SAFE: Benchmarking AI Weather Prediction Fairness with Stratified Assessments of Forecasts over Earth

Nick Masi
Department of Computer Science
Brown University
nicholas_masi@brown.edu

Randall Balestriero
Department of Computer Science
Brown University
randall_balestriero@brown.edu

Abstract

The dominant paradigm in machine learning is to assess model performance based on average loss across all samples in some test set. However, this approach fails to account for the non-uniform patterns of human development and geography that exist across Earth. We introduce Stratified Assessments of Forecasts over Earth (SAFE), a package for elucidating the stratified performance of a set of predictions made over Earth. SAFE integrates various domains of data to perform stratification on different attributes associated with gridpoints: territorial affiliation, global subregion, and gross national income per capita. In this work, we utilize SAFE to benchmark modern artificial intelligence-based weather prediction models, finding that they exhibit disparities in forecasting skill across all of these attributes. We provide a comprehensive benchmark of stratified model forecast fairness at different lead times for the T850 and Z500. To support further work in this direction, the SAFE package is made available at https://github.com/N-Masi/safe.

1 Introduction

Artificial intelligence (AI) weather prediction (AIWP) models, alternatively machine learning weather prediction (MLWP) models or neural weather models (NWM), are becoming increasingly competitive with traditional numerical weather prediction (NWP) models. All of these approaches are typically used in making medium-range weather forecasts (interchangeably, "prediction"). The range of a forecast is determined by its lead time τ. When a weather prediction model is fed the state of variables at time d, its task is to predict the state of those variables (or some subset of them) at time $d + \tau$. There is no consistent definition for medium-range, with the the European Centre for Medium-Range Weather Forecasts (ECMWF) defining it as any prediction made with τ (or $n \times \tau$ if taking an autoregressive rollout of n steps) within 0–15 days [6], while other sources more narrowly define it as 3–7 days [23, 27]. AIWP are seeing increasing adoption in interfaces where they provide these medium-range forecasts, from Google's Weather app [19] to various experimental models at the National Oceanic and Atmospheric Administration (NOAA) [26, 36].

Root mean square error (RMSE) is the preeminent metric used in assessing the quality of AIWP models [29, 32]. The general form of RMSE is shown in Equation 1, where Y is the set of all ground truth variable values that a model is trying to predict, and $\hat{y}$ is the model's prediction for each corresponding $y \in Y$. Every y is the value of some variable (e.g., temperature or wind speed) at some point in time $d \in D$, longitude $i \in I$, latitude $j \in J$, and, for certain atmospheric variables, vertical level $v \in V$.

There are various approaches for how different models handle being able to make predictions at different lead times. The naive approach is to train a model with the ability to predict some fixed $\tau' \in T$, where T is a set of durations. This allows the model to forecast the weather with temporal resolution of τ' (i.e., multiples of τ' after the timestamp of the input variables) through

Preprint.

autoregressive rollout. This is the approach taken by Keisler [14], FourCastNet [24], and the spherical Fourier neural operator (SFNO) [3], all with $\tau' = 6$ hours. Esteves et al. [10] trains with $\tau' \in \{6 \text{ hours}, 3 \text{ days}, 5 \text{ days}\}$ depending on the task. Pangu-Weather [2] trains four different models, each with a different, fixed lead time. This is used in tandem with a greedy algorithm that minimizes the number of autoregressive steps that need to be taken to make a prediction at any given lead time (which must be a multiple of their smallest lead time model). FuXi [8] uses a cascaded set of three different models that cover different ranges of lead times.

$$\sqrt{\frac{\Sigma_{y \in Y}(\hat{y} - y)^2}{|Y|}} \tag{1}$$

The square of RMSE, mean squared error (MSE), frequently referred to as the $L2$ loss, is often used as a training objective. This is the case for Spherical CNN [10] and GenNet [21]. GraphCast [17] and GenCast [28] use weighted MSE loss functions. Keisler takes a weighted sum of MSE values [14]. NeuralGCM [15] has a five-term loss function, each of which is a variation of MSE. FuXi [8] uses the mean absolute error (MAE, the $L1$ counterpart of MSE).

The underlying commonality across all of these functions is that they completely reduce across the spatial dimensions I and J. An issue with spatial averaging as the loss function is the resulting "double penalty" that arises when predictions for high resolution events are even slightly spatially displaced, incurring the penalization for both that faulty prediction and the lack of prediction at the true location [11]. This encourages models to blur their predictions, dropping these highly localized events [17]. However, neglecting to predict these outlier events can have dramatic real-world consequences. For example, improved accuracy of extreme heat predictions has been found to reduce mortality [37]. If it is unknown precisely where models are and are not performing well, then it is impossible to know whether they can be trusted at inference time for a prediction in a given location.

2 Related work

WeatherBench 2 (WB2) [31] is an existing benchmark that assesses the spatially-averaged error of models using weather data from ERA5, ECMWF's most modern eanalysis dataset [12]. It provides functionality to get per-region RMSE, but these regions are limited to being rectangular in shape, making them unusable for the real-world attributes we care about.

Stable equitable error in probability space (SEEPS) [33] is a metric that was introduced to assess the quality of precipitation forecasts in particular. In the original paper [33], the authors perform region-specific analysis of forecasts in South America, Europe, and the extropics. Again, however, the region shapes are defined with crude, rectangular boundaries ([70°W–35°W, 40°S–10°N], [12.5°W–42.5°E, 35°N–75°N], and [above 30°N or below 30°S], respectively).

NeuralGCM also calculated per-region RMSE for T850 and Z500 [17, Supp. Mat. Fig. S14–S16], borrowing region definitions from ECMWF scorecards. There are 20 of these regions, 3 that are hemispheric (North, Tropical, and Southern) and 17 geographic. These regions are overlapping and include oceans, but the geographic regions miss considerable sections of populated landmass (including but not limited to significant portions of Central America, Eastern Africa, Brazil, California, and the island of New Guinea). The hemispheric regions cover the whole globe, with the Tropical region bounded by the $\pm 20°$ latitude lines.

