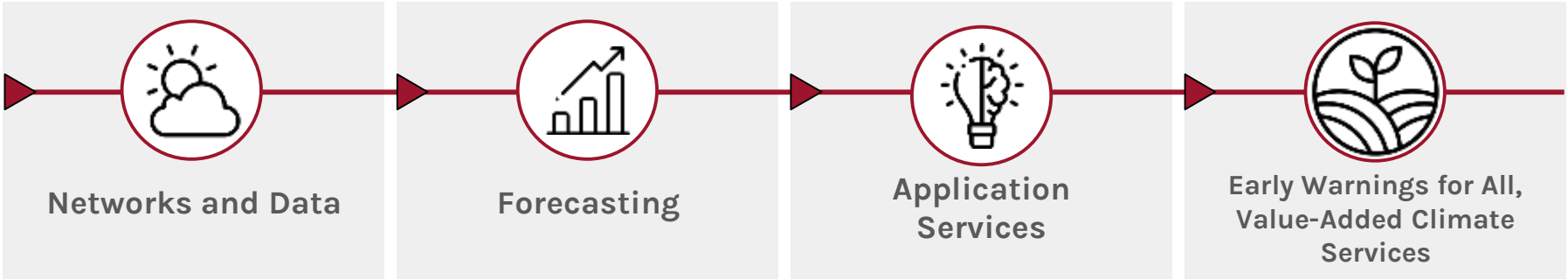
SEP 22-26
AIM for Scale
AI Weather Training Program

AIM FOR SCALE | WEATHER PACKAGE

Operational Benchmarking Goals and Practice

Dr. Genevieve Flaspohler

Partnership to Support NMHS Goals



Rhiza Research is a US-based non-profit that specializes in:

- AI-based forecasting and downscaling
- Weather station analysis and QC
- Technical infrastructure, software, and data systems

and can provide NMHSs technical and capacity support across their end-to-end forecasting system.

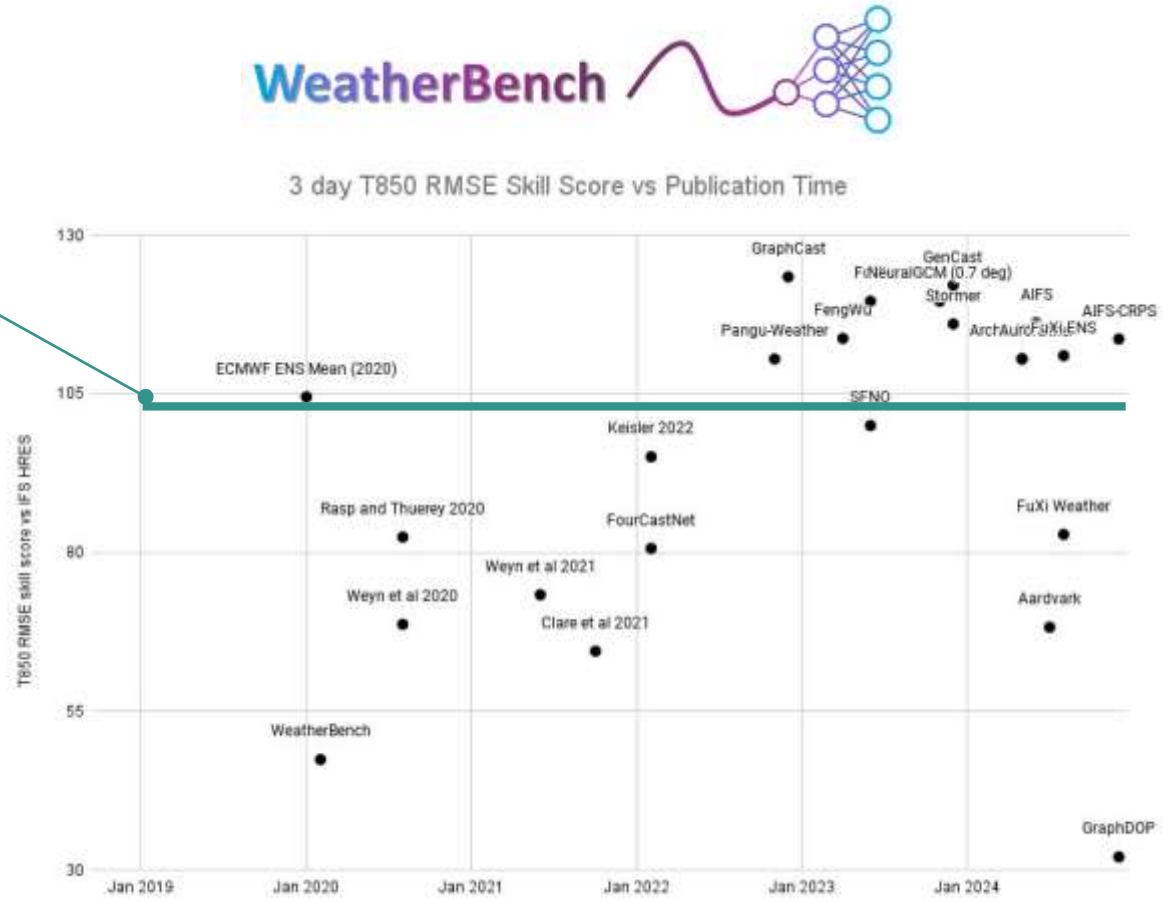
rhiza research

Benchmarking =

Ongoing and historical forecast verification of existing and new forecasting models

In the past 5 years, AI weather prediction (AIWP) models have **matched and then surpassed the skill** of leading numerical weather prediction (NWP) models.

