



A Gentle Introduction to Creating and Evaluating Uncertainty Estimates with Neural Networks for Earth Science Applications

Katherine Haynes, Ryan Lagerquist, Marie McGraw, Kate Musgrave, Imme Ebert-Uphoff
Cooperative Institute for Research in the Atmosphere (CIRA)
Colorado State University, Fort Collins, CO

Presented by **Imme Ebert-Uphoff**
Machine Learning Lead - Cooperative Institute for Research in the Atmosphere (CIRA)
Research Professor - Electrical and Computer Engineering @ Colorado State Univ.

Nov 15, 2022
ECMWF–ESA Workshop on Machine Learning for Earth Observation and Prediction



Katherine Haynes



Ryan Lagerquist



Marie McGraw



Kate Musgrave



Imme Ebert-Uphoff

This presentation is largely based on the following paper:

Katherine Haynes, Ryan Lagerquist , Marie McGraw , Kate Musgrave , Imme Ebert-Uphoff,
Creating and evaluating uncertainty estimates with neural networks for environmental-science applications, *AMS journal Artificial Intelligence for the Earth Systems* (conditionally accepted).

Preprint: <https://doi.org/10.1002/essoar.10512538.1>

Code provided for all methods: https://github.com/thunderhoser/cira_uq4ml

Presentation Overview

1. **Motivation**
2. **Aleatory vs. Epistemic Uncertainty** – it's not as trivial as it seems.
3. **Simple methods to estimate uncertainty.**
How can we estimate uncertainty when using neural network methods for classification or regression?
4. **Simple methods to evaluate uncertainty.**
Once we obtained uncertainty estimates, how do we know whether they are any good?
5. **Illustration of the above for real-world application.**

Intended audience:

- Novices - folks who use NNs in their applications and would like to get uncertainty estimates, but don't know where to get started.
- Intermediate – folks who have tried some uncertainty modeling, but would like to learn more about other methods and evaluation.

To uncertainty experts: now is a good time to check your email, take a nap, etc.

Motivation

- Neural networks are now widely used in weather/climate applications.
- Classic neural networks have no awareness of their own limitations:
 - They deliver a result.
 - They *may* also deliver some sort of “confidence score”, but that is usually not a reliable score.
Ex.: A NN for classification may have a softmax layer that provides a “pseudo-probability” for each class. However, it’s understood to just be an indication, not a true probability for that class.
- In the forecasting world we want a reliable uncertainty estimate, to be delivered along with the prediction of a NN.
- **Some acceptable means to express uncertainty:**
 1. Probability distribution, e.g., parameters of a normal (or other) distribution.
 2. Non-parametric summary statistics, e.g., confidence interval, histogram or quantiles.
 3. Ensemble, i.e. a set of representative samples of the true distribution.

What could that look like?

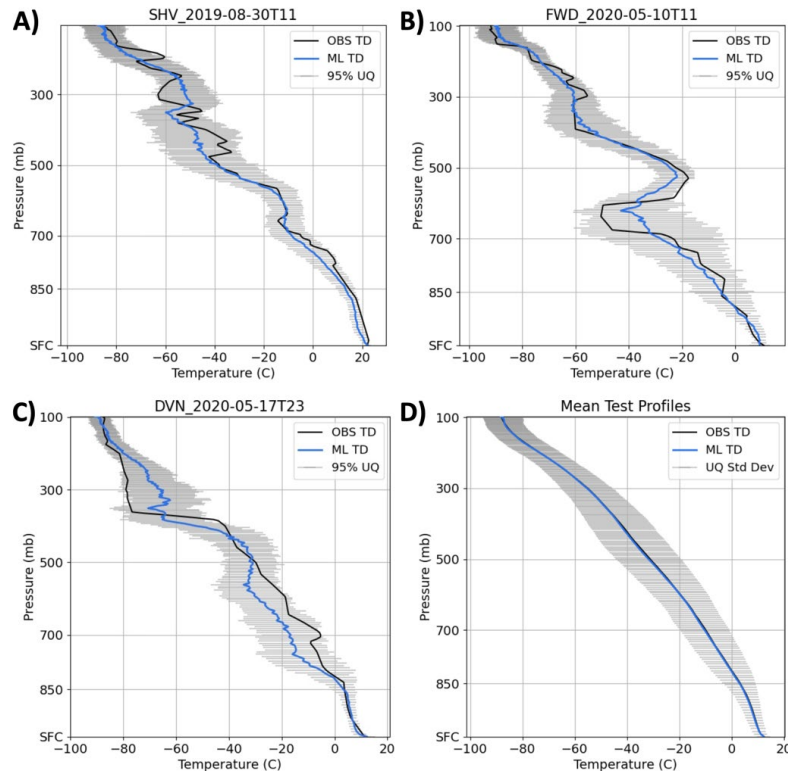


Katherine Haynes

Sample application:

- “ML soundings” project
- Task: Use AI to improve vertical profiles for temperature and dewpoint generated by Rapid Refresh (RAP) model.
- Shown here:
 - estimate of dewpoint,
 - along with estimate of uncertainty (95% confidence interval).

Black:	Observation of dewpoint (by radio sonde)
Blue:	Vertical profiles predicted by AI
Gray lines:	Uncertainty predicted by AI (95% confidence interval)



A), B), C): Individual soundings.
D): Mean over all test samples.

Key Vocabulary

The ML community distinguishes two components of uncertainty:

1. **Aleatory uncertainty**
2. **Epistemic uncertainty**

Caution – concepts of aleatory/epistemic uncertainty are not as simple as they seem!

A word of caution:

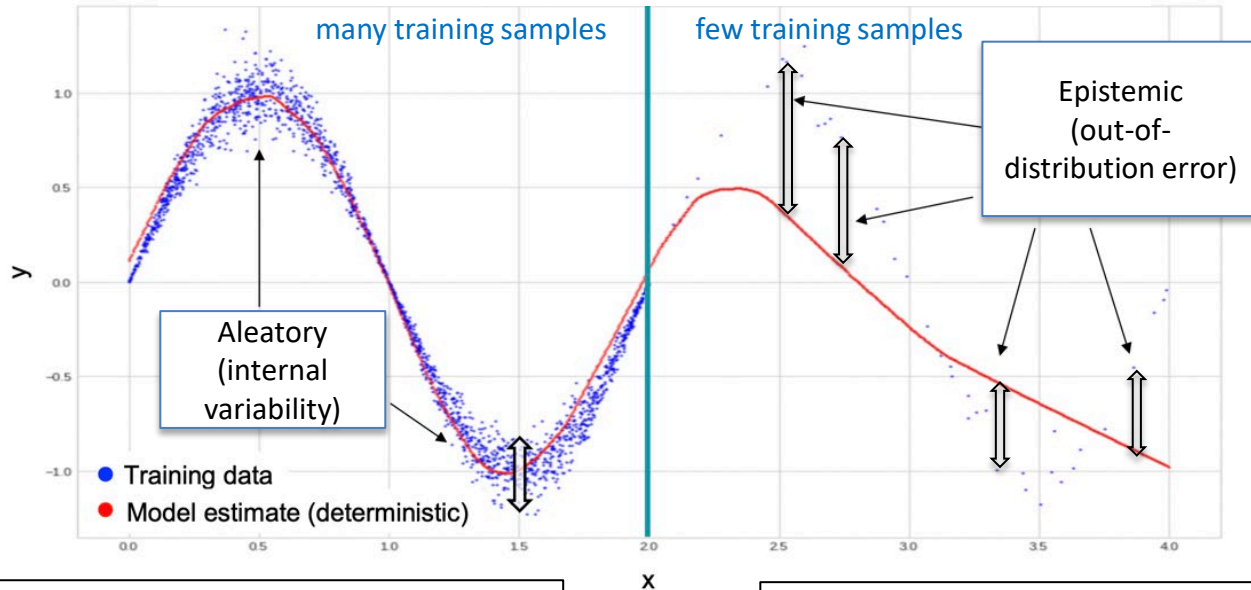
Many ML papers make the concepts of aleatory and epistemic uncertainty sound as if they are unambiguous and straight forward. But they are neither!

- Aleatory and epistemic uncertainty are very slippery concepts – highly dependent on community and context.
- In fact, [Bevan \(2022\)](#) illustrates **four different definitions of the terms aleatory and epistemic uncertainty** that are used in different communities.
- Be careful that you know which definition you're using. Many papers do not make that clear.

Classic example from ML textbooks to explain aleatory vs. epistemic uncertainty

Blue dots = training samples

Red line: model prediction



Aleatory uncertainty:

Given x , there is no unique value for y , because of internal variability of observed system.

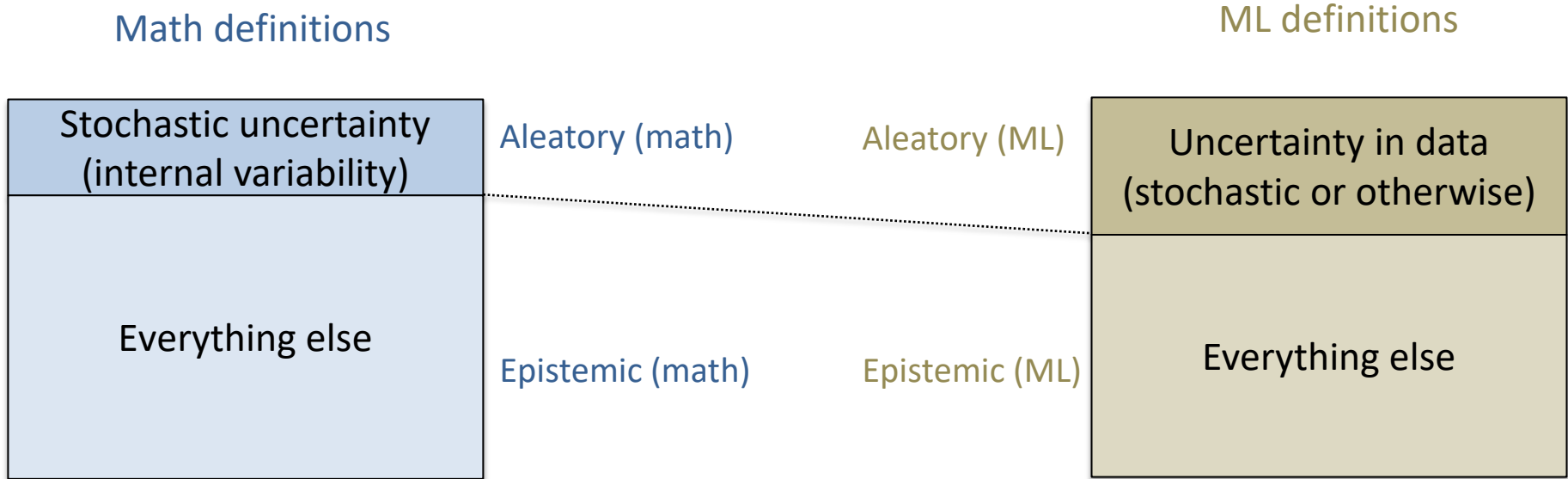
--> Even the best model cannot get it "exactly right".

