Understanding and Evaluating Trust in AI Forecasts

Amy McGovern

Lloyd G. and Joyce Austin Presidential Professor | School of Meteorology and School of Computer Science

Director, NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES)

Director, IDEA Lab | University of Oklahoma

Lead AI and Meteorology Strategist – Advisor, Brightband

amcgovern@ou.edu

Overview

- **AI2ES: Understanding and evaluating trust in AI forecasts**
  - **What is trust and trustworthiness?**
  - Why do forecasters to trust AI guidance?
    - Storm scale
    - Global scale
  - Trustworthy AI development lifecycle
- Brightband: Trustworthy forecast verification for global AI models
  - Extreme Weather Bench

NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES)

AI2ES is developing *novel*, *physically based* AI techniques that are demonstrated to be *trustworthy*, and will directly improve *prediction, understanding, and communication* of high-impact weather and climate hazards, directly improving climate resiliency.

This material is based upon work supported by the U.S. National Science Foundation under Grant No. RISE-2019758

(Re) Conceptualizing trustworthy AI: A foundation for change

Publication

Wirz, C. D., Demuth, J. L., Bostrom, A., Cains, M. G., Ebert-Uphoff, I., Gagne II, D. J., Schumacher, A., McGovern, A., & Madlambayan, D. (2025). (Re) Conceptualizing trustworthy AI: A foundation for change. Artificial Intelligence, 104309. https://doi.org/10.1016/j.artint.2025.104309

“Trustworthy AI” has gained a lot of traction as a frame

Policy

- U.S. executive (Executive Order 14110, 2023; OSTP, 2022) and legislative (GAO, 2021, 2023) branches
- Organisation for Economic Co-operation and Development (OECD, 2019)
- The High-level Expert Group on Artificial Intelligence (HLEG, 2019), and the European Parliament (European Parliament, 2024)

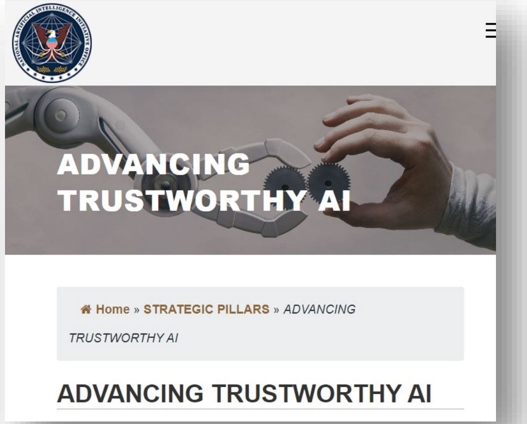
– OECD.AI Policy Observatory

Policies, data and analysis for trustworthy artificial intelligence

Executive Order 13960 of December 3, 2020

Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government

Funders



Researchers

- Sousa et al. (2024)
- Bostrom et al. (2023)
- McGovern et al. (2022)
- Jacovi et al. (2021)
- Toreini et al. (2020)
- Ashoori & Weisz (2019)
- Etc.

Private sector

Deloitte.

Services

Trustworthy AI™

Bridging the ethics gap surrounding AI

As more companies adopt AI, leaders grapple with etł
design and use. Global AI regulations will eventually a
until then the Deloitte AI Institute is working to bridge

“Trustworthy AI” has been used inconsistently and in some cases problematically

(General) status quo for how “trustworthiness” is conceptualized

- Largely **atheoretical** ways that **vary widely** source to source
- Encompasses diverse, **ranging set of subdimensions**
- Emphasis on **model performance** but **not precise** discussions of it
- Tends to refer to “**appropriate use**”

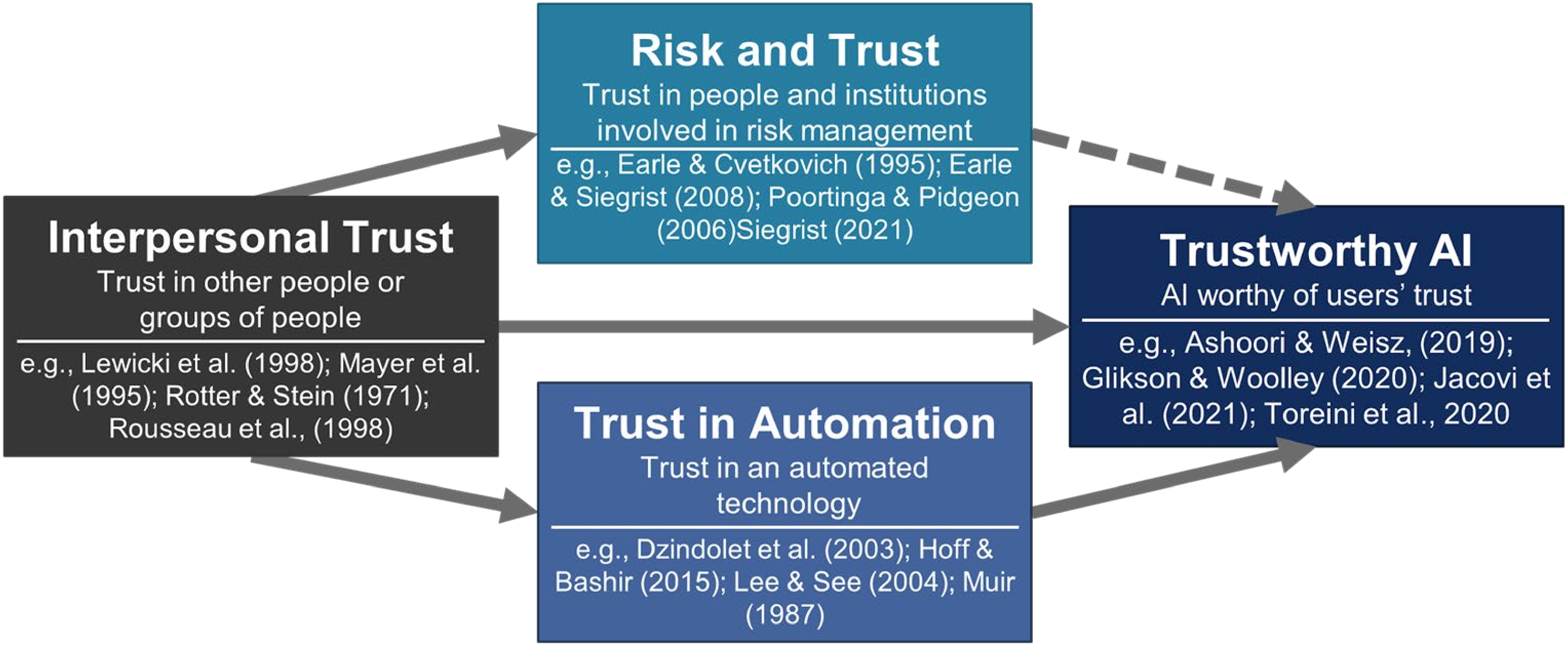
One set of trust and trustworthiness definitions (of many)

Trust: In the presence of **uncertainty**, the **degree** to which someone does or does not **rely on, or put faith in**, someone or something.

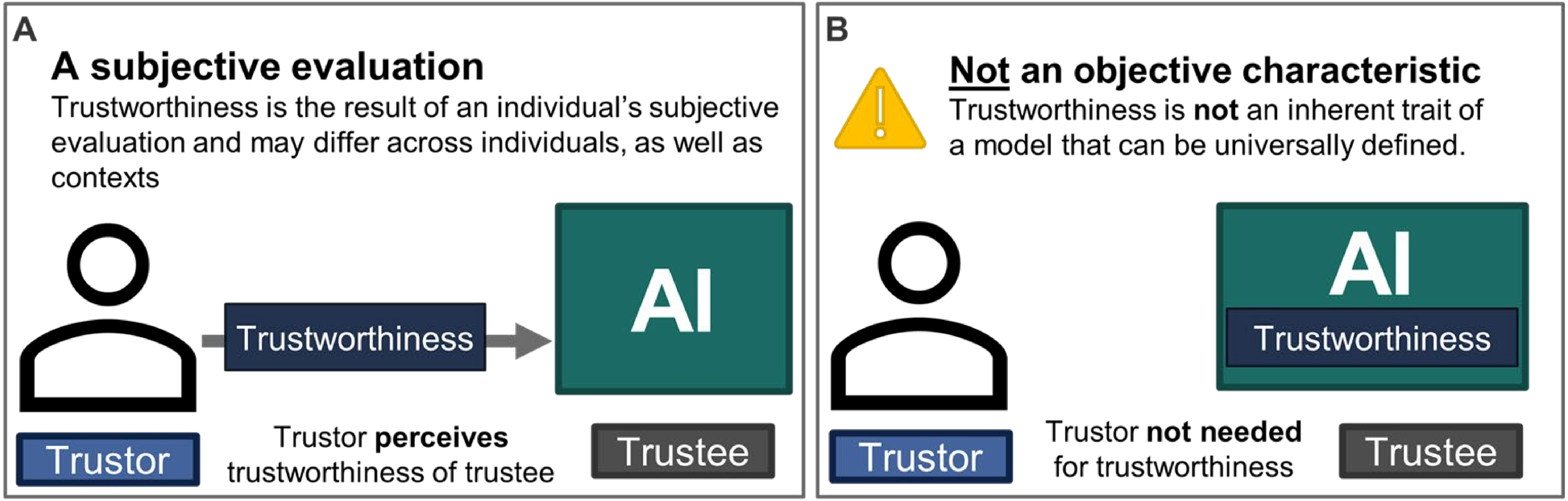
Trustworthiness: An **assessment** of **whether, why, or to what degree** someone or something **should or should not** be **trusted**. (**Wirz et al., 2022**)

We synthesized different literatures on trust to clear up the conceptual ambiguity around trustworthiness

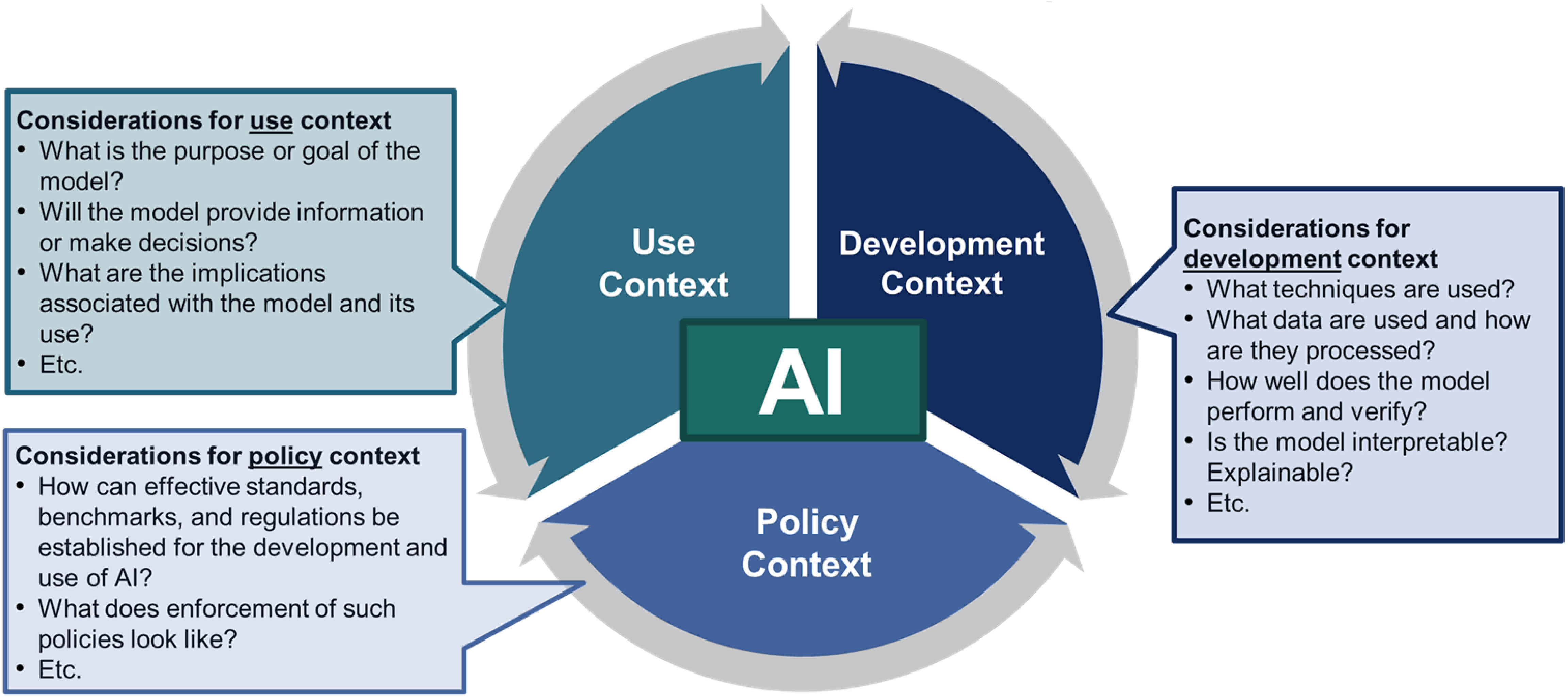
Relationships among reviewed trust and trustworthiness literatures