Lots of important caveats here: coarse resolution, short time scales, focused on temperature evaluation, physics-based reanalysis as input …

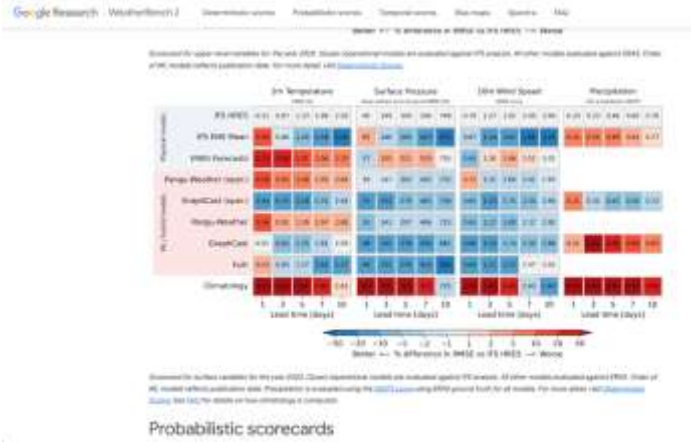


Graph from Stephan Rasp, Google Research

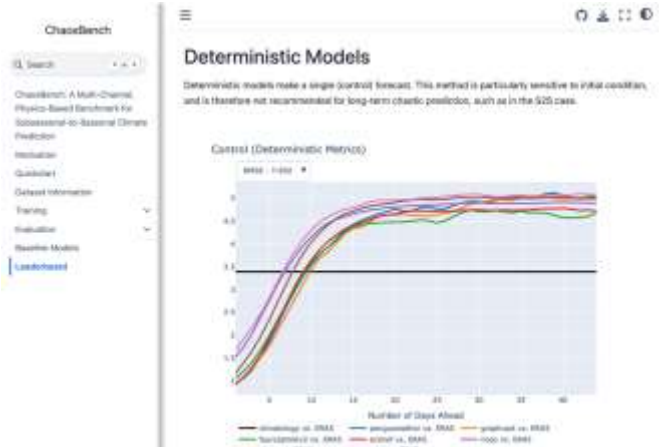
WMO's Dynamical Model Verification

IRI's Seasonal Forecast Verification

Google Research's WeatherBench 2 / X



Columbia and UCLA's ChaosBench



A good benchmark will let you do three important things:

1. Evaluate whether new forecasting models are skillful enough to be operationalized,
2. Choose high performing forecasting models to disseminate,
3. Improve existing models through, e.g., parameter selection.

1. Evaluate whether new forecasting models are skillful enough to be operationalized.

Scenario A: Company X has started to produce a high-resolution, 4km AI-based 15-day forecast. This model has been tested extensively in the US and Europe, but not in your country. Company X claims that the model can outperform existing models, and wants you to operationalize the forecast. You use a country-specific benchmark to compare forecast quality over your country to your current operational standards (ECMWF HRES, WRF, UKMO) and decide whether the new model is ready for operationalization .

2. Choose high performing models to use for high-impact events

Scenario B: Your agricultural sector is requesting a forecast of the start of the MAM rainy season. Your forecasting team looks at the outputs of several forecasting models—ECMWF HRES, UKMO, GEFS, locally run WRF—to generate this advisory. Additionally, your team has recently started to run the novel research AI-forecasting model Y.

Your forecaster knows that some models perform badly during MAM rains, producing non-physical or very extreme rains. She wants to adjust her forecast to use the highest performing models, but it's challenging to remember which models have done well, especially for the new model Y. She makes use of a benchmark to visualize model performance during March over the past 5 years and selects the most skillful models to make her start of the rainy season forecast.

3. Improve existing models through, e.g., parameter selection

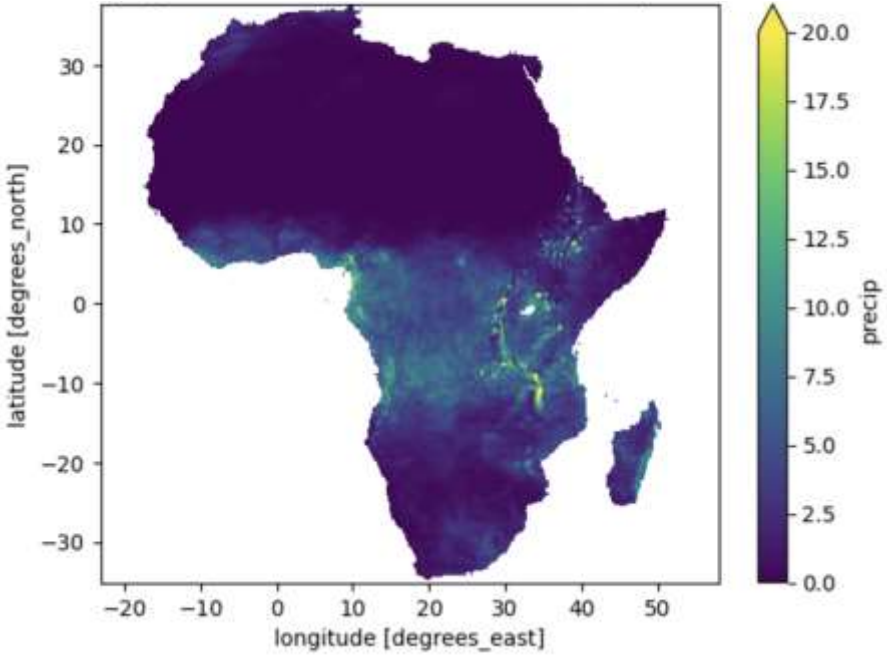
Scenario C: Your NWP team runs a WRF model to forecast 10-day rainfall. You've been running microphysics scheme 6 for years, but you're not sure if that's the best choice over your region. You've recently been collaborating with researchers to run an AI-WRF model that's 10x faster than your current operations, but you're not sure whether the outputs are as good as the old model, and there are several model parameters that can be turned to improve performance. You use a benchmarking tool to compare forecast outputs from several parameterization and select the best.

