EXTENDING TWO EXPLAINABLE ARTIFICIAL INTELLIGENCE METHODS FOR DEEP CLIMATE EMULATORS

Wei Xu*, Rui Qiu, Xihaier Luo, Yihui Ren, Balu T. Nadiga, Luke Van Roekel, Han-Wei Shen, Shinjae Yoo; Brookhaven National Lab, Los Alamos National Lab, The Ohio State University

Climate Prediction Challenges

Deep climate emulators based on Implicit Neural Representations (INRs) can predict and reconstruct complex spatiotemporal fields from sparsely distributed sensors, but their "black-box" nature limits trust and domain insight. To address this, we integrate two Explainable AI (XAI) methods—gradient-based feature attribution and influence functions—to reveal how an INR model learns physical dependencies from both input features and training examples.

Methodology

- **Feature Attribution**
  Compute spatial gradients $\nabla_x u(x,t)$ and latent-code gradients $\nabla_z u(x,t)$ to quantify sensitivity of the predicted field to small perturbations in coordinates and learned context variables.
- **Influence Functions**
  Approximate the effect of up-weighting a single training point (x_j, t_j, u_j) on a target prediction (x_{test}, t_{test}) via

$$I(x_{test}, t_{test}; j) \approx -\nabla_\theta f_{\hat{\theta}}(x_{test}, t_{test})\, H_{\hat{\theta}}^{-1}\, \nabla_\theta L\big(f_{\hat{\theta}}, (x_j, t_j, u_j)\big)$$

  identifying which historical measurements drive each forecast.

Key Results

- **Spatial Sensitivity Map** (*Figure 1*)
  The standard deviation of $\nabla_x D_\phi(z_t, x)$ over N = 1024-time steps highlights "hotspots" (e.g., ocean currents, land–sea boundaries) where small shifts in x cause large output fluctuations—guiding where additional sensors would most reduce uncertainty.
- **Influence Map** (*Figure 2*)
  By plotting $I(x_{test}, t_{test})$ for every training point, we observe tight clusters of highly influential coordinates (red dots) that align with known teleconnections and reveal potential outliers for data-quality checks.

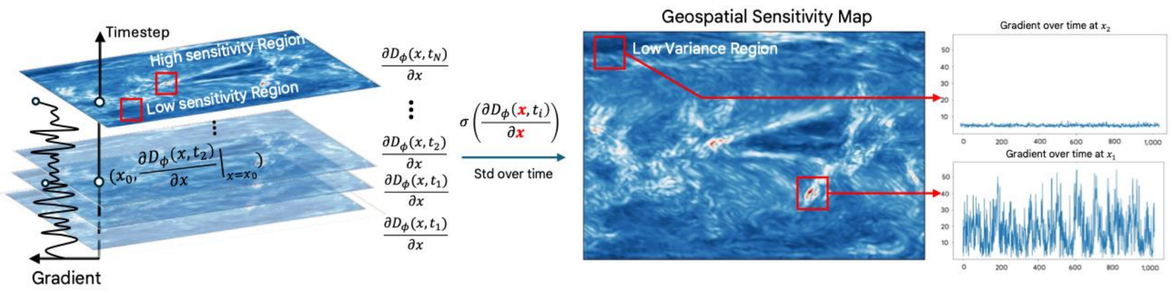


Figure 1: Standard deviation map of spatial gradients $\nabla_x D_\phi(z_t, x)$ over N = 1024-time steps. Higher values (red regions) suggest areas where small changes in x lead to large temporal fluctuations in the predicted field.

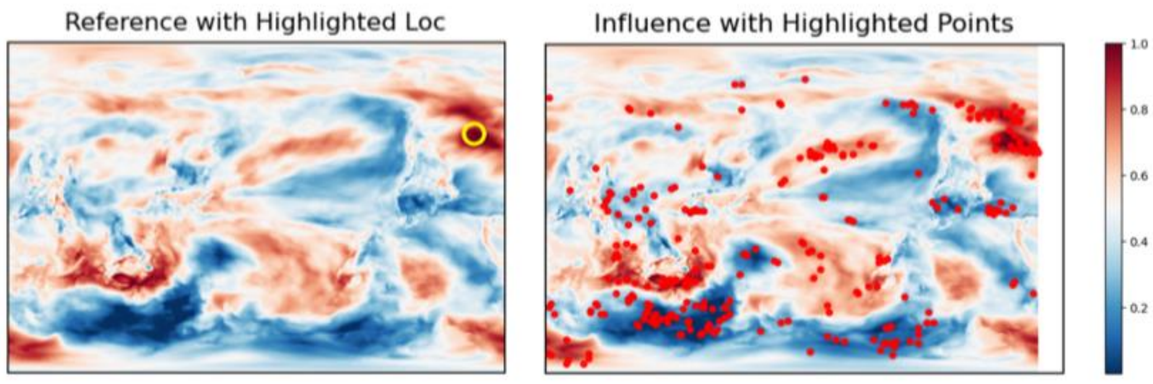


Figure 2. Left: The reference temperature field at a target time step, with a chosen location of interest (yellow circle). Right: Influence map (red dots) indicating which training coordinates exert the greatest positive or negative influence on the chosen prediction.

Conclusions & Next Steps

- Our dual-XAI pipeline—combining gradient-based feature attribution with influence functions—provides both local (input sensitivity) and global (training-example impact) interpretability for INR-based climate emulators.
- Applied to an INR model, spatial-sensitivity maps highlight physically meaningful hotspots (e.g., ocean currents, land–sea boundaries), while influence maps pinpoint the most impactful training points for any target prediction—greatly enhancing trust and diagnostic insight

Next Steps

- Active Learning for Sensor Deployment leverages spatial-gradient variance and influence scores to identify regions of highest uncertainty, guiding the placement of new sensors and iterative data collection.

This paper is supported by Department of Energy's SciDAC ImPACTS and SciDAC RAPIDS2 projects.

U.S. DEPARTMENT of ENERGY

www.bnl.gov

Brookhaven National Laboratory