Trust is relational and trustworthiness is perceptual

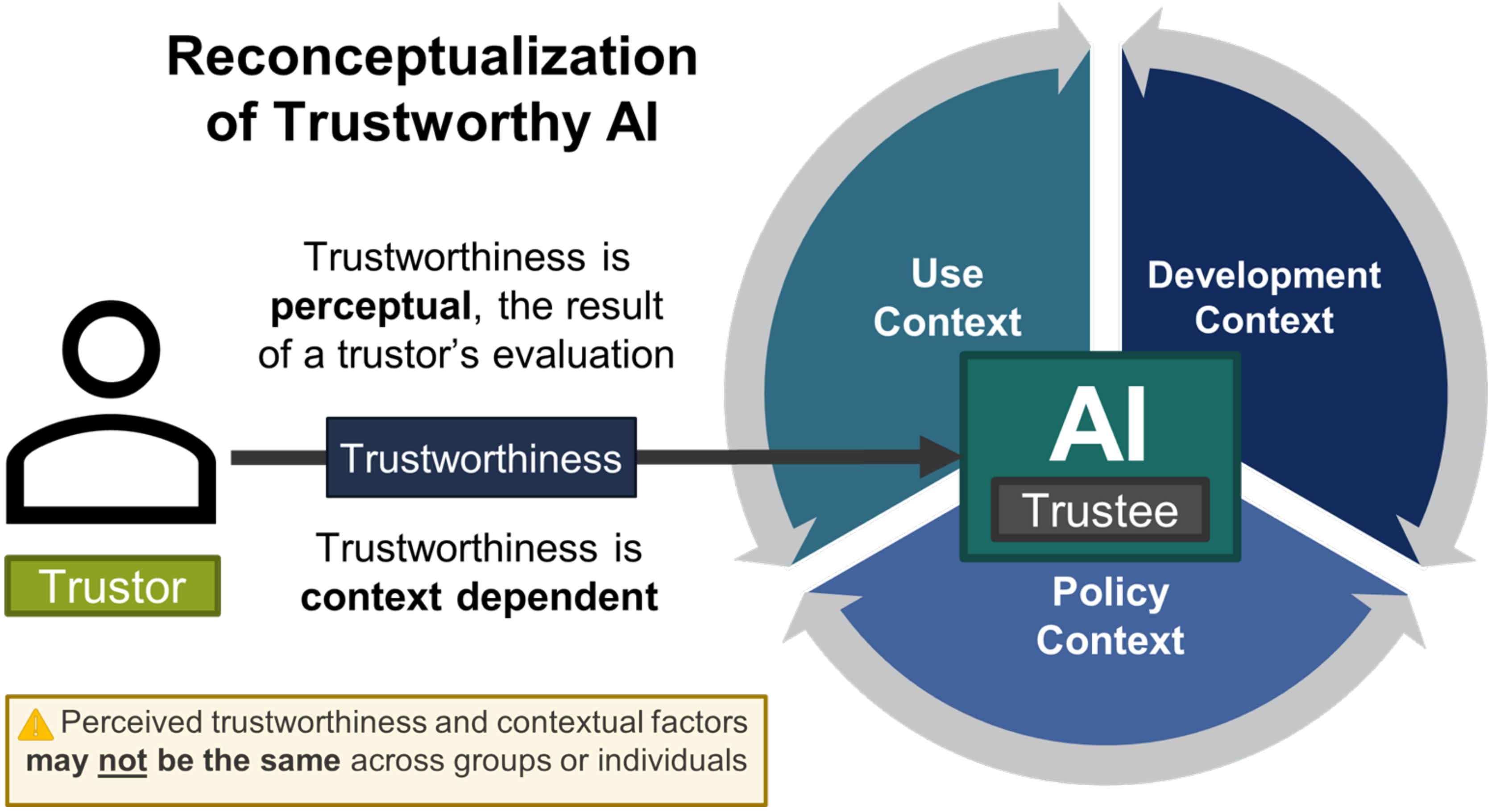


AI is embedded in interrelated contexts that affect its perceived trustworthiness



Conclusion – our reconceptualization of trustworthy AI

Development and policy efforts that **focus both on AI and its potential trustors** are more likely to lead to **AI that is deemed trustworthy, trusted, and used**



Overview

- **AI2ES: Understanding and evaluating trust in AI forecasts**
  - What is trust and trustworthiness?
  - **Why do forecasters to trust AI guidance?**
    - **Storm scale**
    - Global scale
  - Trustworthy AI development lifecycle
- Brightband: Trustworthy forecast verification for global AI models
  - Extreme Weather Bench

Exploring NWS Forecasters' Assessment of AI Guidance Trustworthiness

Publication

Cains, M. G., C. D. Wirz, J. L. Demuth, A. Bostrom, D. J. Gagne, A. McGovern, R. A. Sobash, and D. Madlambayan, 2024: Exploring NWS Forecasters' Assessment of AI Guidance Trustworthiness. Wea. Forecasting, 39, 1219–1241, https://doi.org/10.1175/WAF-D-23-0180.1.

National Weather Service (NWS) Forecasters' perceptions of AI/ML and its use in operational forecasting

Publication

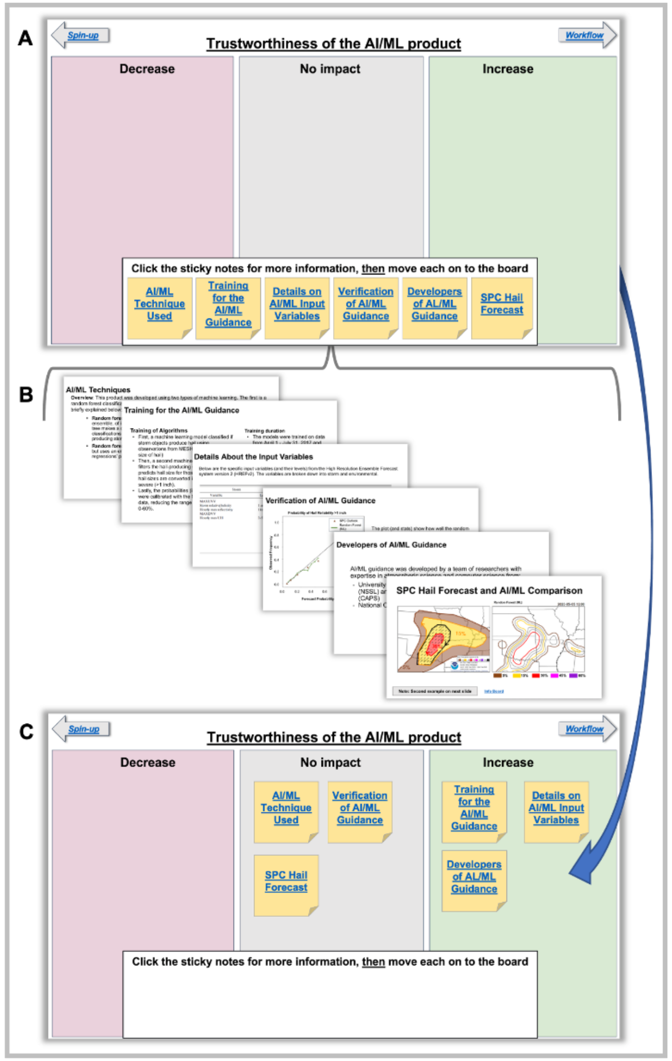





Wirz, C. D., Demuth, J. L., Cains, M. G., White, M., Radford, J., & Bostrom, A. (2024). National Weather Service (NWS) Forecasters' Perceptions of AI/ML and Its Use in Operational Forecasting. *Bulletin of the American Meteorological Society*, 105(11), E2194–E2215. https://doi.org/10.1175/BAMS-D-24-0044.1

Interviews to develop fundamental understanding of human behavior regarding expert decision making and risk assessment

- **We completed 29 structured interviews** with NWS forecasters (generals, leads, and SOOs) from around the U.S. October of 2021- July of 2023
- **Two sets of interviews** one focused on **severe weather** (**16** forecasters)and the other on **coastal fog** (**13** forecasters)
- **Explored prototype AI/ML guidance** for the respective hazards (severe weather and coastal fog)
- **Two versions of each the interview**: 17 mentioned AI in the interview and 12 did not



Cains et al. (2024)

Despite a range of familiarity with AI/ML, forecasters are open to using AI/ML tools operationally

CF02: I think it's great. I'm glad. It's obviously the next step in the evolution of how we process data.

SF12: From a standpoint of job security and those types of things [it's] kind of a scary thought. But also it's an exciting idea to think about how that type of technology may be able to help us.

Highly supportive

Open, but…

Don't know much

Formal training

CF01: I guess I'm just not exposed to it as much. I don't know exactly what has been machine learning. I really don't – I don't know.

CF09: I happened to do machine learning for my master's degree research. So, I'm relatively familiar with

Although forecasters outlined several potential positives with AI/ML, they also had important concerns

Specific applications and implications of AI/ML that forecasters discussed

+ Positives	– Negatives
• Better-performing and enhanced guidance • Bias correction • Limiting forecasters' biases • Guidance that continually improves over time as it 'learns' from more cases • Increased confidence in forecasts and improvements in their ability to message that would come with better and more efficient guidance	• Not being able to catch extreme or rare events given the lack of cases models are trained on • Over-reliance on AI/ML products beyond their application areas or training data • Lack of hands-on experience • Might be too black-boxed for some to feel confident using • Replacing or removing forecasters from the forecasting process

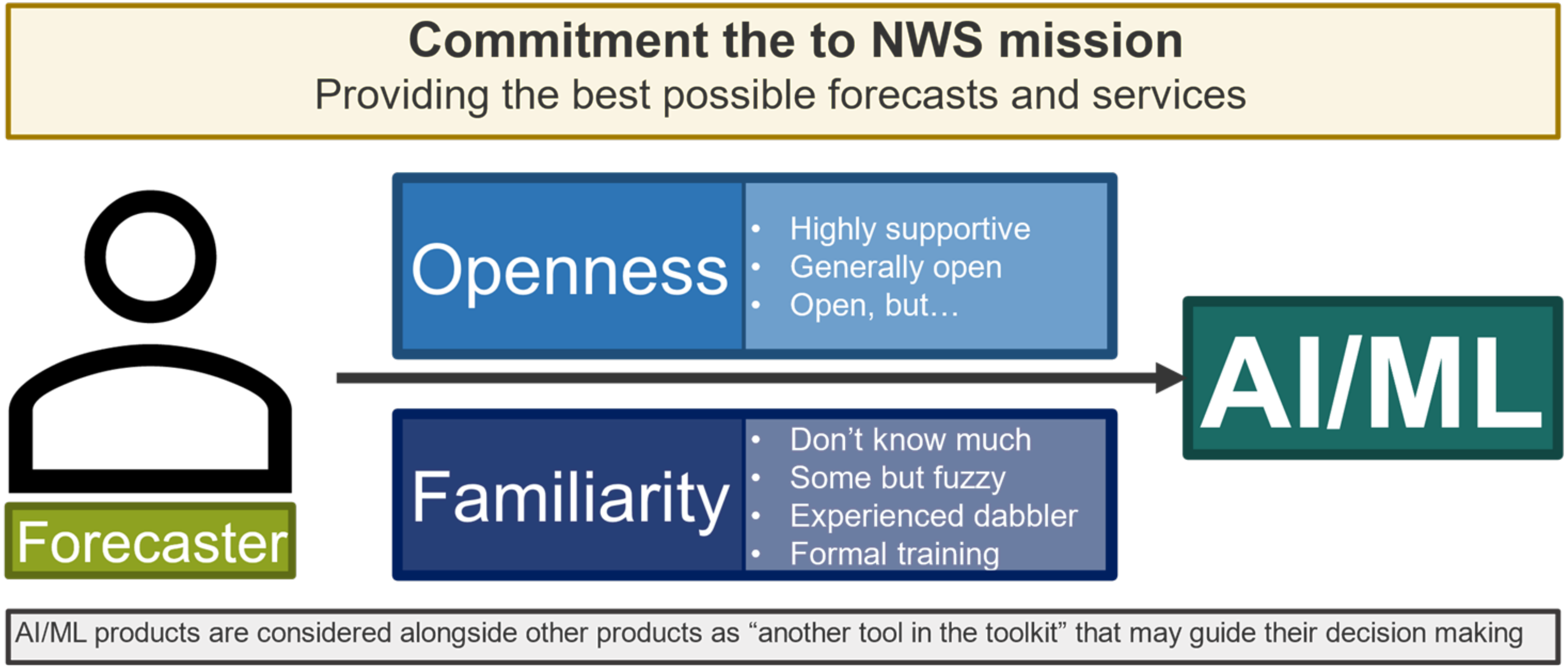
Forecasters expressed a widespread and deep commitment to the best possible forecasts and services to uphold the agency mission

Commitment the to NWS mission
Providing the best possible forecasts and services

SF12: "The mission of the National Weather Service is to **save lives and protect property**, and I think all of us who take that mission to heart, we really do want to **message our forecast with confidence**, especially leading up to these higher-impact events **so that people will take action**.

So if there's **any advantage from new technology, with AI** and others, then **certainly I am all in favor of that** because we want **the public to be more confident in our forecast**, not less confident. [...] And so with that I say **I'm all in favor of ways that we can improve, and AI seems to be a very good possibility to get there**."

Forecasters are generally open to, and sometimes excited about, using AI/ML guidance, if it helps them improve products and services



Although some forecasters see **AI/ML products as the exciting cutting edge of science**, others **care little of the development approach** and more about **how well the product verifies and helps them do their job**.

Forecasters preferred "**machine learning**" over "artificial intelligence" and that labeling a product as being AI/ML did not hurt but **made some more excited**

When looking across all of the data (survey, interview, think aloud):

NWS forecasters' **development of trust** in AI guidance is a **deliberative, dynamic process**. Assessments of AI guidance trustworthiness result from **iterative, intentional engagements** with the guidance.

We saw three phases in which forecasters evaluate new guidance in different ways:

1. initial **exposure and orientation** to new guidance,
2. further familiarization with new guidance through **non-operational information-seeking and interrogation**, and
3. operational experience through **real-time observation of guidance** and potentially use of it for forecasting.

Phases may overlap and are not necessarily a linear progression; each a key part of how forecasters assess trustworthiness.

An Assessment of How Domain Experts Evaluate Machine Learning in Operational Meteorology

Publication

Harrison, D. R., A. McGovern, C. D. Karstens, A. Bostrom, J. L. Demuth, I. L. Jirak, and P. T. Marsh, 2025: An Assessment of How Domain Experts Evaluate Machine Learning in Operational Meteorology. Wea. Forecasting, 40, 393–410, https://doi.org/10.1175/WAF-D-24-0144.1.

An Assessment of How Domain Experts Evaluate Machine Learning in Operational Meteorology

- Research questions:
  - What factors influence a domain expert's decision to trust and implement new products and technologies in their daily procedures?
  - Do these factors differ between AI/ML-derived products and products designed via more traditional methods?

