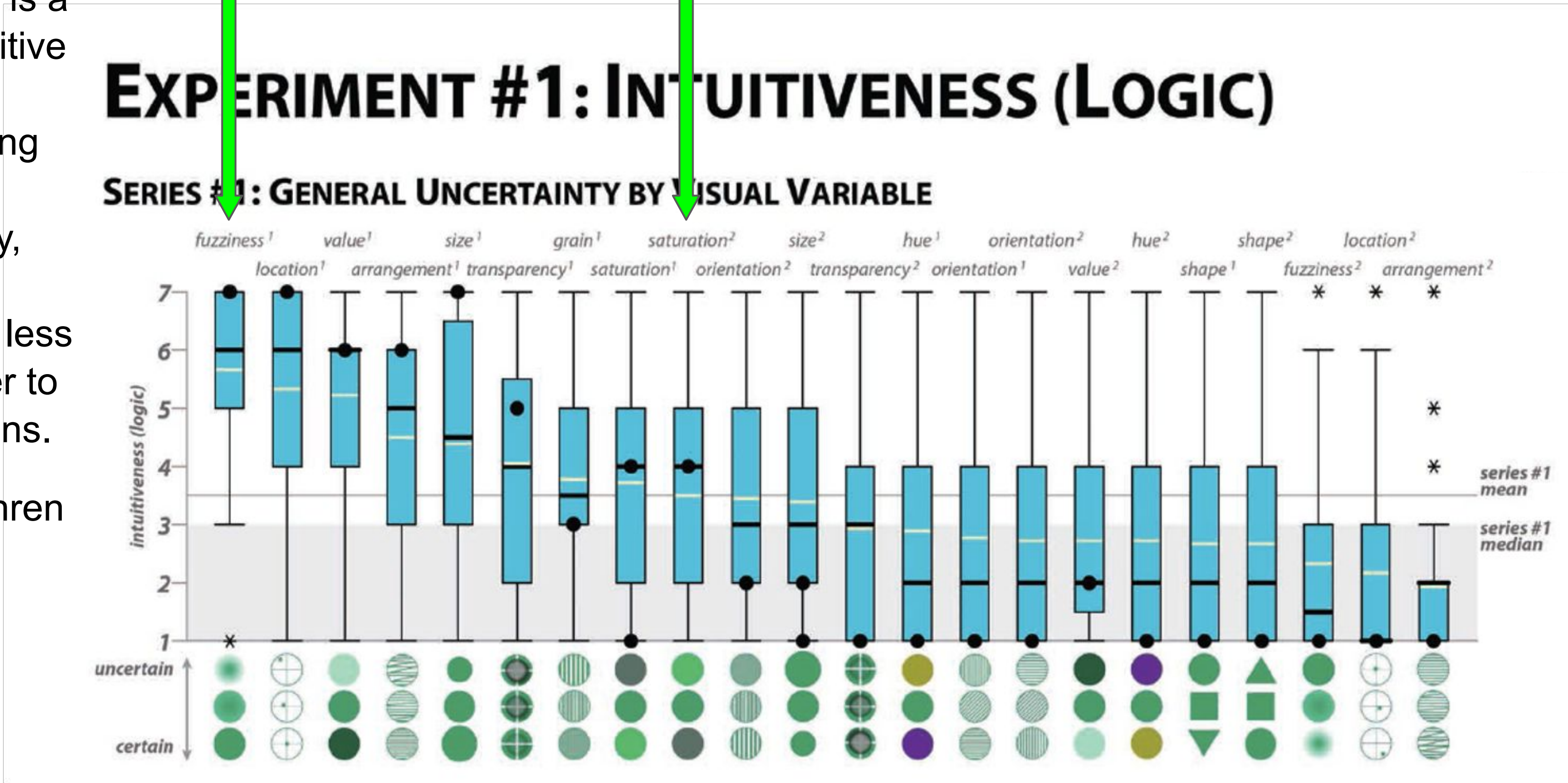
Expressing uncertainty visually: semiotics