In contrast, the regions used within SAFE cover all landmass (including islands) across the Earth and no oceanic landcover. This more aptly captures metrics for where fairness in weather forecasts matters most: the places where people are. Our regions are non-overlapping, except at their borders where gridpoint polygons stretch over the border (this being a result of finite resolution).

3 SAFE

In this paper we create a framework for performing Stratified Assessments of Forecasts over Earth (SAFE). This tool enables stratification by various geographically-related attributes, allowing the user to see the fine-grained quality of a set of predictions when broken down by the different constituent

groups, or strata, of each attribute. We leverage SAFE to benchmark the fairness of existing AIWP models. Despite the life or death impacts of weather forecasts and concrete evidence that existing forecasts provided by the National Weather Service have error rates that vary across the geography of the United States [25], there is little existing work that investigates model error spatially (see: section 2).

3.1 Data sources

Within SAFE, we provide the ability to investigate different attributes: territory, global subregion, and income. The strata within the territory attribute is typically the country which a gridpoint is located within, though there are some sub-national or not universally recognized territories. Territory borders are pulled from the geoBoundaries Global Administrative Database [35]. Global subregions follow the United Nation's classifications over territories [38]. The income stratum of a gridpoint is one of "high income", "upper-middle income", "lower-middle income", or "low-income" as defined by the World Bank's classification for the territory the gridpoint is within [40]; the World Bank uses the gross national income (GNI) per capita of the territory, calculated using the Atlas methodology. The polygons associated with each strata are accessed through the MIT-licensed pygeoboundaries_geolab package [1]. This package is a python wrapper for the geoBoundaries Global Administrative Database [35], which itself is made available under a open license CC-BY 4.0.

3.2 Methods

3.2.1 Stratification

Forecasts made over the Earth are associated with specific (longitude, latitude) coordinates, or "gridpoints" on the Earth. Each pair of coordinates is converted into the polygon that is centered on the gridpoint but which covers all the quadrilateral surface area defined by extending its borders to the midpoint with its neighbors in both the longitude and latitude directions. To unify the coordinate system across all integrated data sources, latitude ranges [-90, 90] with index 0 at -90, and longitude [-180, 180] but with index 0 at 0 and a wraparound from 180 to -180 in the middle. This is because metadata sourced from pygeoboundaries_geolab follows this coordinate system, and it is easiest to bring tabular data into conformance.

The forecasts for a gridpoint's polygon are associated with all of the strata that have any polygon which intersects it. While this will double count some gridpoints towards different strata, measures are taken so that no single gridpoint counts more than once within a given strata. The double counting that does occur is in line with the philosophy of SAFE, as the alternative is that—without high enough resolution—there will be strata for which no data is recorded, rendering them invisible and left out of fairness assessments. In total, there are 230 territory, 23 subregion (see: Appendix B), and 4 income strata. Of the 230 territories, 212 have an associated income strata. 76 are classified as high-income, 57 as upper-middle-income, 45 as lower-middle-income, and 34 as low-income. Subregions vary from having 1 territory (Antarctica) to 25 (Caribbean).

3.2.2 Area weighting

In calculating the loss function for training it is common to weight the (squared if $L2$) difference in variable prediction and ground truth by the area of the gridpoint cell the forecast was made at before averaging. This weight varies with latitude. The reason for latitude weighting is that, when using an equiangular gridding, the gridpoints are closer together near the poles than they are at the equator. This results in a higher density of samples per area at the poles, which left unaccounted for could cause the model to overfit to forecasting polar weather.

Complicating the matter, Earth is an oblate spheroid with an equatorial radius of 6378137m and a slightly smaller polar radius of 6356752m. However, no python library that is known to the authors exists which takes this into account to get the precise surface area of equiangular grids on Earth's surface. The standard solution would be to convert the cells to vector data and get the area of polygons. However, virtually every approach, both training [17, 14, 2, 15, 24, 3] and benchmarking [31], make the simplifying assumption of a perfectly spherical Earth. WB2 takes this approach in computing its metrics as well [31]. As part SAFE, we have provided a utility that can get the surface area of grids

[1] `https://github.com/ibhalin/pygeoboundaries`

of the Earth. We use the equation for getting the surface area of oblate spheroid caps from [5, Eq. 49] which builds on the model developed by [39]. For testing, the total surface area of the Earth was found with the equation for oblate spheroid surface area from [1, p. 131], yielding an approximation of $510065604944206.145\text{m}^2$.

In calculating the RMSE as reported throughout this paper, we use these exact surface areas as weights, but with the important distinction of normalizing them by the mean area. This same normalization is used in WB2 [31] and common in training [24, 3].

3.2.3 Metrics

The main metric utilized in SAFE is the latitude-weighted RMSE, which is averaged temporally by initialization time (the timestamp of the climate variables fed into the model) not lead time (the amount of time into the future for which to forecast the state of climate variables at), and averaged spatially within each strata. Unless otherwise specified, reported RMSE refers to this. The anomaly correlation coefficient (ACC) is another evaluation metric that is often used for cross-model comparison. Like RMSE, ACC is spatially averaged [7] and would thus benefit from stratified assessment. The fact that the most popular metrics employ spatial averaging underscores the need for SAFE. We emphasize RMSE in this work under the same rationale as taken by WeatherBench: the similarity between RMSE and the models' training objectives [32]. In this work we focus on benchmarking deterministic models. Probabilistic, or ensemble, AIWP models have other metrics that can be used such as the continuous ranked probability score (CRPS), but also are commonly evaluated on the RMSE of the ensemble's average prediction.