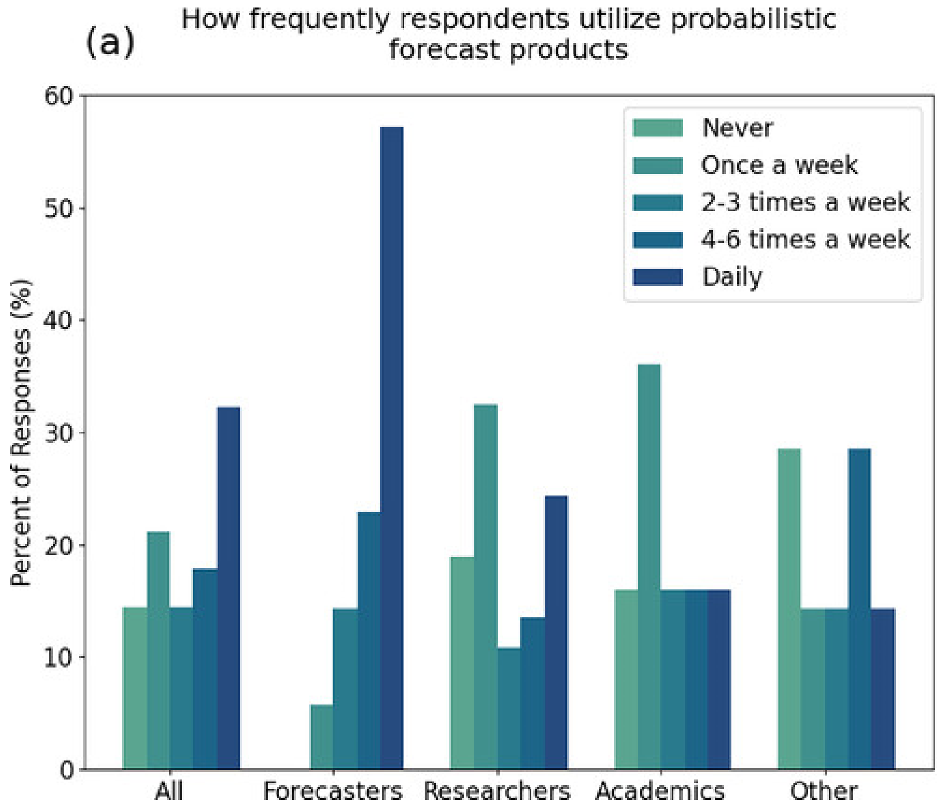
The UNIVERSITY of OKLAHOMA

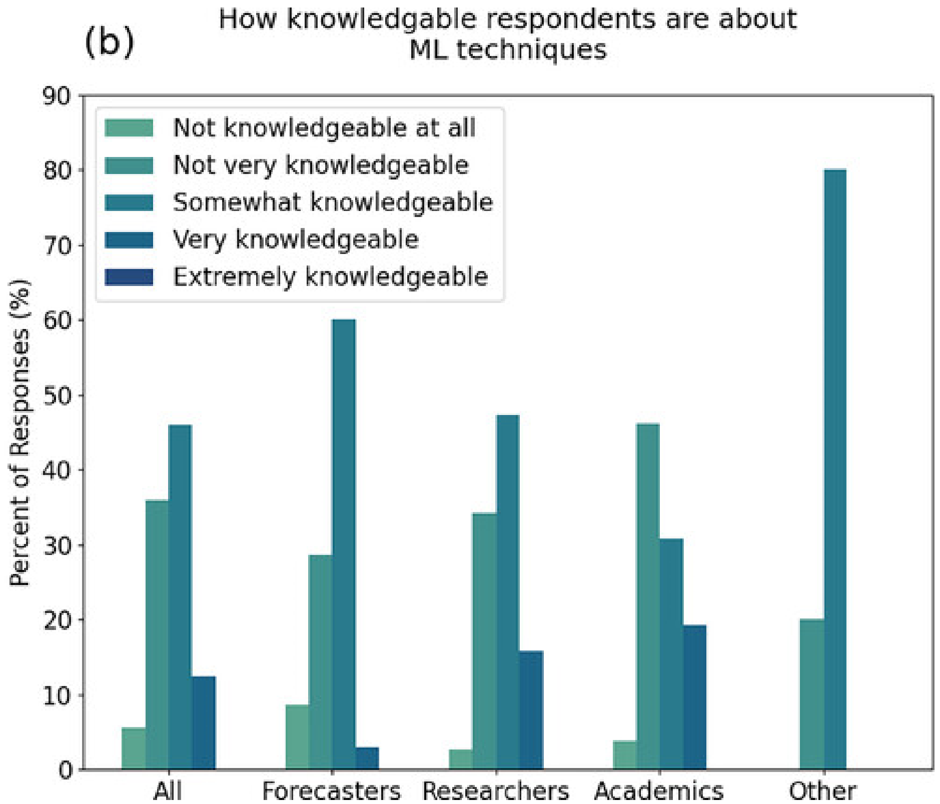
Approach

Survey of participants in the 2021 NOAA Hazardous Weather Testbed

- 133 forecasters, researchers, and students over 5 weeks

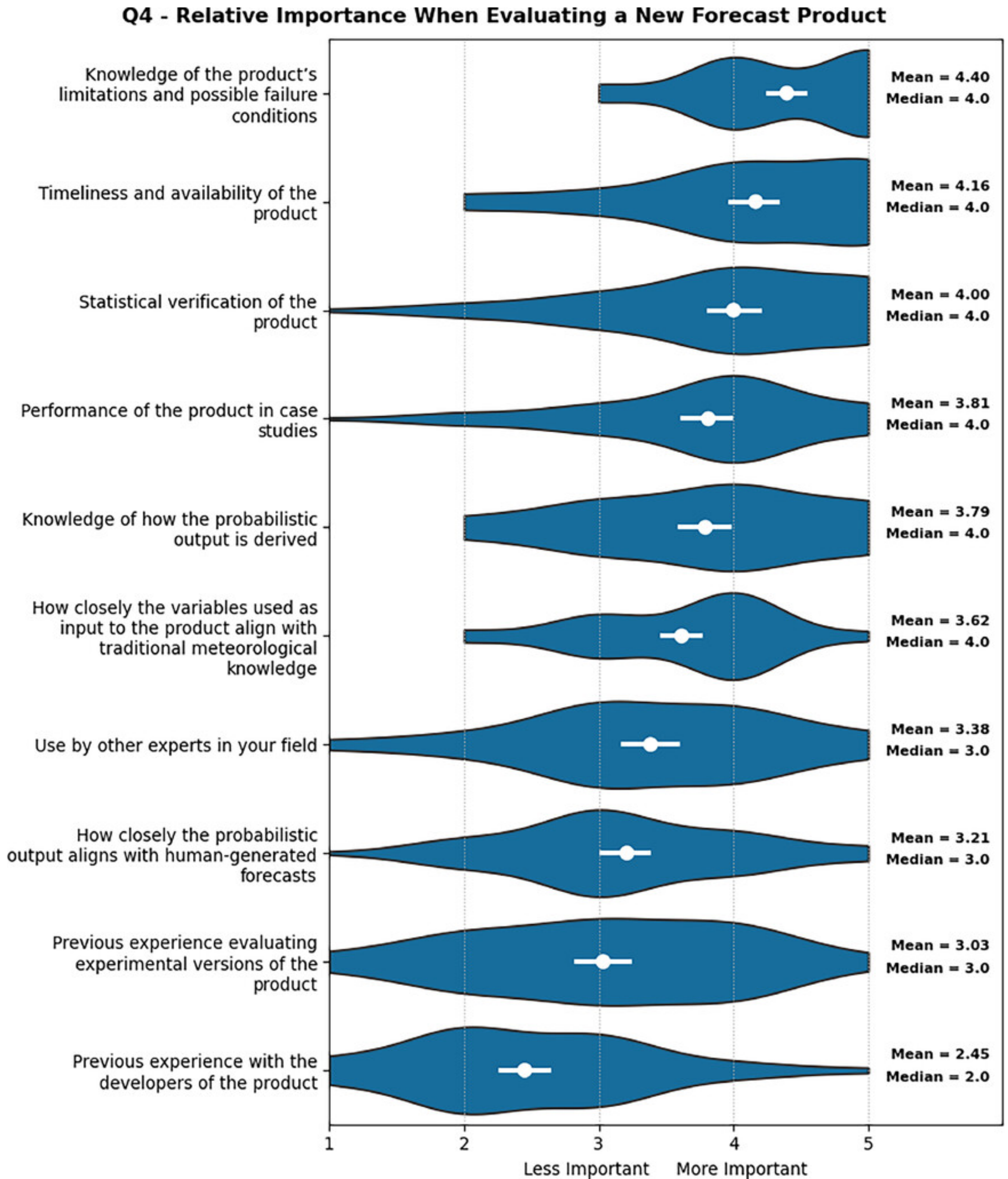
Background	No. of respondents	Percent of respondents (%)
Operational forecaster	36	34
Researcher	38	35
Academic faculty/staff	16	15
Students	10	9
Other	7	7





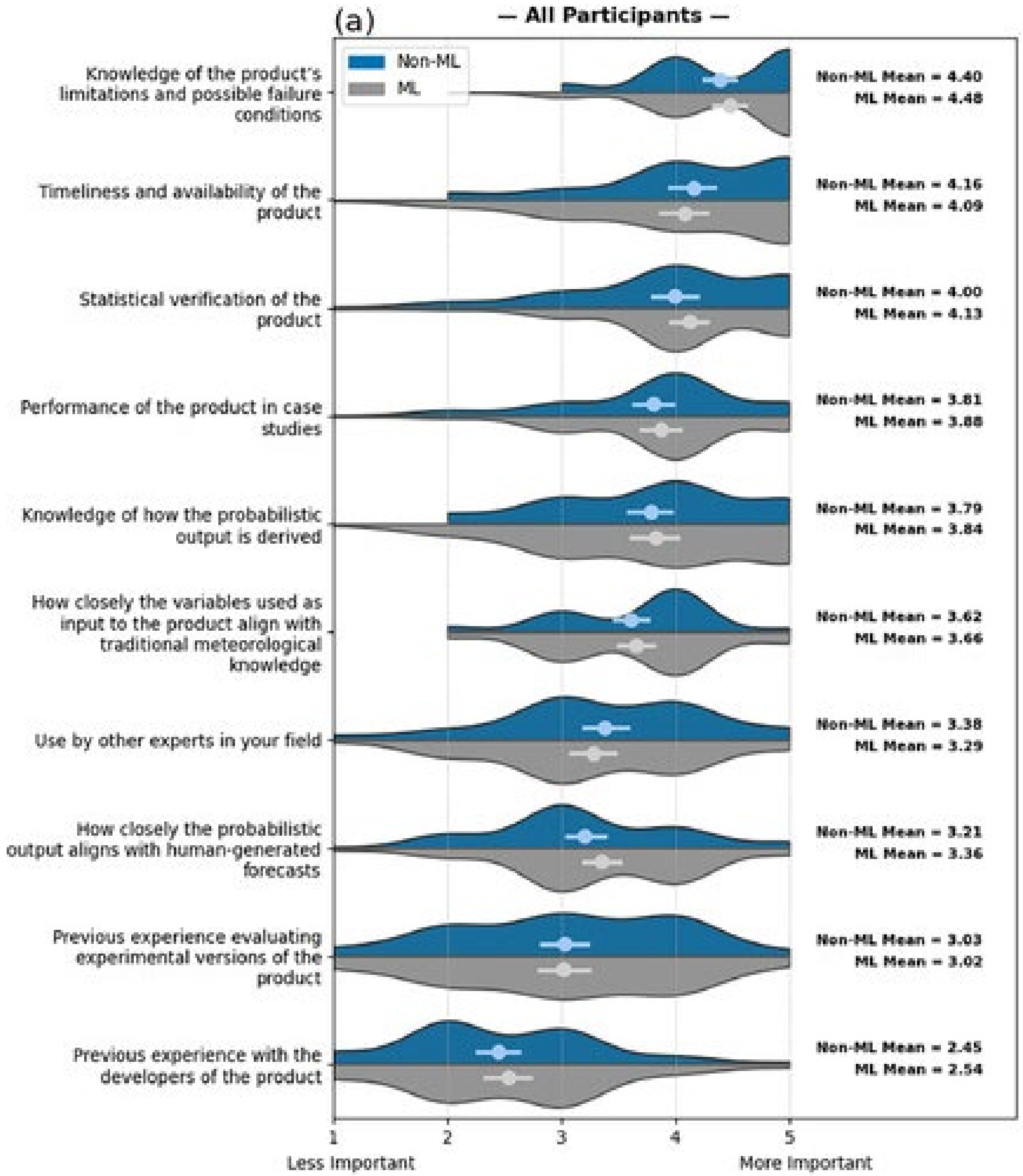
The UNIVERSITY of OKLAHOMA

When evaluating how useful that product might be to your personal forecasting process, how important are each of the following factors?



The UNIVERSITY of OKLAHOMA

How does this change for ML versus traditional products?



