Quantifying the Value of Data in Scientific Machine Learning Models with Likelihood-Weighted Active Learning

Author Names

FUNDING SOURCES or CONTACT INFORMATIO

Sections were chosen selectively to appeal to different audience subsets (a general viewer, machine learning expert, algorithm validation, applications focused). Methods are the primary focus.

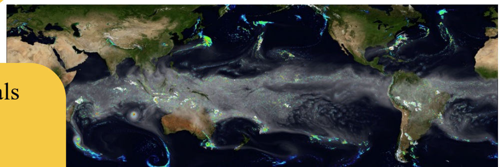
Motivation: Identifying Data Points Most Relevant to Extreme Weather Events

Bullet points with selective emphasis motivate the poster

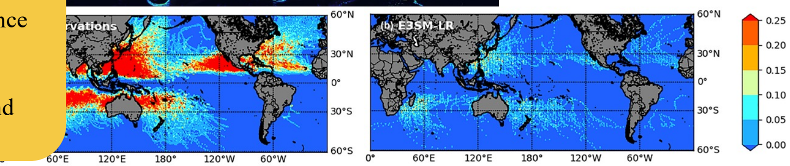
→ Extreme weather events have increased in **severity** and **frequency**.

→ The **broad range of dynamically relevant spatiotemporal scales** in the Earth's atmosphere makes direct numerical simulations computationally expensive and simplified data-driven approaches inaccurate.

→ **Scientific machine learning** methods are a useful substitute but are slow or intractable for large data sets. Active learning can be used to reduce the size of the training data set.



→ Weather forecasting can be improved by **identifying the data points most relevant to the dynamics of extreme events**.



Enticing visuals but might be difficult for a general audience to understand significance (lacking legend detail)

Method: Output-Weighted Active Learning Framework

Ensemble Neural Network (ENN): Train N models with the same architecture and hyperparameters but different weight initializations.

OR

Dropout Neural Network (DNN): Train one model with dropout layers. Make N predictions with different randomly dropped nodes.

Training Data Set → **Probabilistic Model** → **Mean Prediction**: $\overline{y}_N(x) = \frac{1}{N}\sum_{i=1}^{N} y_i(x)$

Probabilistic Model → **Uncertainty**: $\sigma_N^2(x) = \frac{1}{N}\sum_{i=1}^{N}(y_i(x) - \overline{y}_N(x))^2$

Acquisition Function: quantifies value of each point's potential to improve model

$$Q_{IVR-LW}(x) = \frac{\sigma_N^2(x)\, p_x(x)}{p_{\overline{y}_N}(\overline{y}_N(x))}$$

Acquisition Function → **Optimal Data Points** → **Training Data Set**

Goal: Identify points which are useful for predicting extreme events.

In output-weighted uncertainty sampling (or likelihood-weighted integrated variance reduction IVR-LW), we assign value to each data point according to:

1. Its potential to reduce the model's epistemic uncertainty
2. Its high likelihood of occurrence
3. Its high likelihood of producing extreme outcomes (tail of the distribution)

Color coded variables easily link to the text descriptions

Synthetic Problem: Majda-McLaughlin-Tabak (MMT)
A One-Dimensional Model for Dispersive Wave Turbulence

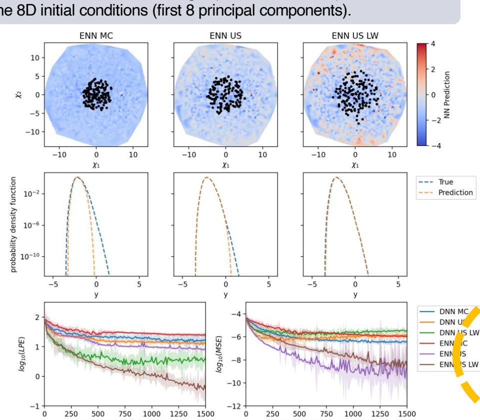
Goal: Predict the maximum future wave height (an extreme event) as a function of the 8D initial conditions (first 8 principal components).

→ We use *multi-dimensional scaling* (mdscale) to visualize the spread of the 8D selected points (black dots) in 2D. We compare uncertainty sampling (US) and likelihood-weighted US (US LW) with Monte Carlo (MC).

→ We measure success with respect to how well the model predicts the *tails of the distribution* via the log-pdf error (LPE).

$$\text{LPE} = \int |\underbrace{\log p_y(y)}_{\text{true pdf}} - \underbrace{\log p_\mu(y)}_{\text{predicted pdf}}| dy$$

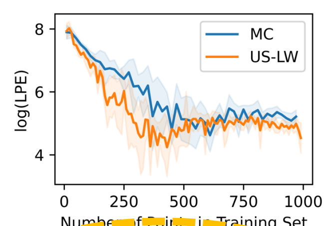
→ The likelihood-weighted uncertainty sampling (US LW) achieves lower error with fewer points than the other selection schemes. ENN requires more training time but outperforms DNN.

