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 SAFE, a package for elucidating the stratified performance of a set of predictions made over Earth. SAFE integrates data from myriad sources to perform stratification on different attributes associated with gridpoints. In this work, we utilize SAFE to benchmark modern artificial intelligence-based weather prediction models, finding that they exhibit disparities in error rates when stratifying by territorial affiliation, global subregion, and gross national income per capita. Among the tested models, FuXi is consistently the most fair at lead times between a week and 10 days. The SAFE package is made available at https://github.com/N-Masi/safe.

1 Introduction

Artificial intelligence 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. As a result, AIWP are seeing increasing adoption in interfaces such as Google's Weather app [16] to various experimental models at the National Oceanic and Atmospheric Administration (NOAA) [21, 28].

Root mean square error (RMSE) is the preeminent metric used in assessing the quality of AIWP models [23]. The general form of RMSE is shown in Equation 1, where Y is the set of all ground truth 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 $i \in I$, and latitude $j \in J$, and vertical level $v \in V$.

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

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 [8] and GenNet [18]. GraphCast [14] and GenCast [22] use weighted MSE loss functions. Keisler takes a weighted sum of MSE values [11]. NeuralGCM [12] has a five-term loss function, each of which is a variation of MSE. FuXi [6] 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

Preprint. Under review.

displaced, incurring the penalization for both that faulty prediction and the lack of prediction at the true location [9]. This encourages models to blur their predictions, dropping these highly localized events [14]. 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 [29]. 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 SAFE

In this paper we create a framework for stratifying Earth predictions by various attributes, and leverage it to benchmark the fairness of existing artificial intelligence (AI) weather forecasting models. Despite the life or death impacts of weather forecasts and concrete evidence that existing forecasts provided by the National Weather Service have error that varies across the geography of the United States [20], there is little existing work that investigates model error spatially (see: subsection 2.4).

2.1 Data sources

Within SAFE, we provide the ability to investigate different attributes: territory, global subregion, and territory 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 [27]. Global subregions follow the United Nation's classifications over territories [30]. The income strata 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 [32]. 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 [27], which itself is made available under a open license CC-BY 4.0.

2.2 Methods

2.2.1 Stratification

Predictions from WeatherBench 2 [25] 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 neighbords in both the longitude and latitude directions. The forecasts for this polygon are associated with all of the strata it intersects with. 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. We find that the double couting 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. In total, there are 230 territory, 23 subregion, and 4 income strata.

2.2.2 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 [14, 11, 2, 12, 19, 3] and benchmarking [25], make the simplifying assumption of a perfectly spherical Earth. WB2 takes this approach in computing its

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

metrics as well [25]. As part SAFE, we have provided a utility that can get the surface area of grids of the Earth. We use the equation for getting the surface area of oblate spheroid caps from [5] which builds on the model developed by [31]. 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 $510065604944206.145m^2$.

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 taken in WB2 [25] and common in training [19, 3].

2.2.3 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 the strata. Unless otherwise specified, reported RMSE refers to this.

2.3 Fairness definition

For each given weather variable (see: subsubsection 2.3.1) and a given attribute, fairness is defined as the inverse of the greatest absolute difference between the area-weighted RMSE of any two strata.

2.3.1 Variables

In line with WB2 [25] 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. These are also defaults assessment variables for model developers to report on in their work. Keisler [11], Pangu-Weather [2], Spherical CNN [8], and NeuralGCM [12] report RMSE on these two variables in particular, FuXi [6] includes them among other variables, and GraphCast [14] primarily reports on Z500.

2.4 Related work

WB2 [25] is an existing benchmark that assesses the spatially-averaged error of models against ECMWF Reanalysis v5 (ERA5) [10], the most modern reanalysis dataset from the European Centre for Medium-Range Weather Forecasts (ECMWF). 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.

NeuralGCM also calculated per-region RMSE for T850 and Z500 [14, 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 geographical. These regions are overlapping, and the geographic regions include oceans while missing 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).

In contrast, the regions used within SAFE cover all landmass across the Earth and no oceans. Regions are non-overlapping, except at their borders where prediction polygons stretch over the border (this being a result of finite resolution).

3 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.