Elements of a human-centered benchmarking system

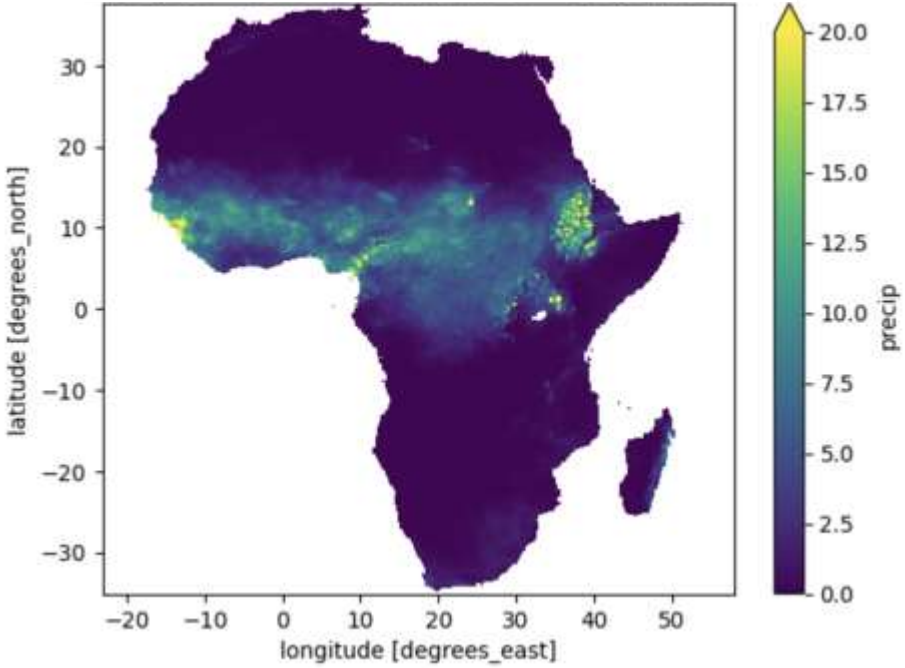
1. **[Events]** What events matter to users?
2. **[Metric]** What metric are you using to assess model performance?
3. **[Data]** What is the verification data?
4. **[Scope]** What locations and time period are you evaluating?

A good benchmark **should** include **baselines**
historical norms = climatology = **the** baseline to beat

1986-2015 Climatology for April 1st



1986-2015 Climatology for August 1st



A good benchmark should include **baselines**
ECMWF IFS: the numerical model to beat

The top-performing physics-based model globally.

Precipitation bias, 2016-2022

Forecasts too wet

Forecasts too dry

User needs should determine event definitions.

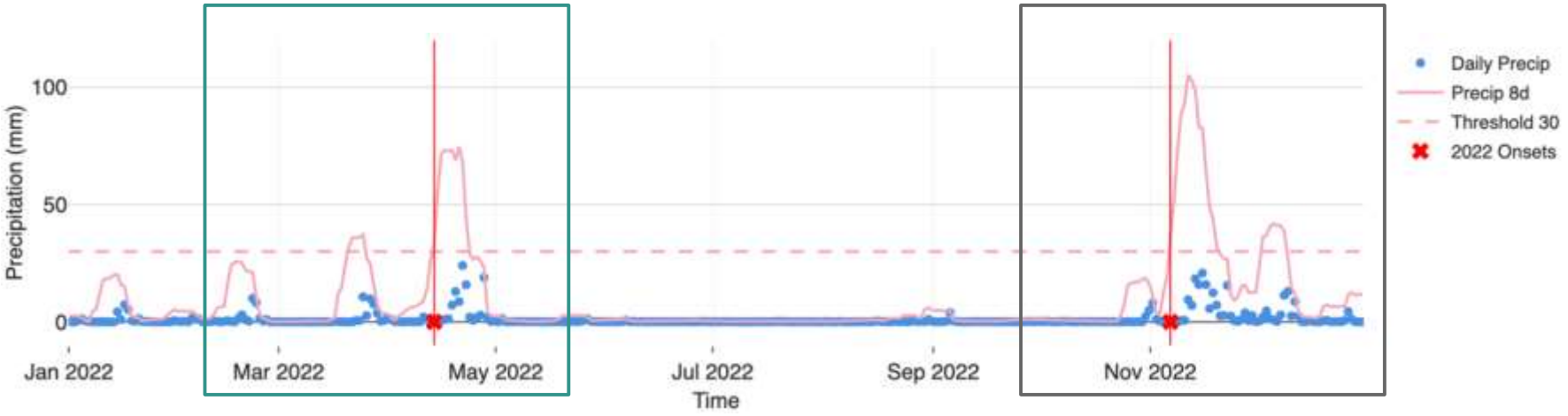
1. [**Event**] What is the event (variables, events, etc.)?

An event is specified as a set of bounds on the average or cumulative values of a meteorological variables.

Rainfall events above 75mm daily rain? Rainfall exceeding 30mm over 11 days? Dropping below 7mm over a week? Temperature above 38C? High winds?

These event should be developed as a collaboration between a meteorologist and a user or organization representing users.

An example: Rainfall error during the MAM rains in Kenya



MAM season

OND season

Your metric should be chosen to match your events.

2. [Metric] What metric are you using to assess model performance?

Metrics can evaluate two types of forecasts:

- **Deterministic**: How well did a single, “best-guess” forecast match what occurred?
- **Probabilistic**: For a forecast that provides several possible outcomes, did it give high probability to the actual occurrence? (sharpness & calibration)

Additionally, metrics can be **value-based** or **event-based:**

- **Value-based:** Measures the gap between the predicted and true event (e.g., 27°C vs 21°C has a 6°C gap)
- **Event-based:** Measures whether a specific event occurred (>10mm of rainfall)

I often find event-based metrics to be more informative in a human-center evaluation.