Overview

- **AI2ES: Understanding and evaluating trust in AI forecasts**
  - What is trust and trustworthiness?
  - **Why do forecasters to trust AI guidance?**
    - Storm scale
    - **Global scale**
  - Trustworthy AI development lifecycle
- Brightband: Trustworthy forecast verification for global AI models
  - Extreme Weather Bench

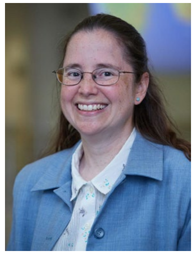
Perceptions and Performance of Global AI Models in the 2024 Hazardous Weather Testbed

Maria Madsen[1,2], David Harrison[2,4,5], Michael Baldwin[4,5], Adam Clark[2,6], Joseph Ripberger[3], Sean Ernst[3,4], Amy McGovern[1,2] , and Aaron Hill[1,2]










[1]NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography, University of Oklahoma

[2]School of Meteorology, University of Oklahoma

[3]University of Oklahoma Institute for Public Policy Research and Analysis

[4]Cooperative Institute for Severe and High-Impact Weather Research and Operations, University of Oklahoma

[5]NOAA/NWS/Storm Prediction Center

[6]NOAA/OAR/National Severe Storms Laboratory

Background

- Interested in measuring accuracy, reliability, and gaps in model performance versus perception of global AI models

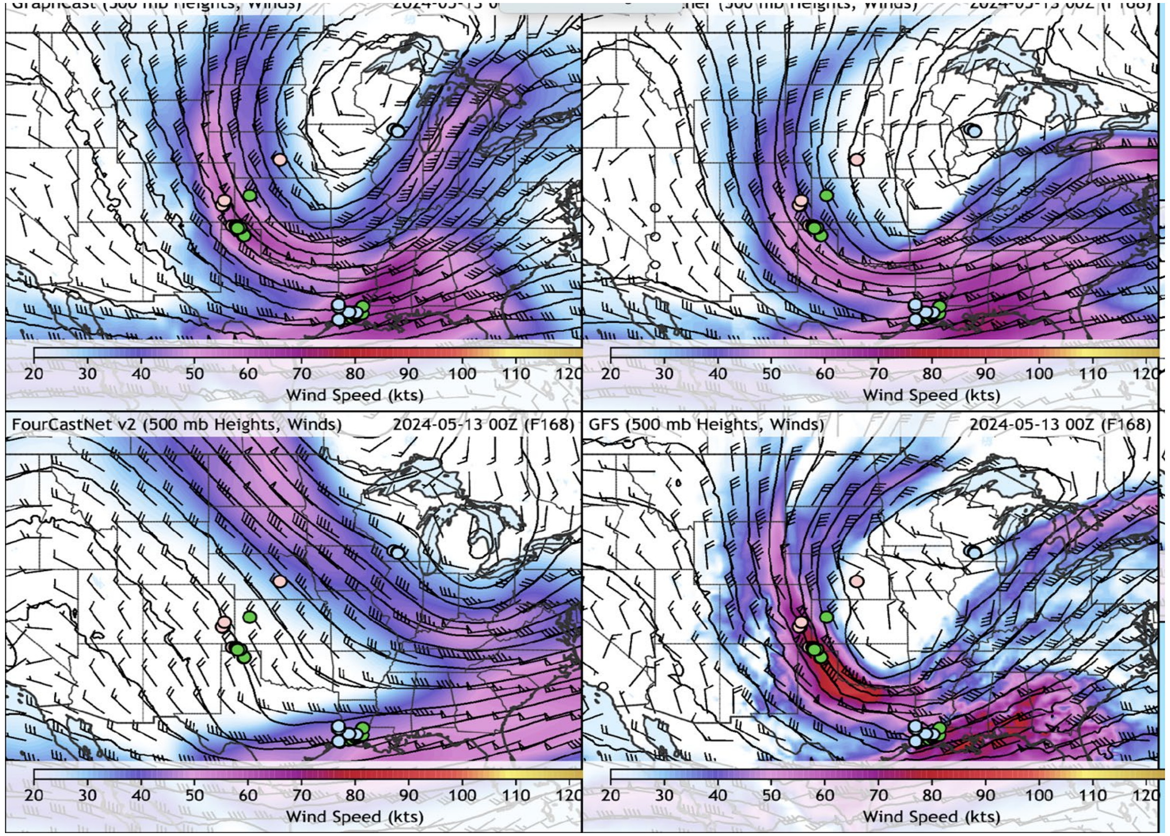
Research Questions:

1. Can global AI models provide value at medium-range timescales?
2. Is there a gap between AI model performance and the perception of performance by end users?
3. Is there a difference in perception between operational forecasters and researchers/model developers?
4. What aspects do users find most important for global AI models?

HWT 2024 Design

- 2024 HWT Spring Forecasting Experiment held April 29-May 31
  - New participants each week
- Each AI model run experimentally at the Cooperative Institute for Research in the Atmosphere (CIRA)
  - GraphCast
  - PanguWeather
  - FourCastNet v2
- Initialized using GFS initial conditions
- AI models compared against GFS

HWT 2024 Design



https://hwt.nssl.noaa.gov/sfe_viewer/2024/ai/

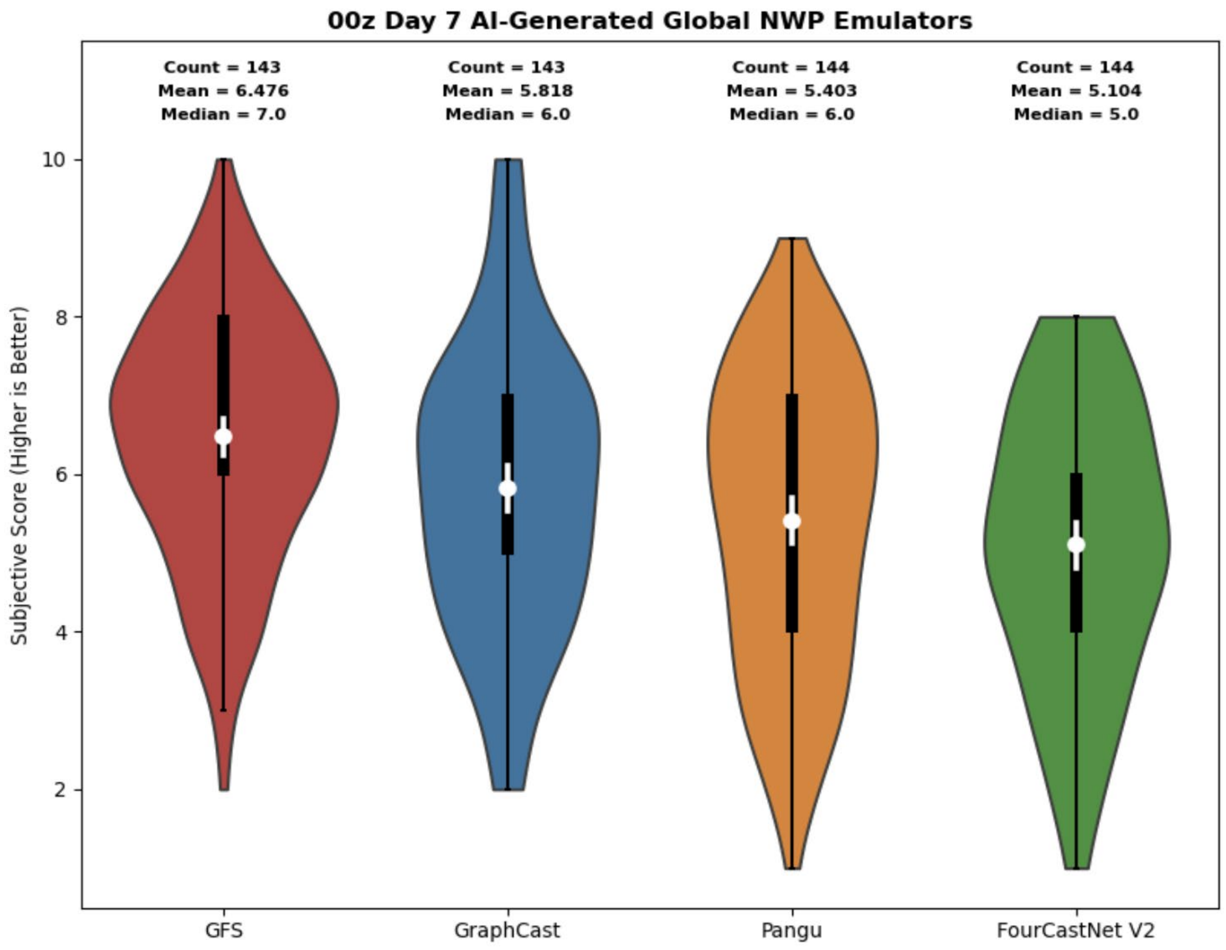
- Projected fields for day 7 (F156-180) for previous day evaluation
  - 500 mb wind & geopotential heights
  - 850 mb wind & geopotential heights
  - 2-m temperature
  - 6-h QPF (GraphCast only)

- Participants focused on area of severe weather

Subjective Results

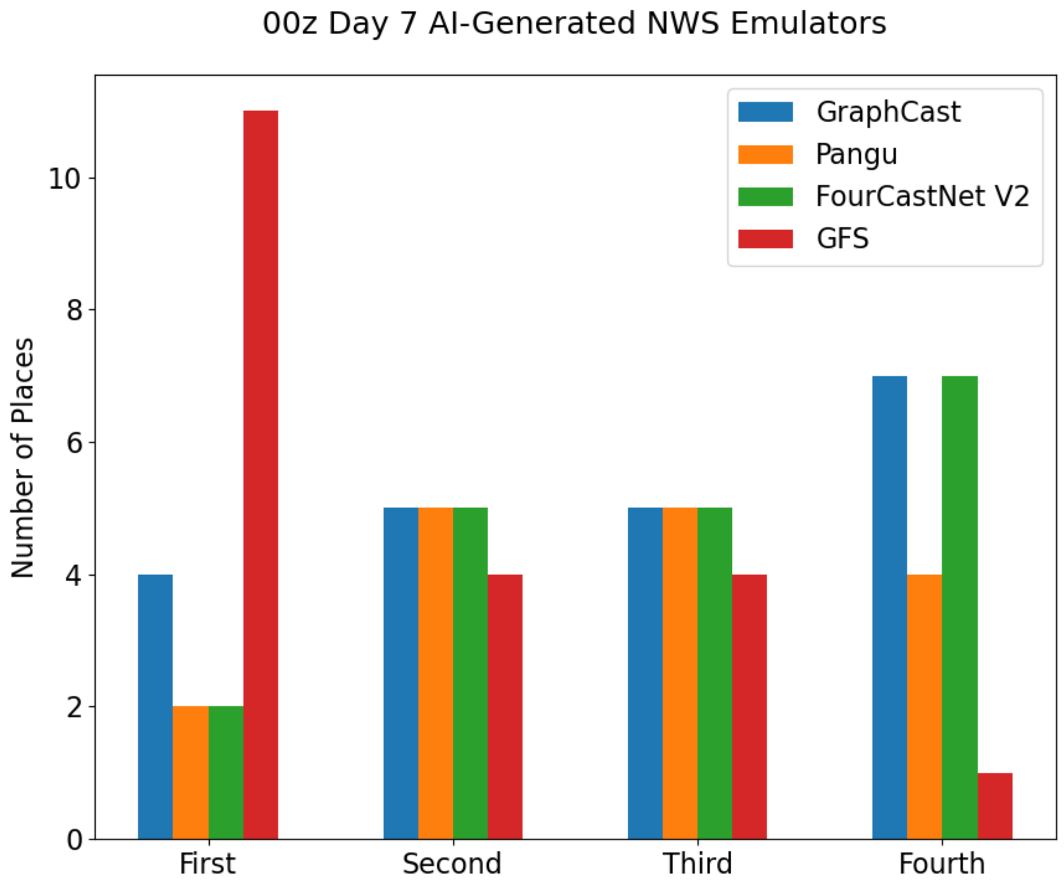
Compared to GFS Analysis at Day 7 lead times:

- GFS rated the highest on average
- GraphCast rated statistically similar to GFS (at 95% confidence interval)
- Pangu and FourCastNet v2 rated statistically lower than GFS



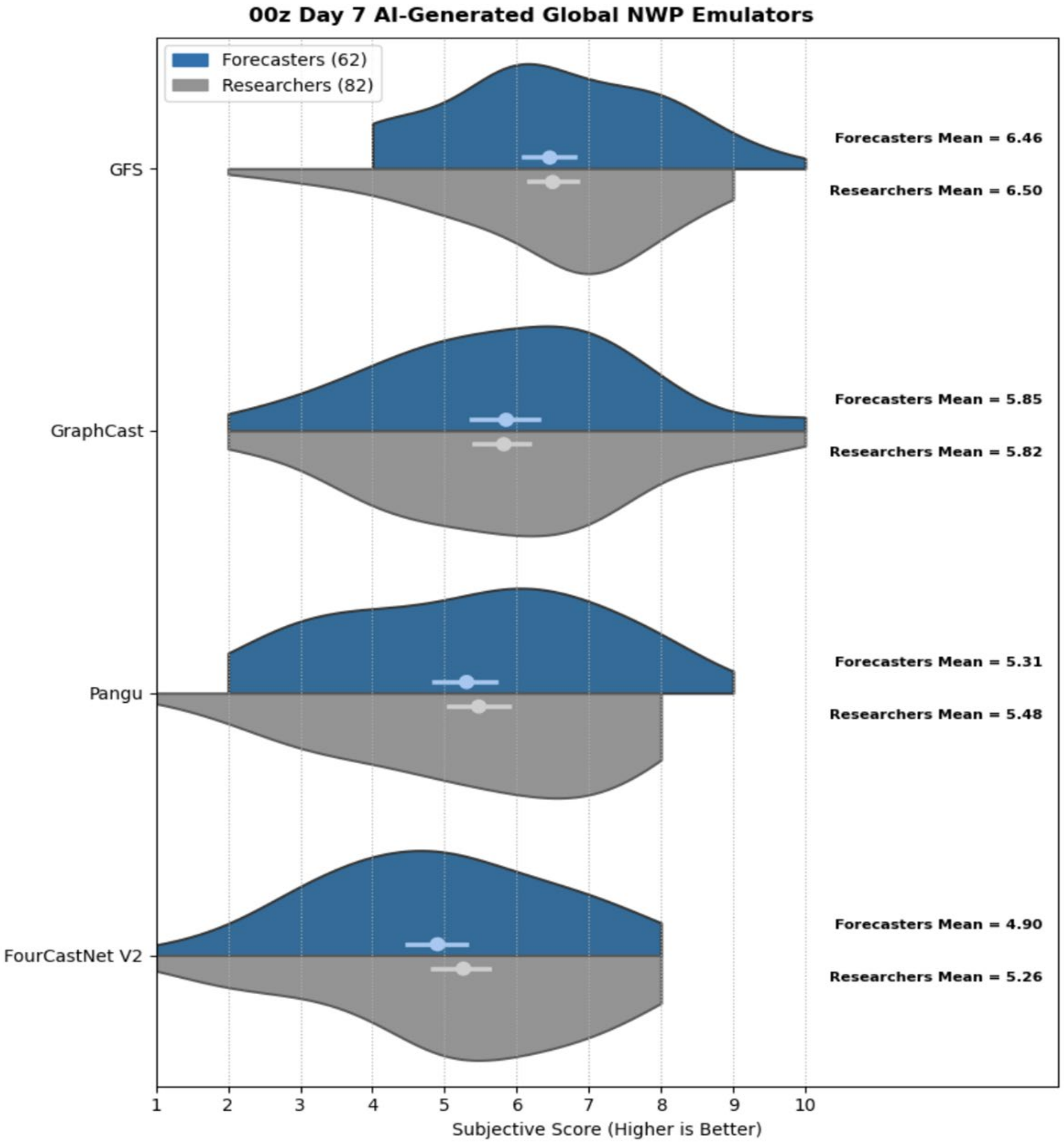
Subjective Results

- GFS rated the best model on 11 of 20 days
- GraphCast rated the best on 4 days
- No clear winner among AI models
- Pangu rated the worst model less often than the other AI models