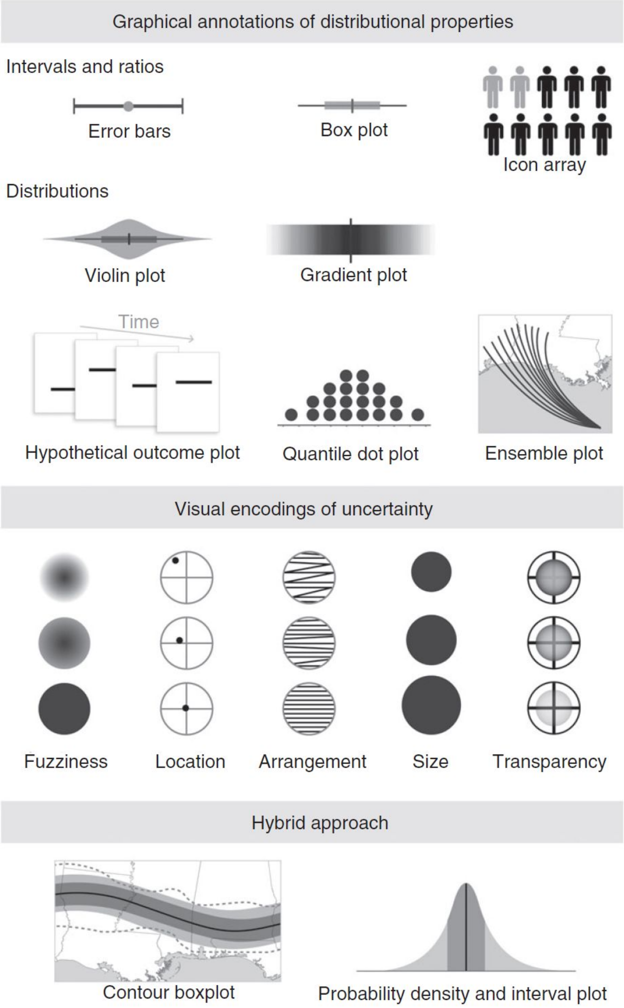
Expressing uncertainty visually: semiotics

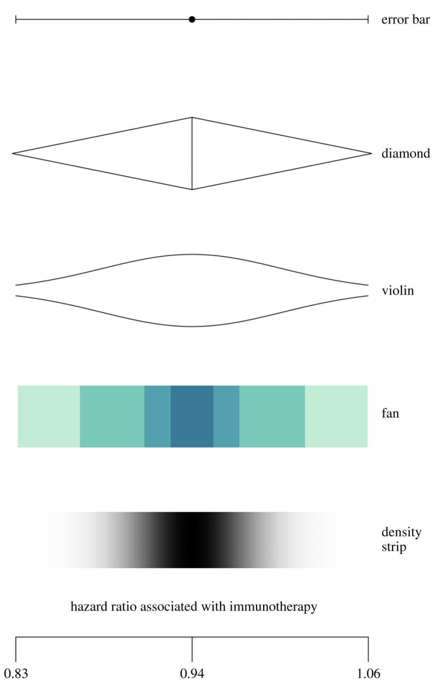
Fuzziness is a highly intuitive way of representing general uncertainty, color saturation less so, counter to expectations.

- MacEachren



Communicating uncertainty visually: statistical graphics





Padilla, L., Kay, M., & Hullman, J. (2014). Uncertainty Visualization. *Wiley StatsRef: Statistics Reference Online*, 1-18.

van der Bles AM, van der Linden S, Freeman ALJ, Mitchell J, Galvao AB, Zaval L, Spiegelhalter DJ. 2019 Communicating uncertainty about facts, numbers and science. R. Soc. open sci. 6: 181870.

Resources on communicating risk and uncertainty

Aven, T., & Renn, O. (2009). On risk defined as an event where the outcome is uncertain. Journal of risk research, 12(1), 1-11.Bostrom, A., Morss, R. E., Lazo, J. K., Demuth, J. L., Lazrus, H., & Hudson, R. (2016). A mental models study of hurricane forecast and warning production, communication, and decision-making. Weather, Climate, and Society, 8(2), 111-129

Bloodhart, B., Maibach, E., Myers, T., & Zhao, X. (2015). Local climate experts: The influence of local TV weather information on climate change perceptions. PloS one, 10(11), e0141526.