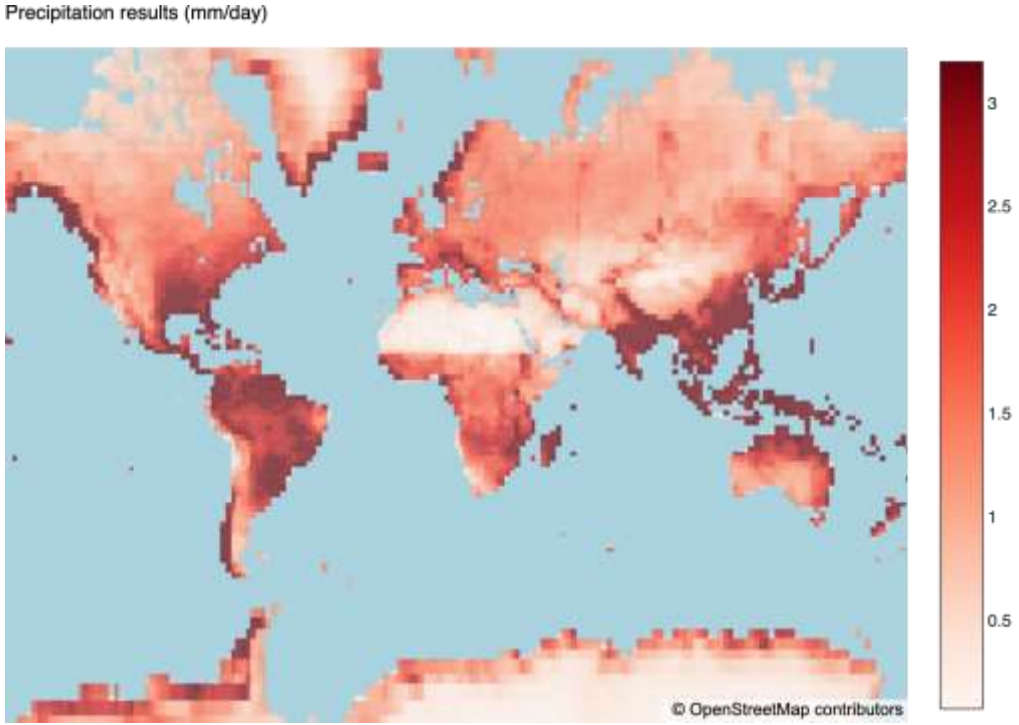
RMSE - Root mean squared error

Deterministic, Value-based

The most popular metric in machine learning applications.

Measures distance between observed and forecast, but heavily penalizes large errors.

Being 10mm off one time is worse than being 1mm off 10 times.



ECMWF IFS Extended Range, Week 3

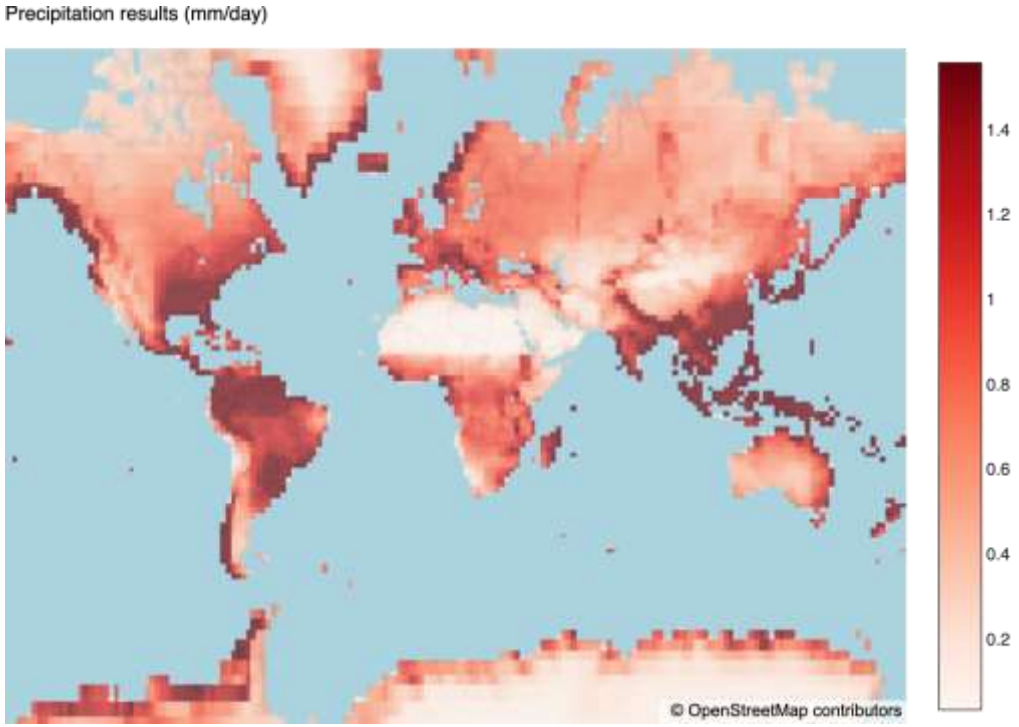
CRPS - Continuous ranked probability score

Probabilistic, Value-based

The most popular metric in machine learning applications for ensemble forecasts.

Measures distance between observed and forecast, weighted by the probability given to that event.

Considers both sharpness and reliability.



ECMWF IFS Extended Range, Week 3

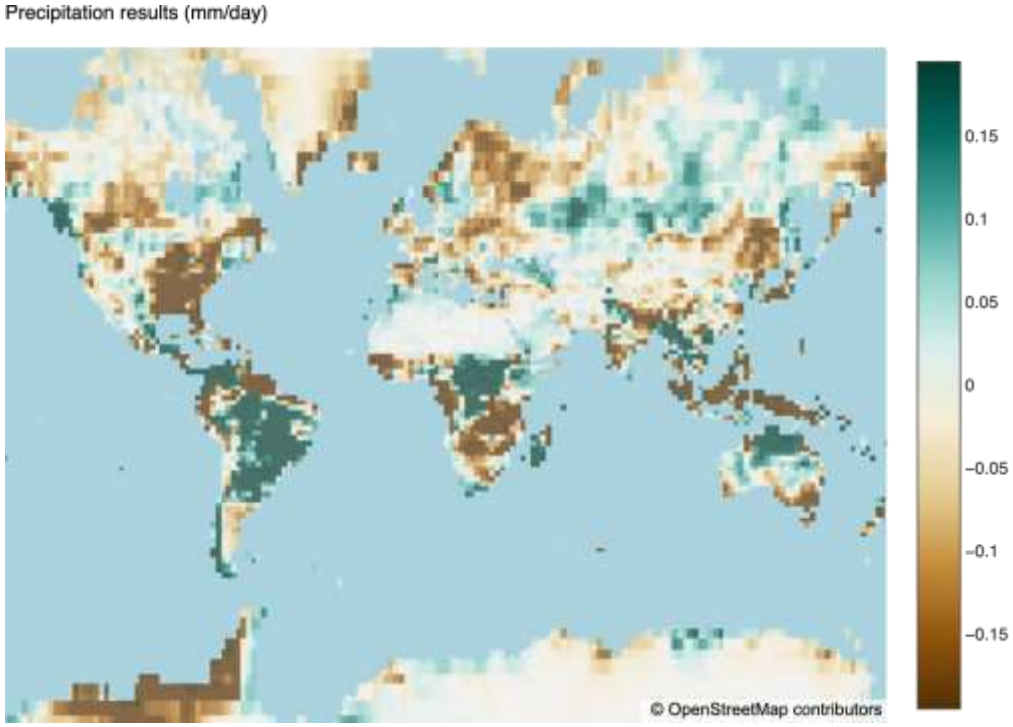
Bias

Deterministic, Value-based

Dynamical and AI-based forecasting models have **systematic** errors.

Provides insight into where a model can be improved.

Sometimes, these can be corrected in post-processing



ECMWF IFS Extended Range, Week 3

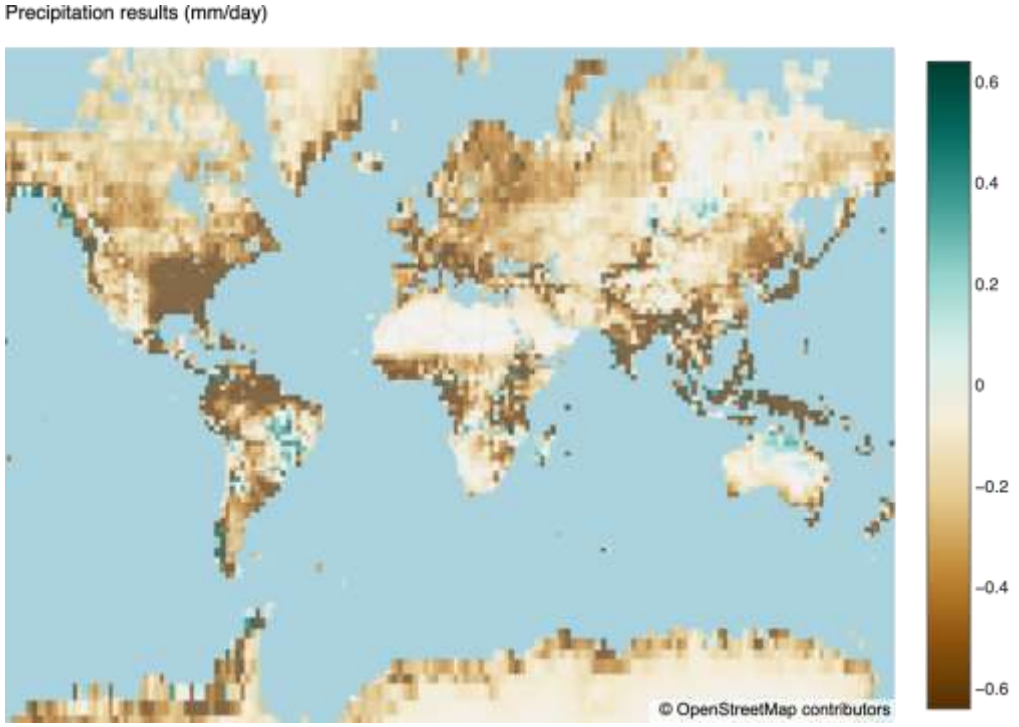
Bias

Deterministic, Value-based

Dynamical and AI-based forecasting models have **systematic** errors.

Provides insight into where a model can be improved.

Sometimes, these can be corrected in post-processing



FuXi S2S, Week 3

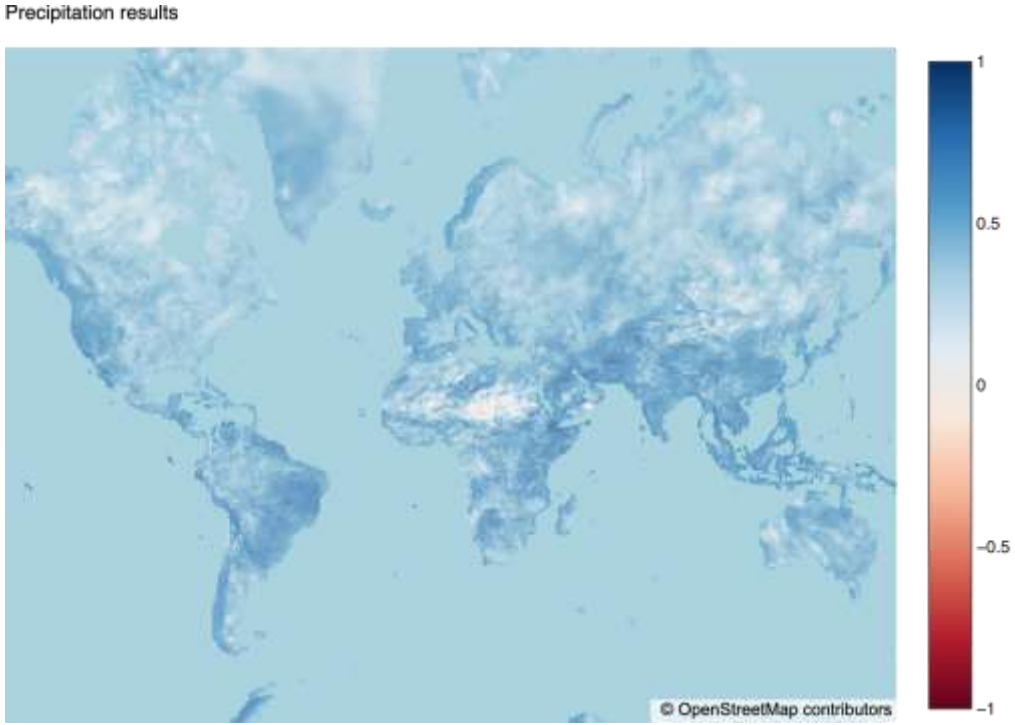
ACC - Anomaly Correlation Coefficient

Deterministic, Value-based

A nicely interpretable score that evaluates forecast anomalies, e.g., deviations from historical norms.