Table 1: Models assessed

Model	Architecture	Parameters
GraphCast [14]	Graph neural network (GNN)	36.7 M
Keisler [11]	GNN	6.7 M
Pangu-Weather [2]	Earth-specific transformer	256 M
Spherical CNN [8]	Spherical convolutional neural network (CNN)	Not reported
FuXi [6]	SwinV2 [17] transformer blocks in U-net [26] arrangement	Not reported
NeuralGCM [12]	Multi-layer perceptrons (MLPs) + CNNs + numerical solver	31.1 M

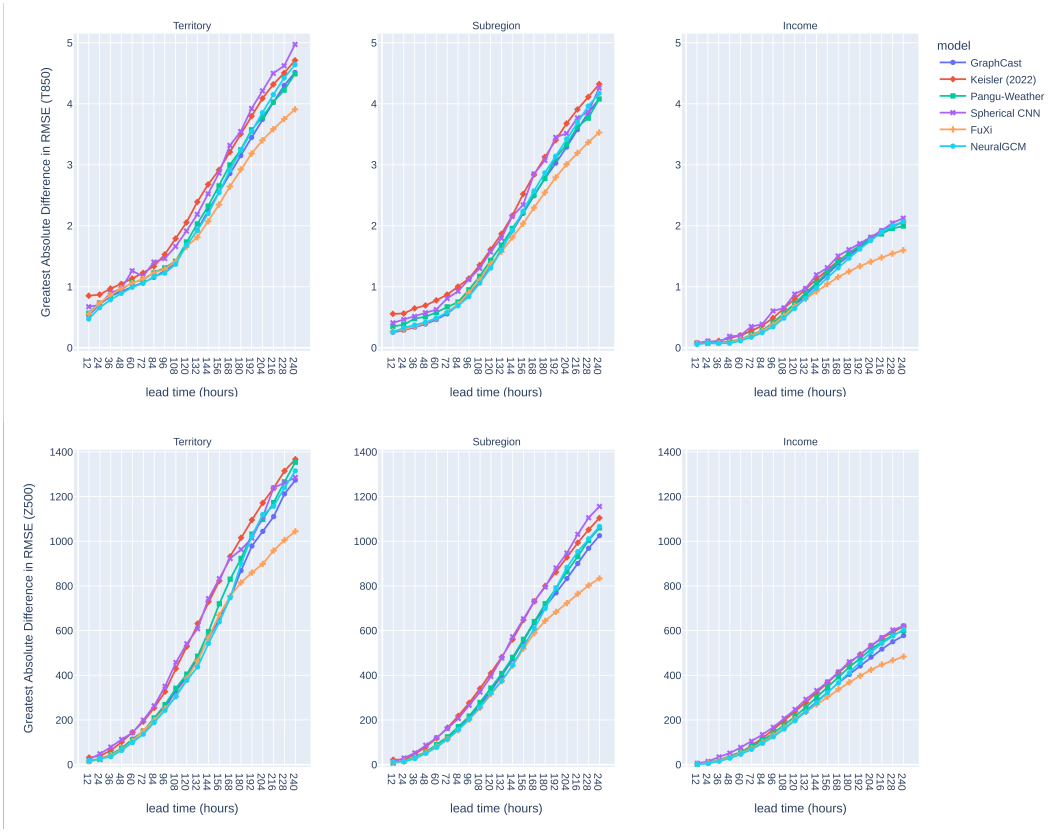

Figure 1: Greatest absolute difference of any two per-strata RMSE for each attribute when predicting T850 and Z500 at different lead times. Lower difference is more fair. Starting at a lead time of one week, FuXi is the most fair model across all attributes and variables.

3.1 Assessed forecasts

In this work with utilize WB2's $1.5°$ resolution equiangular forecasts. Higher resolution forecasts would permit more fine-grained stratification and remediate the double-counting issue discussed in subsubsection 2.2.1. Indeed, WB2 provides higher resolution than this for many of its models. 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 number of longitude by latitude) because it has the most models with predictions available. These models are listed in Table 1. None of the models we assess were trained on ERA5 variable data from 2020. For all the attributes in SAFE, we get the RMSE for every strata and calculate the inverse fairness, where higher is worse. Generating results for all six models took under 8 hours on a single CPU.

3.2 Results

As seen in Figure 1, the fairness of predictions begin to rapidly decline once the lead time gets to around 2 days. 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 all the other models across all attributes.

3.3 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 2.2.1) 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 2. 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 3. 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.

