Whose weather is it? Building a fairness framework for global AI weather models

Knowledge gap

- AI weather models display impressive performance across a range of global and regional standard metrics, potentially improving on baseline physical models (e.g. Rasp et al., 2024).
- But, are those improvements fairly distributed across different regions and demographics? For example, do high and low income regions enjoy a similar share of these improvements?

Defining fairness

- We focus on a narrow, outcome-based definition of fairness, following prevailing ML practices (e.g. Mitchell et al., 2021, Mehrabi et al., 2021).
- We define two key criteria, based on the proportion of grid points enjoying improvements:

1. **Group fairness**: Improvements are equally likely across protected and non-protected groups, e.g.,

$$p_{\text{improved}} = p_{\text{improved\_high\_income}} = p_{\text{improved\_low\_income}}$$

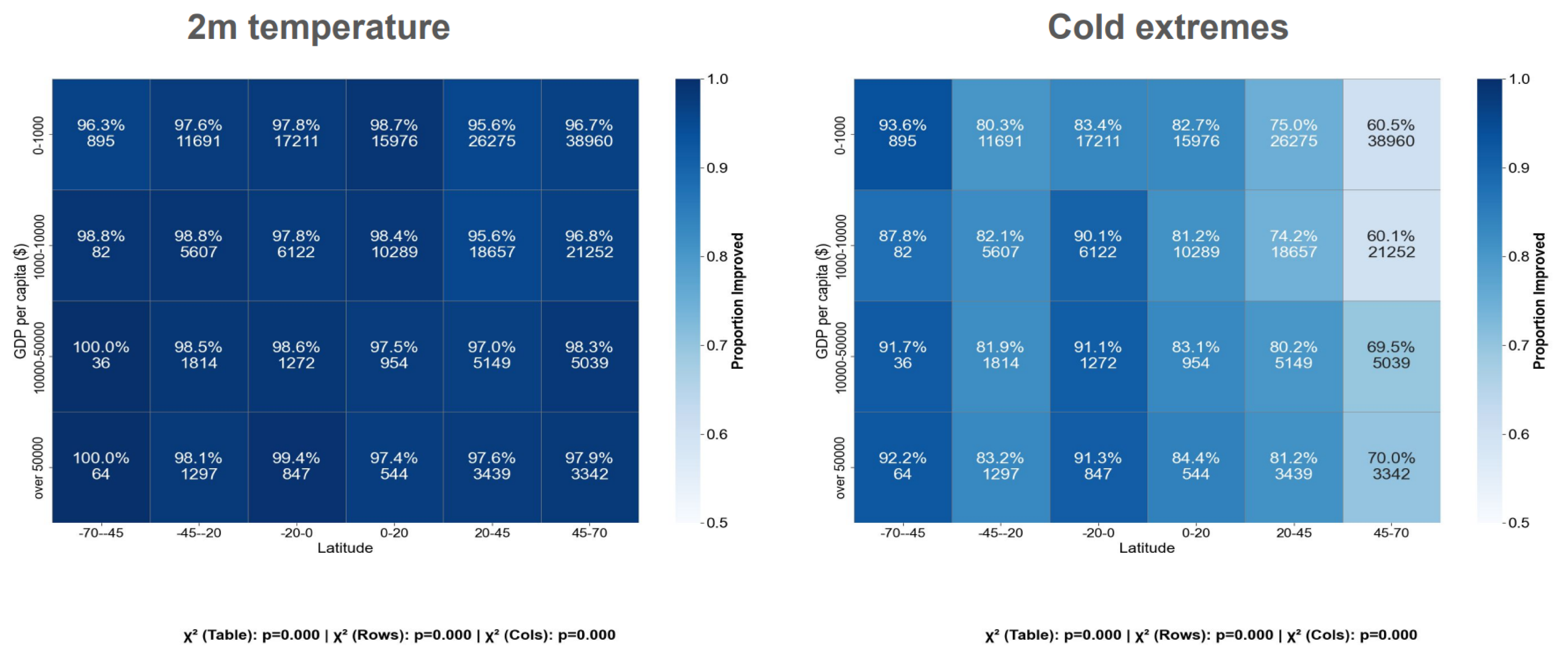
2. **Statistical independence**: Improvements are not predicted by protected attributes, e.g.,

$$\mathbb{E}\left(\Pr(\text{improvement})\right) \perp\!\!\!\perp \text{GDP} \mid Z,$$

where $\perp\!\!\!\perp$ denotes statistical independence, and Z is a set of control variables (e.g., latitude, longitude, elevation).