A score > 0.6 is considered actionable for short-term forecasts.

A negative score means the model performs worse than climatology (not skillful)



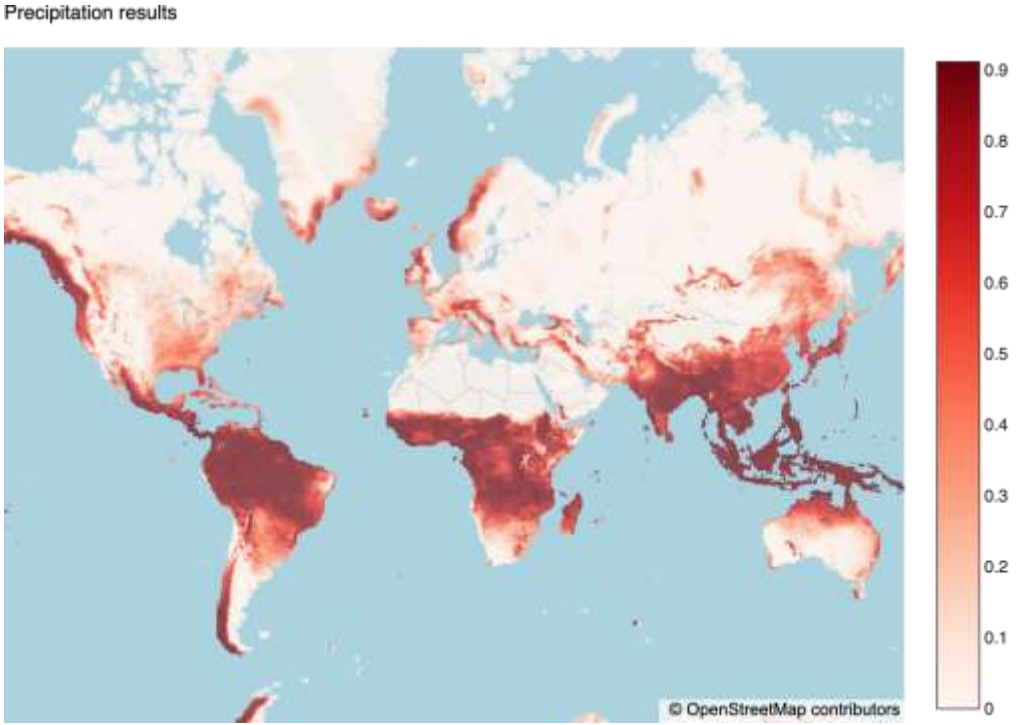
GenCast, Week 2

POD - Probability of Detection

Deterministic, Event-based

What was the probability that a forecast detected an event of rainfall above a specific threshold per day?

Should be used in combination with FAR - False Alarm Rate.



GenCast, Week 2, >5mm

Station-enhanced gridded data are often the best choice.

3. [Data] What is the verification data?

There are many datasets that can be used for verification:

- ERA5 for all meteorological variables
- CHIRPS, TAMSET, IMERG are public gridded rainfall products
- ENACTS is a private, gridded rainfall product
- Public GTS station data
- Privately-held station data (NMHS data, TAHMO data, other networks)
- SMAP is a public, satellite-based soil moisture estimate

Verifying with (**quality controlled**) station data is the industry standard, but also often limited in spatial scope. Gridded datasets that have been calibrated with local station data are often the best practical choice.

You can evaluate gridded data versus stations with the same metrics as forecasts!

East Africa precipitation results (mm/day)

Forecast	Daily	Weekly	Biweekly	Monthly
CBAM	11.55	7.35	6.91	6.26
CHIRPS	8.65	4.90	4.34	3.94
ERA5	7.84	4.97	4.53	4.21
IMERG	7.90	4.47	3.76	3.24

Mean absolute error of data sources vs station data in East Africa.

Verify as specifically as you can in space and time.

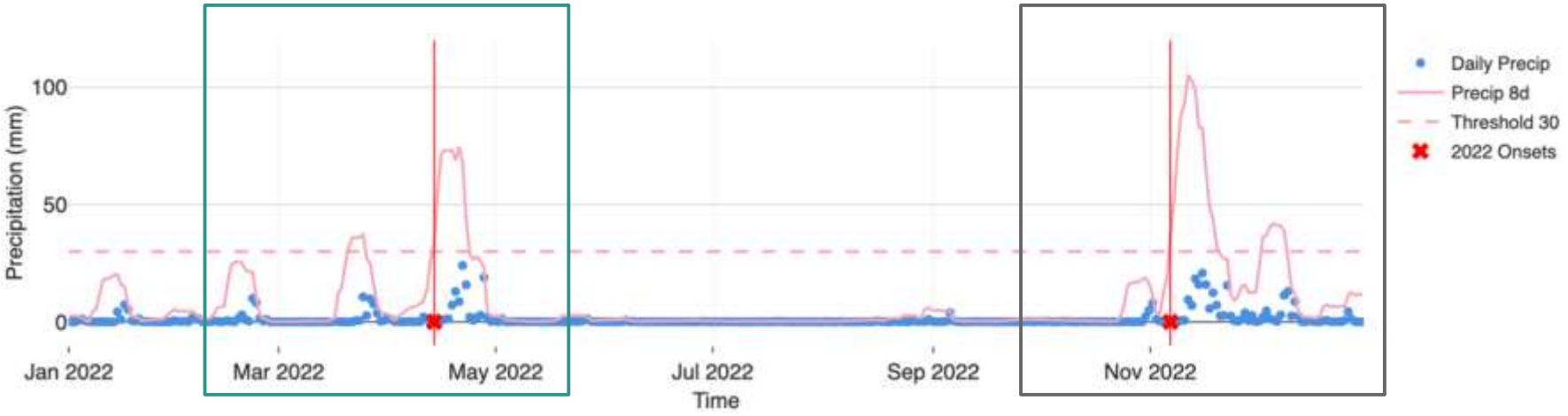
4. [Scope] What locations and time period are you evaluating?