4 Using SAFE to benchmark AIWP forecast fairness

To minimize computational costs, we investigate models with already available predictions. This eliminates the need for model training or inference, reducing the carbon footprint of our research. WB2 provides easily-accessible cloud datasets of ERA5 data and inference runs in the year 2020 for a number of models. Because of the unified access endpoints and resolution, we use the models available through these datasets to begin our investigation. Furthermore, these models are among the most state of the art (by standard metrics such as RMSE and ACC) [30], so it is in fact preferable to study these than retrain our own, potentially inferior models.

4.1 Forecasts assessed

In this work with utilize WB2's 1.5° resolution equiangular predictions on ERA5. Higher resolution forecasts would permit more fine-grained stratification and remediate the double-counting issue discussed in subsubsection 3.2.1. Indeed, WB2 provides higher resolution than this for some of the models in its zoo. However, benchmarking models against one another is only meaningful when performed at the same resolution. Without this, predictions made at higher resolutions may not get assigned to the same strata. We choose the 1.5° resolution (240×121 in terms of longitude by latitude) because it has the most amount of models with provided forecasts at a single common resolution. The forecasts provided are made on ERA5 data from 2020. WB2 retrieved this subset of ERA5 data from ECMWF via the Copernicus Climate Data Store, which makes its products available through an open license.[2] WB2 itself is available through an Apache License 2.0.

The models whose predictions are assessed are listed in Table 1. All of the assessed models were trained on ERA5 data, making it an appropriate common benchmark, and none of them included 2020 in their training set. The set of lead times τ that is common to the provided predictions for all models is every 12 hours up to 10 days, so we assess all models at each of these.

4.2 Variables

In line with WeatherBench [32, 31], we choose as our variables y the atmospheric temperature at 850hPa (T850, unit: m) and geopotential at 500hPa (Z500, unit: m^2s^{-2}) as the main benchmark variables for comparing cross-model performance. Geopotential is the strength of Earth's gravitational field, so predicting the geopotential at a fixed atmospheric pressure level (500hPa) amounts to

[2] `https://apps.ecmwf.int/datasets/licences/copernicus/`

Table 1: Models assessed

Model	Architecture	Parameters
GraphCast [17]	Graph neural network (GNN)	36.7 M
Keisler [14]	GNN	6.7 M
Pangu-Weather [2]	Earth-specific transformer	256 M
Spherical CNN [10]	Spherical convolutional neural network (CNN)	Not reported
FuXi [8]	SwinV2 [20] transformer blocks in U-net [34] arrangement	Not reported
NeuralGCM [15]	Multi-layer perceptrons (MLPs) + CNNs + numerical solver	31.1 M

Table 2: Benchmark of models on the **territory** attribute. Each cell contains the greatest absolute difference in per-strata RMSE of T850 or Z500 for a given model at a given lead time (rounded to the nearest ten-thousandth); lower is better. The RMSE difference for the best model per lead time per variable is in bold.

		Model					
Variable	Lead time (h)	GraphCast	Keisler	Pangu-Weather	Spherical CNN	FuXi	NeuralGCM
T850	12h	0.5301	0.8523	0.5677	0.6726	0.5548	**0.4715**
T850	72h	1.0666	1.2268	1.1138	1.1681	1.1301	**1.0552**
T850	240h	4.5130	4.7116	4.4912	4.9728	**3.9086**	4.6413
Z500	12h	**13.4222**	31.7980	17.1554	23.3231	15.6101	17.7155
Z500	72h	149.7483	192.5330	150.8271	199.7432	150.0260	**136.3233**
Z500	240h	1273.6925	1367.5504	1354.0935	1285.5398	**1044.5271**	1315.3407

predicting the vertical synoptic-scale distribution of pressure in Earth's atmosphere. This knowledge is highly useful in meteorological predictions [17]. T850 is chosen because the variable of temperature and pressure level closer to the surface make it more tangibly impactful [32].

These particular variables are often used by model developers by default in reporting on their work. Keisler [14], Pangu-Weather [2], Spherical CNN [10], NeuralGCM [15], and FourCastNet [24] report RMSE on these two variables in particular, FuXi [8] includes them among other variables, and GraphCast [17] primarily reports on Z500.

4.3 Experimental design

The main data we collect in our experiments is that we get the area-weighted squared difference between the models prediction $\hat{y}$ and the ERA5 ground truth value y at every individual gridpoint, for every lead time $\tau \in \{12\text{h}, 24, ..., 240\text{h}\}$, at every 12 hour interval in 2020. For each τ, we first get the RMSE by averaging over the spatial and temporal (but not lead time) dimensions. This serves as a baseline, and is the RMSE that would often get reported in the weather forecasting literature.

Then, for each of our three attributes and both variables, we calculate the per-strata RMSE (averaged temporally over the whole year) by taking the RMSE when spatially averaging over only the gridpoints within that strata. This allows us to see which stratum the models are performing worst at.

Lastly, for each attribute and variable, we take the greatest absolute difference in per-strata RMSE of any pair of per-strata RMSE with the same attribute and variable. This allows us to quantify the fairness of a model's predictions, where the smaller this difference is, the more fair it is.

Calculating these metrics for all six models took under 8 hours on a single CPU. At least 16GB of free storage is necessary to store intermediate data.

4.4 Results

General fairness. As seen in Figure 1, the fairness of predictions begin to rapidly decline once the lead time gets to around 2 days. That is, the greatest absolute difference in RMSE of any two strata rapidly increases. Across all three attributes and all lead times, Spherical CNN and Keisler are generally the least fair. From a lead time of about a week onwards, FuXi is drastically more fair than every other model across all attributes. At early lead times, GraphCast and NeuralGCM appear to perform most faily. We provide a comprehensive benchmark of the model fairness results on the territory (Table 2), subregion (Table 3), and income (Table 4) attributes.