Criterion 1

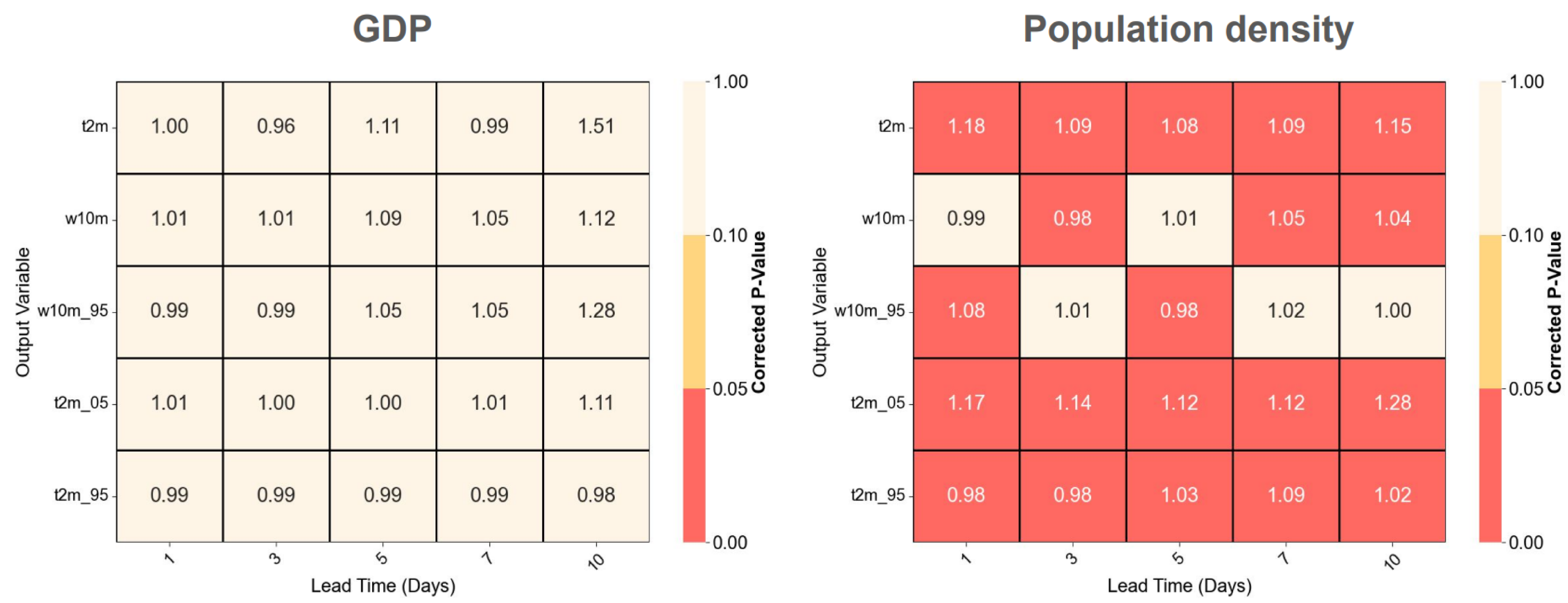
- We compare the performance of ECMWF AIFS to IFS HRES, using ERA 5 as ground truth. Gridded population data from NASA Earth Data and GDP data from Wang and Sun, 2022.
- Does a similar proportion of low, middle and high-income grid points enjoy improved forecasts?



Proportion of grid points with improved forecasts (lower RMSE, AIFS vs HRES), by GDP and latitude

Criterion 2

- Is the probability of improved forecasts at a given grid point independent of GDP and population density?



Standardised effect of GDP and population density on odds of improved forecast for a given output variable, lead time and grid point, estimated through logistic regression. Output varibles: 2m temperature, 10m windspeed, windspeed extremes, cold extremes, hot extremes.

Main conclusions

- AIFS superior to IFS HRES across most regions and demographics - "better forecasts for everyone", but to different extents.
- Neither fairness criterion is fully satisfied. On average, AIFS works best in densely populated areas with high income.
- Much left to be investigated - both for AIFS and other AI models!

Possible solutions?

- Embed fairness criteria in the loss function (fairness through awareness):
  - Weighting schemes to compensate currently disadvantaged grid points.
  - Penalty terms discouraging uneven performance.
- Resources and skill transfers are also a possibility.
- Defining fairness criteria and monitoring their fulfilment are key initial steps.

Leonardo Olivetti[1,2,3], **Gabriele Messori**[1,2,3,4,5]

[1] Dept. of Earth Sciences, Uppsala University; [2] CNDS; [3] CLIMES; [4] Stockholm University; [5] Bolin Centre for Climate Research

Correspondence: Leonardo Olivetti, leonardo.olivetti@geo.uu.se

https://cenae-erc.eu/

ERC grant no. 948309

EGU25-11668