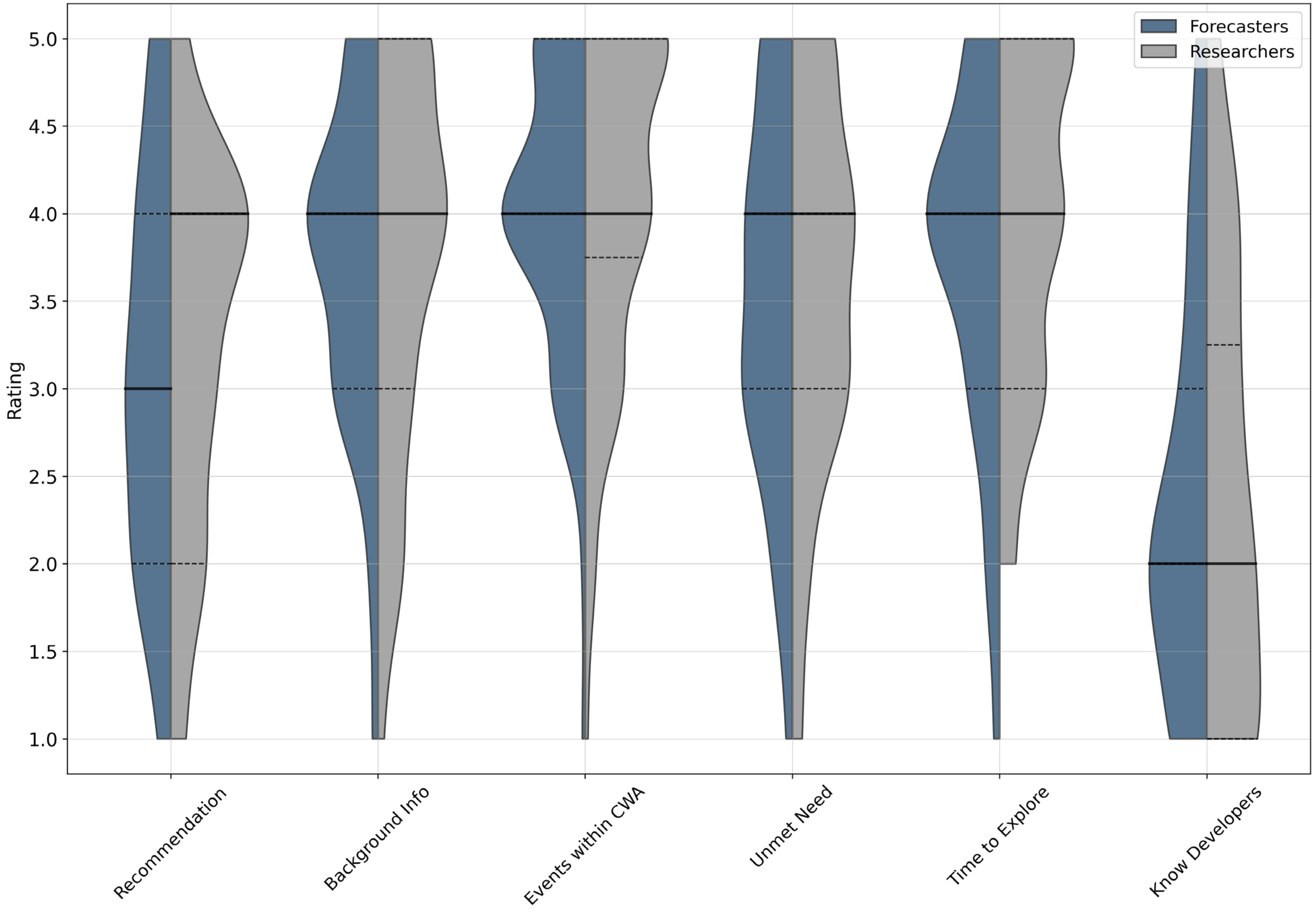
Subjective Results

- GFS rated the best model & GraphCast the best AI model for both forecasters & non-forecasters
- Slight differences in Pangu and FourCastNet v2 between the two groups



Factors that Impact Trust in Global AI Models

- Recommendation was more important for researchers & model developers
- Knowing the developers was least important for both groups
- Similar levels of importance for other factors



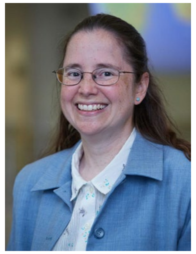
Global AI Model Trust Conclusions

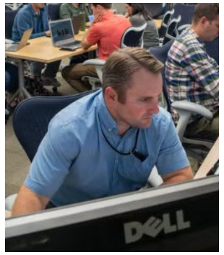
- Although the GFS was rated the highest among participants, GraphCast had the highest AI model subjective rating
  - The “best” AI model varied considerably from day-to-day
- AI models showed less run-to-run and day-to-day consistency than the GFS
- Subjective ratings preferred GFS more often than objective ratings
  - Participants likely prefer more realistic detail in GFS
- The least important factor for AI trust was knowing developers, whereas the other factors were similarly important
- Participants impressed by current state of models, but indicated there’s still work to do before they’re ready for operations

Overview

- **AI2ES: Understanding and evaluating trust in AI forecasts**
  - What is trust and trustworthiness?
  - Why do forecasters to trust AI guidance?
    - Storm scale
    - Global scale
  - **Trustworthy AI development lifecycle**
- Brightband: Trustworthy forecast verification for global AI models
  - Extreme Weather Bench

Leveraging Co-Production to Bridge Research and Operations in Operational Meteorology


Harrison, D. R., A. McGovern, C. D. Karstens, A. Bostrom, I. L. Jirak, and P. T. Marsh, 2025: Leveraging Co-Production to Bridge Research and Operations in Operational Meteorology. To appear in Weather and Forecasting,

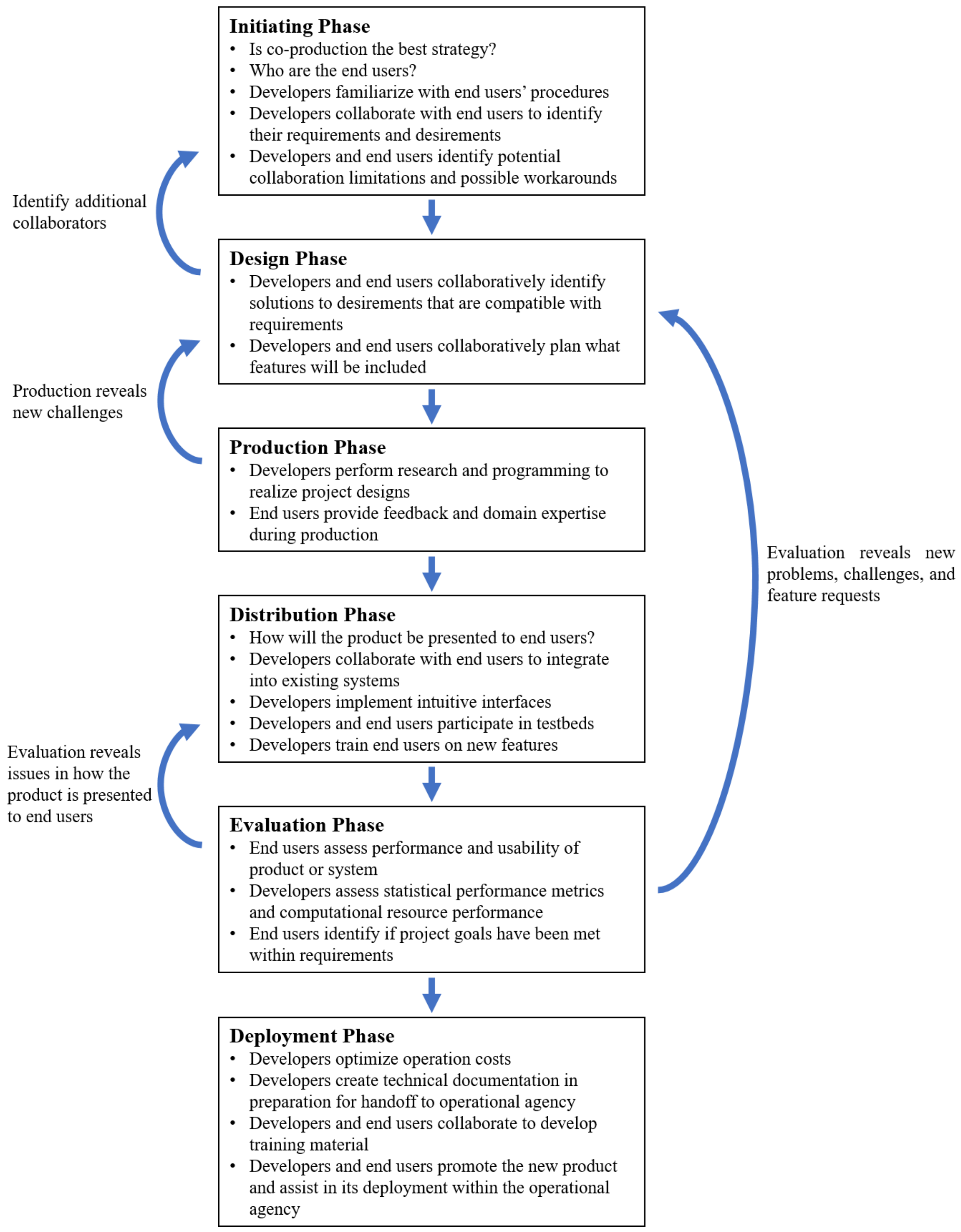
Development is Non-linear

- Existing R2O procedures assume development is *linear*
  - Modern development cycles are **recursive**
- Products may occupy multiple readiness levels simultaneously
- This non-linearity is difficult to define in existing processes
- Revisiting an earlier stage of development may be viewed as backwards or negative progress – potentially disincentivizes iterative development

The UNIVERSITY of OKLAHOMA

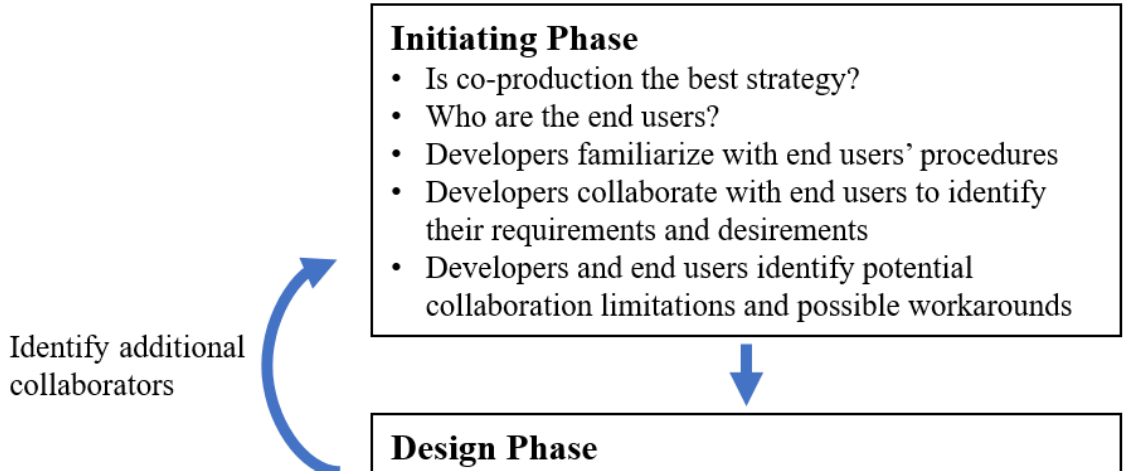
A New R2O Model

- Adapted from NOAA's RLs, Hoffman's Practitioner's Cycles, and modern co-production philosophies
- Main goals:
    - Allow for and encourage iterative product development
    - Require coordination and collaboration with end users throughout the development cycle
    - Project maturity is defined by what goals have been met, rather than what tasks are currently being performed



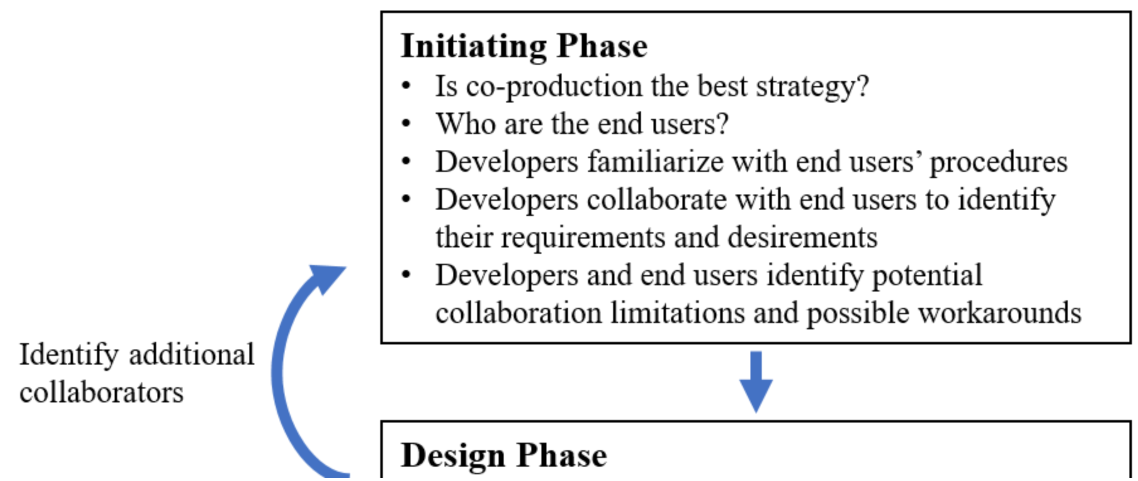
1. Initiating Phase

- **Co-production**: A collaborative process that provides a service or product via an equal, reciprocal relationship between developers and end users
- End user should be an ally and resource of the development process
- Co-production may not always be the best strategy
  - Research that challenges operational norms may benefit from first developing proof of concepts and testing an initial prototype before engaging with operational end users
  - Security requirements, data accessibility, and technical limitations may reduce the effectiveness of collaboration