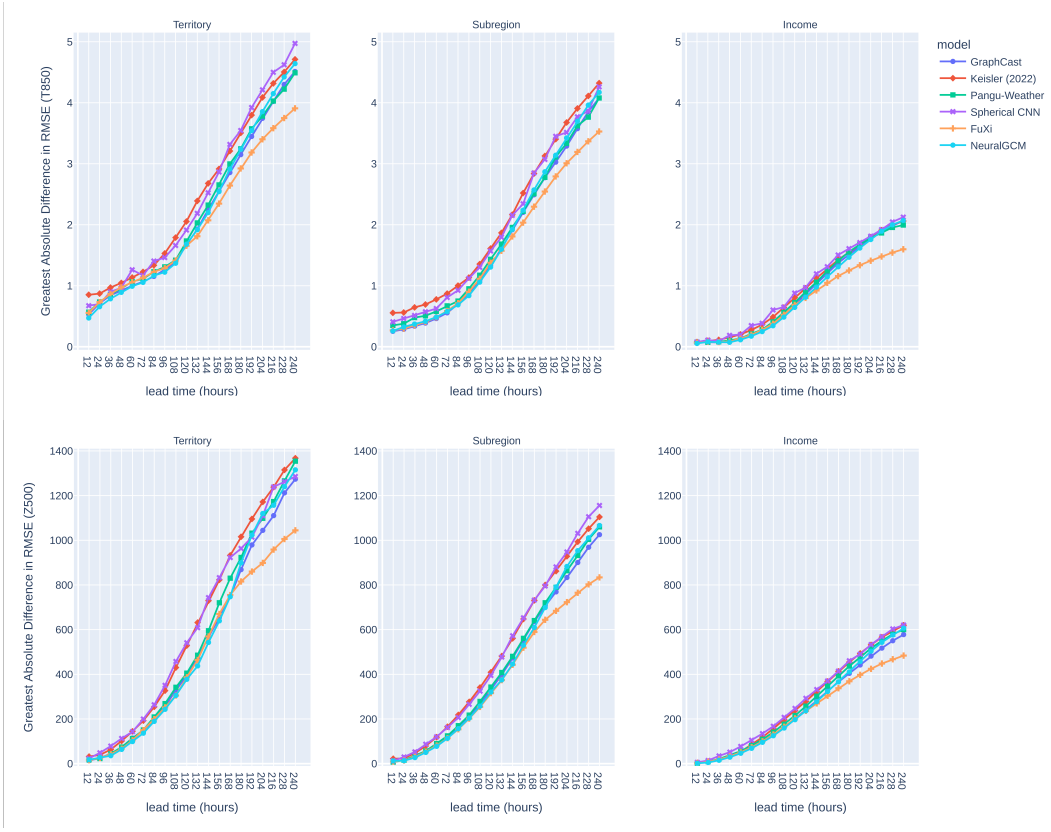

Figure 1: Greatest absolute difference of any two per-strata RMSE for each attribute when predicting T850 and Z500 at different lead times. Lower difference is more fair. Starting at a lead time of one week, FuXi is the most fair model across all attributes and variables.

Table 3: Benchmark of models on the **subregion** attribute. Each cell contains the greatest absolute difference in per-strata RMSE of T850 or Z500 for a given model at a given lead time (rounded to the nearest ten-thousandth); lower is better. The RMSE difference for the best model per lead time per variable is in bold.

		Model					
Variable	Lead time (h)	GraphCast	Keisler	Pangu-Weather	Spherical CNN	FuXi	NeuralGCM
T850	12h	**0.2525**	0.5555	0.3504	0.4085	0.2690	0.2599
T850	72h	**0.5562**	0.8714	0.6702	0.8082	0.5863	0.5778
T850	240h	4.0787	4.3223	4.0785	4.2605	**3.5287**	4.1710
Z500	12h	10.4583	22.0202	**7.0142**	13.0860	9.6233	12.3408
Z500	72h	119.6162	165.2460	124.3231	161.8944	114.0495	**113.2731**
Z500	240h	1025.1756	1104.4001	1060.0999	1155.7330	**833.8870**	1066.3462

Table 4: Benchmark of models on the **income** attribute. Each cell contains the greatest absolute difference in per-strata RMSE of T850 or Z500 for a given model at a given lead time (rounded to the nearest ten-thousandth); lower is better. The RMSE difference for the best model per lead time per variable is in bold.

		Model					
Variable	Lead time (h)	GraphCast	Keisler	Pangu-Weather	Spherical CNN	FuXi	NeuralGCM
T850	12h	0.0620	0.0778	0.0751	0.0774	0.0642	**0.0542**
T850	72h	0.1976	0.2746	0.2127	0.3468	0.1956	**0.1735**
T850	240h	2.0647	2.0616	1.9952	2.1247	**1.5983**	2.0702
Z500	12h	**0.8108**	3.6642	1.6727	5.9048	1.5137	1.6832
Z500	72h	74.6146	84.2772	80.9830	104.9152	73.3257	**68.6393**
Z500	240h	577.7541	619.7738	600.2285	620.6610	**483.7225**	606.3814

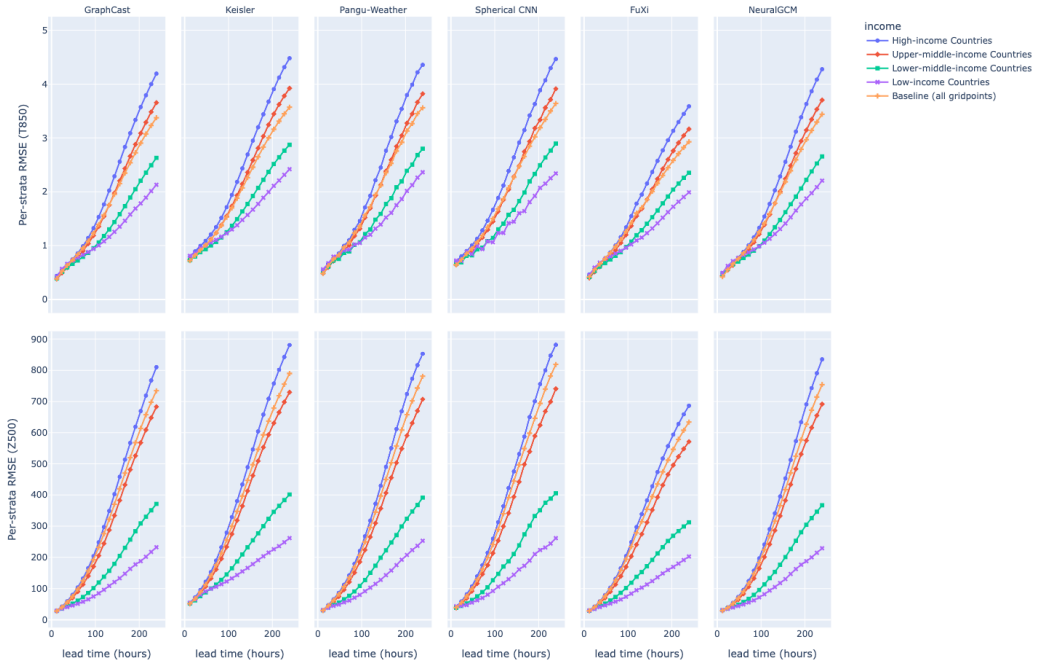

Figure 2: Per-strata RMSE for each of the four income strata across all models and for both variables. Lead time of 12 to 240 hours.

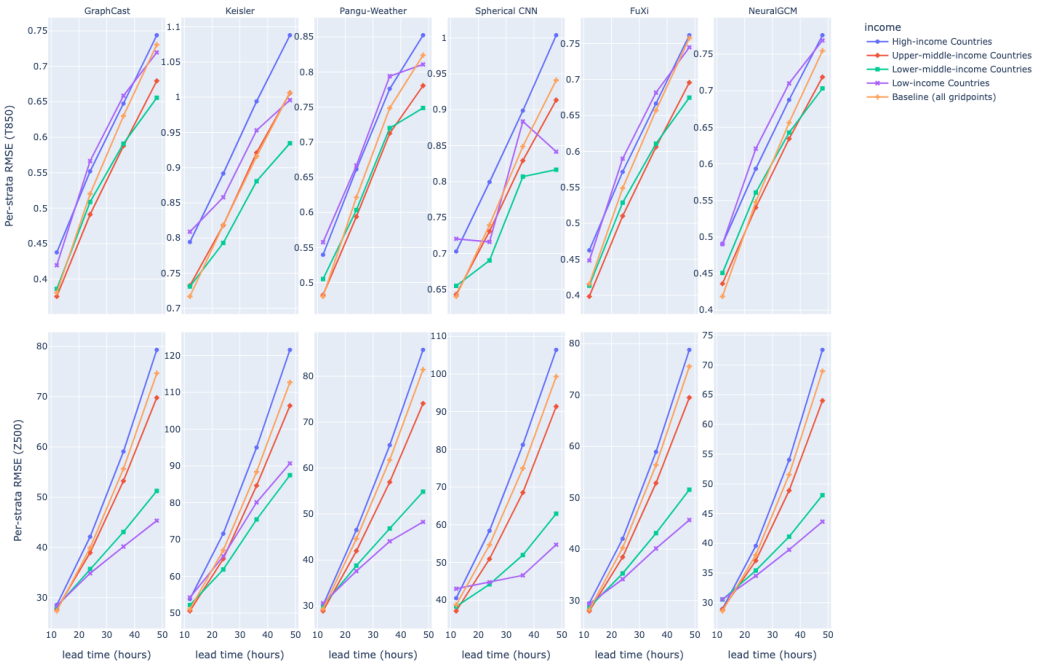

Figure 3: Per-strata RMSE for each of the four income strata across all models and for both variables. Lead time of 12 to 48 hours.

Income attribute. To qualitatively characterize the growing unfairness observed in Figure 1, we take a detailed look at the income attribute. Because it has the least strata, it is easiest to visualize and meaningfully explore. For lead time $\tau = 12$ hours, Keisler, Pangu-Weather, Spherical CNN, and NeuralGCM perform worst at predicting low-income gridpoints for both variables (Figure 3). However, by $\tau = 48$ hours, every model displays the trend for both variables where prediction skill decreases as income increases; this disparity continues to grow with lead time (Figure 2). This is an interesting result, and it shows that lead time is an important dimension to consider, because the disparity observed at one fixed lead time may not hold at another.

4.5 Accounting for outliers

For each model we have assessed, the greatest absolute difference in RMSE for each variable decreases as the number of stratum for each attribute (see: subsubsection 3.2.1) decreases. It is possible that the unfairness phenomenon observed results from rare outliers that appear as the geographic area of the smallest stratum decreases. To account for this, we filtered the list per-strata RMSE for every attribute and removed those with an absolute Z-score greater than or equal to 2. The same figure as Figure 1 was generated except with these filtered values (i.e., excluding outliers), it is in Figure 4. To more easily compare the results when both including and excluding outliers, we graph the largest per-strata RMSE as a percent of the smallest per-strata RMSE in Figure 5. While there is slight differences in the greatest absolute difference in RMSE (as evidenced by the different percentages), the general shape of the curves as a function of lead time holds, while the amplitude has slight differences. This indicates there are consistent trends in unfairness that persist when removing outliers.

This approach in accounting for outliers supports the finding that true disparaties exist across strata. However, we discourage the use of this method in beyond this. Discovering and highlighting disparately treated geographic outliers is the entire aim of this work. To the extent that anyone deserves and benefits from accurate AIWP models, then regardless of how small in size—within reason that is certainly cleared by 1.5° resolution—or count a region is, it and its inhabitants deserve accurate AIWP models too.

5 Future work

An important future direction of work on improving SAFE is incorporating more attributes. First among these is landcover. Datasets such as LandScan Global [9, 18] can provide gridpoint information with the strata of landmass, ocean, and lake. Work with implicit neural representation (INR) models shows that it is important to further consider coastlines and islands as their own strata as well [4]. Additionally, population density as an attribute should be added to SAFE to understand the degree to which AIWP models can be a trusted decision-making tool across different inhabited regions.

Currently, SAFE only operates at inference time. It may prove beneficial to integrate tracking of fairness metrics into the training regimes of models to understand how different training dynamics affect fairness. In general, investigating the underlying causes for why different models had different attribute fairness results is a worthwhile line of future research that is called for by this work.

6 Limitations

The 1.5° resolution used within this paper was beneficial in providing a common resolution to maximize how many models we could benchmark. However, higher resolution ameliorates the double counting issue described in subsubsection 3.2.1, providing more precise stratification. Important future steps involve reproducing the experiments in this work with model forecasts at higher spatial resolution.

Our metric for fairness, the greatest absolute difference in per-strata RMSE is rudimentary. The overwhelming focus of the machine learning fairness community is focused on metrics that apply to binary outcomes, rather than the continuous value we are tracking, and typically in binary (two strata) settings [13, 22]. This meant there was no standard approach for us to take in quantifying fairness as a measure of a continuous outcome that differs across multiple strata per attribute.

7 Conclusion

In this work we created SAFE, a python package that allows the user to assess a set of machine learning predictions made over Earth in terms of stratified fairness. Strata are available for three attributes a gridpoint may have: territorial affiliation, global subregion, and gross national income per capita. This provides developers and decision-makers alike with an important tool to break free from the default approach of spatially averaging. We apply SAFE to a set of state of the art [30] AIWP models, finding that they all display unfair differences in performance across all three attributes. These disparities increase with lead time, particularly starting around 48 hours. These findings justify the approach of capturing more geographically fine-tuned errors and avoiding mere reliance on spatially-averaged RMSE for characterizing AIWP models. We provide a benchmark of current models and their stratified forecast fairness.

Through investigating the income attribute specifically, we found that there are systemic inequalities in AIWP forecast skill across different strata. Notably, models generally perform worse on low-income territories at the shortest lead times, followed by this trend completely reversing at around 2 days, where model forecasts display worse performance as income strata increases.

Organizations like the NOAA are beginning to incorporate ML systems in their work, citing improvements in models such as ECMWF's very own Artificial Intelligence/Integrated Forecasting System (AIFS) [16]. As AIWP models become increasingly relied upon, the results of this work necessitates more careful attention being paid to the stratified performance and fairness of models. By using SAFE to investigate the territory attribute, one is able to find whether a given AIWP is appropriate to leverage in decision making within that territory. This is an important discovery given the life and death consequences that forecasts can impart. The benchmark provided in this work is a first step in this direction. Overall, SAFE empowers deployers to select the model which is most performant for their local application. The visibility provided by SAFE into stratified forecast fairness encourages future development in this direction.

Acknowledgments and Disclosure of Funding

The authors thank Daniel Cai and Philip LaDuca for providing insightful discussion. Part of this research was conducted using computational resources and services at the Center for Computation and Visualization, Brown University.

References

[1] William H Beyer. *Handbook of Mathematical Science*. 6th ed. CRC press, 1987.

[2] Kaifeng Bi et al. "Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast". In: *arXiv preprint arXiv:2211.02556* (2022).

[3] Boris Bonev et al. "Spherical fourier neural operators: Learning stable dynamics on the sphere". In: *International conference on machine learning*. PMLR. 2023, pp. 2806–2823.

[4] Daniel Cai and Randall Balestriero. "No Location Left Behind: Measuring and Improving the Fairness of Implicit Representations for Earth Data". In: *arXiv preprint arXiv:2502.06831* (2025).

[5] Alfredo Calvimontes. "The measurement of the surface energy of solids by sessile drop accelerometry". In: *Microgravity Science and Technology* 30 (2018), pp. 277–293.

[6] European Centre for Medium-Range Weather Forecasts. *Medium-range forecasts*. URL: https://www.ecmwf.int/en/forecasts/documentation-and-support/medium-range-forecasts.

[7] European Centre for Medium-Range Weather Forecasts. *Section 12.A Statistical Concepts: Deterministic Data — Forecast User Guide*. URL: https://confluence.ecmwf.int/display/FUG/Section+12.A+Statistical+Concepts+-+Deterministic+Data#Section12.AStatisticalConceptsDeterministicData-MeasureofSkill-theAnomalyCorrelationCoefficient(ACC) (visited on 05/16/2025).

[8] Lei Chen et al. "FuXi: A cascade machine learning forecasting system for 15-day global weather forecast". In: *npj climate and atmospheric science* 6.1 (2023), p. 190.

[9] Jerome E Dobson et al. “LandScan: a global population database for estimating populations at risk”. In: *Photogrammetric engineering and remote sensing* 66.7 (2000), pp. 849–857.

[10] Carlos Esteves, Jean-Jacques Slotine, and Ameesh Makadia. “Scaling Spherical CNNs”. In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas Krause et al. Vol. 202. Proceedings of Machine Learning Research. PMLR, 23–29 Jul 2023, pp. 9396–9411. URL: https://proceedings.mlr.press/v202/esteves23a.html.

[11] Eric Gilleland et al. “Intercomparison of Spatial Forecast Verification Methods”. In: *Weather and Forecasting* 24.5 (Oct. 2009). Publisher: American Meteorological Society Section: Weather and Forecasting, pp. 1416–1430. ISSN: 1520-0434, 0882-8156. DOI: 10.1175/2009WAF2222269.1. (Visited on 02/16/2025).

[12] Hans Hersbach et al. “The ERA5 global reanalysis”. In: *Quarterly journal of the royal meteorological society* 146.730 (2020), pp. 1999–2049.

[13] Tonni Das Jui and Pablo Rivas. “Fairness issues, current approaches, and challenges in machine learning models”. In: *International Journal of Machine Learning and Cybernetics* 15.8 (2024), pp. 3095–3125.

[14] Ryan Keisler. “Forecasting global weather with graph neural networks”. In: *arXiv preprint arXiv:2202.07575* (2022).

[15] Dmitrii Kochkov et al. “Neural general circulation models for weather and climate”. In: *Nature* 632.8027 (Aug. 2024), pp. 1060–1066. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-024-07744-y. URL: https://www.nature.com/articles/s41586-024-07744-y (visited on 02/16/2025).

[16] Frank Konkel. *Cloud and AI are ‘fundamentally changing’ ability to forecast weather, NOAA chief says*. NextGov. 2024. URL: https://www.nextgov.com/digital-government/2024/12/cloud-and-ai-are-fundamentally-changing-ability-forecast-weather-noaa-chief-says/401453/.

[17] Remi Lam et al. “Learning skillful medium-range global weather forecasting”. In: *Science* 382.6677 (2023), pp. 1416–1421.

[18] V Lebakula et al. “LandScan Silver Edition”. In: *Oak Ridge National Laboratory* (2024).

[19] Lauren Leffer. *AI Weather Forecasting Can’t Replace Humans—Yet*. Ed. by Andrea Thompson. Scientific American. 2024. URL: https://www.scientificamerican.com/article/ai-weather-forecasting-cant-replace-humans-yet/.

[20] Ze Liu et al. “Swin transformer v2: Scaling up capacity and resolution”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 12009–12019.

[21] Ignacio Lopez-Gomez et al. “Global extreme heat forecasting using neural weather models”. In: *Artificial Intelligence for the Earth Systems* 2.1 (2023), e220035.

[22] Ninareh Mehrabi et al. “A survey on bias and fairness in machine learning”. In: *ACM computing surveys (CSUR)* 54.6 (2021), pp. 1–35.

[23] American Meteorological Society. *Glossary of Meteorology*. URL: https://glossary.ametsoc.org/wiki/Medium-range_forecast (visited on 05/15/2025).

[24] Jaideep Pathak et al. “Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators”. In: *arXiv preprint arXiv:2202.11214* (2022).

[25] Washington Post. “We mapped weather forecast accuracy across the U.S. Look up your city". 2024. URL: https://www.washingtonpost.com/climate-environment/interactive/2024/how-accurate-is-the-weather-forecast/ (visited on 02/16/2025).

[26] Corey Potvin et al. *WoFSCast: A GraphCast-based emulator for the Warn-on-Forecast System*. National Severe Storms Laboratory. 2025. URL: https://epic.noaa.gov/wofscast-a-graphcast-based-emulator-for-the-warn-on-forecast-system/.

[27] Weather Prediction Center. *WPC Medium Range Forecasts (Days 3-7)*. National Weather Service. URL: https://www.wpc.ncep.noaa.gov/medr/medr.shtml (visited on 05/15/2025).

[28] Ilan Price et al. “Gencast: Diffusion-based ensemble forecasting for medium-range weather”. In: *arXiv preprint arXiv:2312.15796* (2023).

[29] Jacob T Radford, Imme Ebert-Uphoff, and Jebb Q Stewart. “A comparison of ai weather prediction and numerical weather prediction models for 1–7-day precipitation forecasts”. In: *Weather and Forecasting* 40.4 (2025), pp. 561–575.

[30] Stephan Rasp. *AI-Weather SotA vs Time*. 2024. DOI: https://doi.org/10.6084/m9.figshare.28083515.v1.

[31] Stephan Rasp et al. "Weatherbench 2: A benchmark for the next generation of data-driven global weather models". In: *Journal of Advances in Modeling Earth Systems* 16.6 (2024), e2023MS004019.

[32] Stephan Rasp et al. "WeatherBench: a benchmark data set for data-driven weather forecasting". In: *Journal of Advances in Modeling Earth Systems* 12.11 (2020), e2020MS002203.

[33] Mark J Rodwell et al. "A new equitable score suitable for verifying precipitation in numerical weather prediction". In: *Quarterly Journal of the Royal Meteorological Society* 136.650 (2010), pp. 1344–1363.

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer. 2015, pp. 234–241.

[35] Daniel Runfola et al. "geoBoundaries: A global database of political administrative boundaries". In: *PloS one* 15.4 (2020), e0231866.

[36] Sadegh Sadeghi Tabas et al. "GFS-Powered Machine Learning Weather Prediction: A Comparative Study on Training GraphCast with NOAA's GDAS Data for Global Weather Forecasts". Version 521. In: *NOAA NCEP Office Note* (2025). DOI: 10.25923/xd3y-wy31.

[37] Jeffrey G. Shrader, Laura Bakkensen, and Derek Lemoine. "Fatal Errors: The Mortality Value of Accurate Weather Forecasts". Working Paper. June 2023. DOI: 10.3386/w31361. URL: https://www.nber.org/papers/w31361 (visited on 02/16/2025).

[38] Statistics Division, Department of Economic and Social Affairs, United Nations. *Standard Country or Area Codes for Statistical Use.* Series M, No. 49. Revision 4. 1999. URL: https://unstats.un.org/unsd/publication/SeriesM/Series_M49_Rev4(1999)_en.pdf (visited on 05/14/2025).

[39] Gene Whyman and Edward Bormashenko. "Oblate spheroid model for calculation of the shape and contact angles of heavy droplets". In: *Journal of Colloid and Interface Science* 331.1 (2009), pp. 174–177.

[40] The World Bank. *World Bank Country and Lending Groups*. URL: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups (visited on 05/14/2025).

A LLM usage

LLMs were used in debugging package code. No LLM was used in the writing of this paper.

B Subregion attribute details

The 23 strata included in the global subregion attribute are: Antarctica, Australia/New Zealand, Caribbean, Central America, Central Asia, Eastern Africa, Eastern Asia, Eastern Europe, Melanesia, Micronesia, Middle Africa, Northern Africa, Northern America, Northern Europe, Polynesia, South America, South-Eastern Asia, Southern Africa, Southern Asia, Southern Europe, Western Africa, Western Asia, and Western Europe.