Epistemic uncertainty:

The model is trying to make predictions in an area where few training samples were provided

--> Large errors (out-of-distribution error)

The aleatory-epistemic divide



alea: Latin word, referring to game of dice (random).

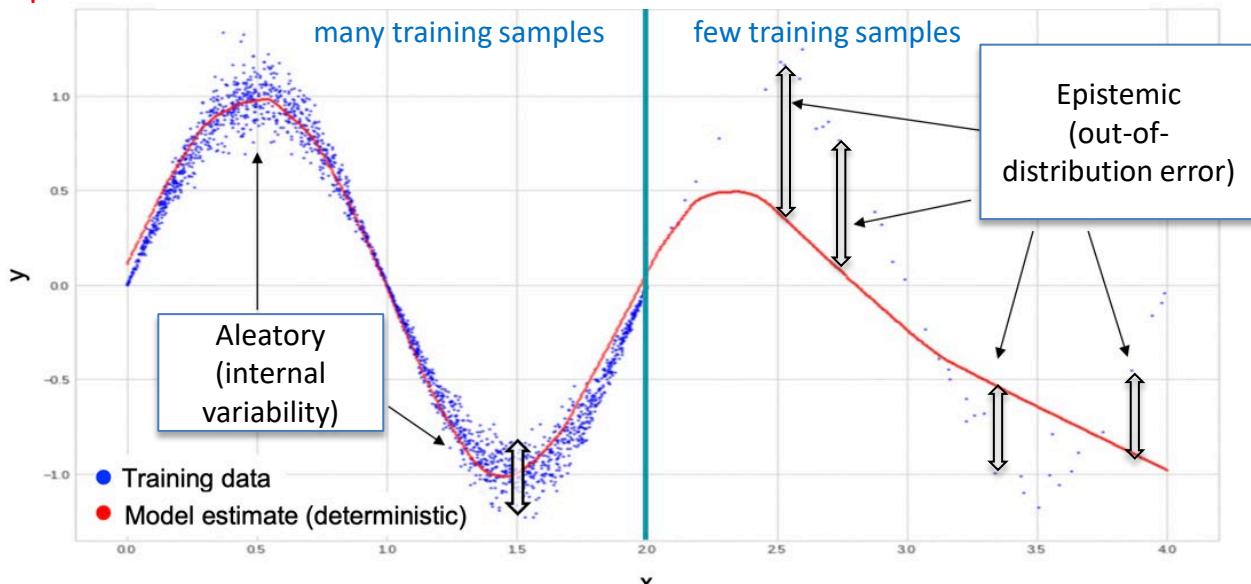
epistēmē: Greek word, meaning knowledge (model).

Dividing lines are different in math vs. ML, but concepts are called the same!
Can get very confusing.

Why didn't we notice that difference in definition in the classic textbook example?

Blue dots = training samples

Red line: model prediction



For this example the math and ML definitions align, because it's an idealized example:

- Only type of data error shown here: internal variability.
- Only type of model error shown here: out-of-distribution error.

No wonder the difference in definition isn't obvious here.

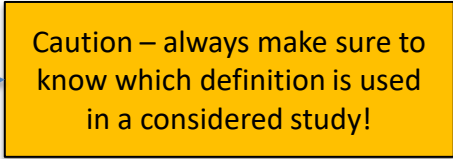
Aleatory vs. Epistemic

Math: Distinction is based on whether source of uncertainty is stochastic.

ML: Distinction is based on whether uncertainty is inherent in the data.

Why is this difference important to understand?

1. Because difference in definitions creates a lot of confusion.
Hard to understand papers sometimes, because of that.
2. ML papers sometimes even use the alternate names
(aleatory=stochastic=irreducible), which creates even more confusion.
3. Using the ML definitions, the **divide between aleatory and epistemic becomes context dependent**.
Example: If you modify your data set, e.g., add more features, some of the aleatory uncertainty can become epistemic uncertainty!



Caution – always make sure to know which definition is used in a considered study!

Recommended reading:

- Hüllermeier, E. and Waegeman, W., 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3), pp.457-506. <https://doi.org/10.1007/s10994-021-05946-3>
- Bevan, L.D., 2022. The ambiguities of uncertainty: A review of uncertainty frameworks relevant to the assessment of environmental change. *Futures*. <https://doi.org/10.1016/j.futures.2022.102919>

Simple Methods to Estimate Uncertainty

We selected four simple methods to derive uncertainty estimates with neural networks:

- **Three non-Bayesian (max likelihood) methods:**

1. Parametric regression
2. Quantile regression
3. Using a Continuous Ranked Probability Score (CRPS) loss

Note: The non-Bayesian methods can only capture aleatory uncertainty.

- **One Bayesian method:**

1. Monte Carlo Dropout

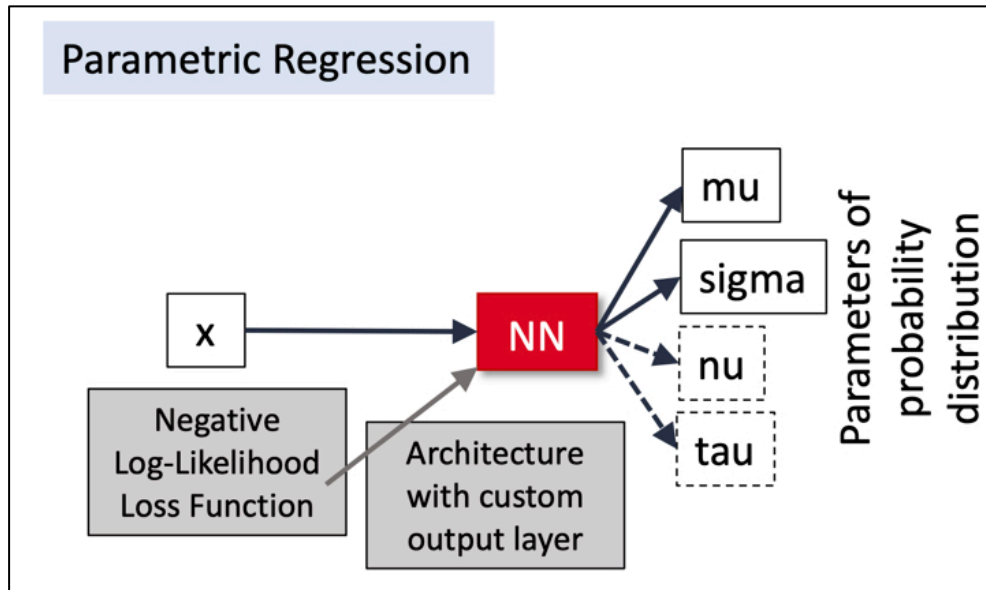
Note: Bayesian methods can – in theory - capture both aleatory and epistemic uncertainty.

Our MC dropout model can only capture epistemic uncertainty due to choice of loss function.

Note:

- Bayesian Deep Learning is very powerful, but is skipped here – focus first on simple methods.
- We only include Monte Carlo Dropout, which can be considered a special case of Bayesian Deep Learning.

1. Parametric regression (Non-Bayesian method)



Key idea:

- Train NN to estimate the parameters of a probability distribution

How to add this to an existing NN model:

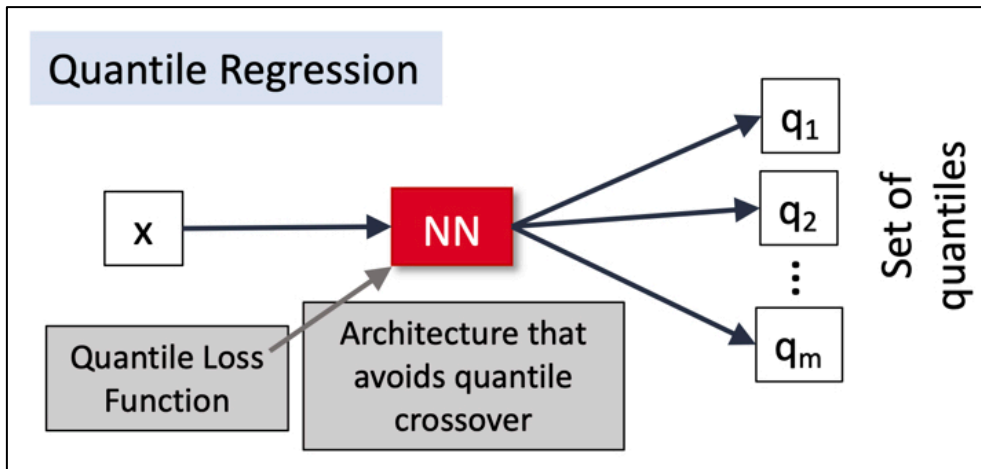
- First choose form of distribution (normal, exponential, etc.)
 - provides parameters to be estimated
 - replace existing output layer with custom layer
- Loss function:
 - Replace by log-likelihood loss function

Examples: [Barnes et al. \(2021\)](#),
[Regression Notebook](#)

Can be used for	
Classification	No
Regression	Yes

Can be used to capture	
ML-Aleatory	Yes
ML-Epistemic	No

2. Quantile regression (Non-Bayesian method)



Key idea:

- Train NN to estimate a set of quantiles

How to add this to an existing NN model:

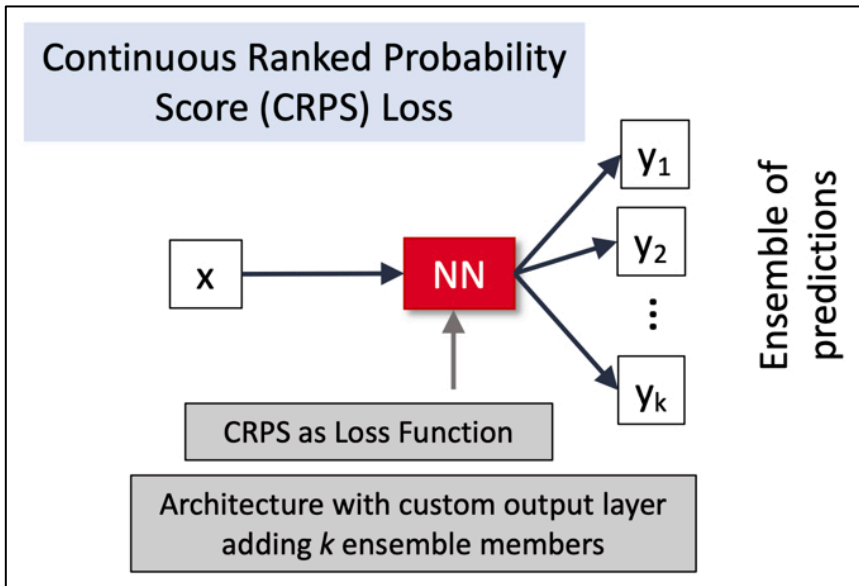
- Choose number of quantiles, m .
→ replace existing output layer with custom layer
- Need to add component to architecture that *prevents quantile cross-over*
- Loss function:
Replace by quantile loss function

Examples: [Quantile Regression Notebook](#)

Can be used for	
Classification	Yes
Regression	Yes

Can be used to capture	
ML-Aleatory	Yes
ML-Epistemic	No

3. Using a CRPS loss (Non-Bayesian method)



Key idea:

- Train NN to estimate an ensemble

How to add this to an existing NN model:

- Choose number of ensembles, k .
→ replace existing output layer with custom layer
- Loss function:
Replace by CRPS loss function

Examples: [CRPS Notebook](#),
[Regression Notebook](#)

[Scher and Messori \(2021\)](#), [Rasp and Lerch \(2018\)](#), [Brey \(2021\)](#)

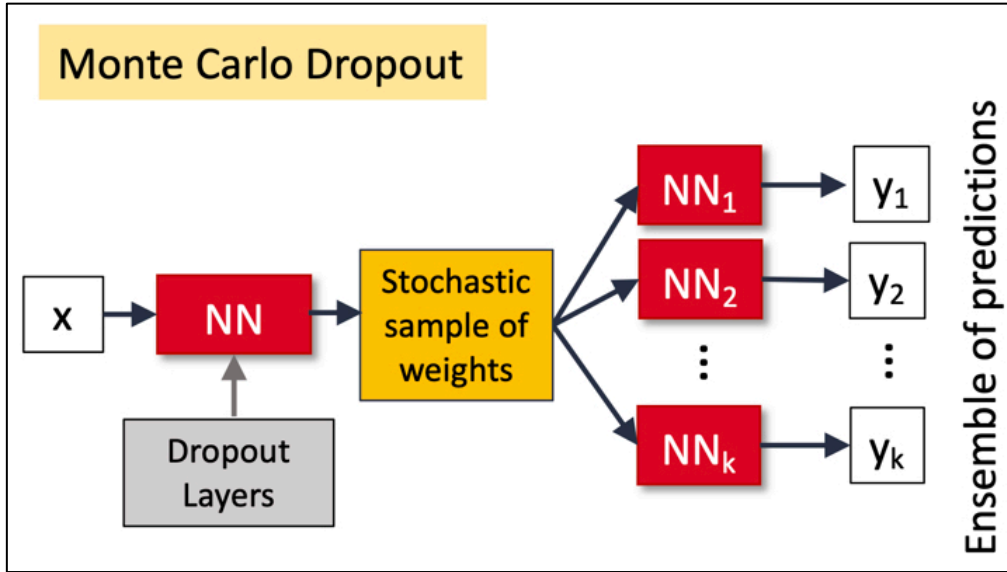
Can be used for

Classification	No
Regression	Yes

Can be used to capture

ML-Aleatory	Yes
ML-Epistemic	No

4. MC Dropout and Bayesian Deep Learning (Bayesian method)



Key idea (shown for MC Dropout, but largely also true for Bayesian Deep Learning):

- Each NN weight is probability distribution (not single number)
- For each prediction:
 - Model randomly selects weights k times
→ generate ensemble of k models
 - Each model provides one prediction
→ generate ensemble of k predictions

How to add MC dropout to existing NN model:

- Add dropout layers
- Use custom loss function

Note: Computationally expensive at run-time, because of all the required sampling.

Examples: [Monte Carlo Notebook](#)
[Regression Notebook](#)

Can be used for	
Classification	Yes
Regression	Yes

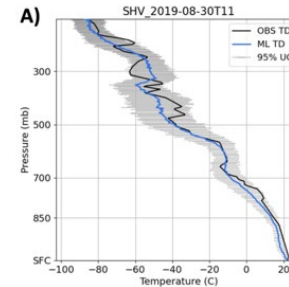
Can be used to capture	
ML-Aleatory	Depends
ML-Epistemic	Yes

Now we have:

- 4 methods to choose from to calculate uncertainty estimates.



Can use those to generate uncertainty estimates



But how do we know which estimates are good?

→ **Need methods to evaluate uncertainty estimates!**

Selected methods for uncertainty evaluation

1. **Spread-skill plot** [Delle Monache et al. \(2013\)](#)
2. **Probability integral transform (PIT) histogram** [Hamill \(2001\)](#)
3. **Discard test** [Barnes and Barnes \(2021\)](#)

Plus – don't forget to evaluate central prediction again, as errors might have changed when uncertainty evaluation was added to the model.

Sample method:

4. **Attributes diagram** [Hsu and Murphy \(1986\)](#)

[Code for all four is included in notebooks as well.](#)

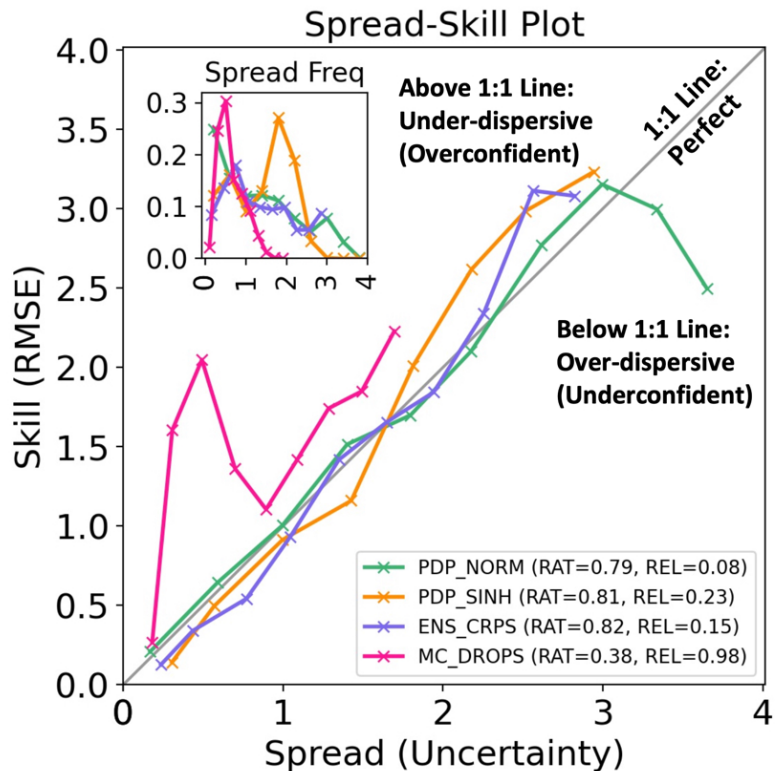
1. Spread-skill plot

Can be used for

Classification Yes

Regression Yes

Sample plot for four models (one color each). Disregard for now which models are used.



Key idea:

- Question: For a given predicted model spread, what is the actual model error?
- Plot the relationship between predicted uncertainty and actual RMSE of prediction.

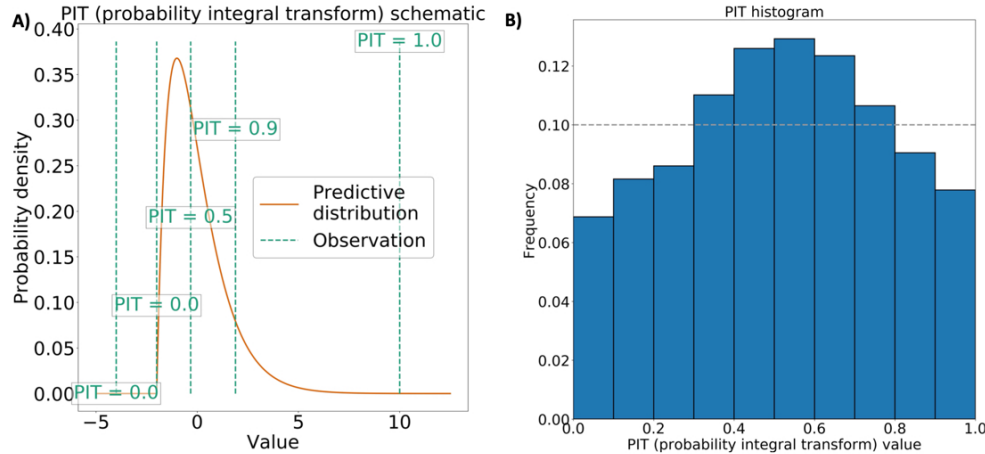
How to read a spread-skill plot:

- Diagonal is ideal: predicted uncertainty matches actual error of prediction.
- Above diagonal: uncertainty estimate is too low (model overconfident).
- Below diagonal: uncertainty estimate is too high (model underconfident).

2. Probability Integral Transform (PIT)

Can be used for	
Classification	No
Regression	Yes

Sample plot for one model



Note: a uniform PIT histogram is a *necessary* but *not sufficient* condition for calibrated uncertainty.

Key idea:

- PIT is the cumulative distribution function (CDF) of the predicted distribution, evaluated at the observed value.
- This can also be interpreted as the quantile of the predictive distribution where the observed value occurs.
- Generalization of “rank histogram” (aka “Talagrand diagram”)

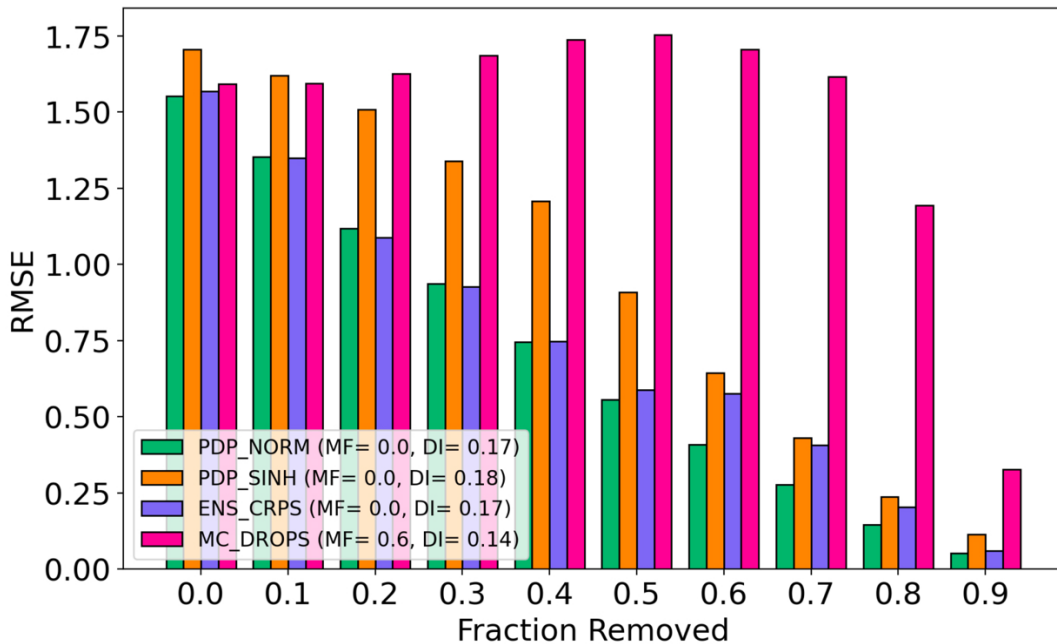
How to read a PIT histogram:

- If uncertainty perfectly calibrated, then PIT histogram is flat.
- If histogram higher at center: model is underconfident (uncertainty too low).

3. Discard test

Can be used for	
Classification	Yes
Regression	Yes

Sample plot for four models (one color each). Disregard for now which models are used.



Key idea:

- Calculate uncertainty estimate.
- Discard the samples with highest X% of uncertainty estimated.
- Calculate model error (RMSE) for only the remaining samples.

How to read a discard test plot:

- Bars for each model (single color) should decrease monotonically from left to right → uncertainty is well calibrated.

Comments on uncertainty evaluation methods

1. Spread-skill plot
 2. Probability integral transform (PIT) histogram
 3. Discard test
- All three methods evaluate **total uncertainty**. They do not distinguish between aleatory and epistemic uncertainty.
Methods exist to split total uncertainty into ML-aleatory and ML-epistemic components. Ortiz et al. (2022) show how to do that for satellite applications (extra step).
 - **Caveat: How much ML-epistemic uncertainty is detected strongly depends on the choice of the test set!**
 - Out-of-distribution error is a key component of the ML-epistemic uncertainty.
 - But we see out-of-distribution error reflected in these tests *only if the test data is chosen to have samples that are very different from the training data!*
 - **Yet to figure out:** how to deal effectively with that caveat.

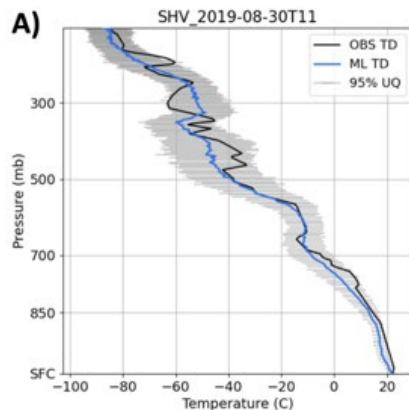
Putting it all together for our sample application



Katherine Haynes

Sample application:

- “ML soundings” project
- Task from before



Calculate uncertainty estimates using 4 methods:

1. **Parametric regression, normal distribution**
(PDP_Norm)
2. **Parametric regression, SHASH distribution:**
(PDP_Shash) SHASH = \sinh -arcsinh
3. **Using CRPS loss to create ensemble**
(ENS_CRPS)
4. **Monte Carlo Dropout**
(MC_DROPS)

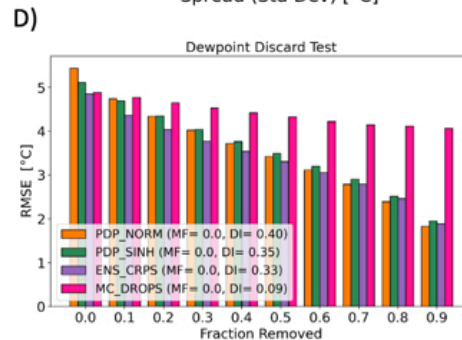
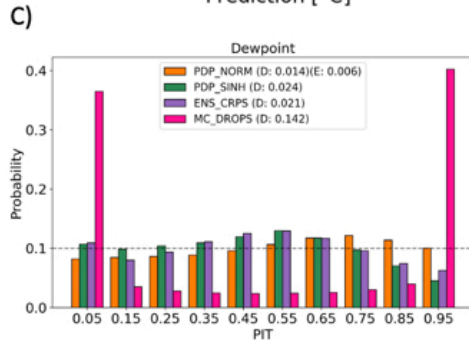
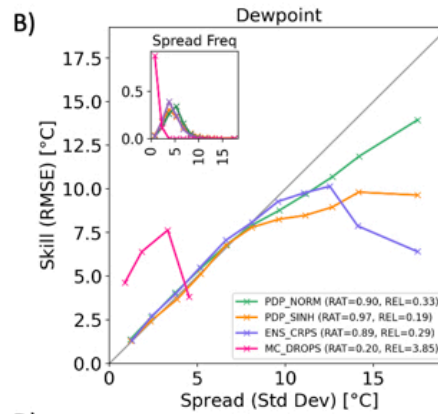
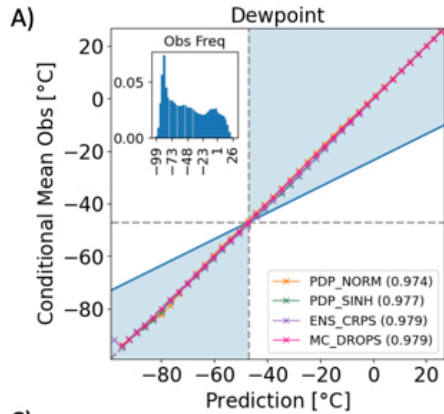
(No quantile regression model here.)

Then evaluated all of them with the three tests:

1. Spread-skill plot
 2. PIT
 3. Discard test
- Plus: attributes diagram (to evaluate mean pred.)

Results on next slides.

What does uncertainty evaluation tell us?



A. Attributes diagram:

- Mean prediction is pretty good with all four methods.

B. Spread-skill plot:

- MC dropout is performing poorly
- Other models are well calibrated for small uncertainty values, then become underconfident.

C. PIT histogram:

- MC dropout seems much too overconfident.
- Other models are doing ok, tend to be underconfident.

D. Discard test:

- All models are doing well, except for MC dropout.

Note: Our version of MC dropout performs poorly in these tests, because it captures only ML-epistemic, not ML-aleatory uncertainty. And since we did not put much emphasis on creating out-of-distribution samples in the test set, MC dropout can't shine here.

Summary and Discussion

- Many simple methods exist to **derive** uncertainty estimates. (Sample code provided)
- Many simple methods exist to **evaluate** uncertainty estimates. (Sample code provided)
- We suggest to always use several evaluation methods, because each one tells you something different. Differences are discussed in more detail in paper.
- We hope these resources are useful for the community to speed up integration of such UQ methods into applications.

Discussion and Future Work:

- Definitions of aleatory and epistemic uncertainty are inconsistent between disciplines. We need to be more concise in our use of these terms!
- IMHO: implementation of these methods for uncertainty estimation and evaluation is not the challenge. The challenge is their proper use and interpretation - that is the hard part!
- Many questions remain:
Ex.: How do we choose the test set to ensure that it contains sufficient out-of-distribution samples to provide representative epistemic error of a model?
- Maybe some of you have already figured all of this out? Would love to hear everyone's thoughts.

References

- **Barnes, E., and R. Barnes, 2021:** Controlled abstention neural networks for identifying skillful predictions for regression problems. *Journal of Advances in Modeling Earth Systems*, 13 (12), e2021MS002 575, <https://doi.org/10.1029/2021MS002575>.
- **Barnes, E., R. Barnes, and N. Gordillo, 2021:** Adding uncertainty to neural network regression tasks in the geosciences. arXiv e-prints, 2109 (07250), <https://arxiv.org/abs/2109.07250>.
- **Bevan, L.D., 2022.** The ambiguities of uncertainty: A review of uncertainty frameworks relevant to the assessment of environmental change. *Futures*. <https://doi.org/10.1016/j.futures.2022.102919>
- **Brey, S., 2021:** Ensemble. GitHub, <https://github.com/TheClimateCorporation/ensemble>.
- **Delle Monache, L., F. Eckel, D. Rife, B. Nagarajan, and K. Searight, 2013:** Probabilistic weather prediction with an analog ensemble. *Monthly Weather Review*, 141 (10), 3498–3516, <https://doi.org/10.1175/MWR-D-12-00281.1>.
- **Haynes, K., Lagerquist, R., McGraw, M., Musgrave, K., and Ebert-Uphoff, I., 2022:** Creating and evaluating uncertainty estimates with neural networks for environmental-science applications, *AMS journal AIES* (conditionally accepted). Preprint: <https://doi.org/10.1002/essoar.10512538.1>. Code: https://github.com/thunderhoser/cira_uq4ml
- **Hamill, T., 2001:** Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129 (3), 550–560, [https://doi.org/10.1175/1520-0493\(2001\)129<0550:IORHFV>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2).
- **Hsu, W., and A. Murphy, 1986:** The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *International Journal of Forecasting*, 2 (3), 285–293, [https://doi.org/10.1016/0169-2070\(86\)90048-8](https://doi.org/10.1016/0169-2070(86)90048-8).
- **Hüllermeier, E. and Waegeman, W., 2021:** Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3), pp.457-506. <https://doi.org/10.1007/s10994-021-05946-3>
- **Ortiz, P., M. Orescanin, V. Petković, S. Powell, and B. Marsh, 2022:** Decomposing satellite-based classification uncertainties in large earth science datasets. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–11, <https://doi.org/10.1109/TGRS.2022.3152516>.
- **Rasp, S., and S. Lerch, 2018:** Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, 146 (11), 3885–3900, <https://doi.org/10.1175/MWR-D-18-0187.1>.
- **Scher, S., and G. Messori, 2021:** Ensemble methods for neural network-based weather forecasts. *Journal of Advances in Modeling Earth Systems*, 13 (2), <https://doi.org/10.1029/2020MS002331>.

Related talk

- D.J. Gagne is going to present on Wed afternoon:

Explainable Uncertainty in Machine Learning for Weather Prediction

Some key ideas:

- Explores **evidential neural networks**.
- Those can be used to derive estimates of both epistemic and aleatoric uncertainty.
- One can then apply XAI methods to models to see how changes in the inputs affect the uncertainty estimates.



Questions or suggestions?

Imme Ebert-Uphoff
iebert@colostate.edu