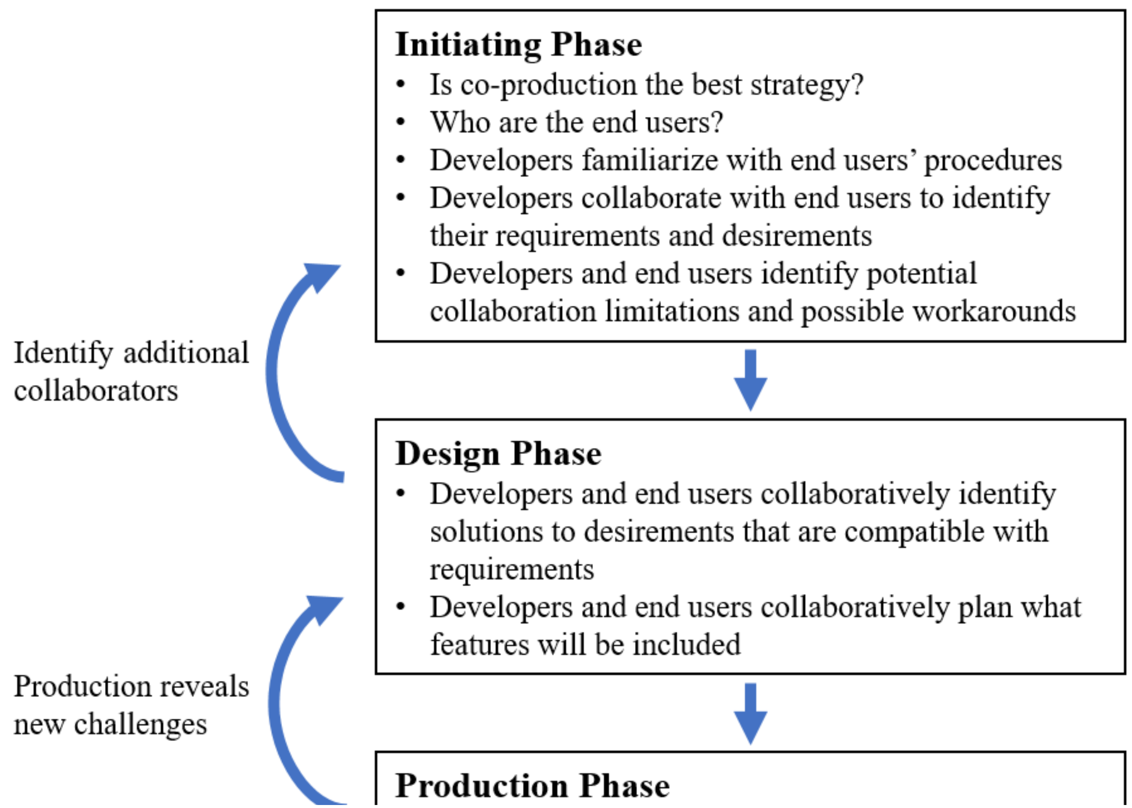
1. Initiating Phase

- Developers should become familiar with operational workflows and procedures (shadow a shift, take forecaster training, etc.)
- Forecasters should use this opportunity to describe what the proposed project must do to be useful (requirements) and what they would like it to do if possible (desirements).
- Ideally, one or two forecasters could sponsor a research project and commit to providing regular feedback throughout the project's development
- Finally, identify any limitations to the collaborative development and begin looking for possible workarounds



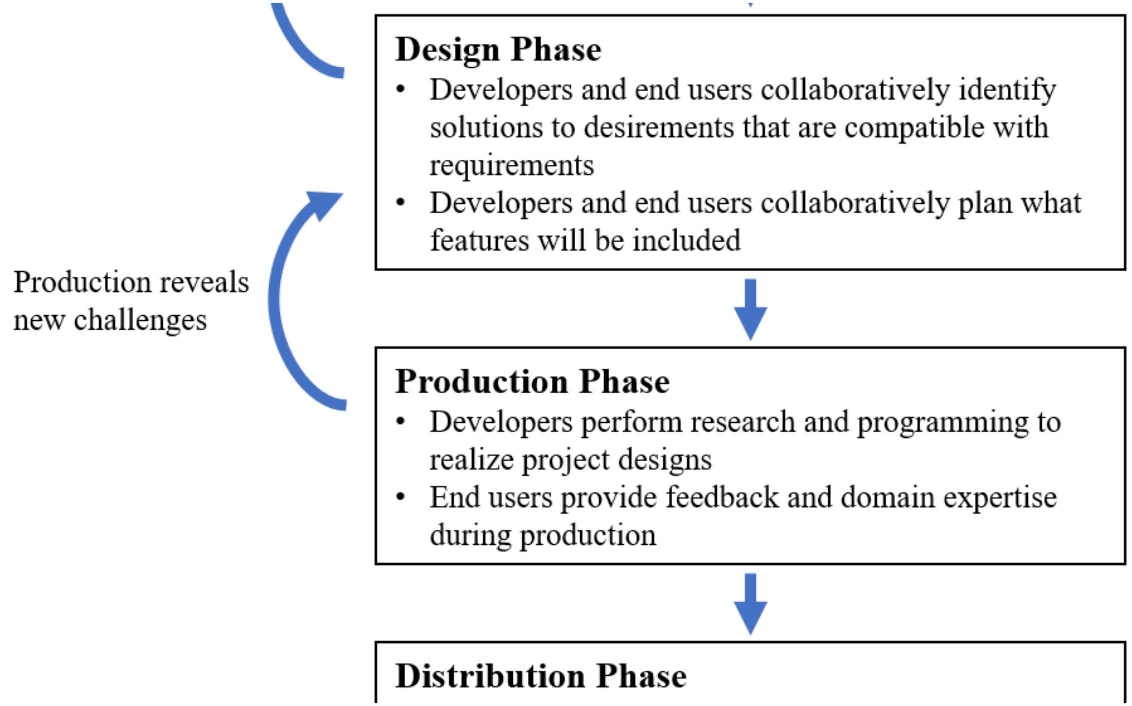
2. Design Phase

- Developers and forecasters work together to define the goals of the development and a rough plan of how to accomplish them
  - Goals should incorporate as many forecaster desirements as possible, but requirements must be given priority
- This process is recursive, and additional requirements and desirements will be identified as development progresses. Project goals and development plans should update accordingly.



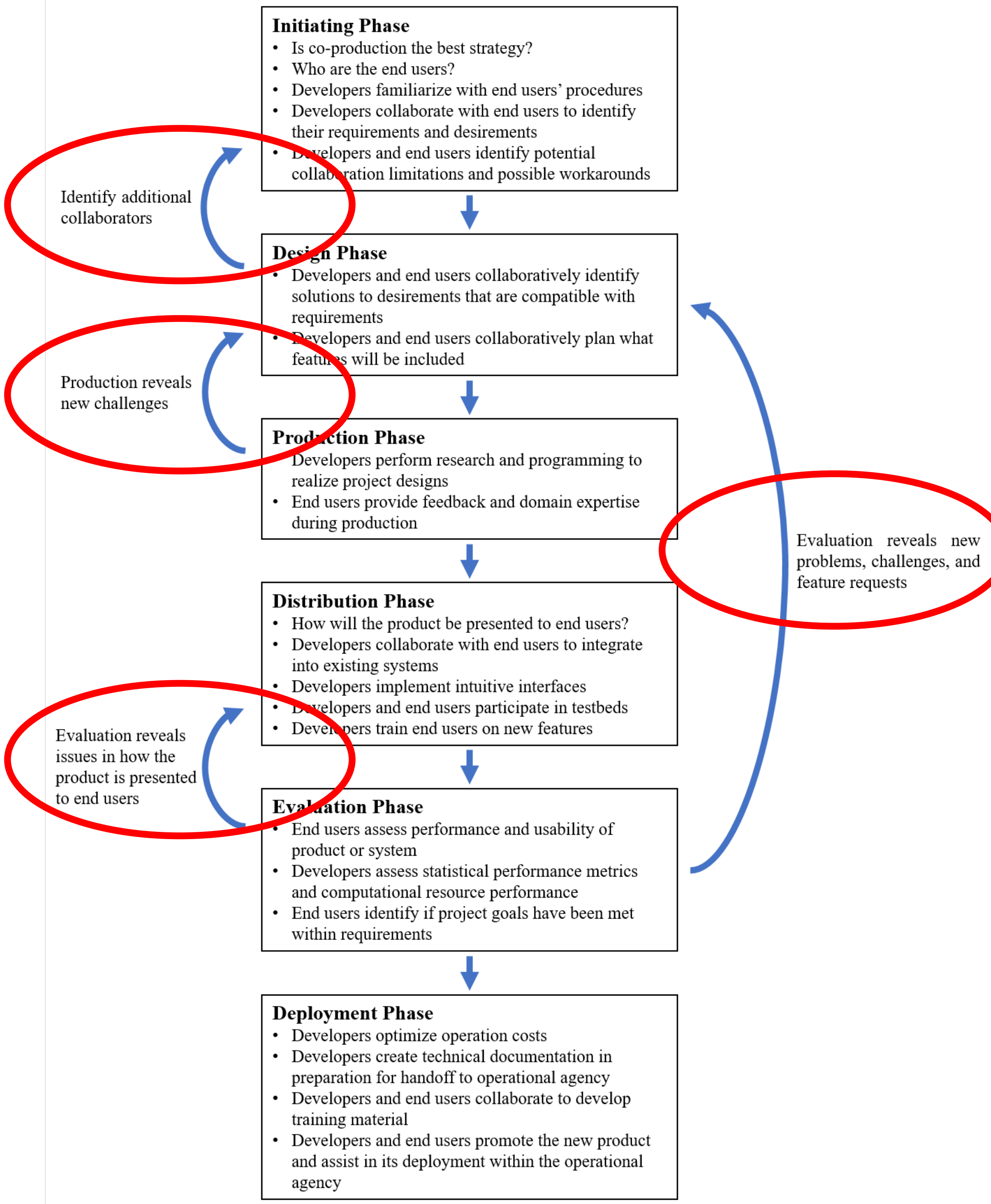
3. Production Phase

- Developers produce a mockup, prototype, or deployable system.
  - Generally less collaboration required at this stage, but regular progress updates are recommended
- Developers may use this phase to familiarize forecasters with development techniques (e.g., machine learning)
- Forecasters may also guide or lead research activities that fall within their domain expertise as time and resources allow



Built-in Recursion

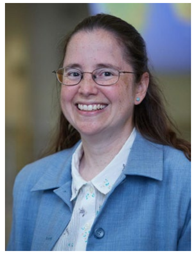
- Developers and forecasters may decide they need additional domain expertise during the design phase.
- Technical or scientific limitations may arise during the production phase that requires changes to project goals established in the design phase
- Forecasters and developers may identify interface or dataflow issues during evaluation that require modifications to how the product is presented
- Evaluation reveals performance issues, new desirements, and new requirements which are addressed in the next product iteration



Overview

- AI2ES: Understanding and evaluating trust in AI forecasts
  - What is trust and trustworthiness?
  - Why do forecasters to trust AI guidance?
    - Storm scale
    - Global scale
  - Trustworthy AI development lifecycle
- **Brightband: Trustworthy forecast verification for global AI models**
  - **Extreme Weather Bench**

Extreme Weather Bench





Amy McGovern
Daniel Rothenberg
Taylor Mandelbaum
Nicholas Loveday, BoM
Linus Magnusson, ECMWF

Brightband

Brightband Mission

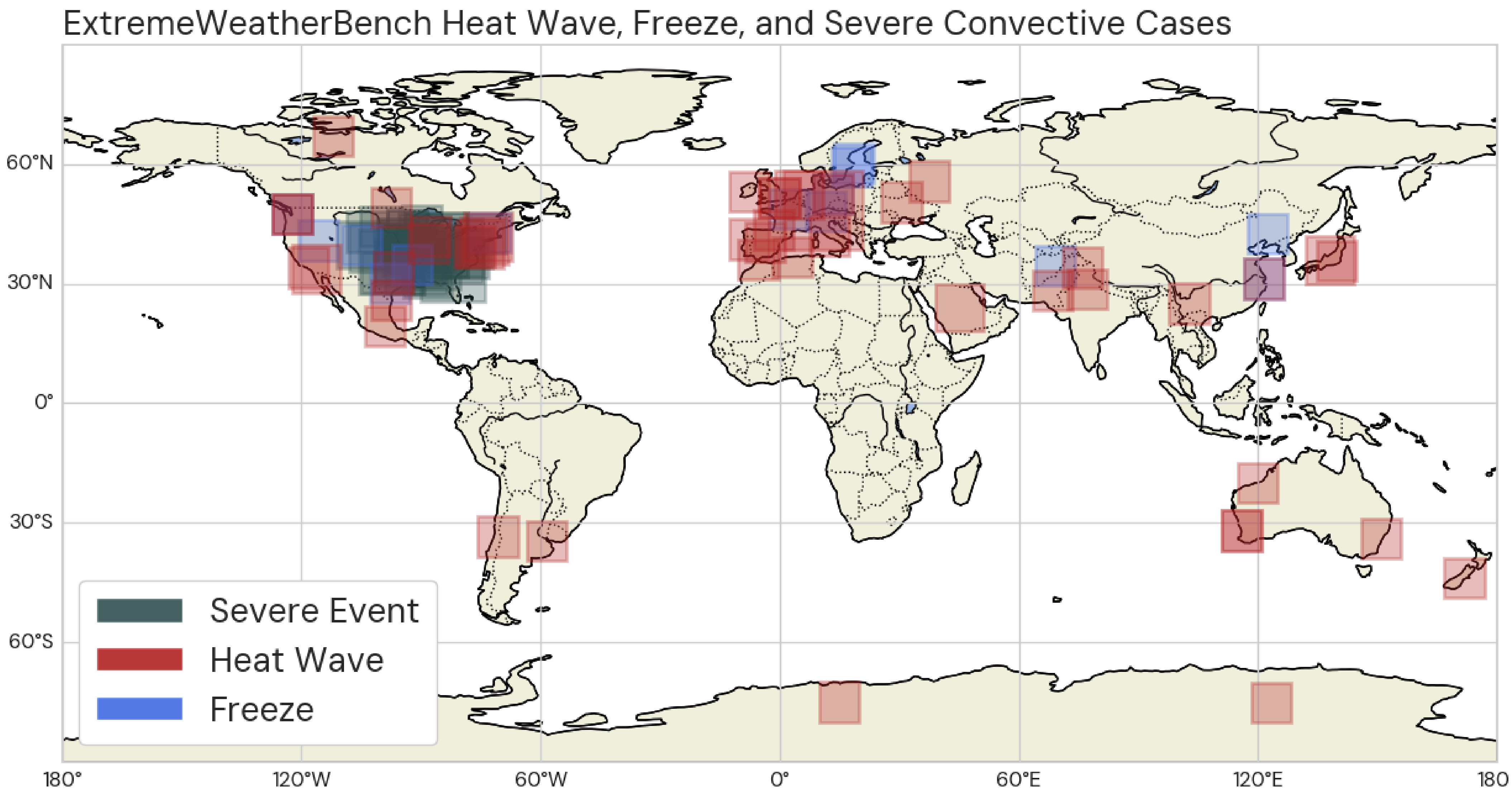
Make AI weather forecasting tools available to all, to help humanity adapt to increasingly extreme weather