C Auxiliary figures

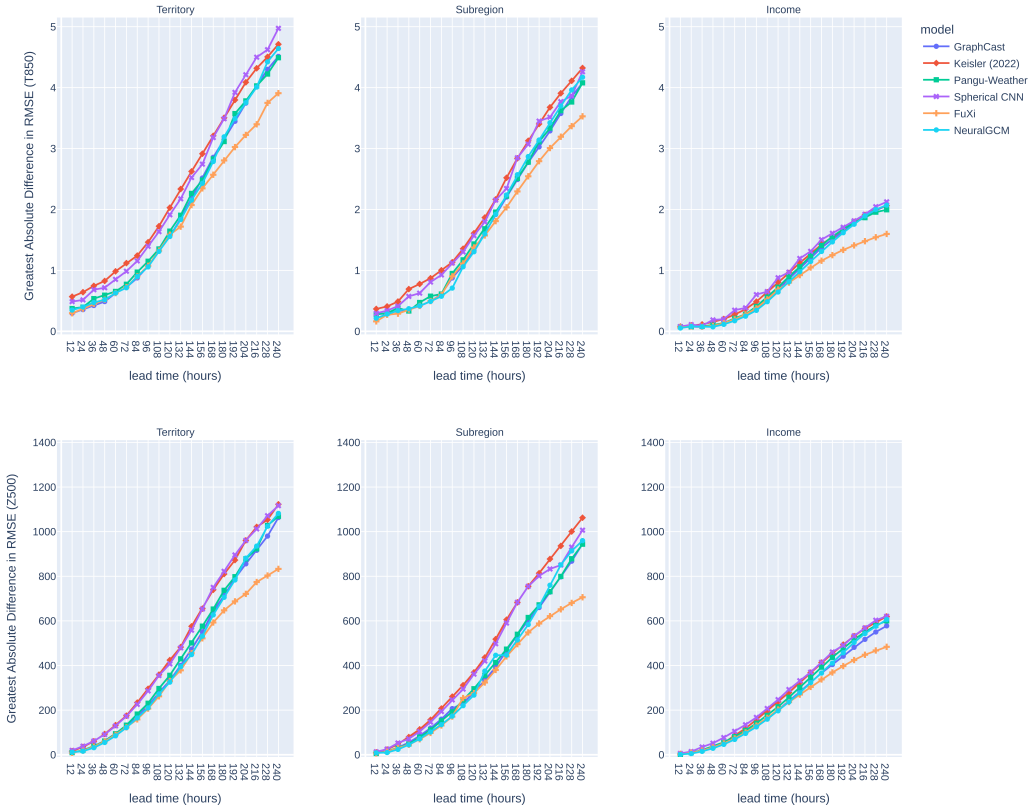

Figure 4: Greatest absolute difference of any two per-strata RMSE for each attribute when predicting T850 and Z500 at different lead times. Lower difference is more fair. Outlier RMSE values have been removed. Starting at a lead time of one week, FuXi is still the most fair model across all attributes and variables.

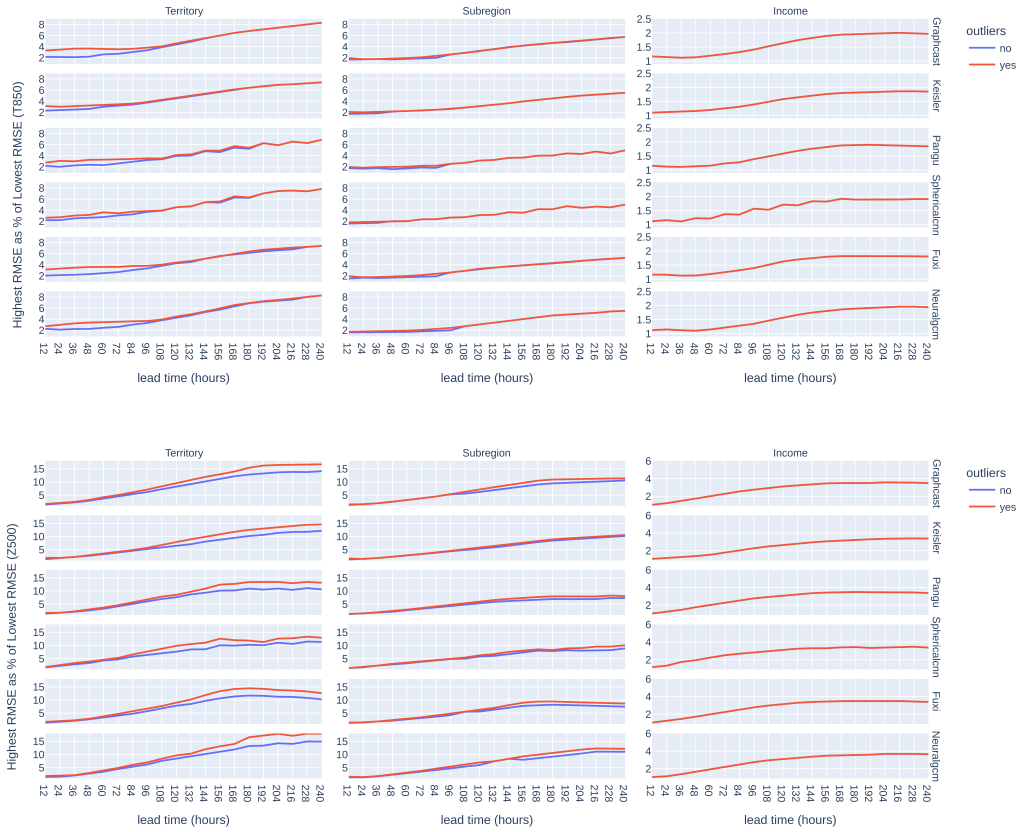

Figure 5: Highest per-strata RMSE as a percent of the lowest per-strata RMSE with and without outliers included.