

SEP 22-26

AIM for Scale
AI Weather Training Program



WEATHER PACKAGE

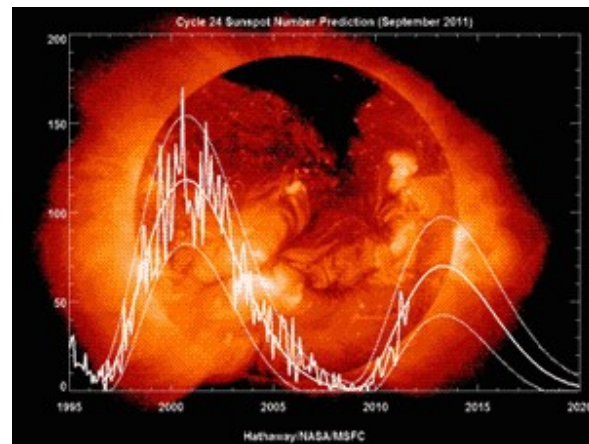
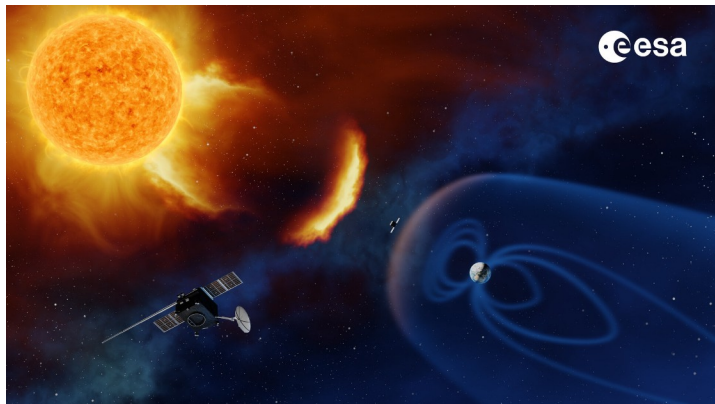
Benchmarking: Motivations and Ecosystem

Katie Kowal

Why human-centered operational benchmarks?

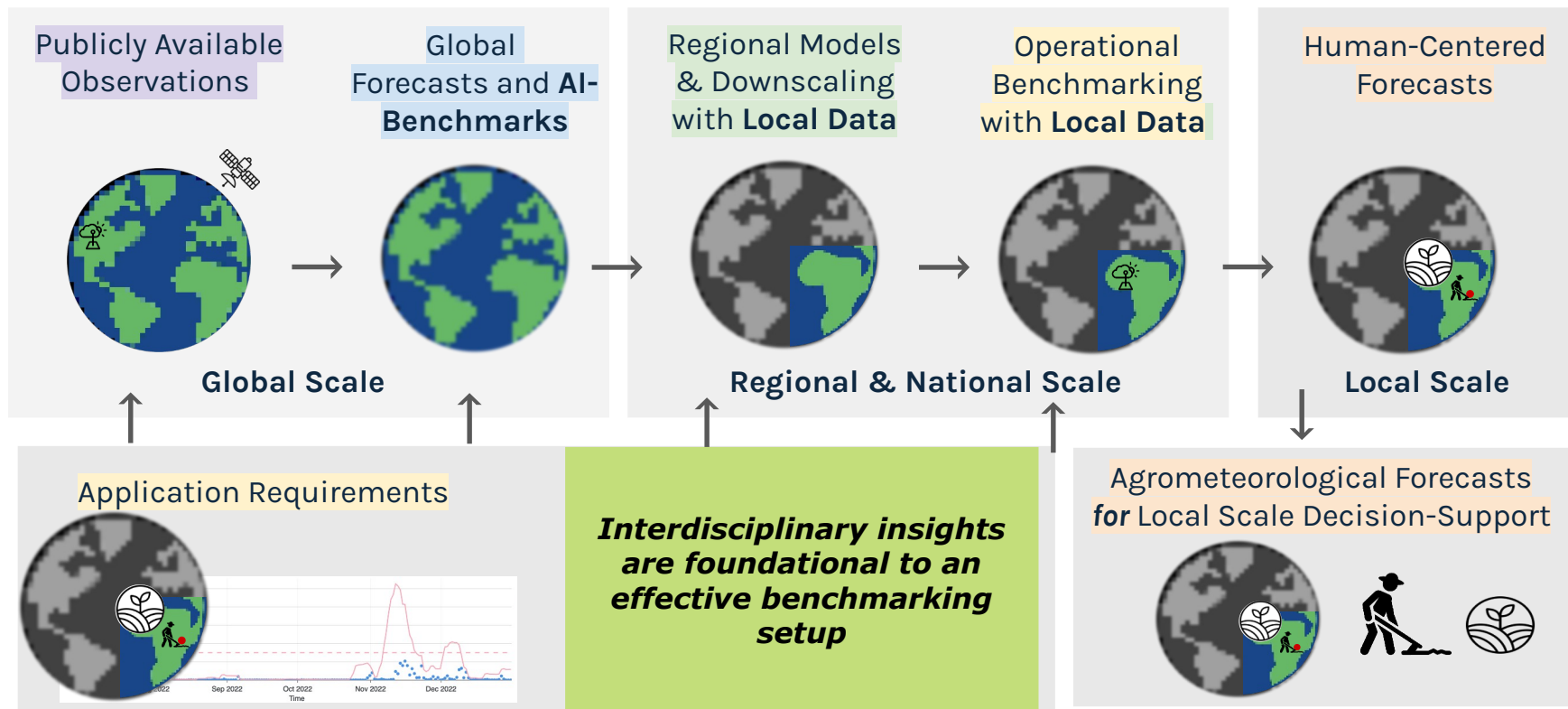
There are often two types of failures in making models useful