Broad, K., A. Leiserowitz, J. Weinkle and M. Steketee, "Misinterpretations of the “cone of uncertainty” in Florida during the 2004 hurricane season", Bulletin Amer. Meteorological Soc., vol. 88, no. 5, pp. 651-667, 2007.

Brown, N. R., & Siegler, R. S. (1993). Metrics and mappings: a framework for understanding real-world quantitative estimation. Psych. review, 100(3), 511.

Earle, T. C. (2010). Trust in risk management: A model‐based review of empirical research. Risk Analysis: An International Journal, 30(4), 541-574.

Freudenburg, W. R., Gramling, R., & Davidson, D. J. (2008). Scientific certainty argumentation methods (SCAMs): science and the politics of doubt. Sociological Inquiry, 78(1), 2-38.

Gedminas, L. (2011). Evaluating hurricane advisories using eye-tracking and biometric data. East Carolina University.

Gigerenzer, G., & Galesic, M. (2012). Why do single event probabilities confuse patients? Statements of frequency are better for communicating risk. BMJ, 344(7839).

IPCC (2013). Climate change 2013: The physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change. Cambridge, UK: Cambridge University Press.

Joslyn, S., & Savelli, S. (2021). Visualizing uncertainty for non-expert end users: The challenge of the deterministic construal error. Frontiers in Computer Science, 2, 590232.

MacEachren, A. M., Roth, R. E., O'Brien, J., Li, B., Swingley, D., & Gahegan, M. (2012). Visual semiotics & uncertainty visualization: An empirical study. IEEE transactions on visualization and computer graphics, 18(12), 2496-2505.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *Academy of Management Review. Academy of Management*, *20*(3), 709. https://doi.org/10.2307/258792

Millet, B., Carter, A. P., Broad, K., Cairo, A., Evans, S. D., & Majumdar, S. J. (2020). Hurricane risk communication: visualization and behavioral science concepts. Weather, climate, and society, 12(2), 193-211.

Millet, B., Majumdar, S. J., Cairo, A., Diaz, C., Ding, Q., Evans, S. D., & Broad, K. (2021, September). End-user Preference for and Understanding of Hurricane Forecast Graphs. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 65, No. 1, pp. 606-610). Sage CA: Los Angeles, CA: SAGE Publications.

Morgan, M. G. Dowlatabadi, H., Henrion, M., Keith, D., Lempert, R., McBride, S., Small, M. and Wilbanks, T., (2009)..Best practice approaches for characterizing, communicating and incorporating scientific uncertainty in climate decision making: Synthesis and assessment product 5.2 Report (Vol. 5). US Climate Change Science Program.

Resources on communicating risk and uncertainty (continued)

Oreskes, N., & Conway, E. M. (2011). Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming. Bloomsbury Publishing USA.

Padilla, L., Kay, M., & Hullman, J. (2014). Uncertainty Visualization. Wiley StatsRef: Statistics Reference Online, 1-18.

Padilla, L. M., Creem-Regehr, S. H., Hegarty, M., & Stefanucci, J. K. (2018). Decision making with visualizations: a cognitive framework across disciplines. Cognitive research: principles and implications, 3(1), 1-25.

Padilla, L.M., Ruginski, I.T. & Creem-Regehr, S.H. Effects of ensemble and summary displays on interpretations of geospatial uncertainty data. Cogn. Research 2, 40 (2017). https://doi.org/10.1186/s41235-017-0076-1

Peters, E. Levin, I. P. (2008). Dissecting the risky-choice framing effect: Numeracy as an individual-difference factor in weighting risky and riskless options. Judgment and Decision Making, 3(6), 435-448.

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. Psychological science, 17(5), 407-413.

Qin, C., S. Joslyn, S. Savelli, J. Demuth, R. Morss, and K. Ash, The Impact of Probabilistic Tornado Warnings on Risk Perceptions and Responses. J. of Experimental Psychology-Applied. (under review) NOAA OAR Grant #NA17OAR4590197

Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. Psychological bulletin, 135(6), 943.

Severtson, D. J., & Myers, J. D. (2013). The influence of uncertain map features on risk beliefs and perceived ambiguity for maps of modeled cancer risk from air pollution. Risk Analysis, 33(5), 818-837.

Tversky, B. (2005). “Visuospatial reasoning,”,The Cambridge handbook of thinking and reasoning, pp. 209–240.

Siegrist, M. (2021). Trust and risk perception: A critical review of the literature. Risk analysis, 41(3), 480-490.

Slovic, P. (1987). Perception of risk. Science, 236(4799), 280-285.

Spiegelhalter, D. (2017). Risk and uncertainty communication. Annual Review of Statistics and Its Application, 4(1), 31-60

Thompson, M. A., Lindsay, J. M., & Gaillard, J. C. (2015). The influence of probabilistic volcanic hazard map properties on hazard communication. Journal of Applied Volcanology, 4(1), 1-24.

van der Bles, A. M., van der Linden, S., Freeman, A. L., & Spiegelhalter, D. J. (2020). The effects of communicating uncertainty on public trust in facts and numbers. Proceedings of the National Academy of Sciences, 117(14), 7672-7683. (earlier version published as an abstract in 2018 here)

van der Bles AM, van der Linden S, Freeman ALJ, Mitchell J, Galvao AB, Zaval L, Spiegelhalter DJ. 2019 Communicating uncertainty about facts, numbers and science. R. Soc. open sci. 6: 181870.

Zikmund-Fisher, B. (2013) The Right Tool Is What They Need, Not What We Have: A Taxonomy of Appropriate Levels of Precision in Patient Risk Communication. Med Care Res Rev DOI: 10.1177/1077558712458541

Resources on UQ paper

Methods to Quantify Uncertainty provided by Neural Networks and their Evaluation for Environmental Science Applications

Authors: Katherine Haynes , Ryan Lagerquist, Marie McGraw, Kate Musgrave, Imme Ebert-Uphoff

Paper:

- **Draft version** (June 29, 2022): released for TAI4ES summer school participants
- **arXiv version** (more refined): to come soon, link will be posted here.
  Please check back here soon for that more official version (also will be submitted to journal AIES soon).

Github repo:

- https://github.com/thunderhoser/cira_uq4ml

So how do we communicate this type of information to users?

An example from AI2ES

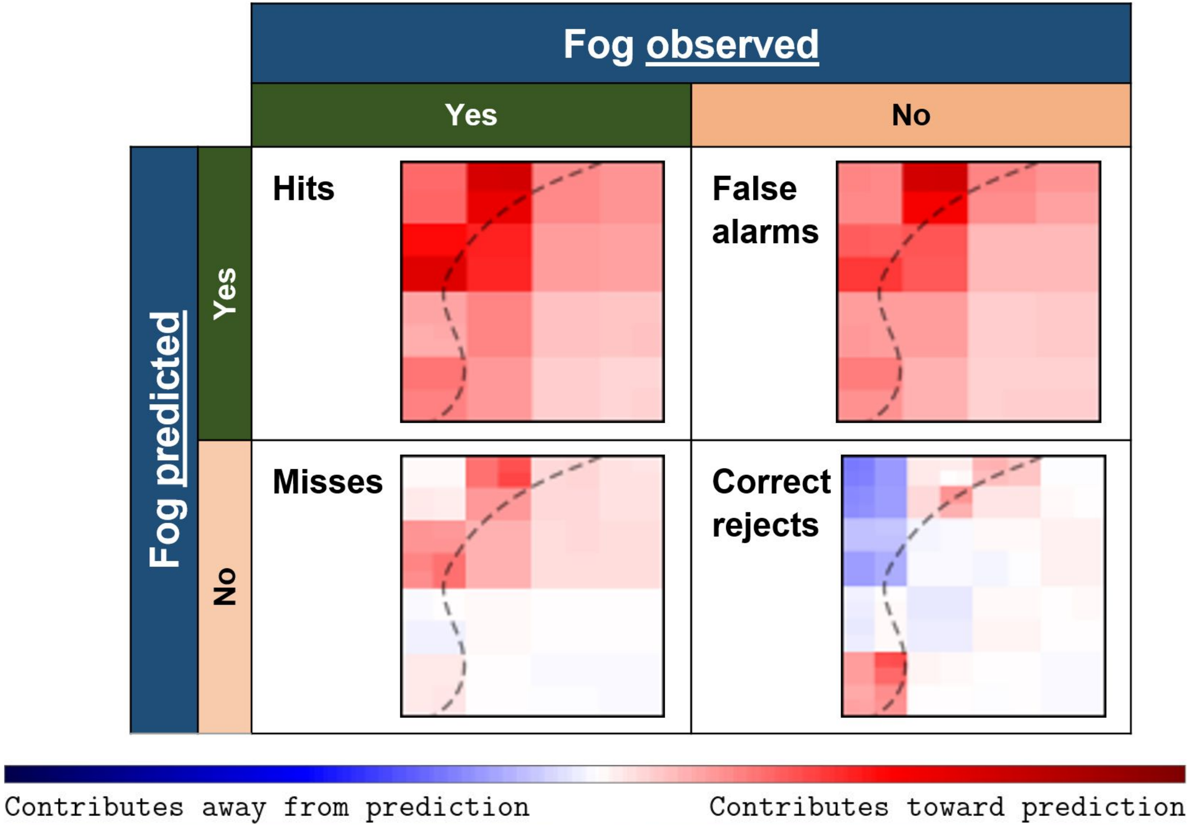
Understanding what the AI/ML guidance is ‘using’

The developers also examined which locations in the overall training area contributed the most to different outcomes. The dashed line in the figure symbolizes the coastline.

The contingency table on the right shows which areas the guidance draws on most heavily.

“Contributes **toward** prediction” means the guidance is relying on that area for the given outcome (hits, misses, false alarms, or correct rejects).

“Contributes **away** from prediction” means the area is contributing to the opposite prediction of the given outcome.



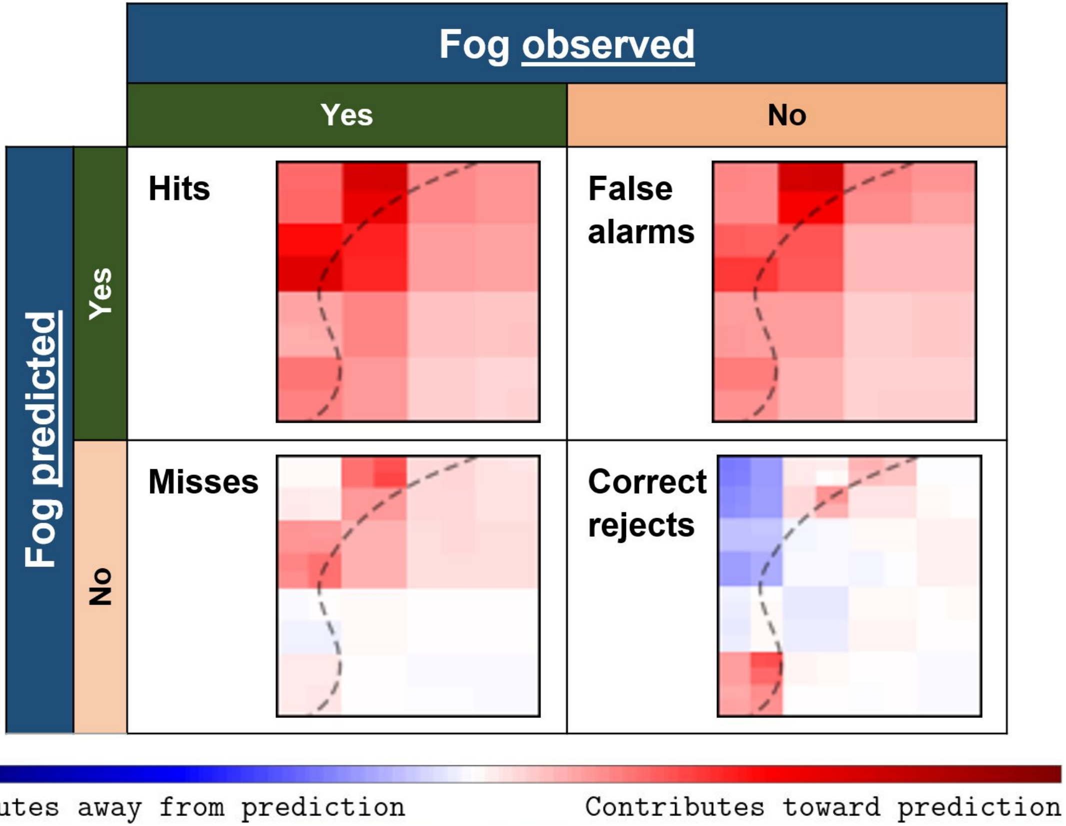
Understanding what the AI/ML guidance is ‘using’