- Building an end-to-end probabilistic forecasting system
- Open-sourcing benchmark datasets, models and metrics to spur innovation in AI weather forecasting
- AI weather forecasting tools to all

Extreme Weather Bench (EWB)

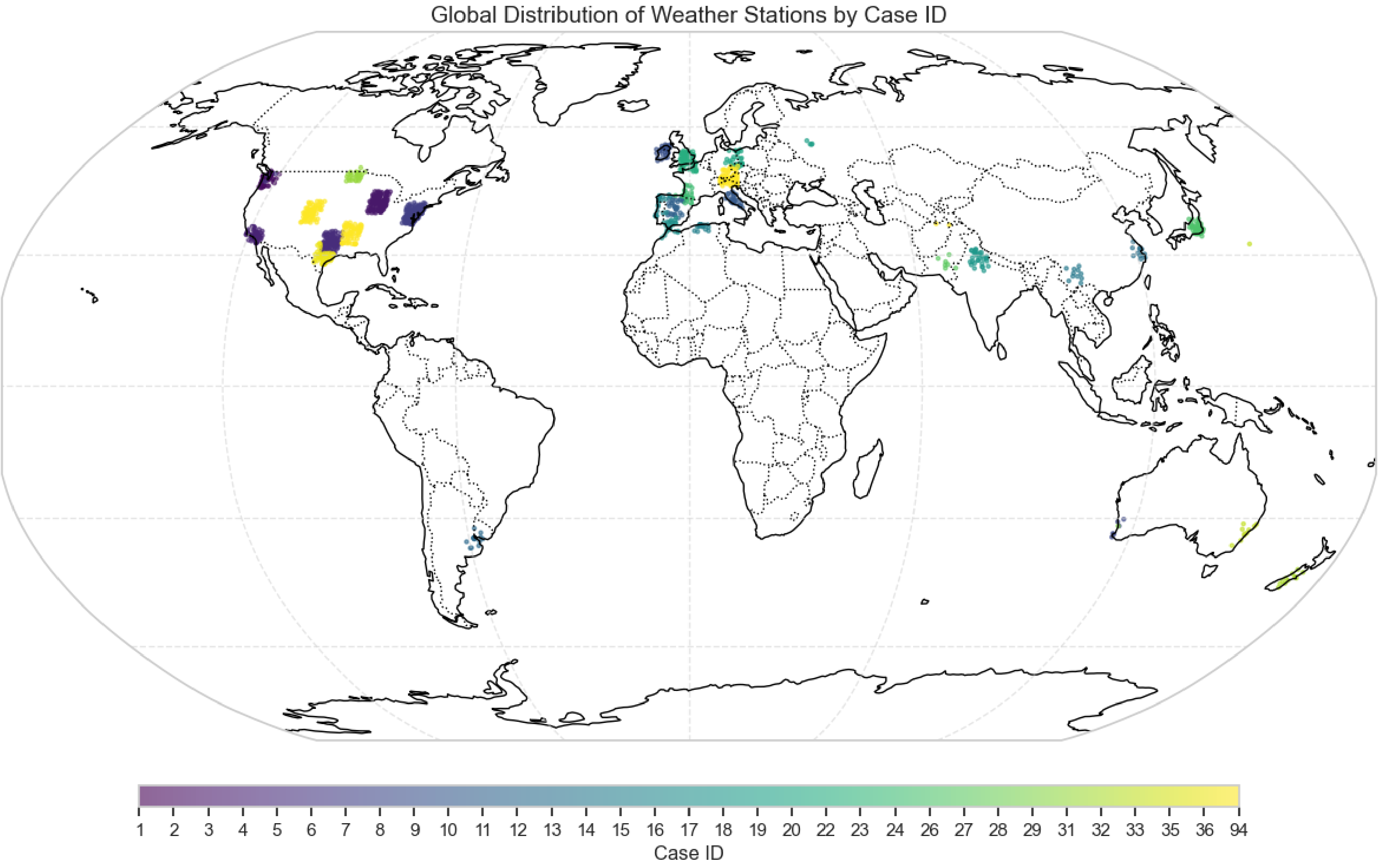
- Standardized set of global high-impact weather events, data, metrics and code
  - **Evaluate both across events and dive deeply into an event or set of events**
- EWB provides
  - Information about the event
  - Data (observations if available)
  - Standard impact-based metrics
- Community driven
  - We want community input, feedback, new data, case studies, and metrics!

EWB Case Studies



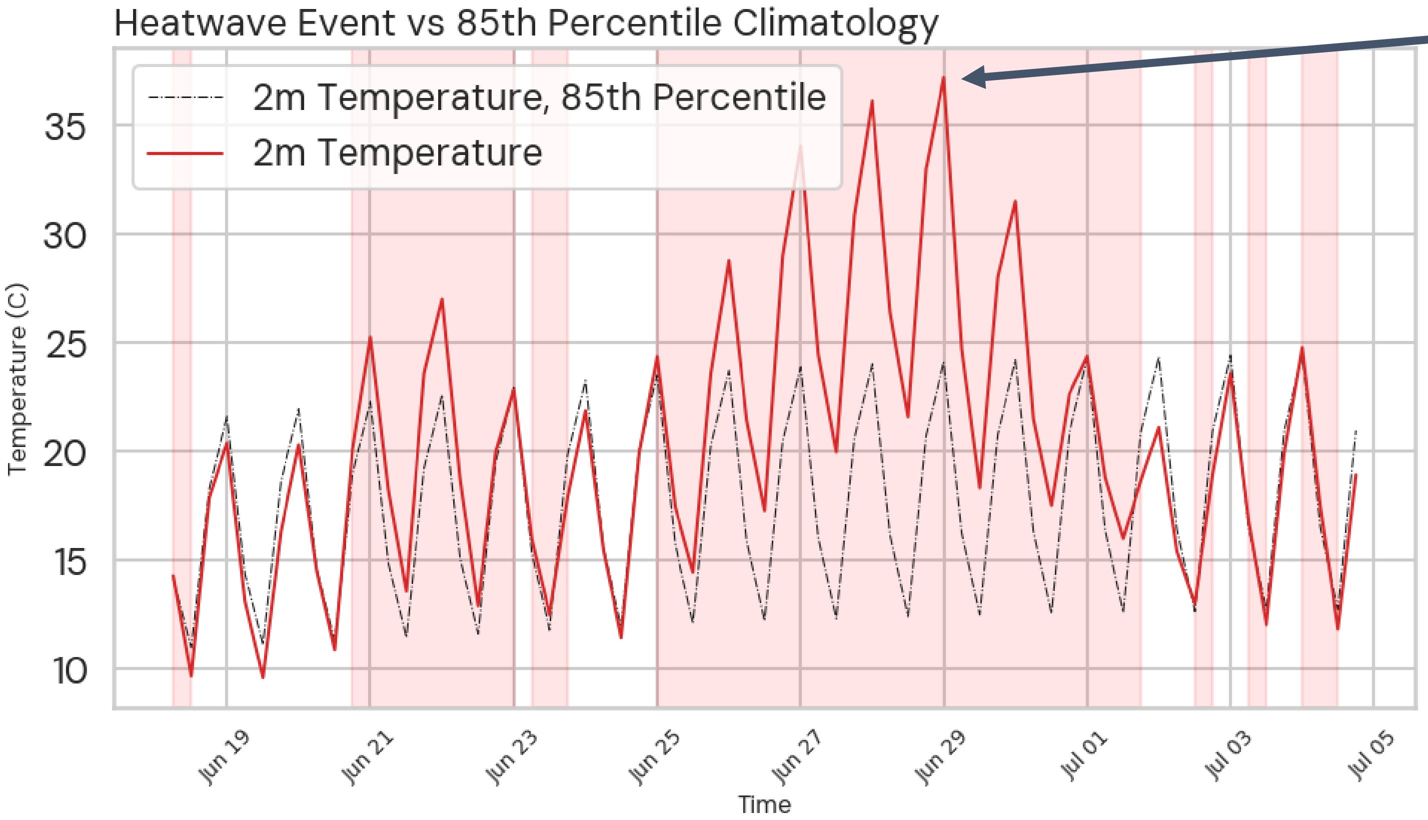
Coming soon: Tropical cyclones and atmospheric rivers

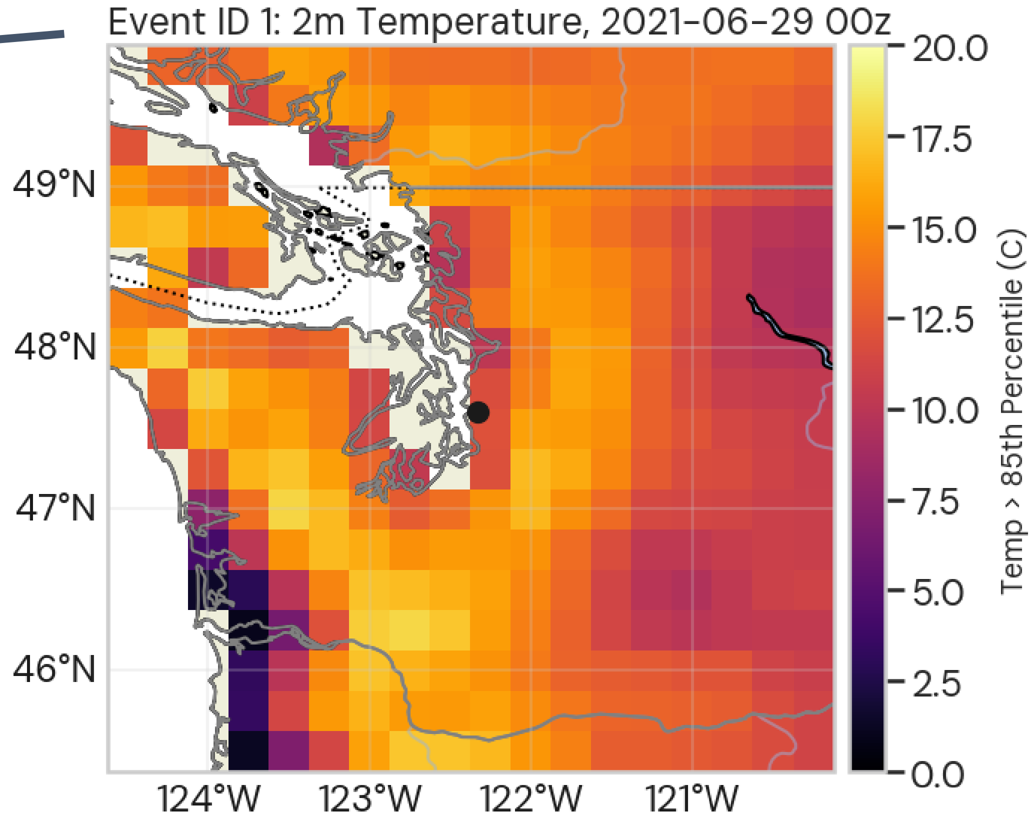
EWB observations (heat/cold only)



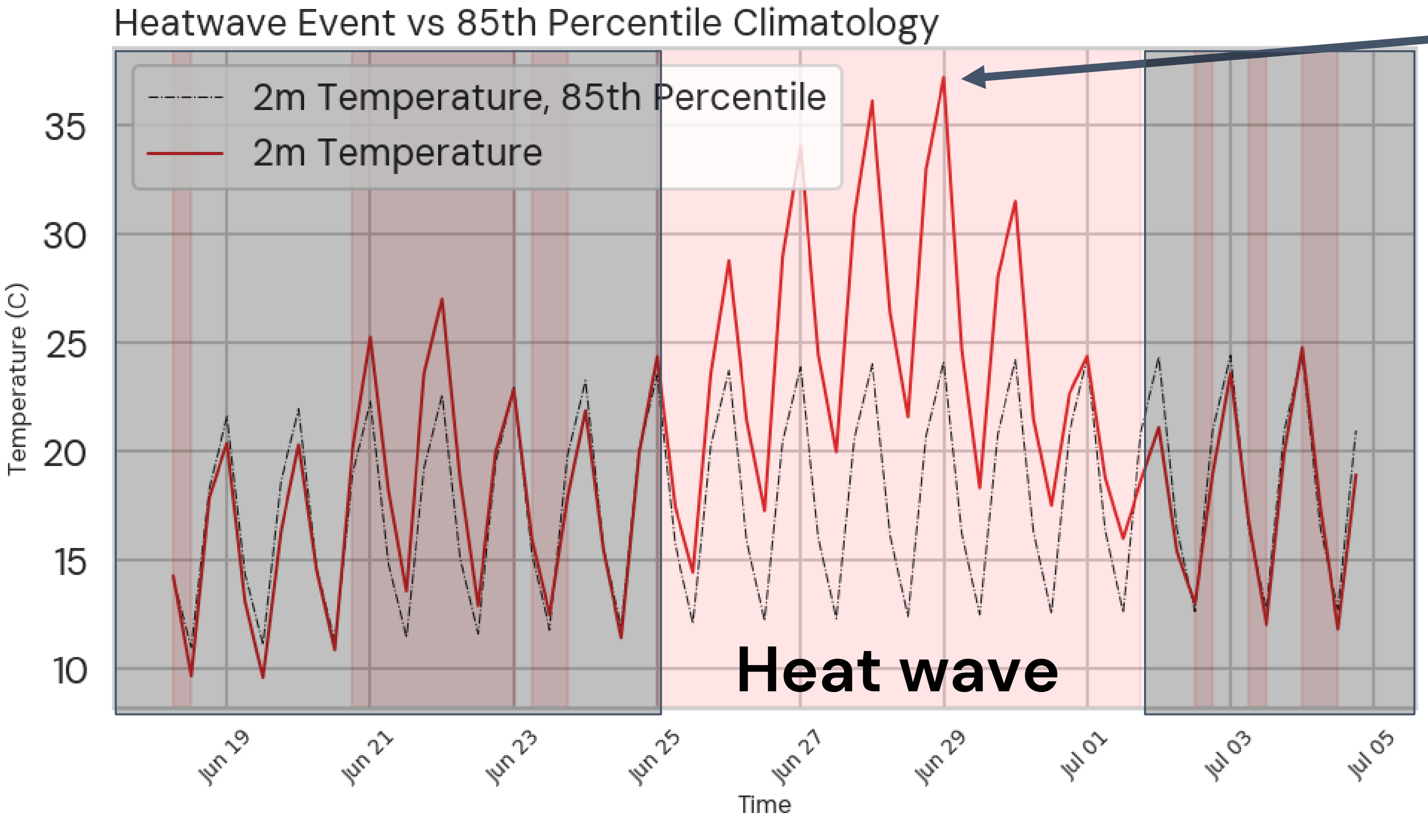
Example: ERA-5 versus point observations

