Deep neural networks for total organic carbon prediction and data-driven sampling

Everardo González
Ewa Burwicz-Galerne

EGU Session - ESSI1.15
Towards SMART Monitoring
6.5.2020

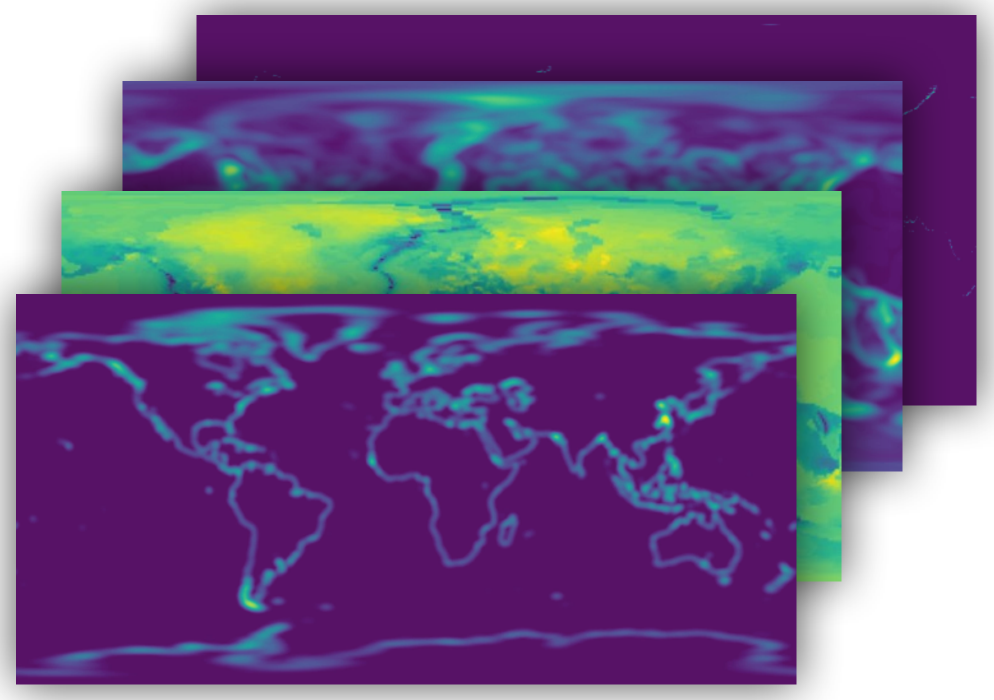
DIGITAL EARTH

1. Introduction:

An idealised Machine Learning Workflow according to a computer scientist

Idealised ML Workflow:

1. Find a set of labels and predictors for your problem



Predictors: ~600 global feature girds

https://zenodo.org/record/1471639#.XPTToCaxXGr

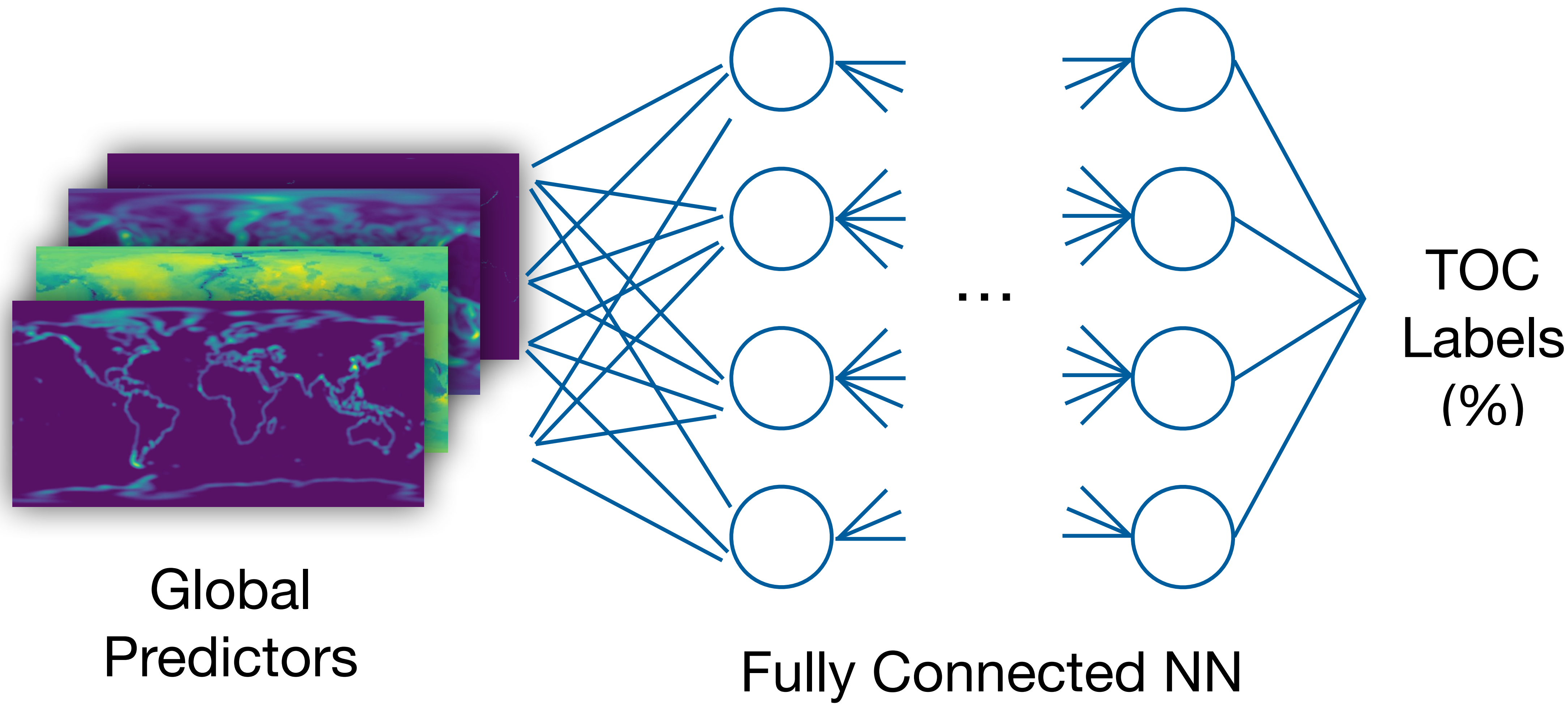
Event	Latitude	Longitude	Elevation [	Depth [m]	TOC [%]
3-14	-28.3315	-20.9410	-4343	0.0000	0.20
8-73	-1.9097	-137.4687	-4387	0.0300	0.10
10103-1B	36.1600	20.4800	-2880	0.0500	0.20
10103-8K	36.1600	20.4800	-2895	0.0150	0.30
108-663	-1.1978	-11.8785	-3706	0.0000	0.19
108-663A	-1.1978	-11.8785	-3706	0.0000	0.19
108-664	0.1073	-23.2275	-3807	0.0000	0.20
108-664B	0.1073	-23.2275	-3806	0.0000	0.20
11BC39	1.9548	-22.7830	-4210	0.0300	0.38
12BC47-2	0.0067	-22.9968	-3858	0.0300	0.58
13BCP56	-2.1083	-23.0050	-4950	0.0300	0.90
159-959	3.6276	-2.7355	-2091	0.0000	1.22
159-959C	3.6277	-2.7353	-2091	0.0250	1.22
159-962	3.2512	-3.1820	-4637	0.0000	1.22
159-962B	3.2511	-3.1820	-4637	0.0500	1.22
167-1011	31.2803	-117.6096	-2021	0.0500	1.82