4.10-12 Go to sli.do and use the code TAI4ES

The developers also examined which locations in the overall training area contributed the most to different outcomes. The dashed line in the figure symbolizes the coastline.

The contingency table on the right shows which areas the guidance draws on most heavily.

Q1: What areas contribute most to ‘Hits’
Q2: What areas lead the model astray?
Q3: What areas contribute most to accurate predictions?



As a developer, what can you do with this information?

What about as a user?

Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School

Day 4: Agenda

- 9:00 Uncertainty quantification methods (Part 1)
- 10:00 *Short brain & bio break*
- 10:10 Uncertainty quantification methods (Part 2)
- 10:45 *Short brain & bio break*
- 10:55 Communicating uncertainty (Part 3)
- **11:55 Lecture series wrap up!**

Time for any open questions!

Questions?

https://app.sli.do/event/1zumy91n

Or go to sli.do and use the code TAI4ES

NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

LEAP

NOAA CENTER FOR ARTIFICIAL INTELLIGENCE

Radiant Earth Foundation EARTH IMAGERY FOR IMPACT

Lecture wrap-up!

- **We hope you have learned a lot about trust in AI especially for environmental science applications!**
- Our recordings and notebooks will stay available on our website permanently
  - https://www.ai2es.org
  - Click on education to find all past recordings and courses
- Want to learn more?
  - Keep up with AI2ES on twitter and our webpage!
  - Many of our site-wide meetings are open to the public - contact us if you want to join a meeting
- Want to collaborate?
  - Talk to us!

Thanks to all the lecture series speakers!

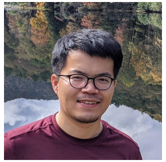

Amy McGovern (OU)


David John Gagne (NCAR)


Imme Ebert-Uphoff (CSU)


Ann Bostrom (UW)


Christopher Wirz (NCAR)

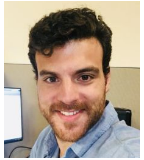

Douglas Rao (NOAA)


Andrea Schumacher (CIRA)


Montgomery Flora (NOAA)


Mariana Cains (NCAR)


Randy Chase (OU)


Antonios Mamalakis (CSU)


Marie McGraw (CSU)


Ryan Lagerquist (CSU)

Thank you!

- This material is based upon work supported by the National Science Foundation under Grant No. ICER-2019758.
- This summer school is being supported by NCAR/UCAR
- Thank you to:
    - **Taysia Peterson** and the multi-media team @ NCAR
    - **Susan Dubbs** @ OU
    - Our sponsors! NCAR/UCAR, Google cloud, LEAP, Radiant Earth
    - All of our guest speakers
    - All of you for coming and participating!

Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School

Reminder for trust-a-thon participants:

The closing fireside chat is today at 3pm MT!

Please take the evaluation survey!!! It'll come in an email soon

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

LEAP

NOAA CENTER FOR ARTIFICIAL INTELLIGENCE

Radiant Earth Foundation
EARTH IMAGERY FOR IMPACT