Industry standard benchmarks and forecast verification are usually run over the whole globe, land and sea, for 5-10 years of hindcasts. These results can be misleading.

If you care about model performance in Eastern Kenya during the MAM rains, you should perform verification in that region in the months of March, April and May.

However, ensure that your evaluation is broad enough to test for model overfitting: generally 7+ years for subseasonal forecasts, 2+ years for short-term forecasts, regional vs per-grid cell evaluation.

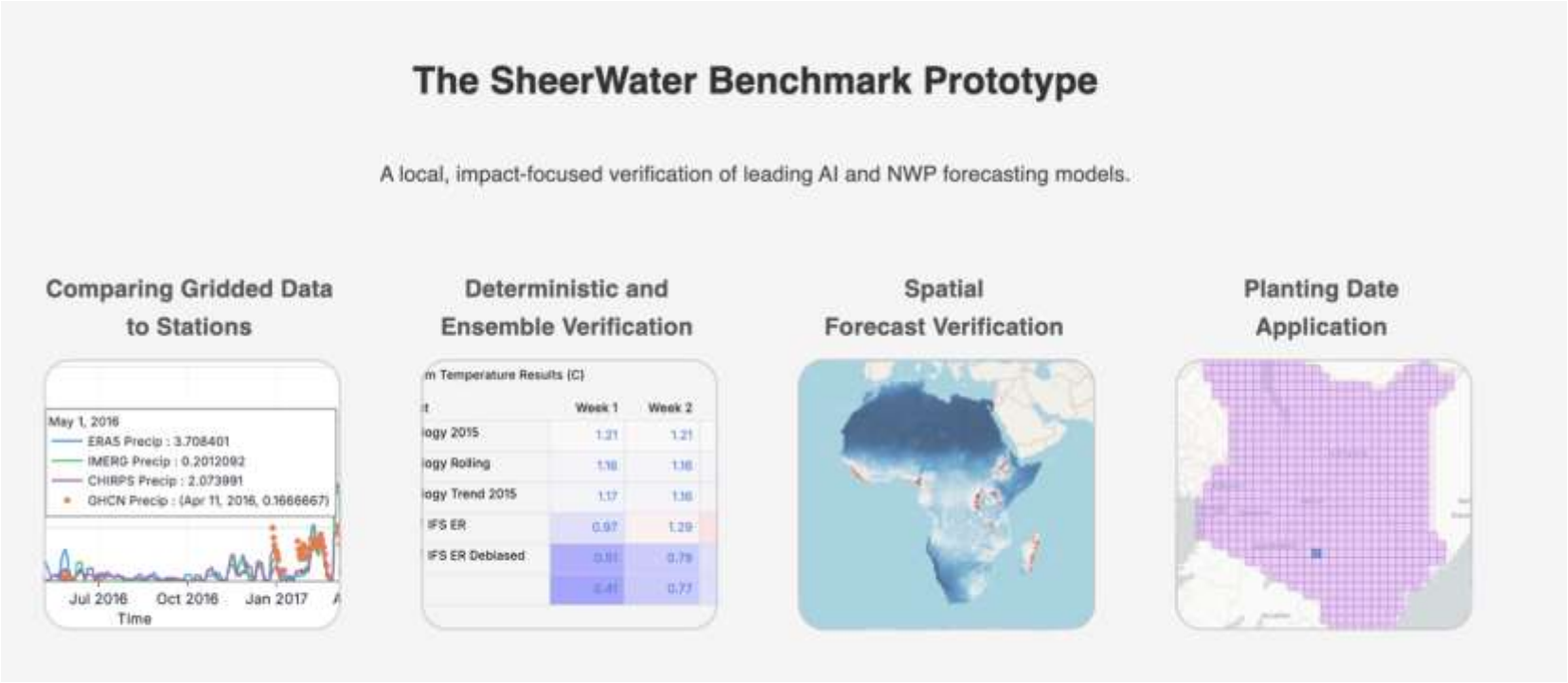
An example: Rainfall error during the MAM rains in Kenya



MAM season

OND season

The SheerWater Benchmark

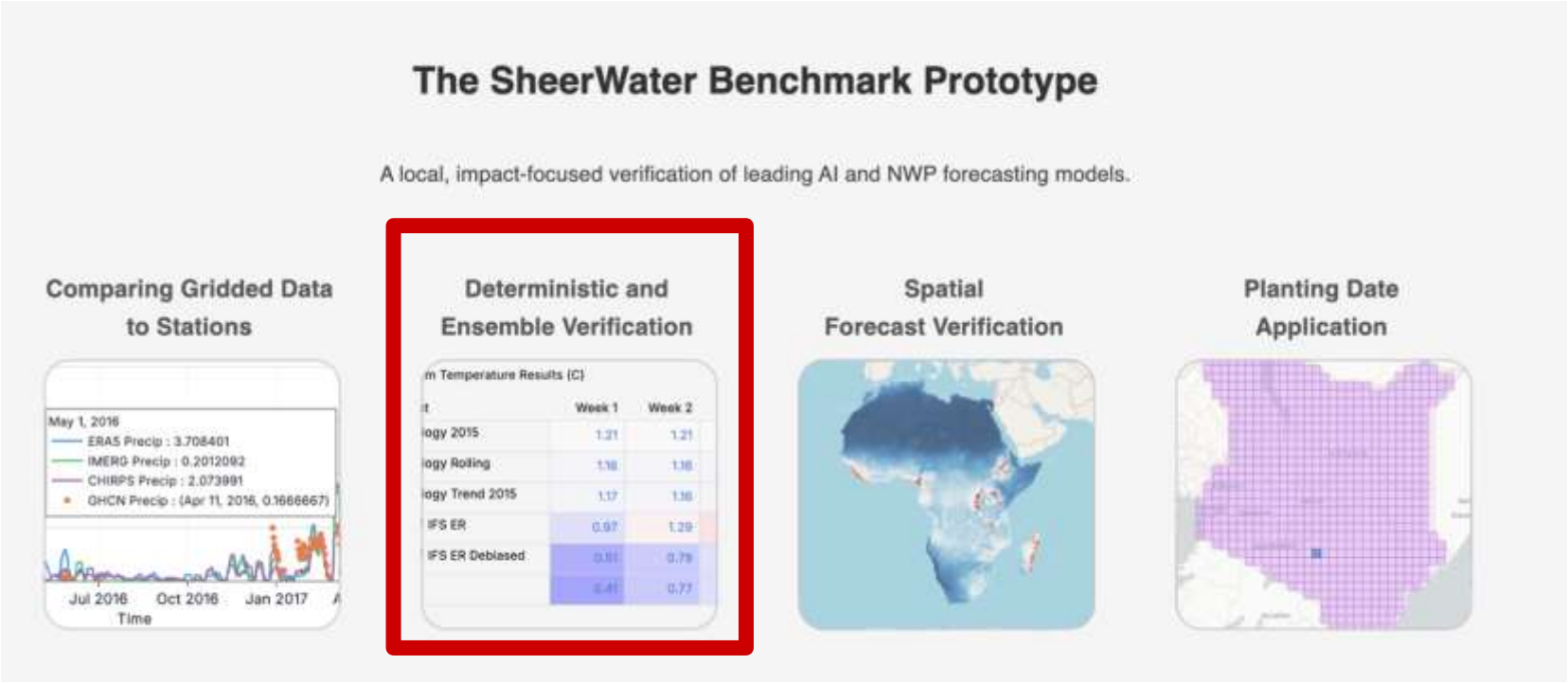


https://benchmarks.sheerwater.rhizaresearch.org

Username: aimforscale

Password: aimforscale2025

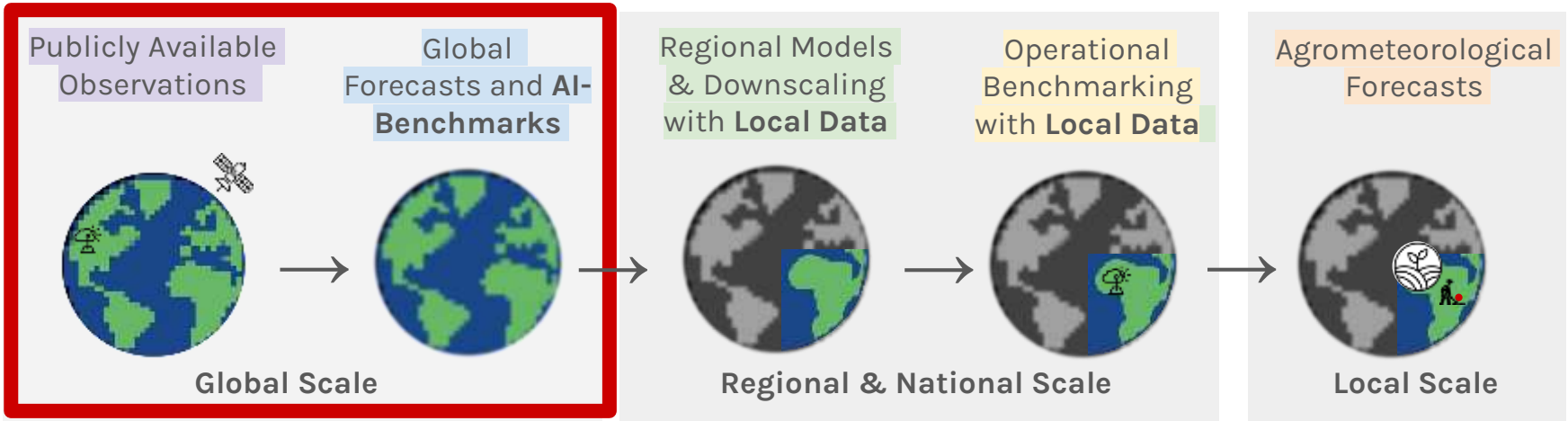
Demos


The SheerWater Benchmark Prototype
A local, impact-focused verification of leading AI and NWP forecasting models.
Comparing Gridded Data to Stations
May 1, 2016
ERA5 Precip : 3.708401
IMERG Precip : 0.2012092
CHIRPS Precip : 2.073991
GHCN Precip : (Apr 11, 2016, 0.1666667)
Jul 2016
Oct 2016
Jan 2017
Time
Deterministic and Ensemble Verification
m Temperature Results (C)
Week 1
Week 2
ogy 2015
1.21
1.21
ogy Rolling
1.16
1.16
ogy Trend 2015
1.17
1.16
IFS ER
0.97
1.29
IFS ER Debiased
0.91
0.79
0.41
0.77
Spatial Forecast Verification
Planting Date Application

Project Nimbus: AI-Enhanced Tropical Meteorology

Through Project Nimbus, Rhiza is supporting an end-to-end operational system for **agrometeorology monitoring**, enhanced by **AI-based forecasting** and downscaling, and supported by scientific **benchmarking** and verification practices.

Please reach out to genevieve@rhizaresearch.org with any questions.



Project Nimbus: AI-Enhanced Tropical Meteorology

Through Project Nimbus, Rhiza is supporting an end-to-end operational system for **agrometeorology monitoring**, enhanced by **AI-based forecasting** and downscaling, and supported by scientific **benchmarking** and verification practices.

Please reach out to genevieve@rhizaresearch.org with any questions.