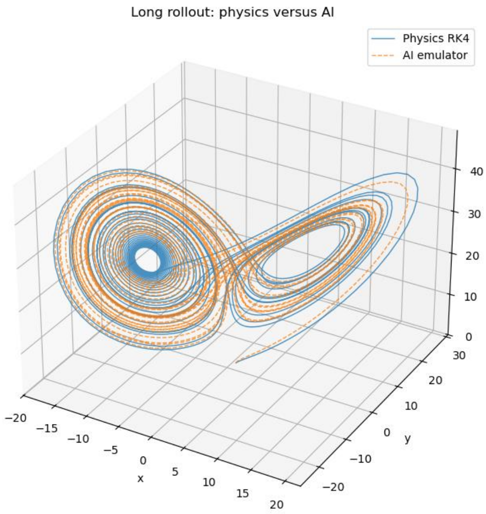
From Lorenz to AI Weather Models: A Minimal Testbed for Predictability, Stability, and Generalization



Data-driven weather and climate models are increasingly used as fast emulators of complex atmospheric dynamics. However, strong short-term prediction skill does not necessarily imply physically meaningful long-term behaviour. A model may perform well for one-step or short-range forecasts, yet drift away from realistic dynamics when rolled out autoregressively.

In this challenge, participants will use the Lorenz system as a controlled testbed for studying neural emulators of chaotic dynamics. Although simple, the Lorenz system captures several key properties that are central to weather and climate prediction: nonlinear evolution, sensitivity to initial conditions, uncertainty growth, regime dependence, and long-term stability. This makes it a useful minimal setting for asking questions that also arise in modern AI-based weather and climate models.

Teams will train machine-learning models to predict Lorenz trajectories and evaluate them not only using standard forecast errors, but also through dynamical diagnostics. These may include sensitivity to initial conditions, long-term rollout stability, attractor reconstruction, state distributions, and generalization to perturbed system parameters.

The challenge connects a simple chaotic system to core questions in AI-based Earth-system modeling: predictability, model drift, physical consistency, uncertainty growth, and climate-like stability. Advanced teams may explore recurrent models, neural ODEs, probabilistic emulators, hybrid physics–ML approaches, or physics-informed losses. The goal is not only to build an accurate emulator, but also to understand when and why a learned model behaves like, or fails to behave like, the underlying dynamical system.

Participants will be encouraged to progress through levels of difficulty, starting with a simple neural emulator and moving toward more advanced diagnostics and model design choices.

- Module 1: Build a baseline emulator
- Module 2: Evaluate beyond RMSE
  - Short-term forecast error, Long-term rollout stability, Attractor reconstruction.
- Module 3: Sensitivity and ensemble behaviour
  - Initialize two nearby states and compare divergence in the physical Lorenz model and the AI emulator.
- Module 4: Generalization under regime shift
  - Train on one Lorenz parameter setting, for example standard $\rho=28$, then test on altered parameters: $\rho=24, 30, 35$
  - Or teams may train on only part of the Lorenz attractor and test whether the model behaves reasonably in previously unseen regions of state space.