The team collaborated with the AI for Climate (AICE) initiative to package a research benchmark developed by University of Chicago climate researchers for forecasting the onset of the Indian monsoon, an event that determines farming decisions for more than 38 million farmers across India.
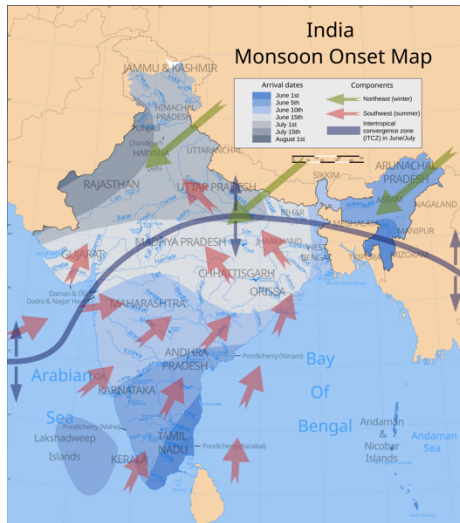


Figure 1 Indian monsoon onset map showing forecast target.

While recent advances in AI weather models have shown promise on top of decade-long advances in physical models, the sheer volume of emerging architectures meant that scientists lacked a standardized, reproducible way to compare their performance.

To address this, the team transformed a research-grade benchmarking workflow into a modular, open-source Python package, enabling climate researchers to evaluate and visualize model skill with minimal setup.

Legacy command-line scripts, originally built for research with repetitive code and hard-coded parameters, were re-engineered into an object-oriented API, featuring flexible data loaders, increased customizability and organization within metrics, and automated generation of model-comparison visualizations and onset maps.

This standardized benchmarking package facilitates rigorous evaluation of, and hopefully eventual well-founded trust, in AI-driven weather prediction, ensuring that improvements in modeling meaningfully support the millions of Indian farmers whose success depends on accurate forecasting of the Indian monsoon.

Next, the team plans to integrate probabilistic metrics (e.g. Brier and Ranked Probability Scores), build an interactive GUI for benchmarking and visualization, and enable state-level comparisons to support more localized, human-centered forecasting across India.
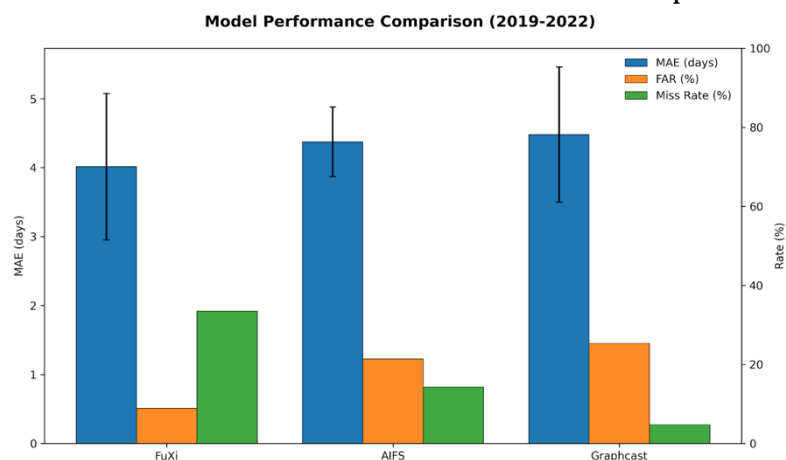


Figure 2 Automated model comparison output from the new benchmarking package.