Failures of coordination



Failures of imagination

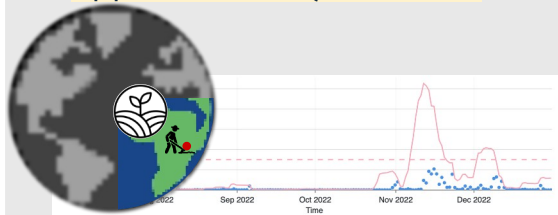
Operational benchmarking / verification: Helping NMHSs decide which novel forecasting models and outputs (if any) to operationalize.



In a holistic benchmarking ecosystem, several ingredients needed to improve forecast usefulness for agricultural decision-support

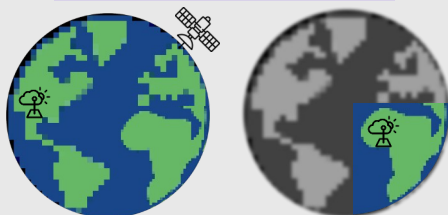
Set **task** for forecast requirements

Application Requirements



Check **data quality** given task

Global and Local Data



Identify **metrics** to select models
Identify **baselines** to evaluate models

Global Forecasts



Regional Forecasts



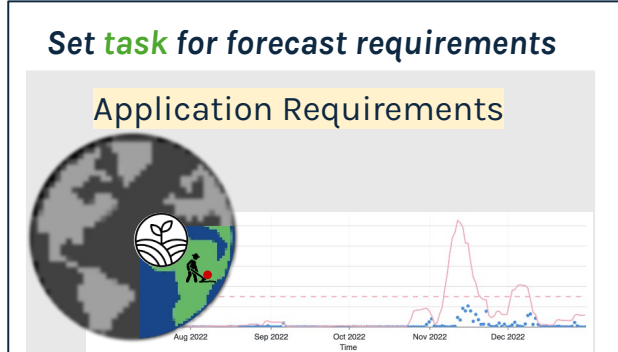
Agrometeorological Forecasts for Local Scale Decision-Support



Task-setting example: planting timing

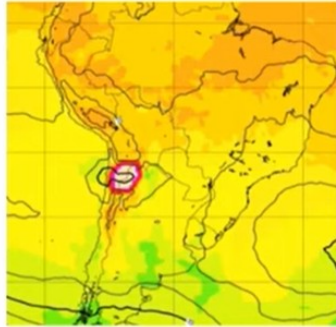
Task setting components

- **who** needs the forecast?
 - individuals in rainfed vs. irrigated zones
 - extension agent with more training
 - a government strategizing resources
- **what** information do they require?
 - wet season onset
 - extreme weather alerts
- **where** is the forecast needed?
 - geography considerations
- **when** is the forecast needed?
 - lead times for planning
 - time of year
- **how** do they need the forecast?
 - communication channels (e.g. radio, WhatsApp)
 - agroclimate advisory bulletins



***Task requirements
can affect how a
forecast is generated,
verified and improved***

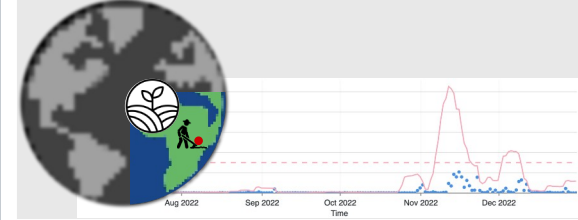
Examples from recent AIFS ENS discussion at ECMWF - Known issues



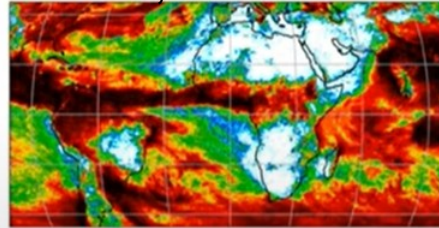
Unphysical pressure and temperature values can develop in mountainous regions

Set **task** for forecast requirements

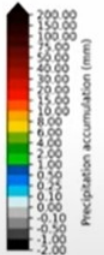
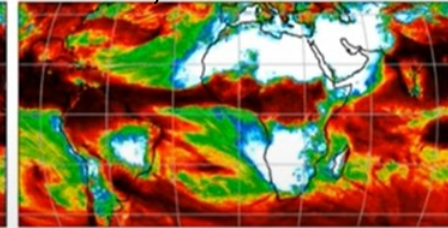
Application Requirements



AIFS, Forecast 240h



IFS, Forecast 240h



Very small values (e.g. $< 0.1\text{mm}/6\text{h}$) of precipitation can occur in arid regions. Particularly noticeable when looking at long accumulation periods

Source: Recent ECMWF developer talk on AIFS ENS - talks can be a great way to keep up with the latest advances and opportunities for improvements - see <https://ai4farmcast.ai/materials.html> where we have started to archive these kinds of talks for reference

Feedback can inform model development - if you spot an issue, can report it at <https://confluence.ecmwf.int/display/FCST/Known+AIFS+ENS+Forecasting+Issues>

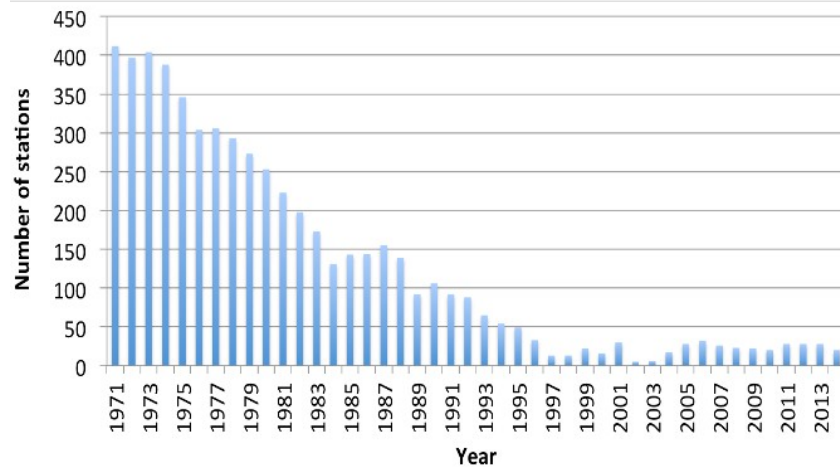
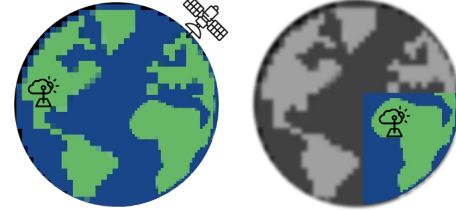
Data quality is not one size fits all

Data quality considerations

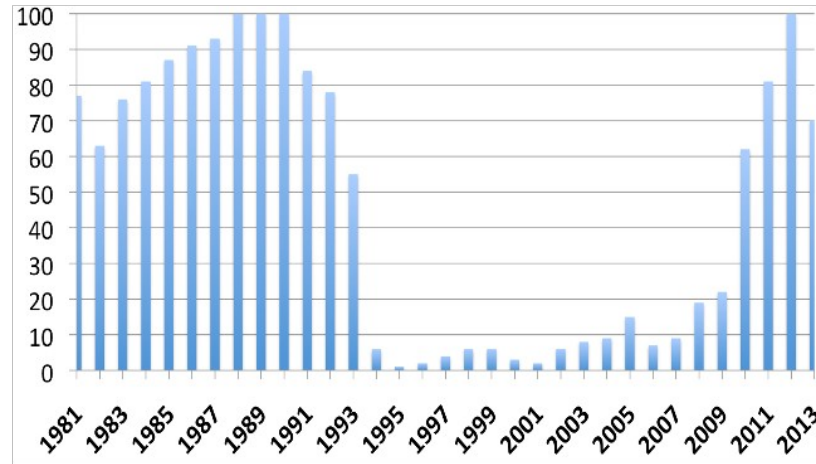
- Quality varies based on time
- Ideal scenario of perfect data is unrealistic - how to fill this gap requires a multi-tool approach

Check **data quality** given task

Global and Local Data



Declining investment in observation infrastructure

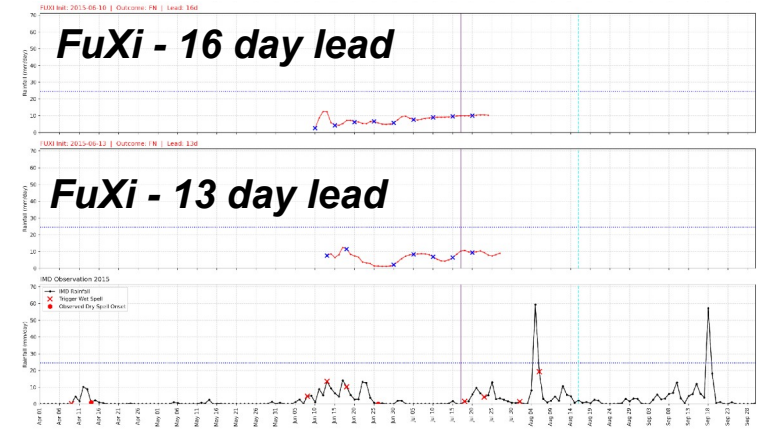
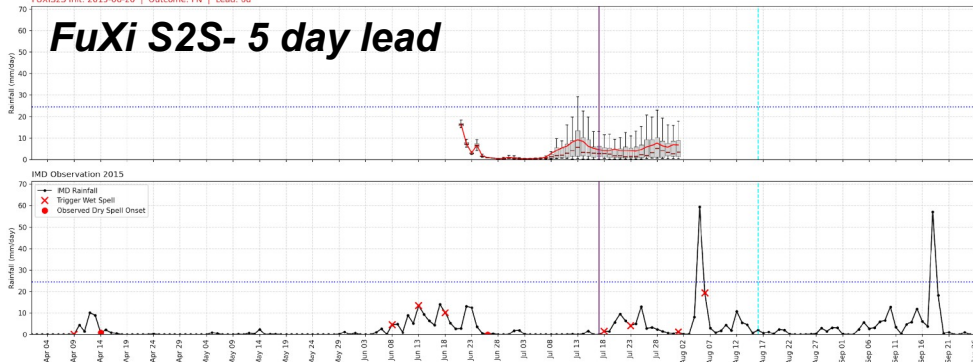
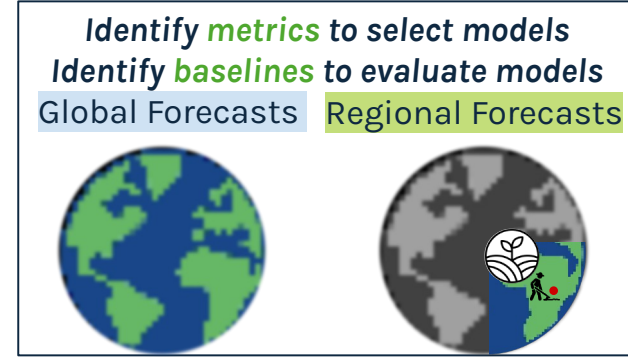


Conflict/unrest: distraction and disruption

Multiple tests of skill can reveal different aspects of model performance

Models may be wrong for different reasons - how they are wrong matters for potential to bias correct their results

Some example cases with some misses in dry spell detection with FuXiS2S and FuXi



Benchmarking 5- to 30-day Forecasts of Indian Monsoon Onset

Comparing AI, NWP, and
Climatology

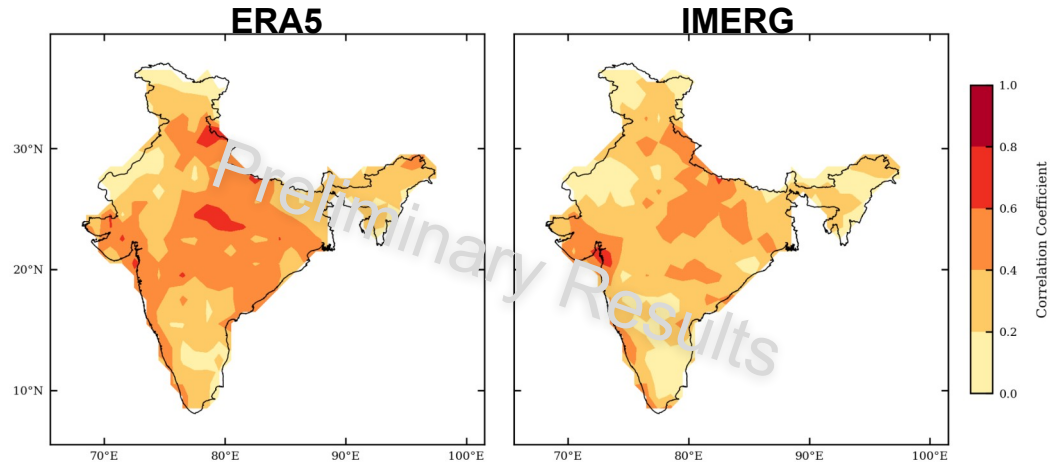
Baseline & NWP	Years
Climatology (history)	1901-2023
IFS/IFS S2S (NWP)	2004-2024

AI models	Training	Fine-tuning	Testing
AIFS	1979-2020	2019-2020 (IFS HRES)	2021-2023, pre 1979
GenCast*	1979-2018	None	2019-2023
GraphCast	1979-2017	2016-2021 (IFS HRES)	2022-2023, pre 1979
NeuralGCM (IMERG)*	2001-2018	None	2019-2023, pre 1979
FuXi	1979-2015	2016-2017	2018-2023, pre 1979
FuXi S2S*	1950-2016	None	2017-2021

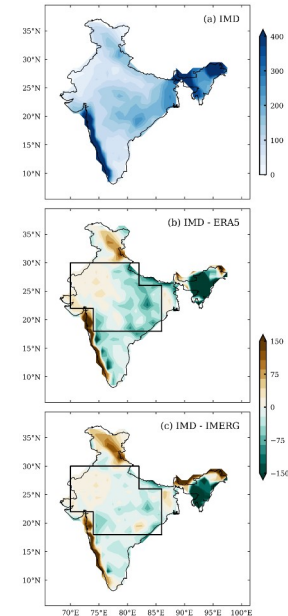
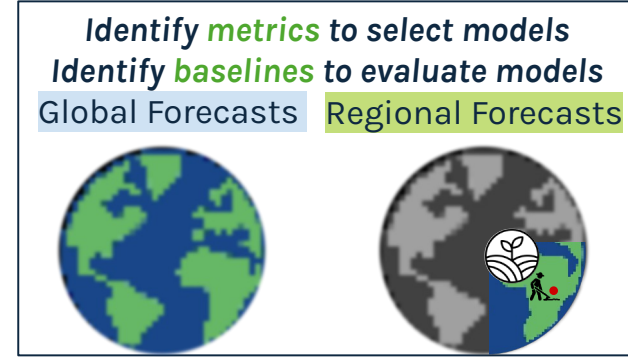
- Small test sample size is a major challenge for subseasonal-to-seasonal (S2S) benchmarking
- Models vary in data they need for operation, cost, forecasted variables, etc. (e.g., soil moisture in only one model)

Metrics and baselines matter

- **Baselines matter** for effective comparison (few use a bias corrected version of IFS to compare)
- **Different metrics** can reveal different things

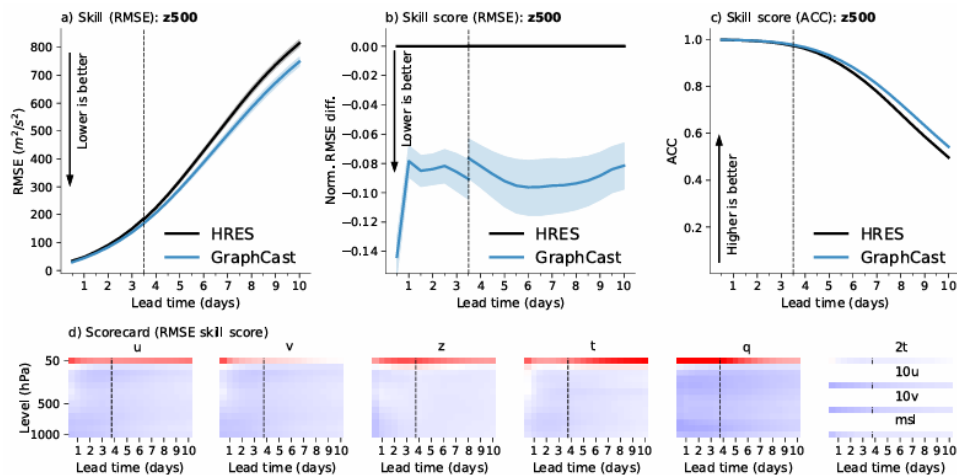


Example - when our team tested ERA5 and IMERG daily rainfall estimates in June vs IMD rain gauge data using ACC - ERA5 had higher scores over core monsoon zone



But when we examined the biases, IMERG had smaller biases compared to IMD gridded 0.25' rainfall data on average

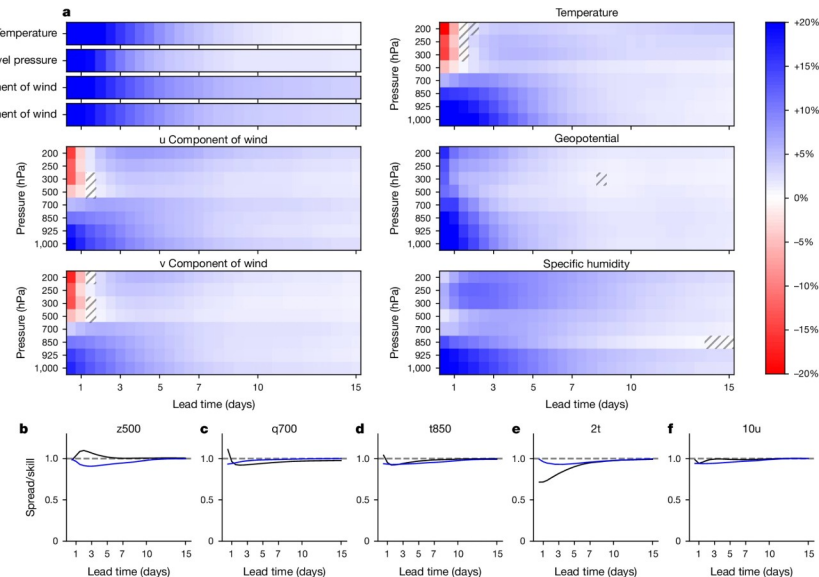
Graphcast demonstrates skill relative to ECMWF HRES



Graphcast

Lam et al. 2023 (Science): 300M parameters (graph neural nets)

Gencast shows skill relative to ECMWF ENS



Gencast

Price et al. 2024 (Nature): (conditional diffusion model)

Training goals - understanding common AI model scorecard approaches to set ourselves up to close more gaps between scientific and operational benchmarking

Benchmarking Questions

What metrics do you use to evaluate models today?

What is your process for assessing a rainfall output? A temperature forecast? wet season onset spell?