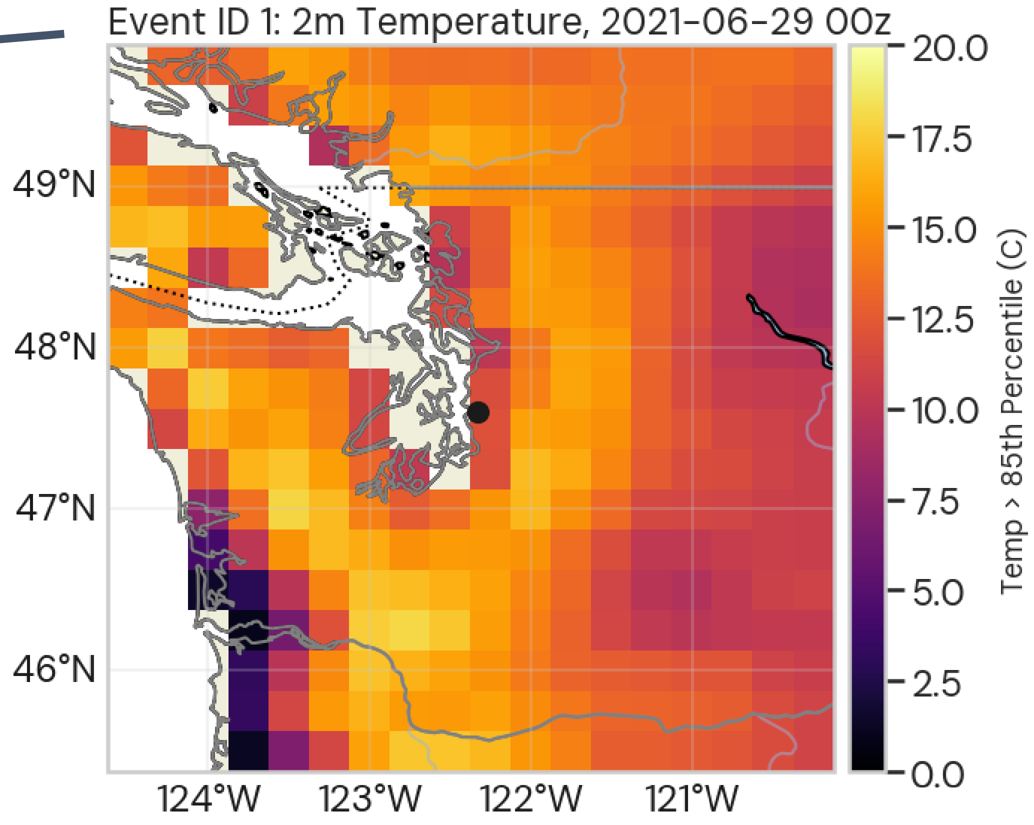
EWB Example: 2021 Pacific Northwest (PNW) Heat Wave



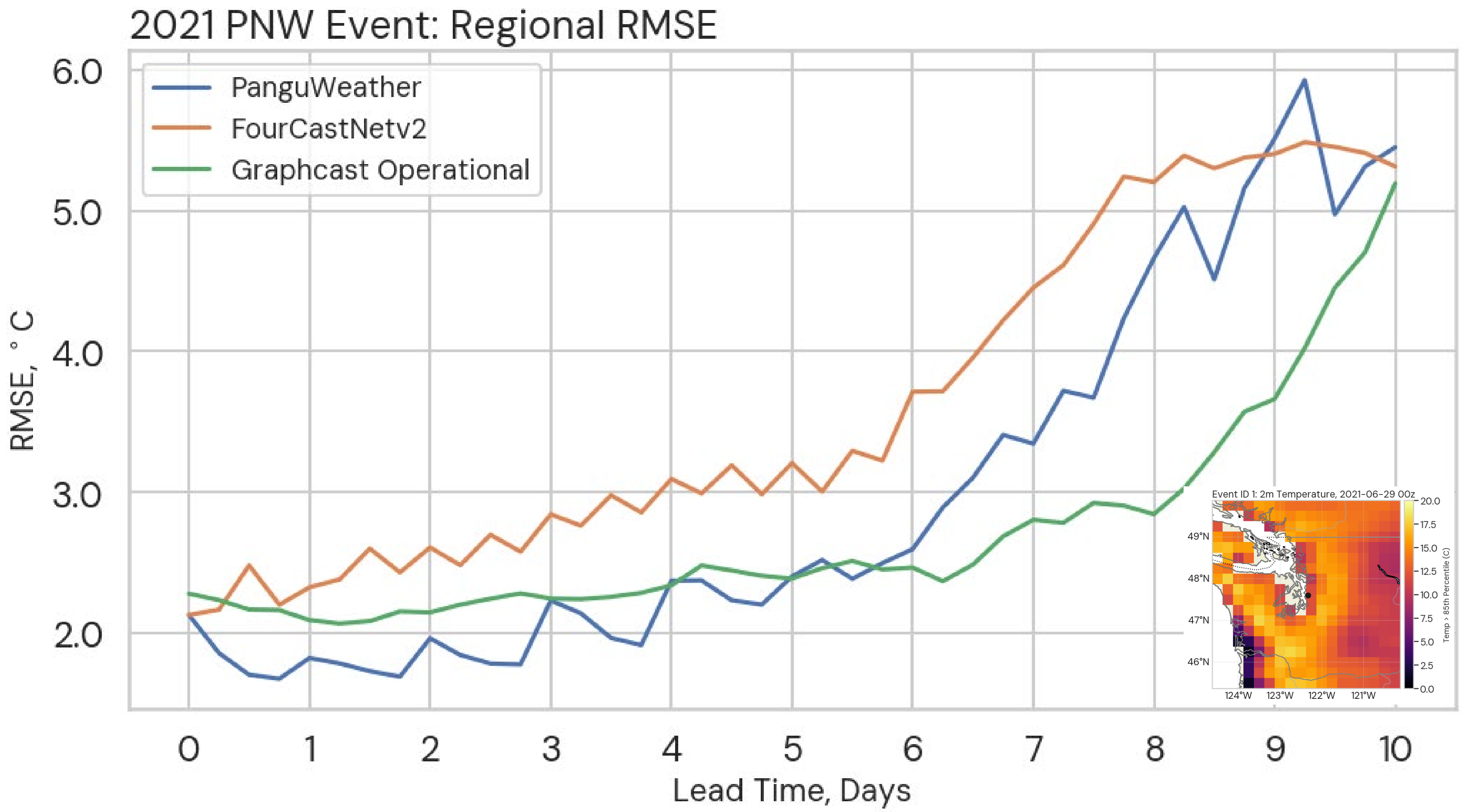


EWB Example: 2021 Pacific Northwest (PNW) Heat Wave

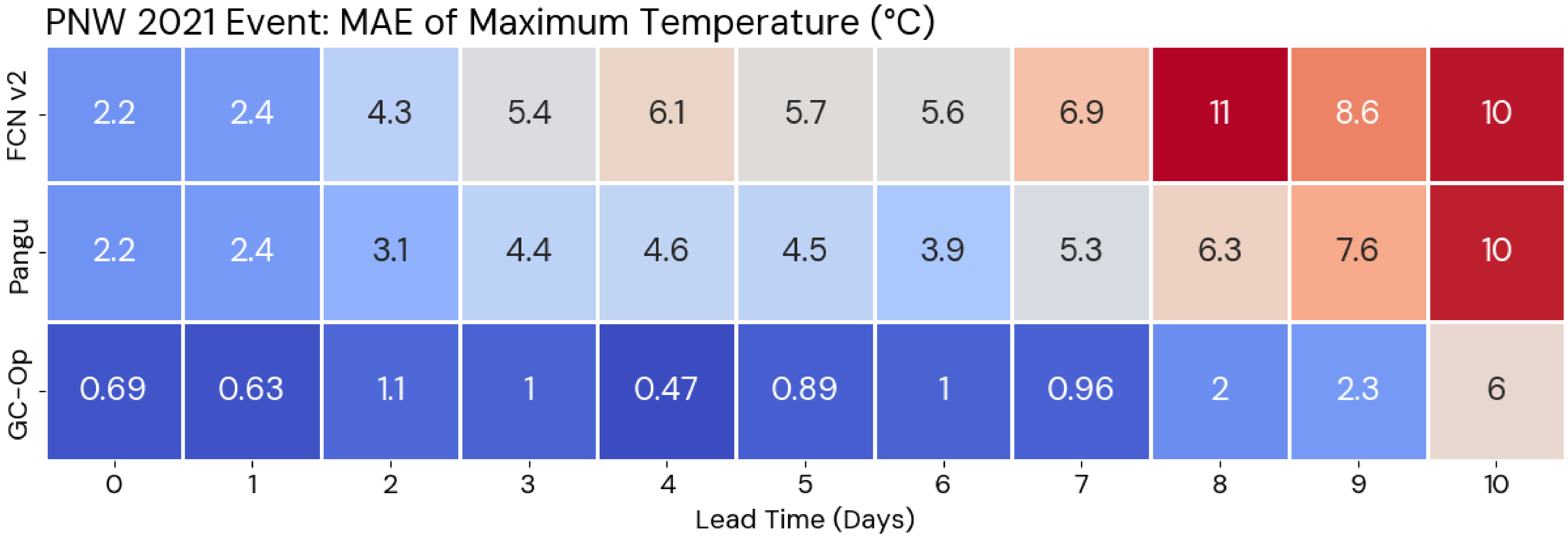




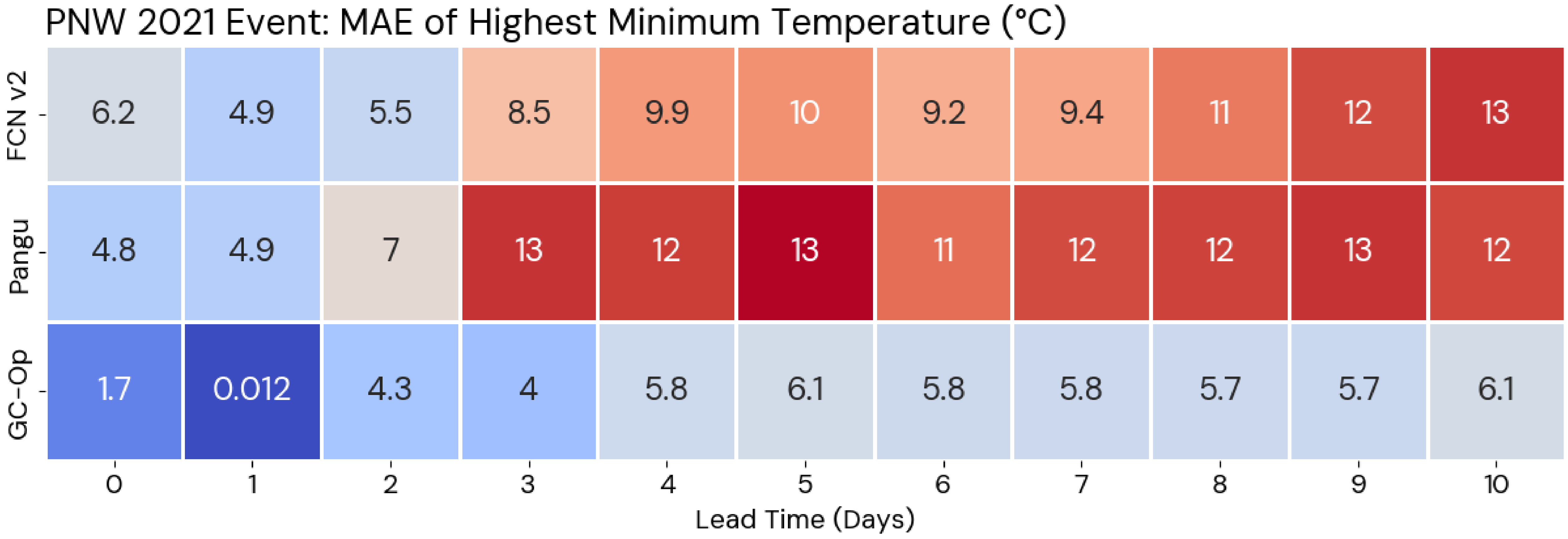
EWB Example: 2021 PNW Heat Wave



EWB Example: 2021 PNW Heat Wave



EWB Example: 2021 PNW Heat Wave



The UNIVERSITY of OKLAHOMA

AI for Weather is Transformational

We need a strong focus on ensuring the AI we create for weather is trustworthy, ethical, and responsible.

This material is based upon work supported by the U.S. National Science Foundation under Grant No. RISE-2019758.

Publication QR codes: