Building Trust in AI models for Extreme Weather

Amy McGovern
Lloyd G. and Joyce Austin Presidential Professor, School of Meteorology and School of Computer Science, University of Oklahoma
Director, NSF AI2ES
Lead AI and Meteorology Strategist - Advisor, Brightband

Outline

- **Building community benchmark suites**
    - Extreme Weather Bench

- **Sharing common datasets of AI models**
    - WMO AI MIP archive
    - Brightband is considering an AI model archive

- **Is AI ever not useful? Can we overtrust it?**

- **Where is AI going in the future for extreme prediction?**

Extreme Weather Bench

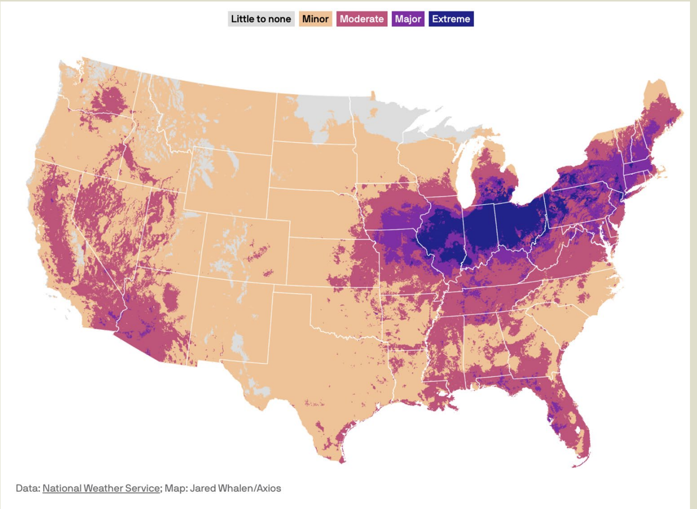
Community driven set of case studies, data, metrics, and code to evaluate your models on the cases

Amy McGovern
Taylor Mandelbaum
Daniel Rothenberg
Nicholas Loveday, BoM
Corey Potvin, NOAA NSSL
Montgomery Flora, The Weather Company
Linus Magnusson, ECMWF
John Allen, CMU

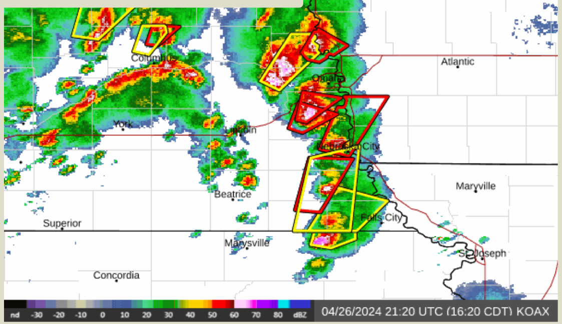
Motivation

- **Weather models should be *useful***
  - Motivated by WeatherBench but not affiliated
- **EWB provides:**
  - A way to compare AI and NWP models on a common set of high-impact events
  - Community-driven impact-based metrics
- **EWB pushes the science forward**

EWB: ExtremeWeatherBench



Data: National Weather Service; Map: Jared Whalen/Axios



04/26/2024 21:20 UTC (16:20 CDT) KOAX

Images from NPR, wikipedia, Axios

Extreme Weather Bench (EWB)

- **Standardized set of global high-impact weather events, data, metrics and code**
  - Evaluate across event categories
  - Dive deeply into a single event or groups or regions
- **EWB provides**
  - Information about the event
  - Data (validated observations if available)
  - Standard impact-based metrics
  - Open-source python codebase on GitHub
- **Community driven**
  - We want community input, feedback, new data, case studies, and metrics!

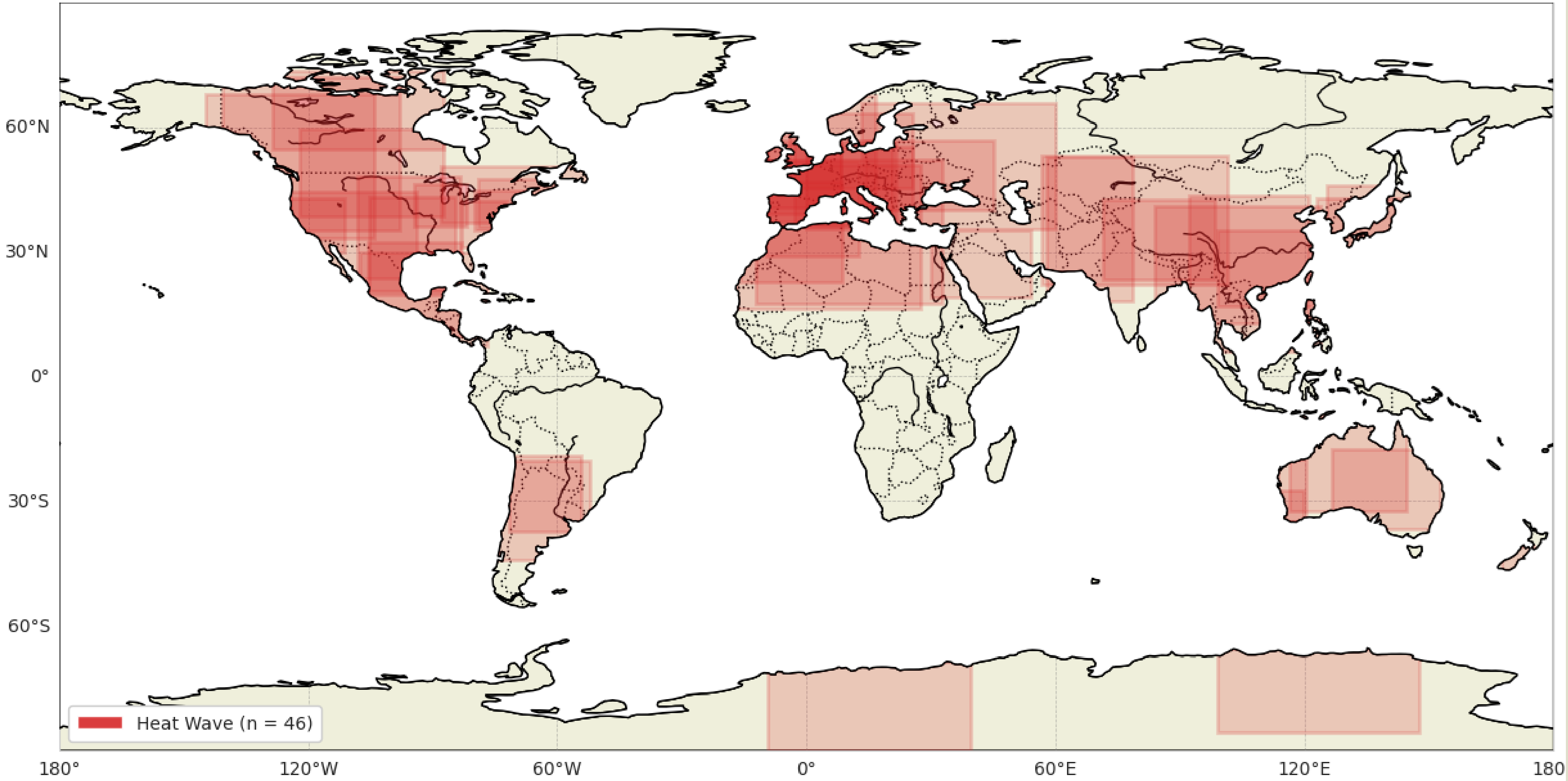
EWB: ExtremeWeatherBench

Choosing cases for EWB

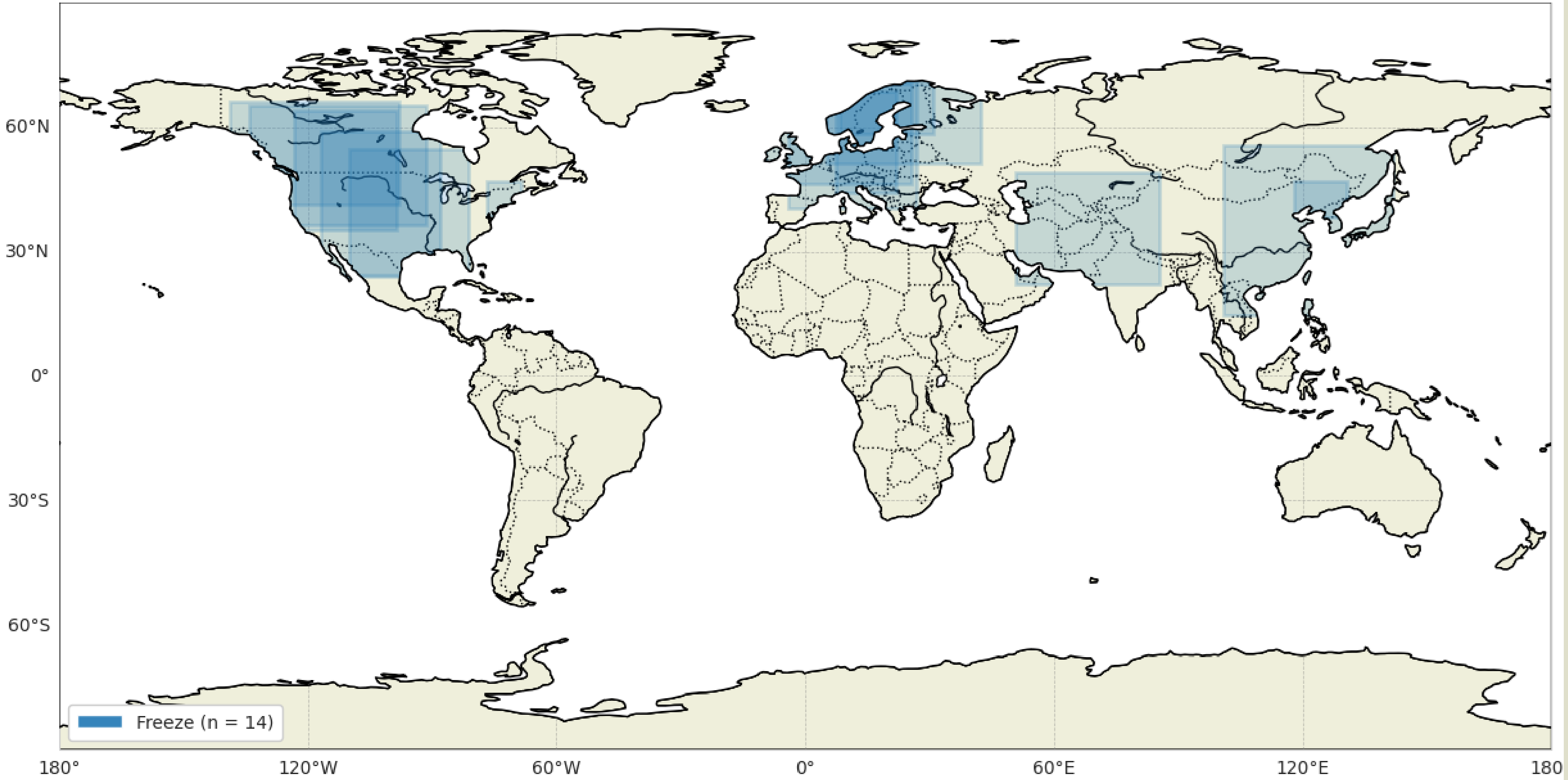
- **EWB provides a curated set of extreme events**
  - Events from 2020-2024
  - Chosen to represent geographical and impact diversity (US and global: e.g. TC basins, etc)
  - Chosen based on impact (e.g. TC must make landfall)
  - Goal of > 30 cases per category (not always possible within 5 years)
- **You can easily extend to create your own set of events**
  - E.g. add recent hurricanes, etc

EWB: ExtremeWeatherBench

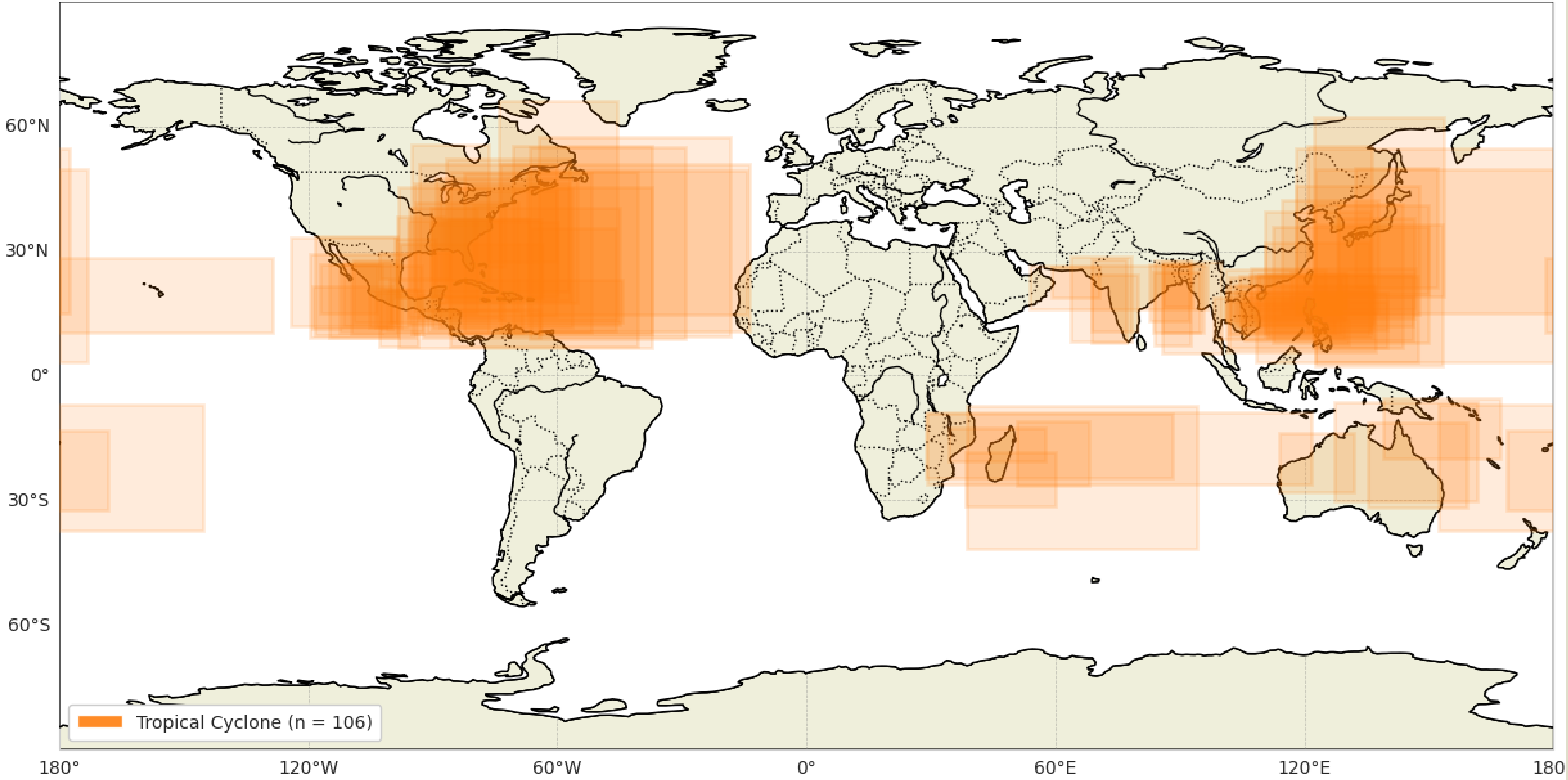
ExtremeWeatherBench Cases: Heat Wave (n = 46)



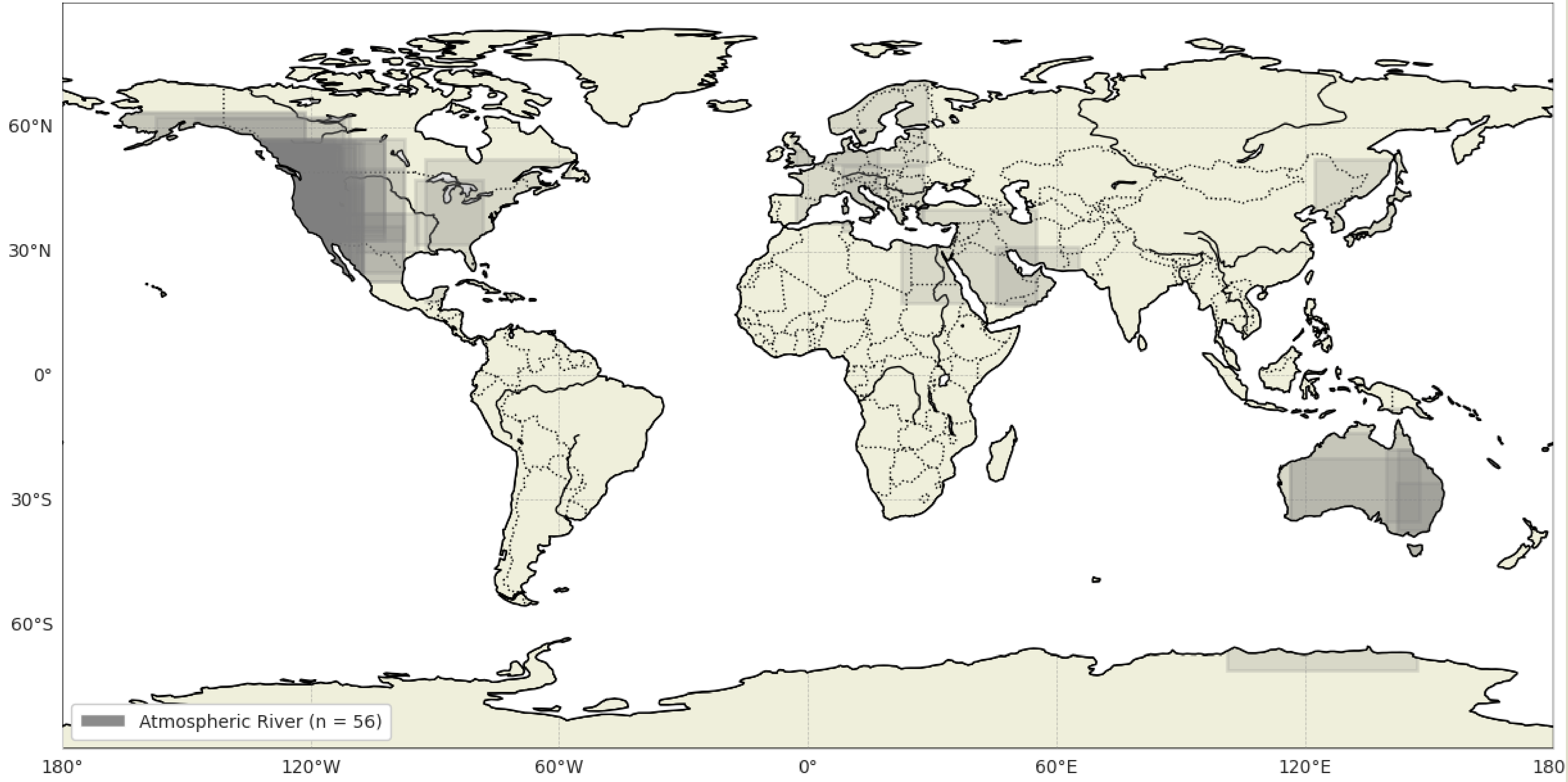
ExtremeWeatherBench Cases: Freeze (n = 14)



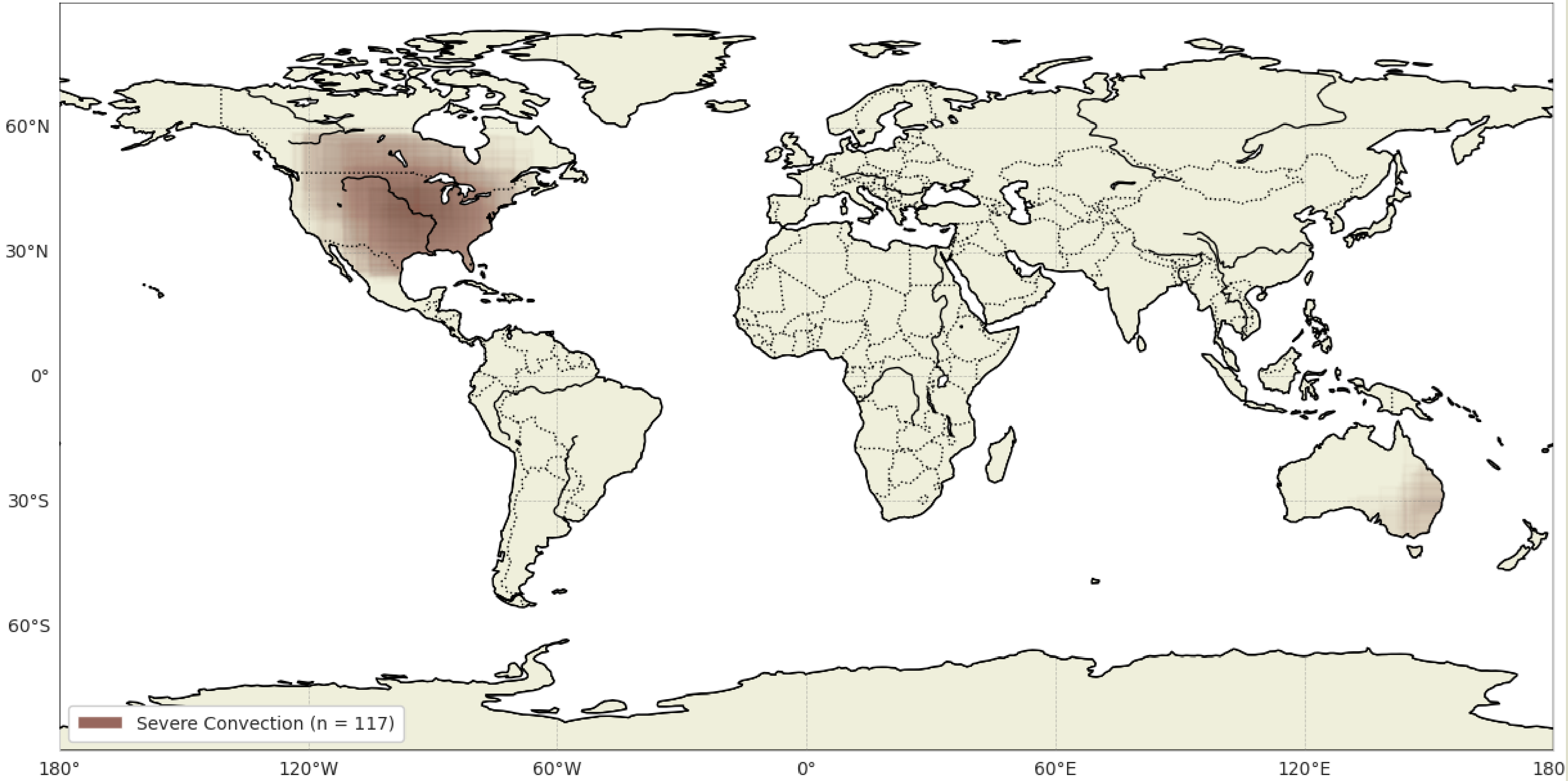
ExtremeWeatherBench Cases: Tropical Cyclone (n = 106)



ExtremeWeatherBench Cases: Atmospheric River (n = 56)



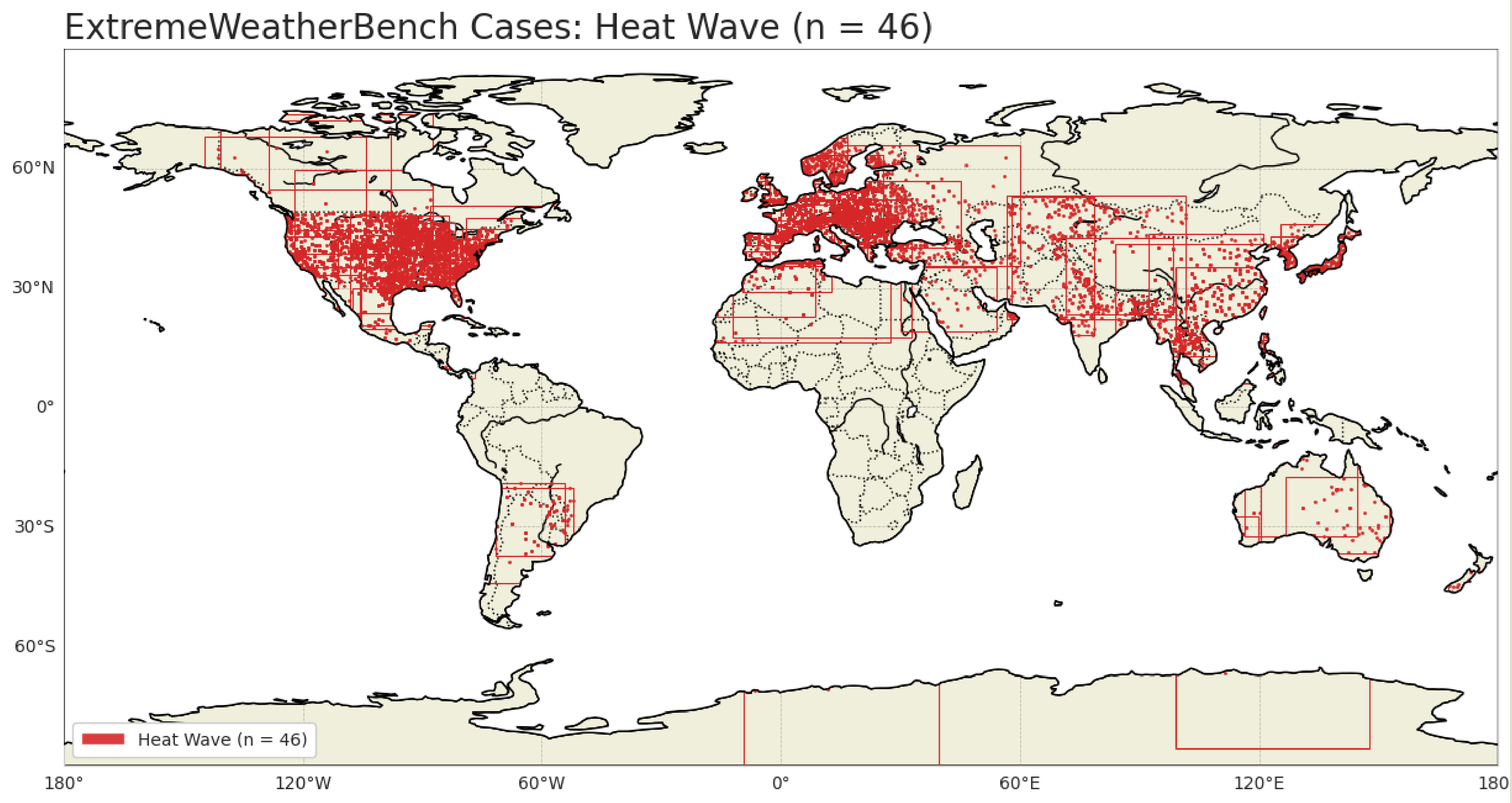
ExtremeWeatherBench Cases: Severe Convection (n = 117)



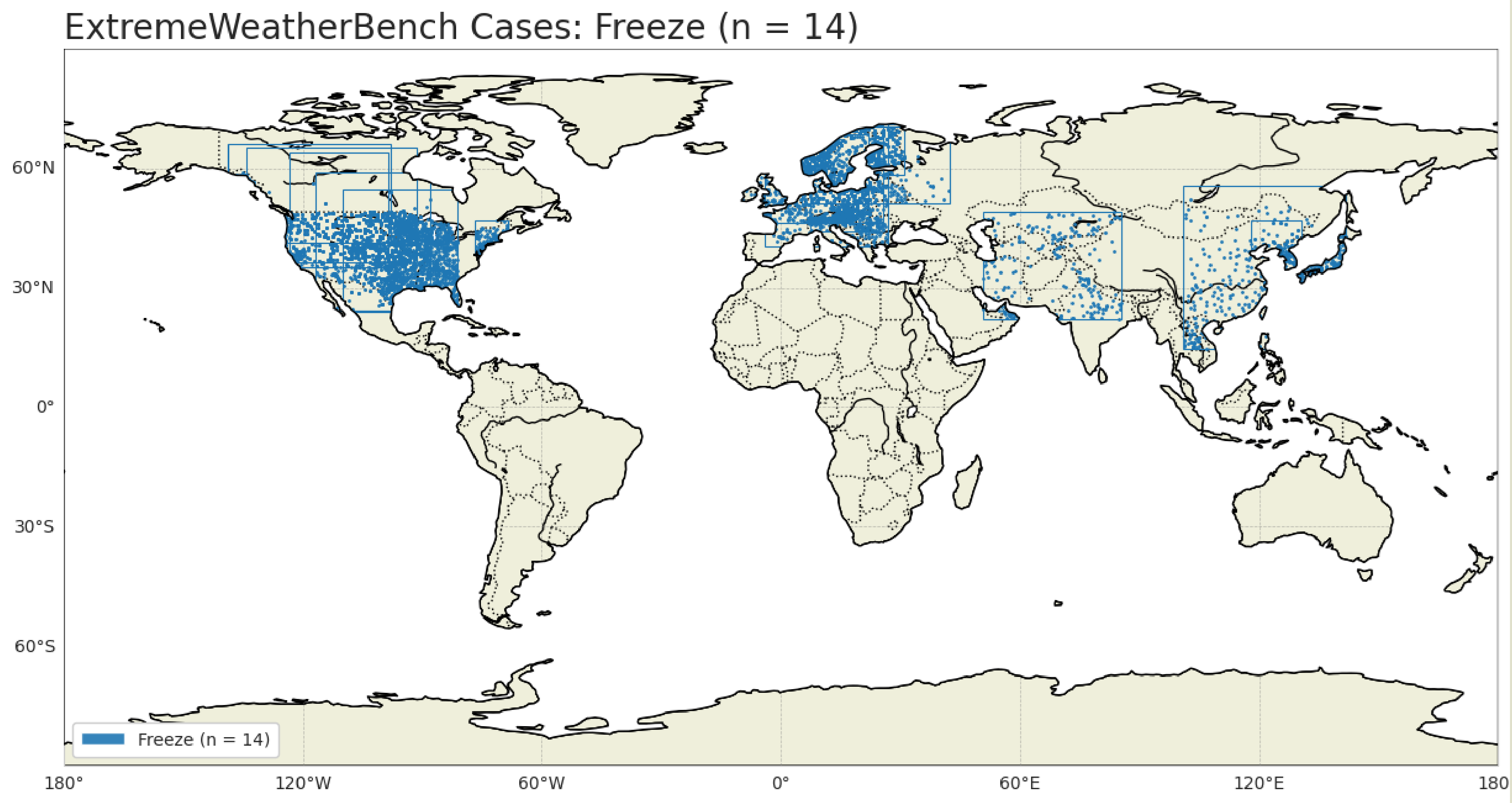
EWB Methodology: Evaluation

- **Evaluation is based on *targets***
  - A target is a verified observation, report, or reanalysis product
  - ERA-5 is the fallback when ground-observations are missing
  - Target name comes from WeatherBench
- **Different event types have different targets:**
  - Tropical Cyclones: IBTrACS
  - Severe Convection: Local Storm Reports and Practically Perfect Hindcasts
  - Hot/Cold: Global historical climatology network
  - ARs: ERA-5

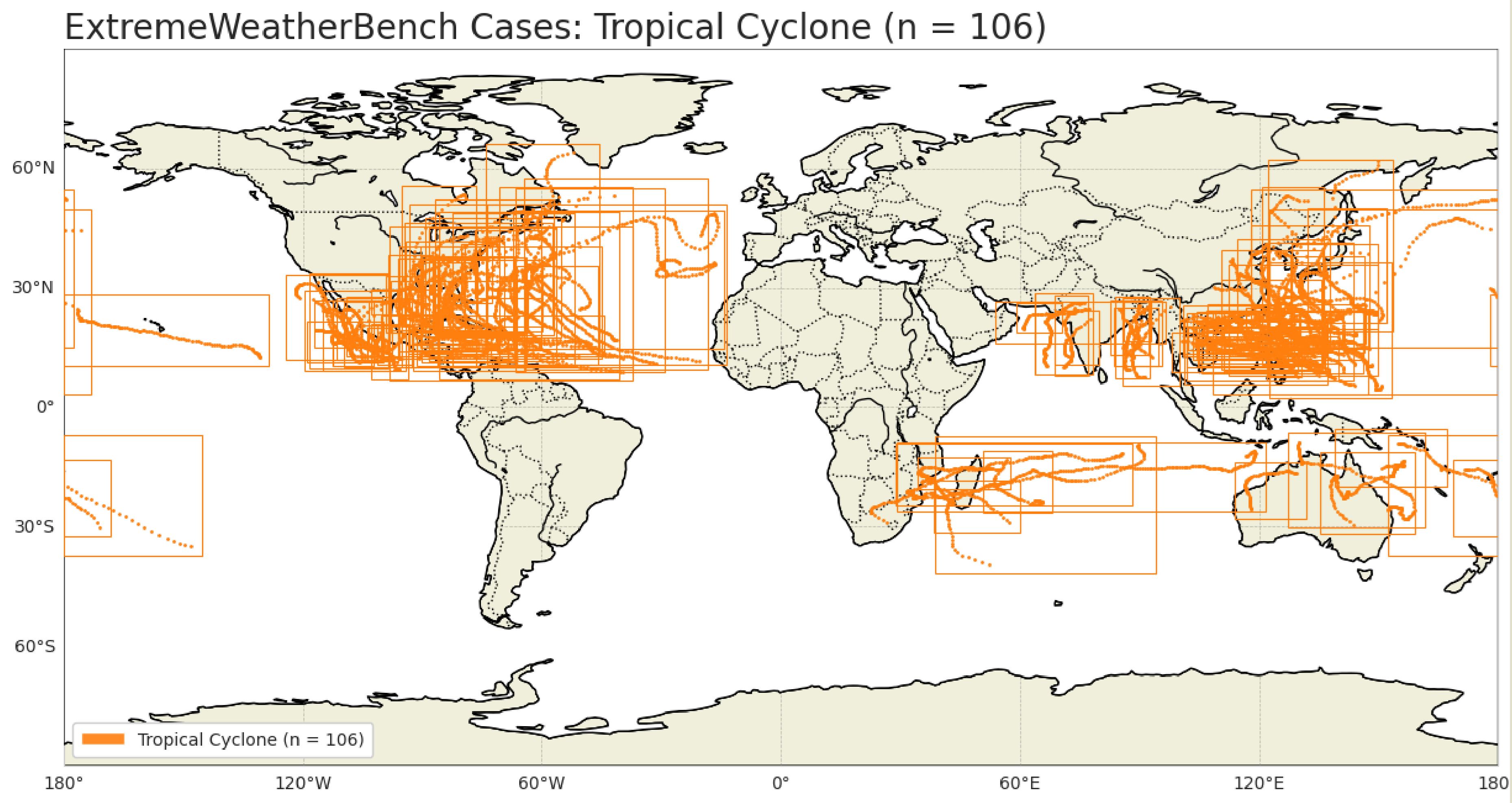
EWB: ExtremeWeatherBench

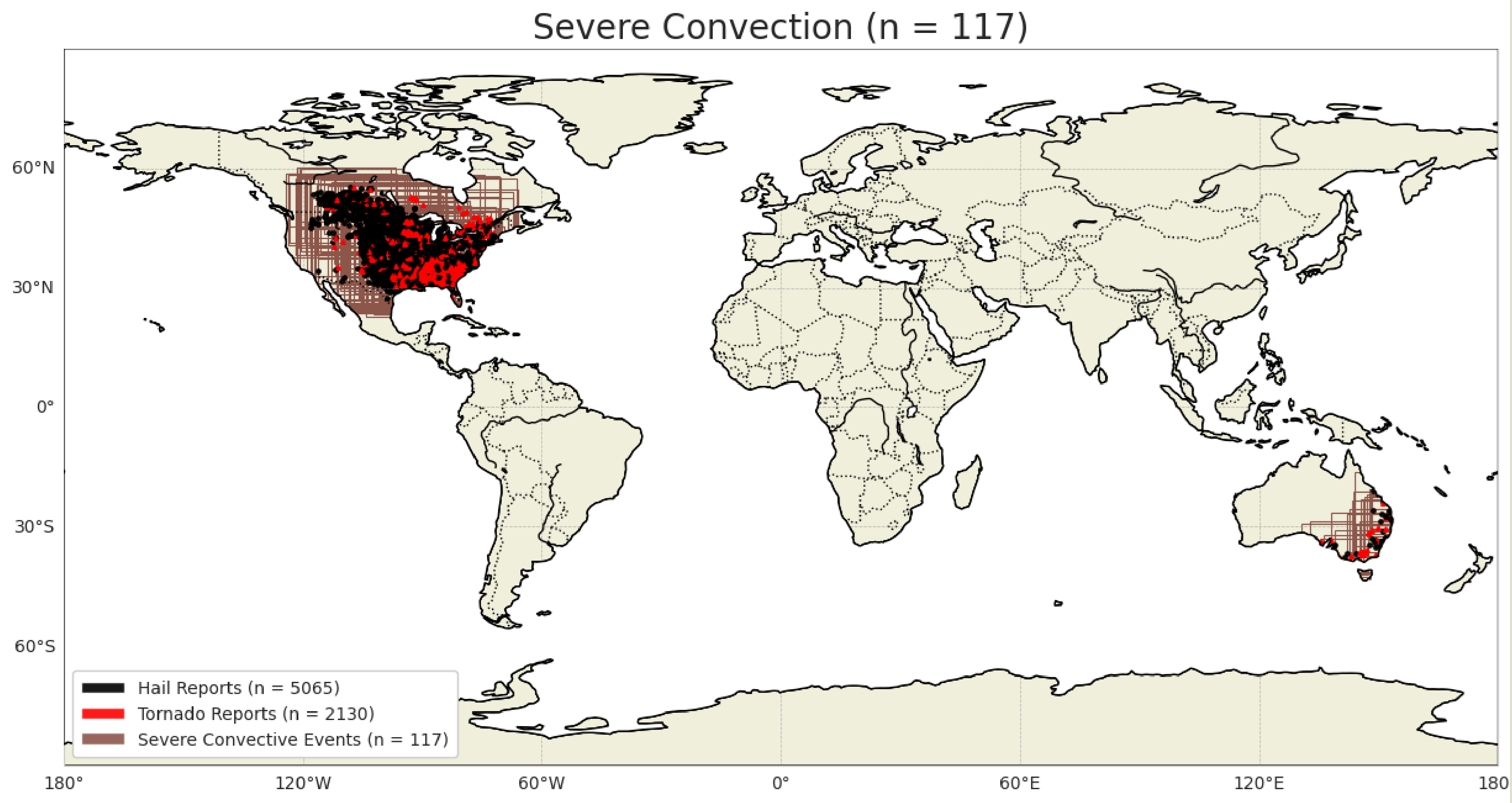


ExtremeWeatherBench Cases: Heat Wave (n = 46)



ExtremeWeatherBench Cases: Freeze (n = 14)

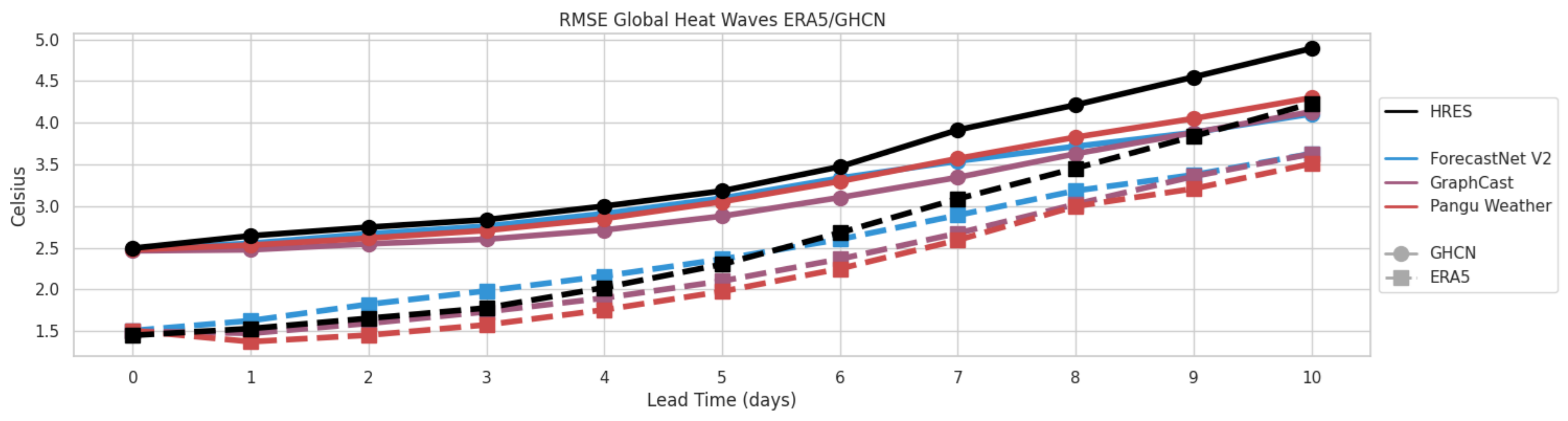

ExtremeWeatherBench Cases: Tropical Cyclone (n = 106)
60°N
30°N
0°
30°S
60°S
180°
120°W
60°W
0°
60°E
120°E
180
🟧 Tropical Cyclone (n = 106)



Severe Convection (n = 117)

How do targets change evaluation?

The models are not estimating the true intensity of the event if they only use ERA5



EWB: ExtremeWeatherBench

Brightband

EWB Methodology: Metrics

- Metrics chosen based on:
  - Use by forecasters as familiar metrics will be easier to verify
  - Used to answer specific questions
    - How well did the model predict the extreme?
    - How early did the model predict the extreme?
    - What is the spatial error?
- Used to capture how good the models are for different use cases
  - Captures the hazard and the user needs
- Used to measure the performance of predicting extremes and hazards
- Many of our metrics use the scores package (Loveday)

Case Study: Heat Waves

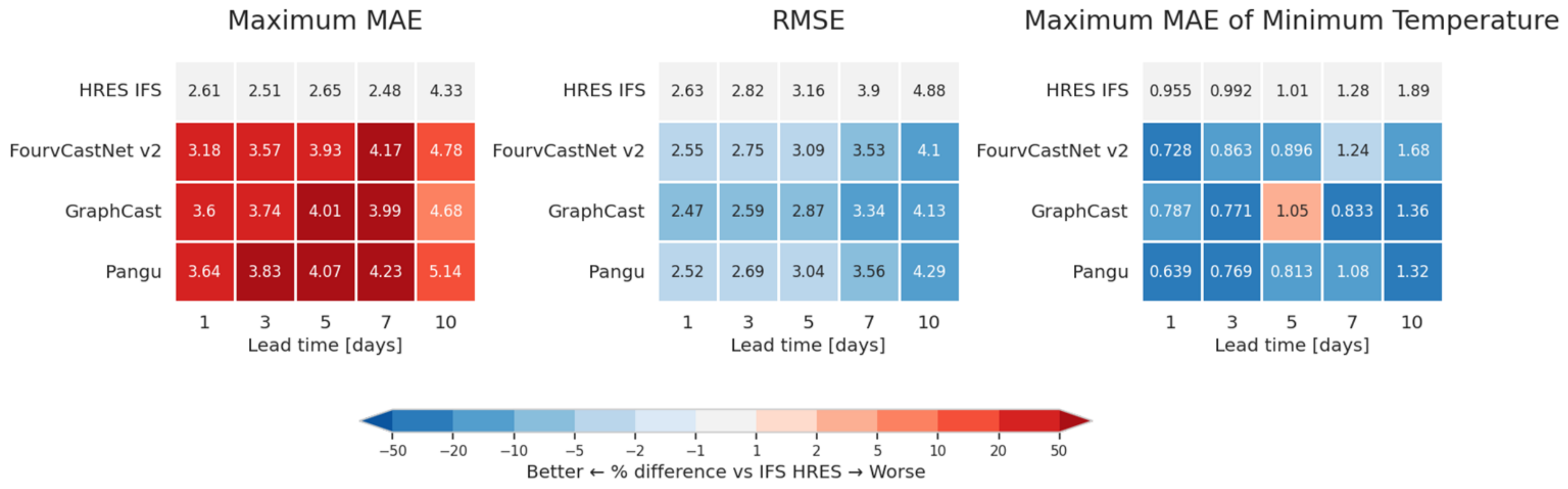
Brightband

EWB Metrics: Heat/Freeze

- **What is the (aggregate and daily) error on the maximum or minimum temperature?**
  - ***MAE and RMSE over the event region (aggregated and daily)***
- **What is the (aggregate and daily) error on the predicted highest low temperature?**
  - ***MAE and RMSE over the event region (aggregated and daily)***
- How far in advance can the model predict a major heat wave or cold spell?
  - *Lead time of heat/freeze event*
- What is the error on the duration of the event? And how does this change as the event gets closer?
  - *Predicted duration of the event shown by days in advance of the event as well as during the event (to know when it ends)*

EWB: ExtremeWeatherBench

Global Heat Evaluation



Maximum MAE

	1	3	5	7	10
HRES IFS	2.61	2.51	2.65	2.48	4.33
FourvCastNet v2	3.18	3.57	3.93	4.17	4.78
GraphCast	3.6	3.74	4.01	3.99	4.68
Pangu	3.64	3.83	4.07	4.23	5.14

Lead time [days]

RMSE

	1	3	5	7	10
HRES IFS	2.63	2.82	3.16	3.9	4.88
FourvCastNet v2	2.55	2.75	3.09	3.53	4.1
GraphCast	2.47	2.59	2.87	3.34	4.13
Pangu	2.52	2.69	3.04	3.56	4.29

Lead time [days]

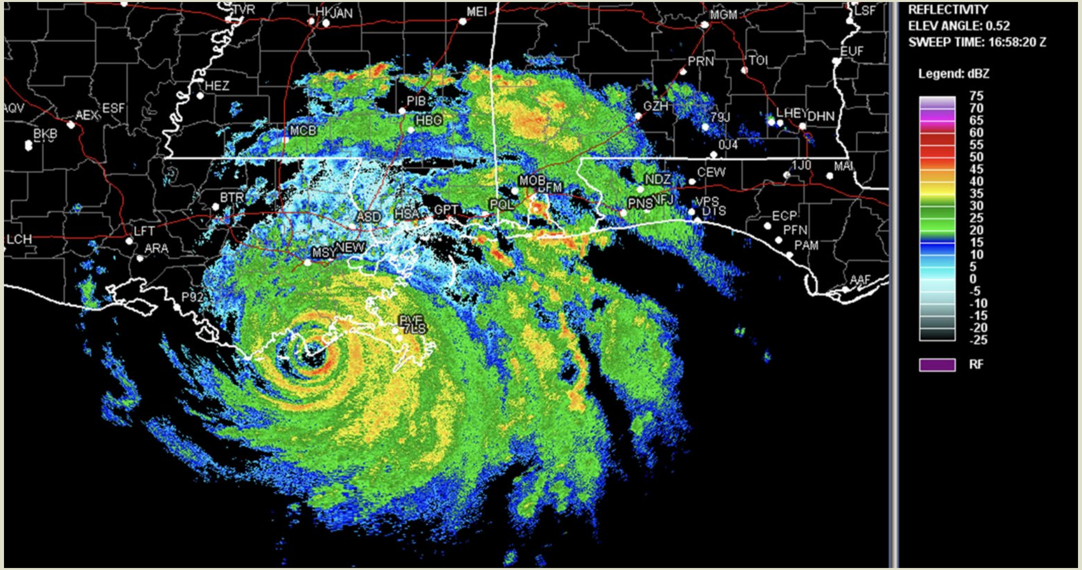
Maximum MAE of Minimum Temperature

	1	3	5	7	10
HRES IFS	0.955	0.992	1.01	1.28	1.89
FourvCastNet v2	0.728	0.863	0.896	1.24	1.68
GraphCast	0.787	0.771	1.05	0.833	1.36
Pangu	0.639	0.769	0.813	1.08	1.32

Lead time [days]

Case Study: Tropical Cyclones

EWB Example: TCs (Ida)

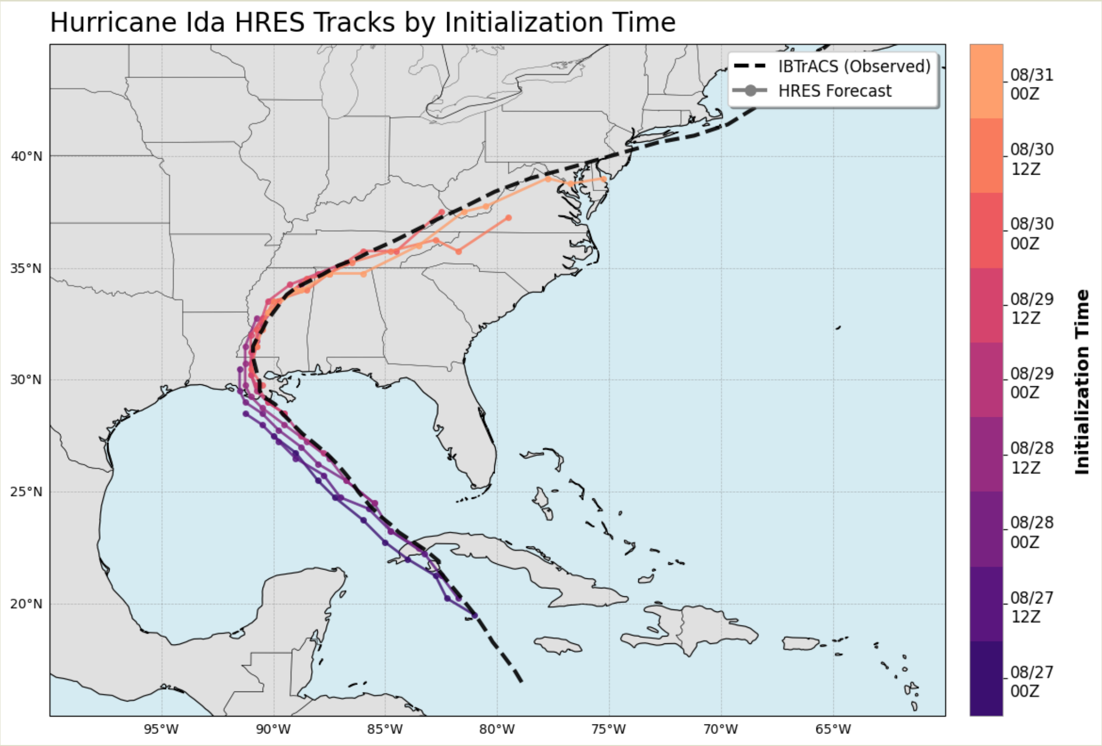


- Genesis in southern Caribbean 2021 Aug 26
- Made 3 landfalls, twice in western Cuba, once in southern US (Louisiana)
- Max wind speed: 130 kts, lowest pressure: ~929mb[1]
- 112 total direct/indirect deaths (US + Venezuela)[1]
- ~$77 billion USD in damage[1]
- Second only to Hurricane Sandy in Northeast US flood damage costs

[1] https://www.nhc.noaa.gov/data/tcr/AL092021_Ida.pdf

EWB Example: TCs (Ida)

*Note: linear interpolation of landfall for IBTrACS and HRES between pre and post-landfall coordinates
Landfall selected to be US (Louisiana) landfall



Landfall Displacement MAE (km)

Model Init Time	HRES	FCNv2	Pangu
2021-08-28 00:00:00	133	33.1	72.7
2021-08-28 12:00:00	58.6	33.1	33.1
2021-08-29 00:00:00	7.72	7.71	79.1

Landfall Time MBE (hours)

Model Init Time	HRES	FCNv2	Pangu
2021-08-28 00:00:00	12.9	4.96	6.10
2021-08-28 12:00:00	5.09	1.96	1.82.1
2021-08-29 00:00:00	1.82	1.82	5.22

EWB: ExtremeWeatherBench

ML models from CIRA archive

Brightband

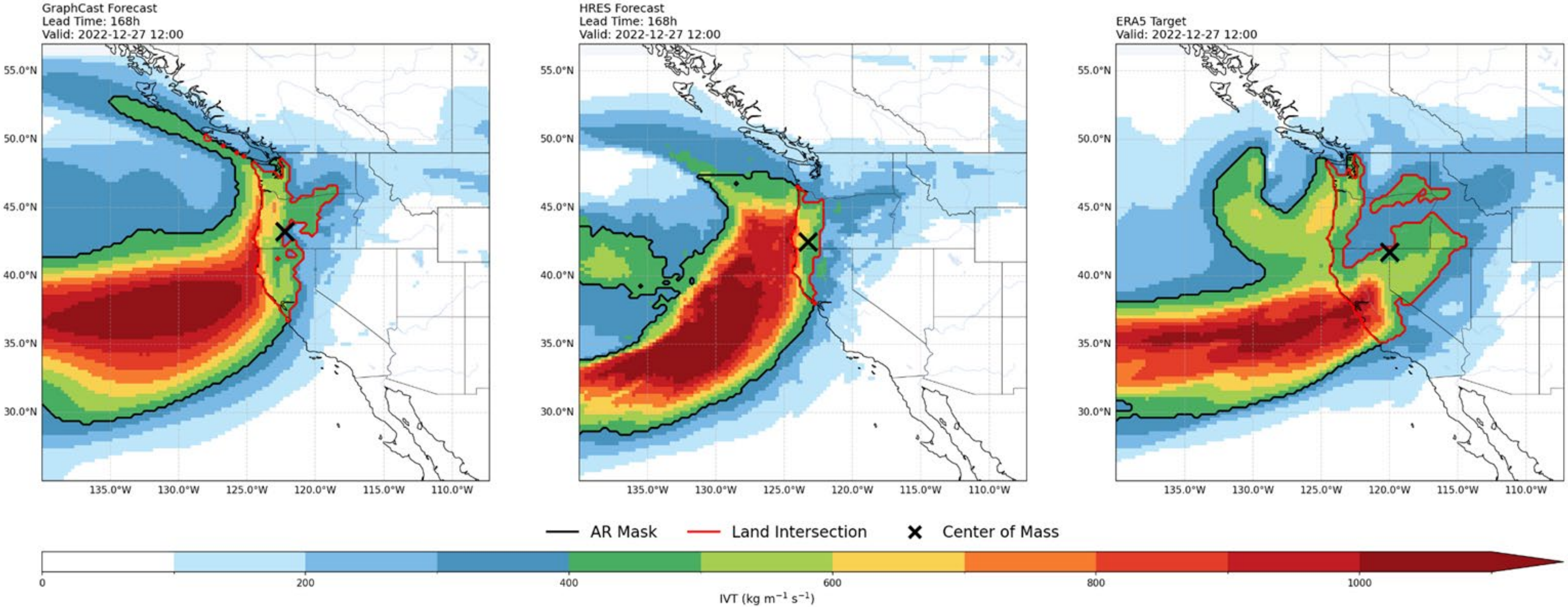
Case Study: Atmospheric Rivers

EWB Metrics: ARs

- How far in advance can a major AR be predicted?
  - Lead time of AR when the AR first intersects land
- What is the spatial error of the predicted AR on land?
  - Spatial displacement of the center of mass
- How well is the predicted area of IVT matched to the area where the AR actually landed?
  - IOU on the predicted versus actual AR
- What is the error on the total precipitation predicted within the area where we know the AR made landfall? (Valid only if model predicts precipitation)
  - Regional MAE on rainfall on points where the AR intersects land
- What is the error for 24 hour predicted totals within the area where we know the AR made landfall? (Valid only if model predicts precipitation)
  - Regional MAE on rainfall on points where the AR intersects land

EWB: ExtremeWeatherBench

AR Case Study: December 2022



EWB: ExtremeWeatherBench

AR Case Study: December 2022

Valid **2022-12-27 12Z**

Intersection over Union (IoU)

Lead Time	HRES	Graphcast	Pangu
0	0.40	0.40	0.40
24	0.37	0.49	0.50
96	0.28	0.35	0.44
168	0.15	0.33	0.01

Spatial Displacement (km)

Lead Time	HRES	Graphcast	Pangu
0	218	218	218
24	225	168	168
96	257	257	251
168	281	249	372

Lead Time Detection (hr)

HRES	Graphcast	Pangu
216	180	180

EWB: ExtremeWeatherBench

Extreme Weather Bench (EWB)

Call to action: Encourage national and regional organizations to use EWB and to give us additional data to enable global analysis of extreme events (Pillar 2 of Early Warnings for All)

```
pip install
git+https://github.com/brightbandtech/ExtremeWeatherBench.git
```

Community driven set of case studies, data, metrics, and code to evaluate your models on the cases

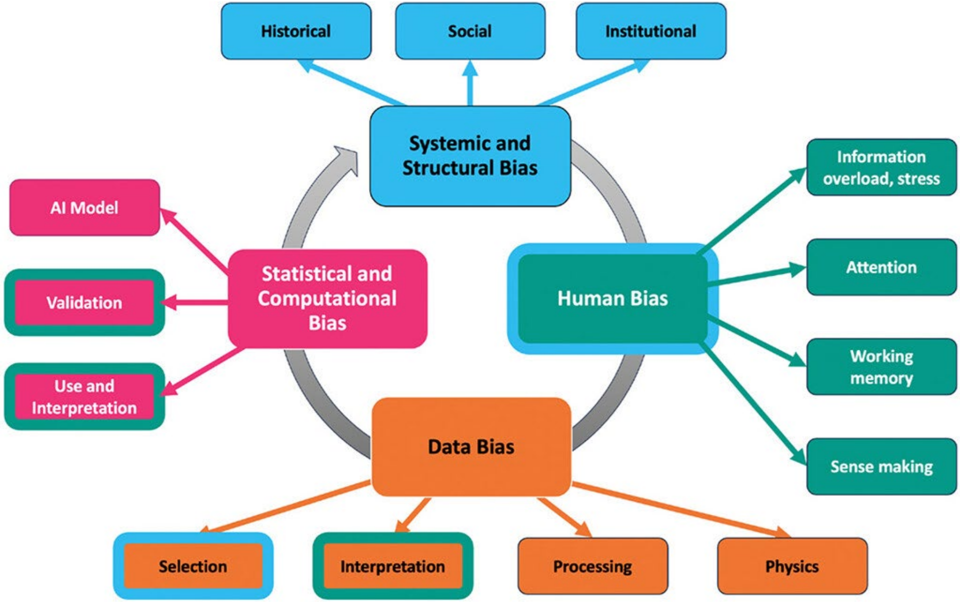
EWB: ExtremeWeatherBench

AI Model archive

- **An AI model archive enables people to dive deeply into AI performance across extremes**
  - WMO has an AI MIP providing an operational archive of forecasts for one year
  - https://www.wcrp-esmo.org/activities/wp-mip
- Brightband is considering publishing a similar archive, extending it in several key ways:
  - Publish data in analysis-ready, cloud-optimized Zarr format
  - Upgrade from 2x runs per day to 4x and extend forecasts to 15 days
  - Incorporate contemporary AI models (FourCastNet-v3, GenCast[/FGN], AIFS-Ens - more TBD!)

When is AI not useful?

- “All models are wrong but some are useful” – Box
    - AI is not magic. It will not “solve” weather prediction
- We need to learn when an AI model is good and what the limitations are
- We need to guard against over trusting a model (any model!)
- AI models must be developed ethically and responsibly



McGovern, A., et al 2024: Identifying and Categorizing Bias in AI/ML for Earth Sciences. Bull. Amer. Meteor. Soc., 105, E567–E583, https://doi.org/10.1175/BAMS-D-23-0196.1.

Future of AI for Extreme Events

- Exploring how we can do hyper-personalized forecasts

- Can we use AI to predict downstream impacts of extreme weather?

- How can we use AI to make humanity more resilient to the growing extremes?

Images from online news sources