Labels: ~6000 Total Organic Carbon (TOC) measurements

https://doi.pangaea.de/10.1594/PANGAEA.199835

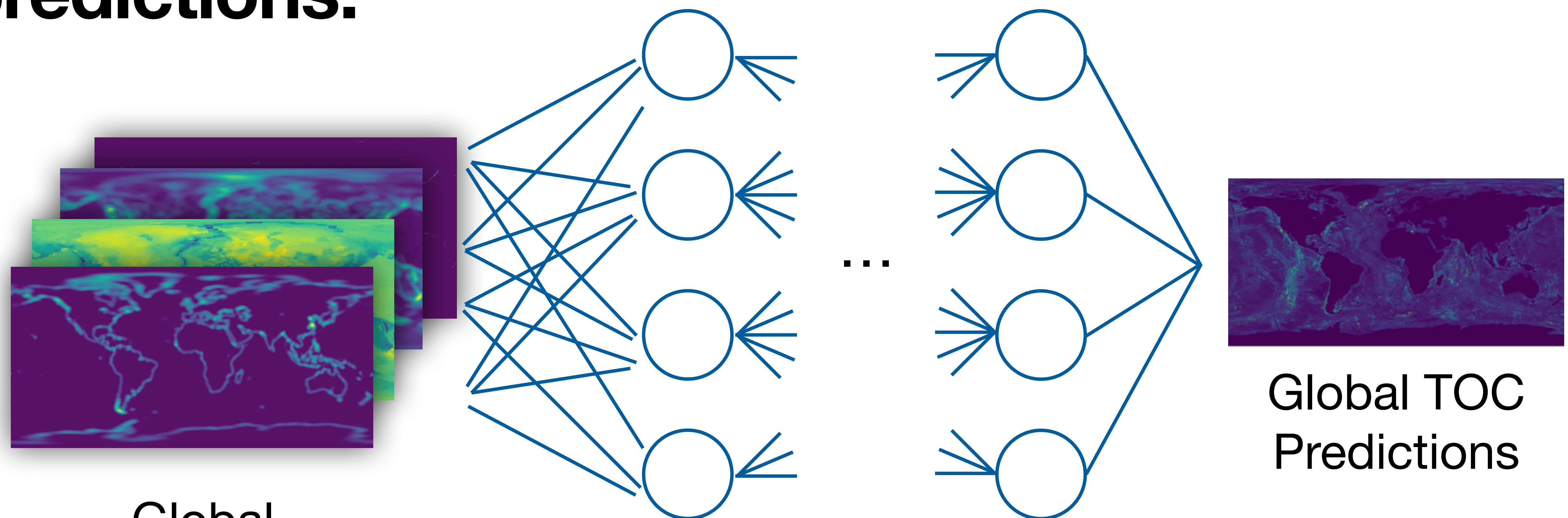
Idealised ML Workflow:

2. Use the predictors and labels to train an ML model



Idealised ML Workflow:

3. Achieve great model accuracy in test and validation datasets and make global predictions:



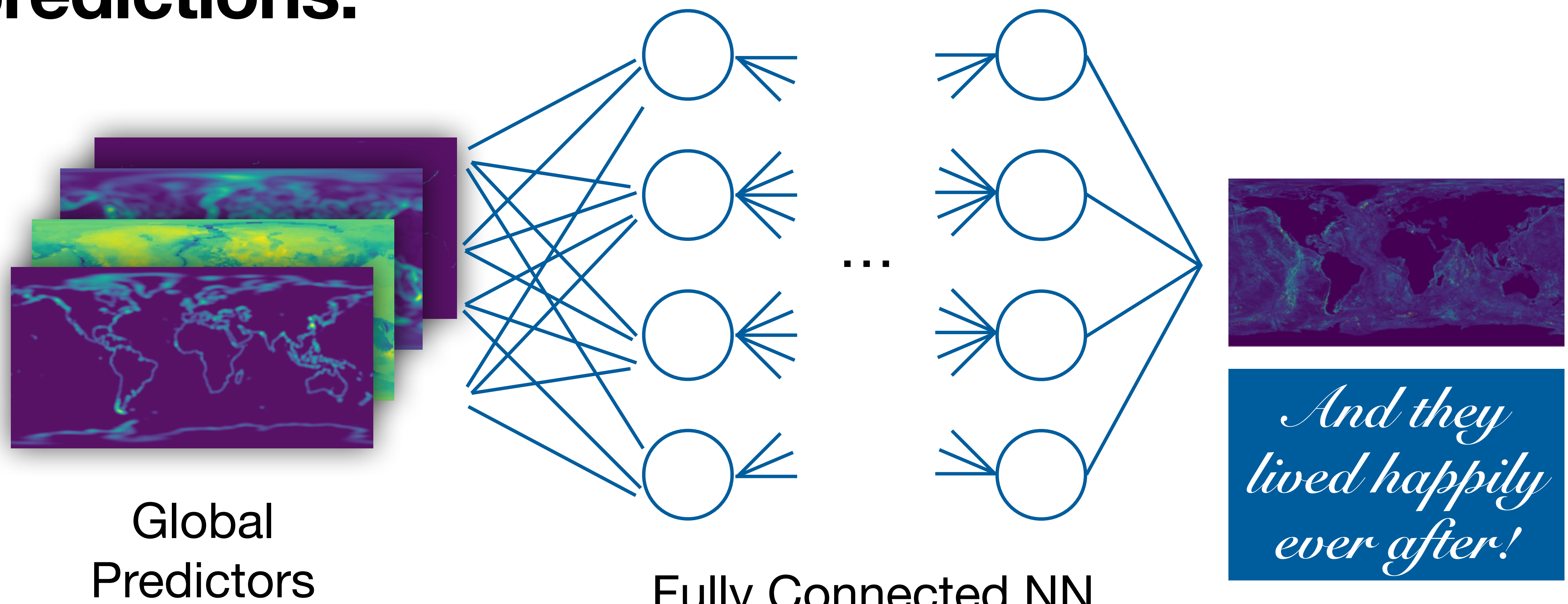
Global Predictors

Fully Connected NN

Global TOC Predictions

Idealised ML Workflow:

3. Achieve great model accuracy in test and validation datasets and make global predictions:



Idealised ML Workflow:

This idealised workflow contains many inaccuracies and oversimplifications incompatible with the application of machine learning to science.

Idealised ML Workflow:

This idealised workflow contains many inaccuracies and oversimplifications incompatible with the application of machine learning to science.

The following slides will discuss what's probably the most problematic one:

HELMHOLTZ
RESEARCH FOR GRAND CHALLENGES

DIGITAL EARTH

Idealised ML Workflow:

This idealised workflow contains many inaccuracies and oversimplifications incompatible with the application of machine learning to science.

The following slides will discuss what's probably the most problematic one:

"Achieve great model accuracy, make global predictions…"

… and they lived happily ever after!

2. Data-Driven Sampling:

Neural Networks and Information Gain

Neural Networks and Information Gain

Computer scientist often see model accuracy as the end goal of machine learning.

Neural Networks and Information Gain

Computer scientist often see model accuracy as the end goal of machine learning.

However, in science, an accurate ML model is often just the begging of the discussion

Neural Networks and Information Gain

Computer scientist often see model accuracy as the end goal of machine learning.

However, in science, an accurate ML model is often just the begging of the discussion

For example, a question that good model accuracy is unable to answer is that of "data-driven sampling"…

Neural Networks and Information Gain

Computer scientist often see model accuracy as the end goal of machine learning.

However, in science, an accurate ML model is often just the begging of the discussion

For example, a question that good model accuracy is unable to answer is that of "data-driven sampling"…

… that is: *assuming this model is correct, where should one sample next?*

Neural Networks and Information Gain

Information Theory can provide an answer using "Information Gain":

Entropy: $$H(P) = -\sum_{x \in X} p(x) \cdot log(p(x))$$ **"Information"**

Cross Entropy: $$H(P,Q) = -\sum_{x \in X} p(x) \cdot log(q(x))$$ **"Message Length"**

KL-Divergence: $$D_{KL}(P \,||\, Q) = H(P,Q) - H(P)$$ **"Information Gain"**

Neural Networks and Information Gain

Information Theory can provide an answer using "Information Gain":

probability distributions

Entropy: $$H(P) = -\sum_{x \in X} p(x) \cdot log(p(x))$$ **"Information"**

observed distribution predicted distribution

Cross Entropy: $$H(P,Q) = -\sum_{x \in X} p(x) \cdot log(q(x))$$ **"Message Length"**

KL-Divergence: $$D_{KL}(P||Q) = H(P,Q) - H(P)$$ **"Information Gain"**

HELMHOLTZ
RESEARCH FOR GRAND CHALLENGES

DIGITAL EARTH

Neural Networks and Information Gain

Assuming sampling as a prediction with absolute certainty:

$$D_{KL}(P\,||\,Q) = H(P,Q) - H(P)$$

$$= -\sum_{x \in X} \cancel{p(x)} \cdot log(q(x)) + \sum_{x \in X} \cancel{p(x)} \cdot \cancel{log(p(x))}$$

Sampling!

$$= -\,log(q(x_i))$$

Neural Networks and Information Gain

Assuming sampling as a prediction with absolute certainty:

$$D_{KL}(P\,||\,Q) = H(P,Q) - H(P)$$

$$= -\sum_{x \in X} \cancel{p(x)} \cdot log(q(x)) + \sum_{x \in X} \cancel{p(x)} \cdot \cancel{log(p(x))}$$

Sampling!

$$= -\,log(q(x_i))$$

"Information-Gain from sampling is the highest where the prediction probability is the lowest"

Neural Networks and Information Gain

… however, TOC estimation is a regression problem, and we need probability distributions:

We have: 0 50 100 %

probability distributions

We want: $$H(P,Q) = -\sum_{x \in X} p(x) \cdot log(q(x))$$

Neural Networks and Information Gain

… however, TOC estimation is a regression problem, and we need probability distributions:

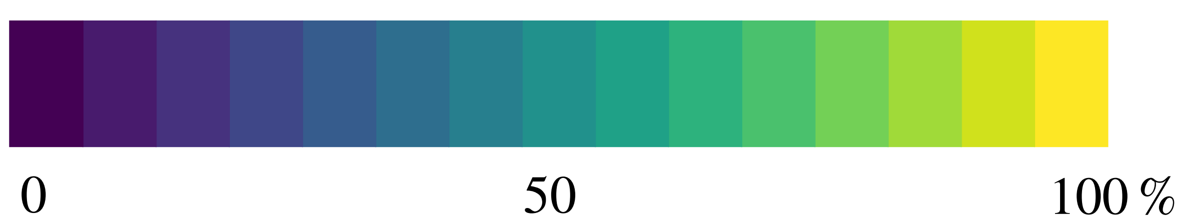
We have:

We want:

probability distributions

$$H(P, Q) = -\sum_{x \in X} p(x) \cdot log(q(x))$$

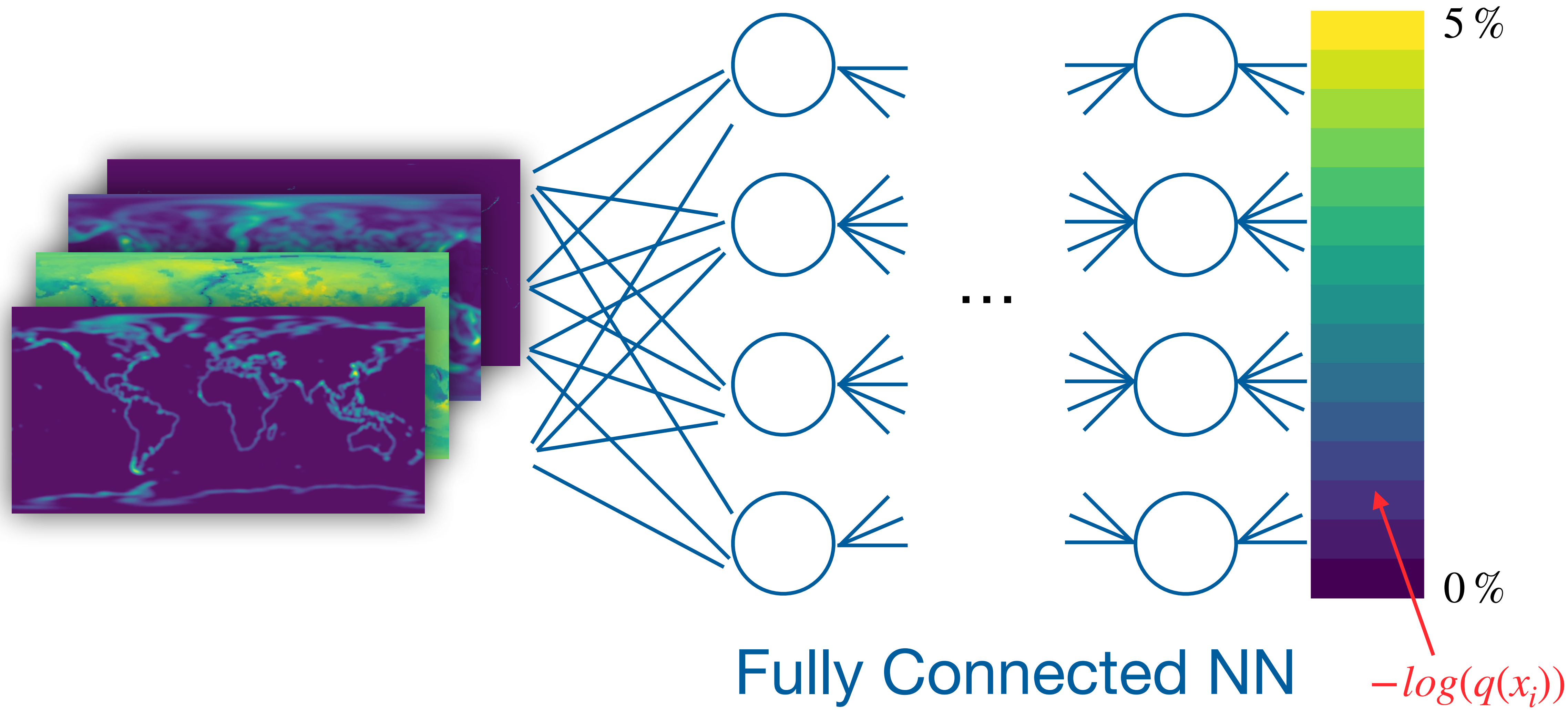
Solution:



This can be solved by digitising the output range…

Neural Networks and Information Gain

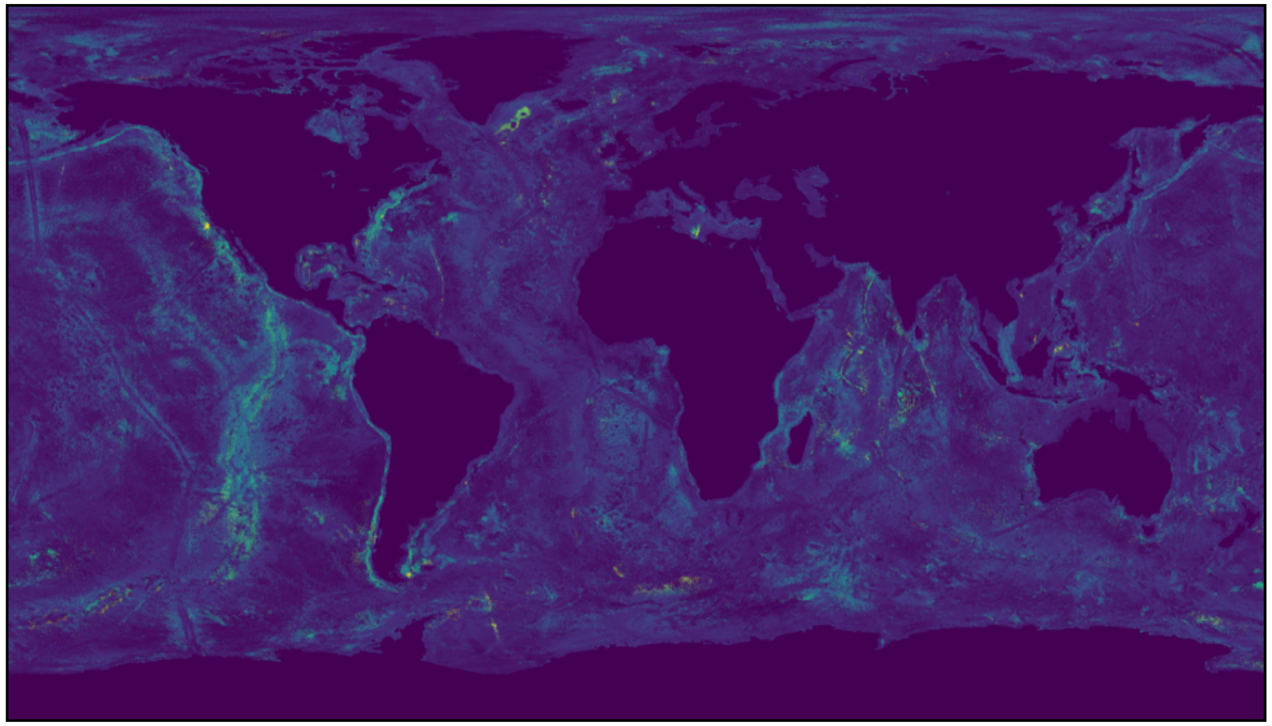
… and applying it to a *softmax* activation layer at the the of the fully connected NN:



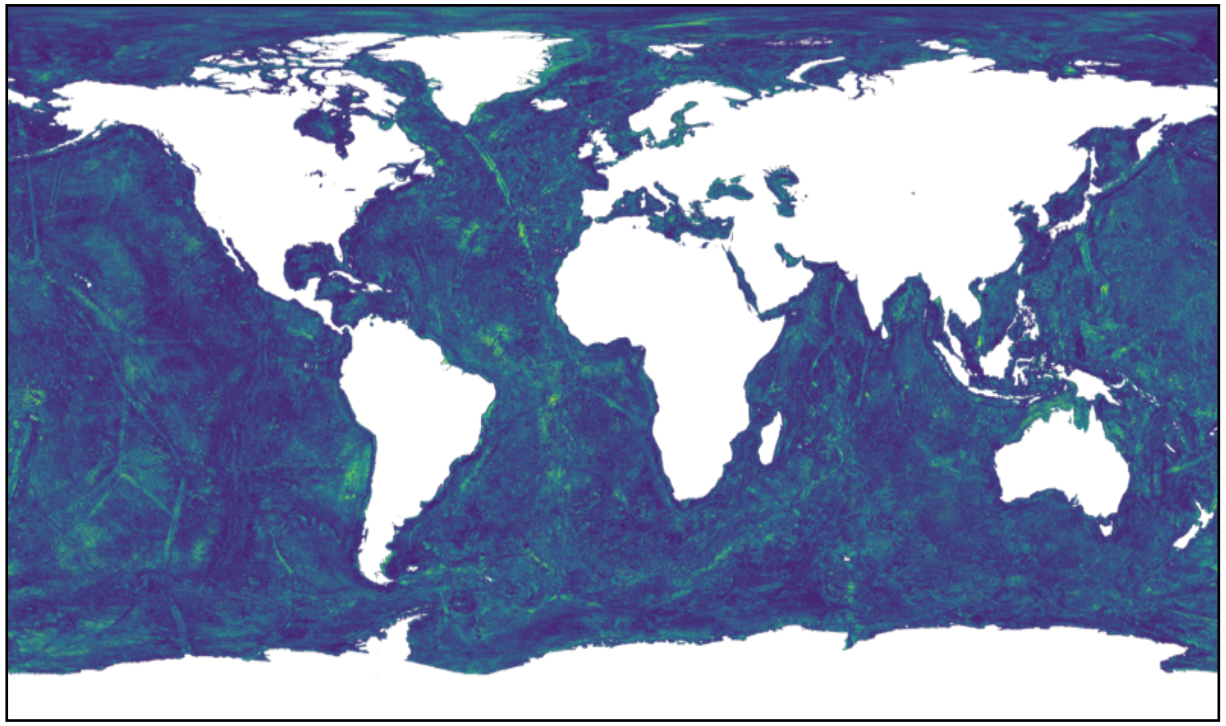
Neural Networks and Information Gain

Such a model produces a prediction along with a probability for it:

Global TOC Predictions



Global Information Gain Predictions



3. Monte Carlo Dropout

Neural Networks and Uncertainty Quantification

Uncertainty Quantification

Back to the question of data-driven sampling:

assuming this model is correct, where should one sample next?

Uncertainty Quantification

Back to the question of data-driven sampling:

***assuming this model is correct*, where should one sample next?**

**This is an uncomfortable assumption to make!
A quantification of model uncertainty would be much preferable.**

Uncertainty Quantification

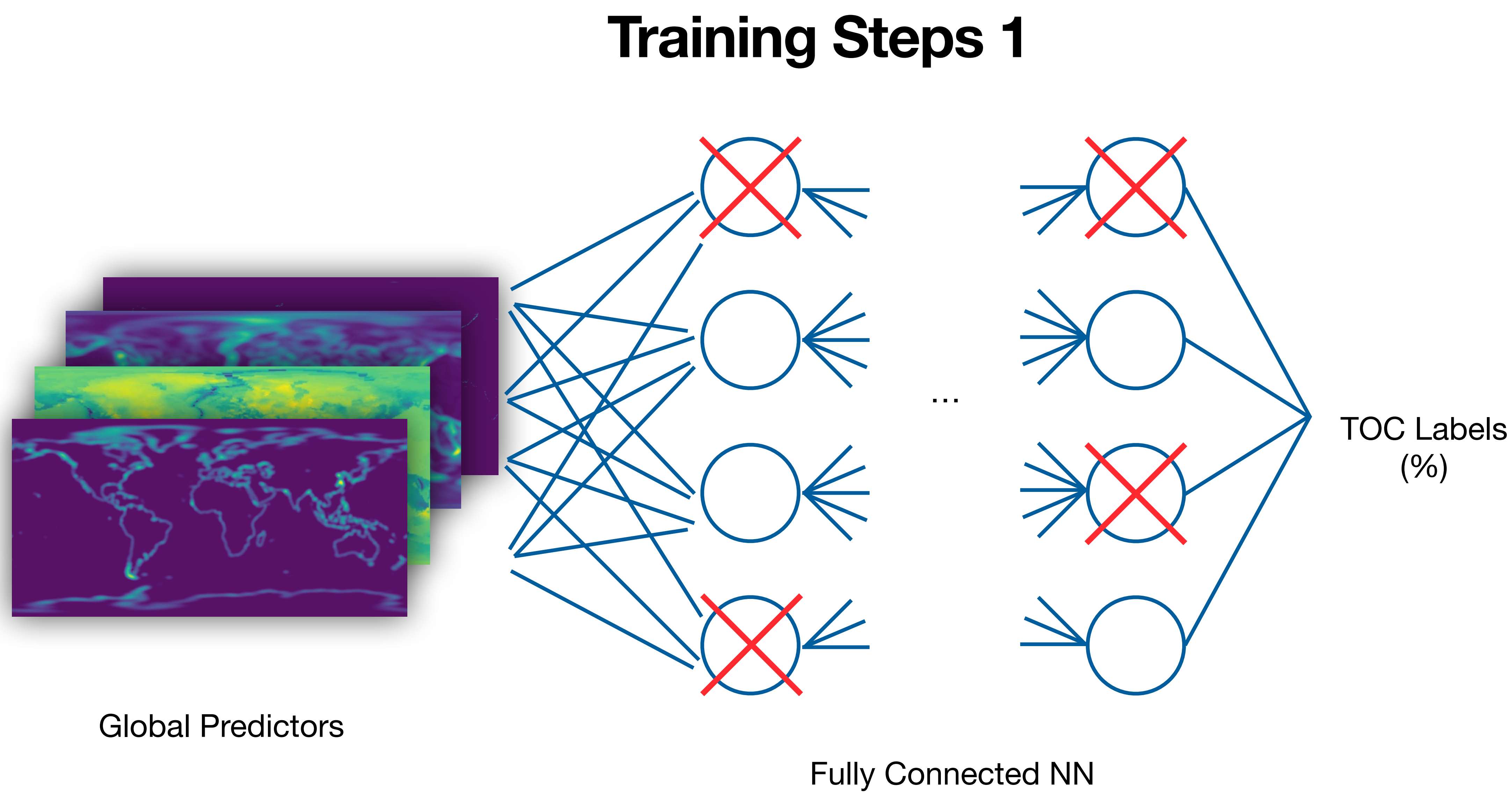
“Dropout[1]” is a standard technique for training neural networks.

It avoids overfitting by randomly deactivating connections between nodes of a neural networks during the training process.

1 https://www.cs.toronto.edu/~hinton/absps/JMLRdropout.pdf

Uncertainty Quantification

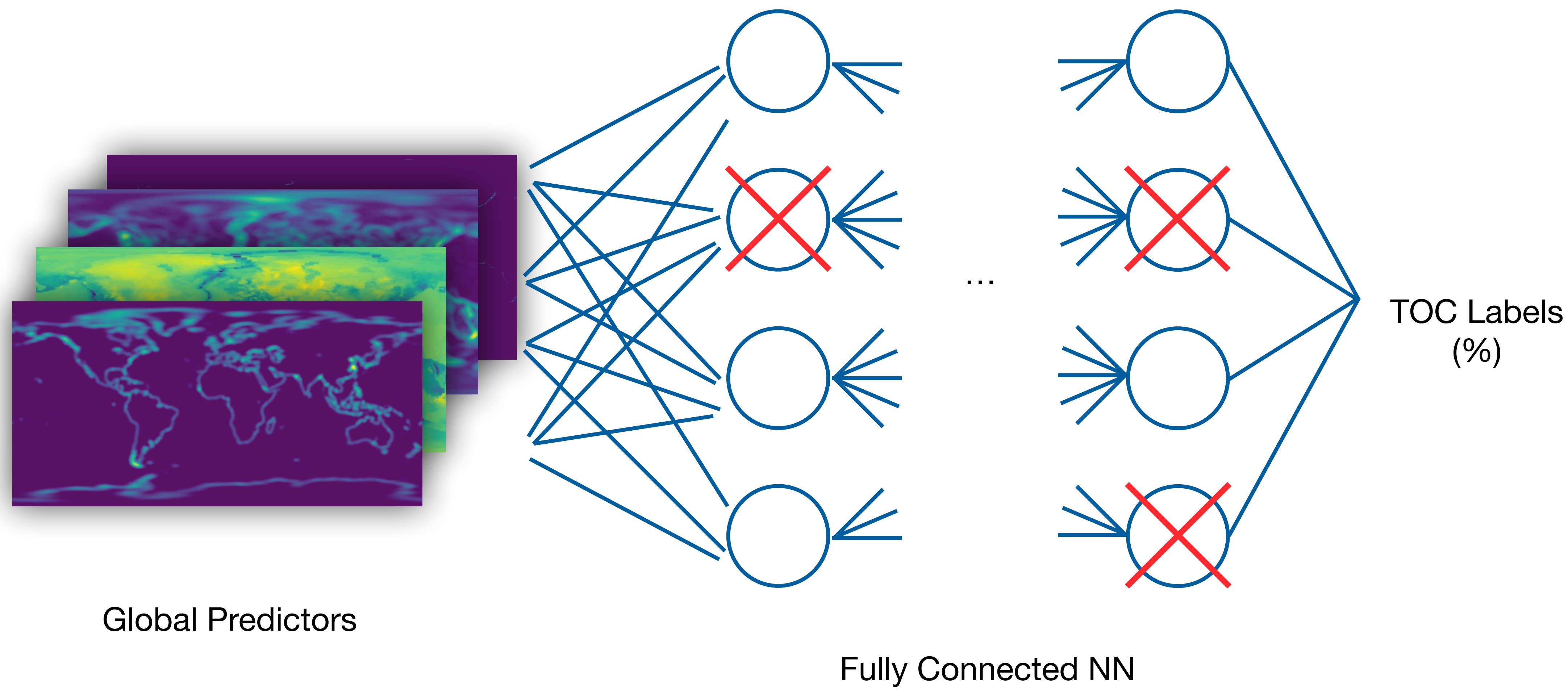
Dropout:



Uncertainty Quantification

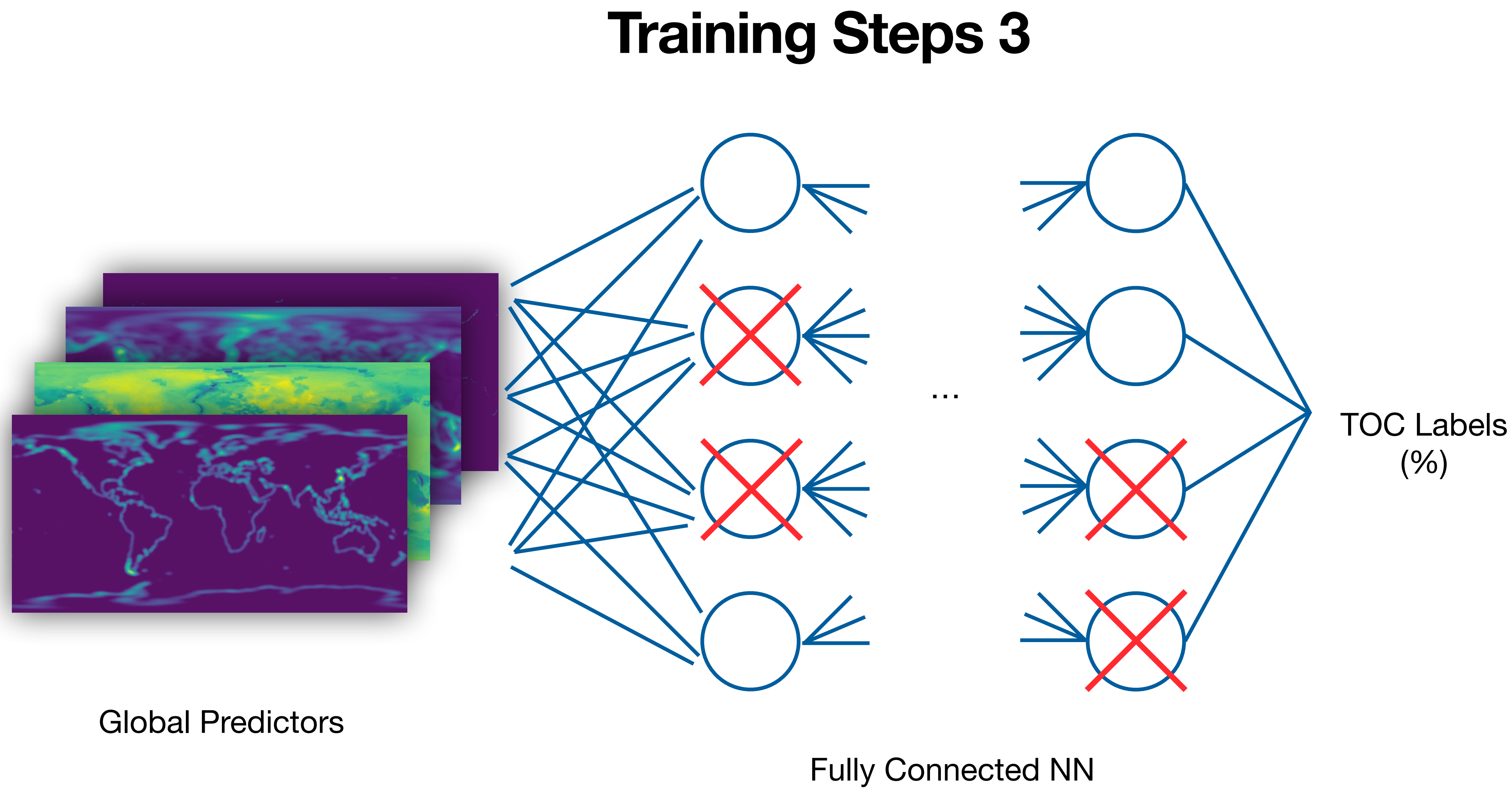
Dropout:

Training Steps 2



Uncertainty Quantification

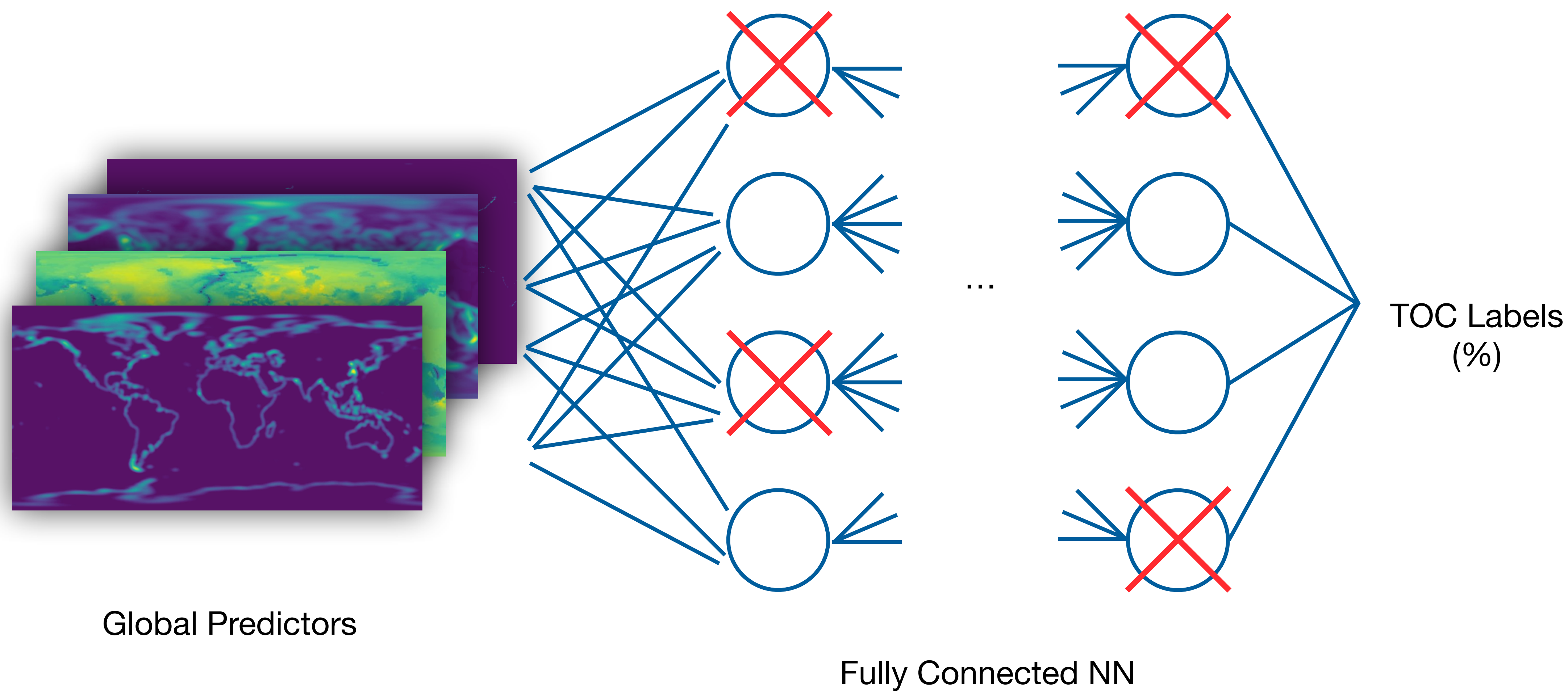
Dropout:



Uncertainty Quantification

Dropout:

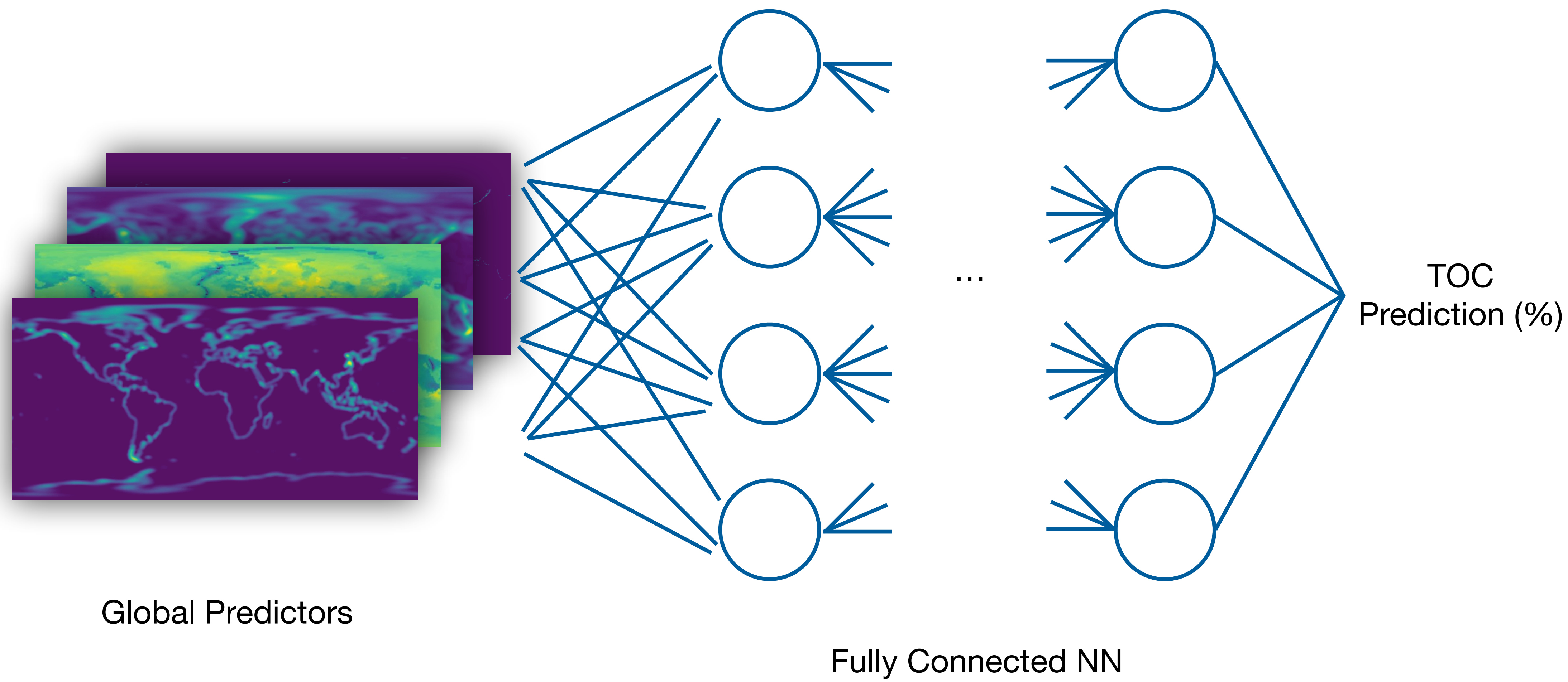
Training Steps n



Uncertainty Quantification

Dropout:

Inference Step



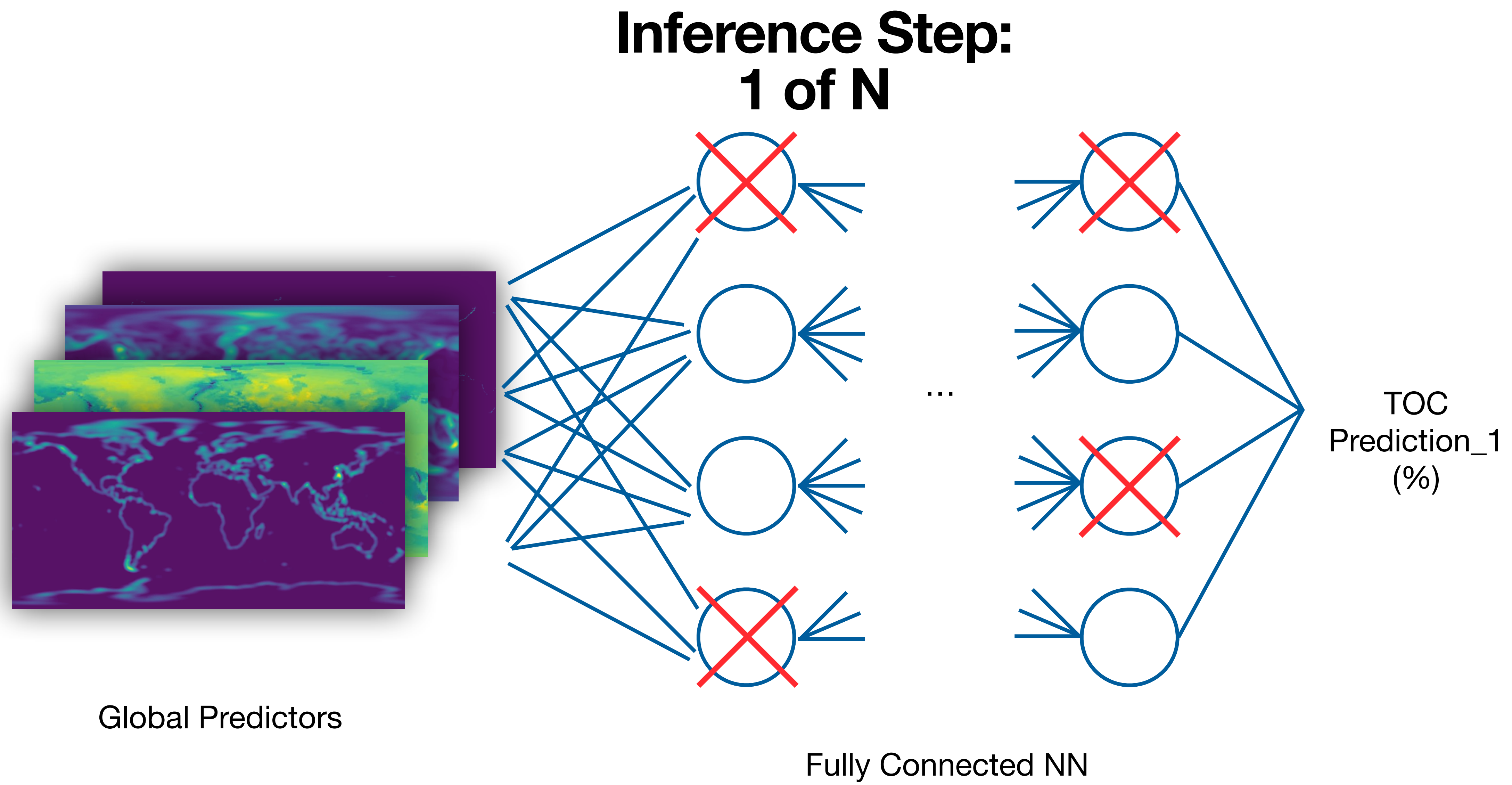
Uncertainty Quantification

Monte Carlo Dropout[1] applies dropout to several repetitions of the inference step, and averages the results to obtain a more robust estimate which also provides a way to quantify uncertainty.

1 https://arxiv.org/pdf/1506.02142.pdf

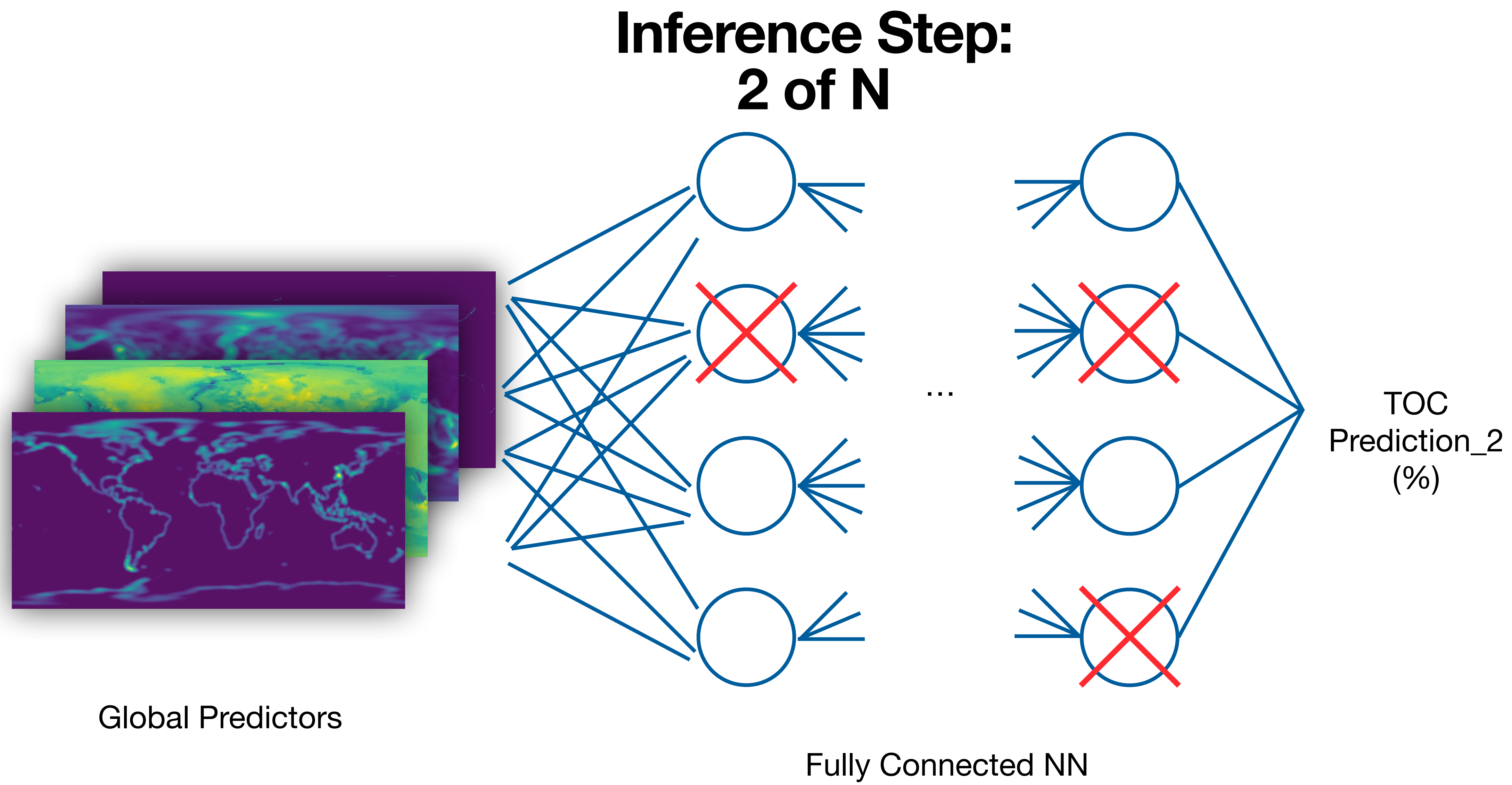
Uncertainty Quantification

Monte Carlo Dropout:



Uncertainty Quantification

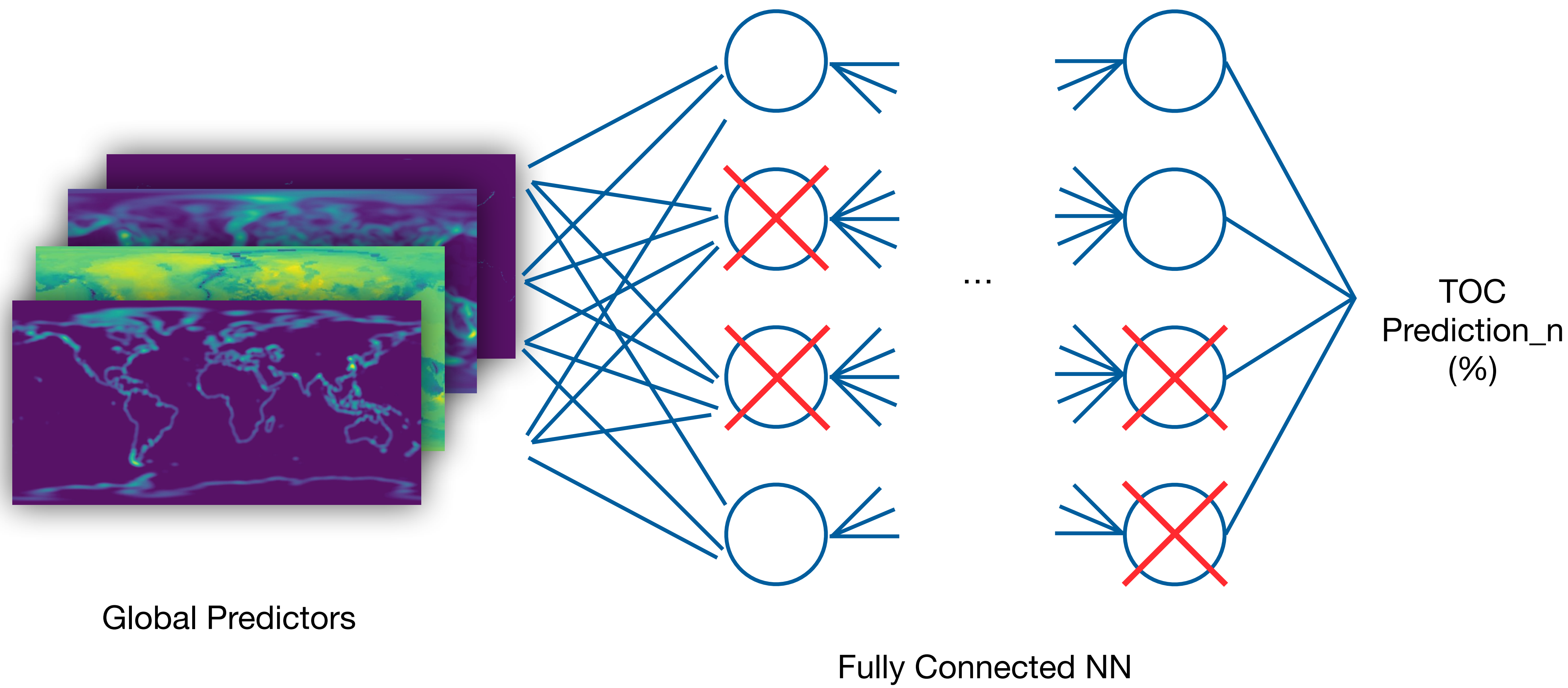
Monte Carlo Dropout:



Uncertainty Quantification

Monte Carlo Dropout:

Inference Step:
n of N



Uncertainty Quantification

Monte Carlo Dropout:

$$Pred\_val = 1/N \sum_{n \in N} Pred\_val_n$$

4. Conclusion

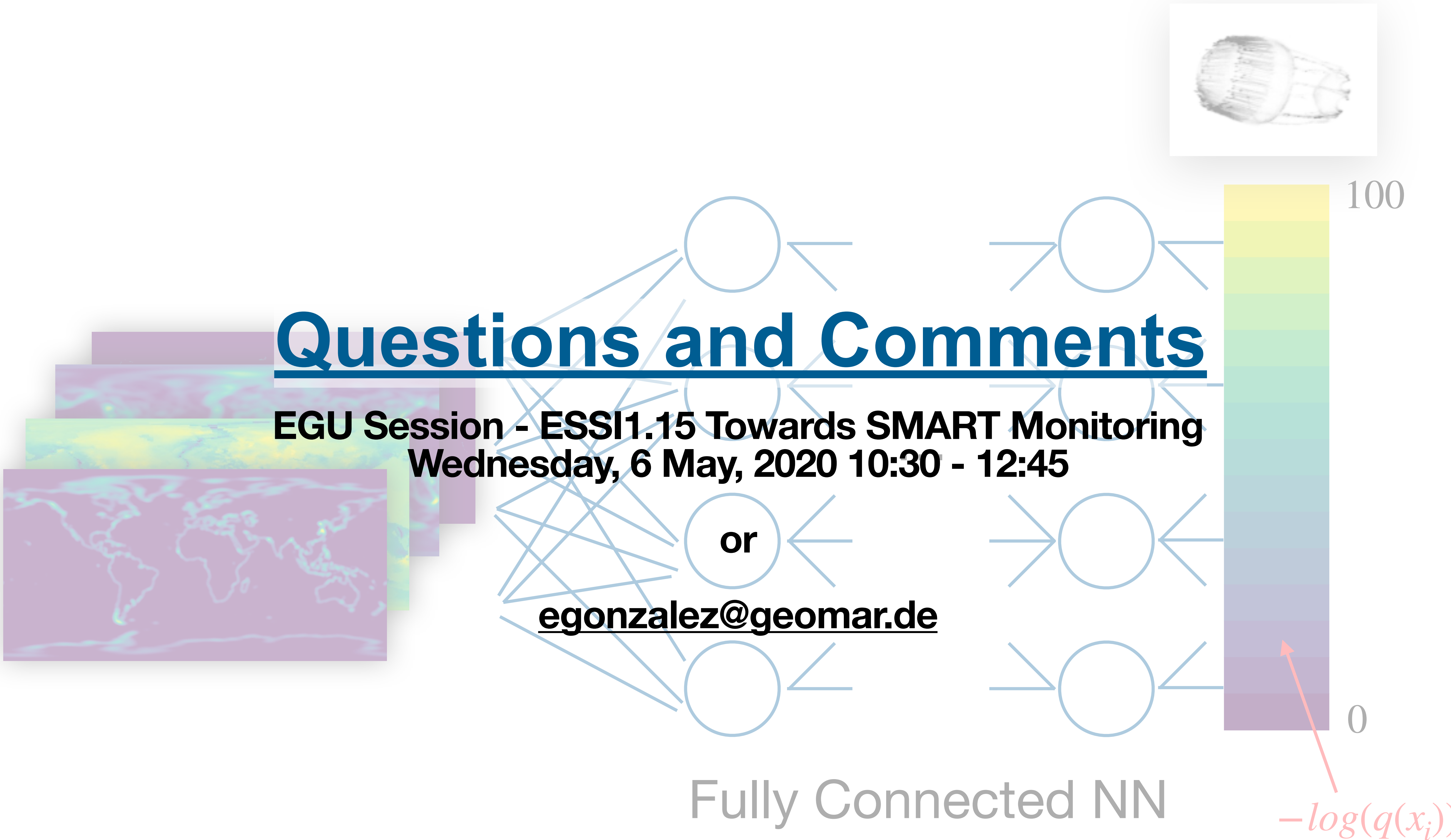
- **A measure for information gain was obtained by applying Softmax distributions to a discretised regression problem.**
- **A measure for model uncertainty was obtained by using MonteCarlo Dropout.**
- **Both techniques, alone or combined, are suitable to answer scientific questions other traditional neural network architectures can not.**

Questions and Comments

EGU Session - ESSI1.15 Towards SMART Monitoring
Wednesday, 6 May, 2020 10:30 - 12:45

or

egonzalez@geomar.de



Questions and Comments

EGU Session - ESSI1.15 Towards SMART Monitoring
Wednesday, 6 May, 2020 10:30 - 12:45

or

egonzalez@geomar.de

Questions and Comments

EGU Session - ESSI1.15 Towards SMART Monitoring
Wednesday, 6 May, 2020 10:30 - 12:45

or

egonzalez@geomar.de