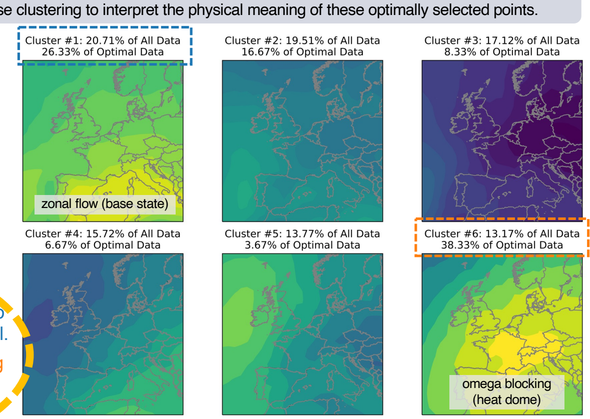


The author was not at this poster the whole time. The layered levels of information with a clear visual difference between the "quick take-away" and details.

Real-World Problem: Correction Operators for Coarse-Scale Climate Models
Mapping Low-Resolution E3SM to High-Resolution ERA5 Reanalysis

Goal: Use the output-weighted active learning framework to find the subset of points that most optimally map low-resolution climate models to high-resolution climate models. Use clustering to interpret the physical meaning of these optimally selected points.





Most of the points in the full data set (or points that are randomly selected) belong to Cluster #1, corresponding to standard zonal.

Most of the optimally selected points belong to Cluster #5, corresponding to omega blocking (extreme weather events).

Results are color coded and boxed to help the viewer read this section

A references section or link to published results could be useful for the audience to learn more

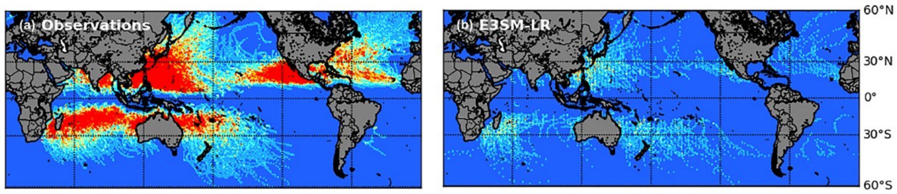
Catchy title draws in the audience

Overcoming Fear Of Missing Out (FOMO)

Active Selection of Training Points to Predict Extreme Weather Event Statistics

Author Names

The author stood next to this poster the entire time it was presented. They leverage the poster to be mostly visuals and rely on talk through more detailed information.



Climate modeling with machine learning can be sped up and improved by using ...elihood-weighted active selection (US-LW)** to choose a subset of the data for training.

Ensemble Neural Network (ENN): Train N models with the same architecture and hyperparameters but different weight initializations.

$$\bar{y}_N(x) = \frac{1}{N}\sum_{i=1}^{N} y_i(x)$$

$$\sigma_N^2(x) = \frac{1}{N}\sum_{i=1}^{N} (y_i(x) - \bar{y}_N(x))^2$$

Initial Data Points → **Training Data Set** → **Probabilistic Model** → **Mean Prediction** / **Uncertainty** → **Acquisition Function:** quantifies the value of each point's potential to improve the model → **Selected Optimal Data Points** → **Training Data Set**

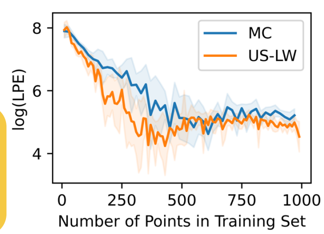
Section divisions are clean and not distracting with colors/lines that don't hold meaning

$$q_{US-LW}(x) = \frac{\sigma_N^2(x)p_x(x)}{p_{\bar{y}_N}(\bar{y}_N(x))}$$

Value is assigned to each data point according to its

1. Potential to reduce the model's uncertainty
2. High likelihood of occurrence
3. High likelihood of producing extreme outcomes

US-LW achieves a **lower error** with **fewer training points**.



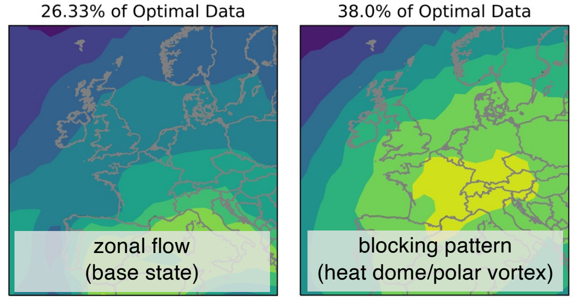
$$\mathrm{LPE} = \int |\underbrace{\log p_y(y)}_{\text{true pdf}} - \underbrace{\log p_\mu(y)}_{\text{predicted pdf}}|\mathrm{d}y$$

Background color to emphasize key contribution

With no prior knowledge of climate physics, US-LW **identifies points relevant to extreme weather events**.

Cluster #1: 20.82% of All Data
26.33% of Optimal Data

Cluster #6: 13.16% of All Data
38.0% of Optimal Data



Compared to the more detailed version of this poster, only the key visuals have been selected.

Obeyed formatting requirements for the competition

MITMECHE