3.4 SAFE utilization beyond weather

While the analyses in this paper have focused on weather forecasting models, another primary contribution is the creation of a general package that allows for benchmarking any model which makes predictions across the globe for which there are ground truth values. That is the collected attribute data can be used to assess the fairness of any set of predictions that are made over the Earth and are associated with a set of gridpoints.

4 Discussion

All of the models evaluated were trained on ERA5, developed by ECMWF. The Member States of ECMWF includes primarily Western, Northern, and Southern Europe territories. Western and Northern Europe are the subregions with the 2 lowest errors across all models (see: Figure 5), and Southern Europe is among the lower half of subregions. Organizations like the NOAA are increasingly utilizing ML systems in their forecasting, citing improvements in models such as ECMWF's very own Artificial Intelligence/Integrated Forecasting System (AIFS) [13]. The results of this work indicate the need for more global efforts in weather data collection and forecasting. This benchmark empowers deployers to select the model which is most performant for their local application. The visibility provided by SAFE into model fairness encourages future development in this direction.

4.1 Future work

Incorporating more attributes within SAFE. First among these is landcover. We assessed the same six models tested in section 3 by getting the average RMSE for landcover strata of land, ocean, and lake areas. Landcover data came from the LandScan Global dataset [7, 15]. This data was available at 0.5° resolution. In this assessment, the predictions of models were treated as points rather than polygons, and so the landcover strata assigned was the landcover of whatever $0.5^\circ \times 0.5^\circ$ cell overlapped with the center of the $1.5^\circ \times 1.5^\circ$ prediction cell. Results are in Figure 4.

This experiment should be reconducted with the majority vote algorithm and using polygons rather than points. The results should then be reduced by model by taking the greatest absolute difference in

per-strata RMSE to find any patterns in unfairness. Additionally, existing work with implicit neural representation (INR) models shows that it is important to consider coastlines and islands and their own strata as well [4].

Next, population density as an attribute should be added to SAFE to understand the degree to which AIWP models should be a trusted decision-making tool for different settled regions. Lastly, SAFE currently operates at the inference-level of AIWP models only. It may prove beneficial to integrate tracking of fairness metrics into the training regimes of models to understand how different training dynamics affect fairness.

5 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 main attributes a gridpoint may have: territorial affiliation, global subregion, and territorial income. 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 [24] 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.

Acknowledgments and Disclosure of Funding

Masi wrote the text of the paper, authored the package code, and performed experiments, collecting and visualizing experimental data. Balestriero gave guidance on research directions and experiments. 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] 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.

[7] 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.

[8] 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.

[9] 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).

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

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

[12] 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).

[13] 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/.

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

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

[16] 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/.

[17] 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.

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

[19] 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).

[20] 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).

[21] 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/.

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

[23] 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.

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

[25] 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.

[26] 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.

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

[28] 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.

[29] 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).

[30] 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).

[31] 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.

[32] 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 Supplemental figures

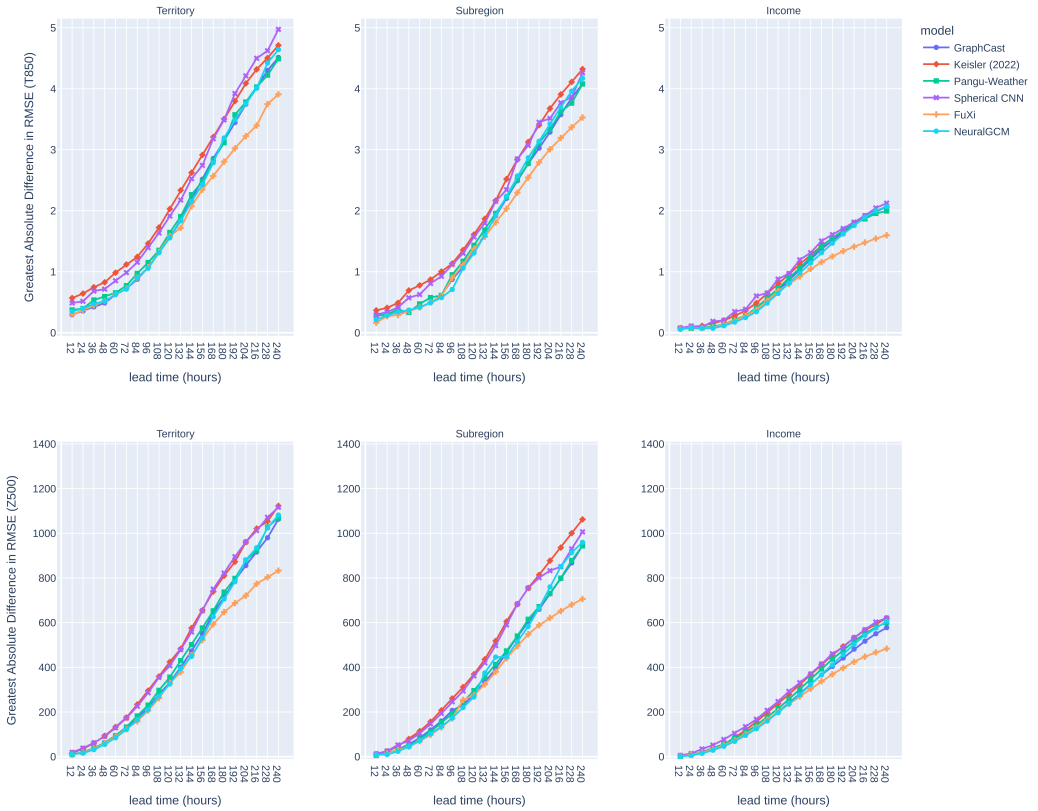

Figure 2: Greatest absolute difference of any two per-strata RMSE for each attribute when predicting T850 and Z500 at different lead times. Lower difference is more fair. Outlier RMSE values have been removed. Starting at a lead time of one week, FuXi is still the most fair model across all attributes and variables.

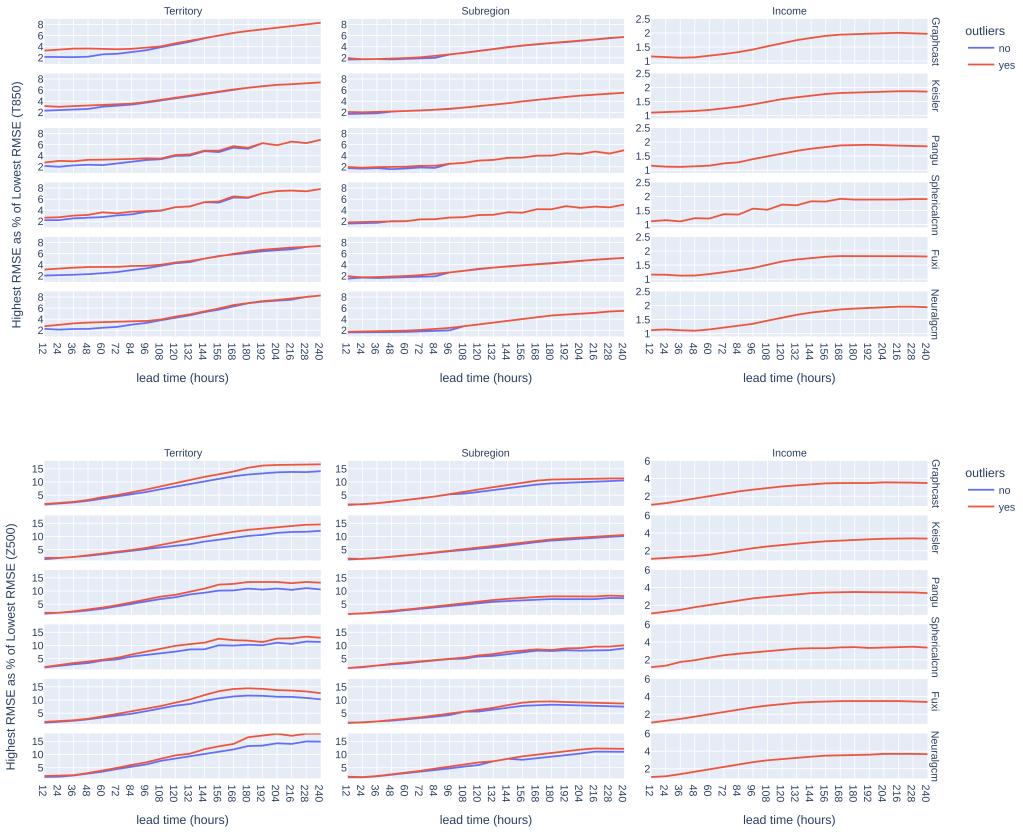

Figure 3: Highest per-strata RMSE as a percent of the lowest per-strata RMSE with and without outliers included.

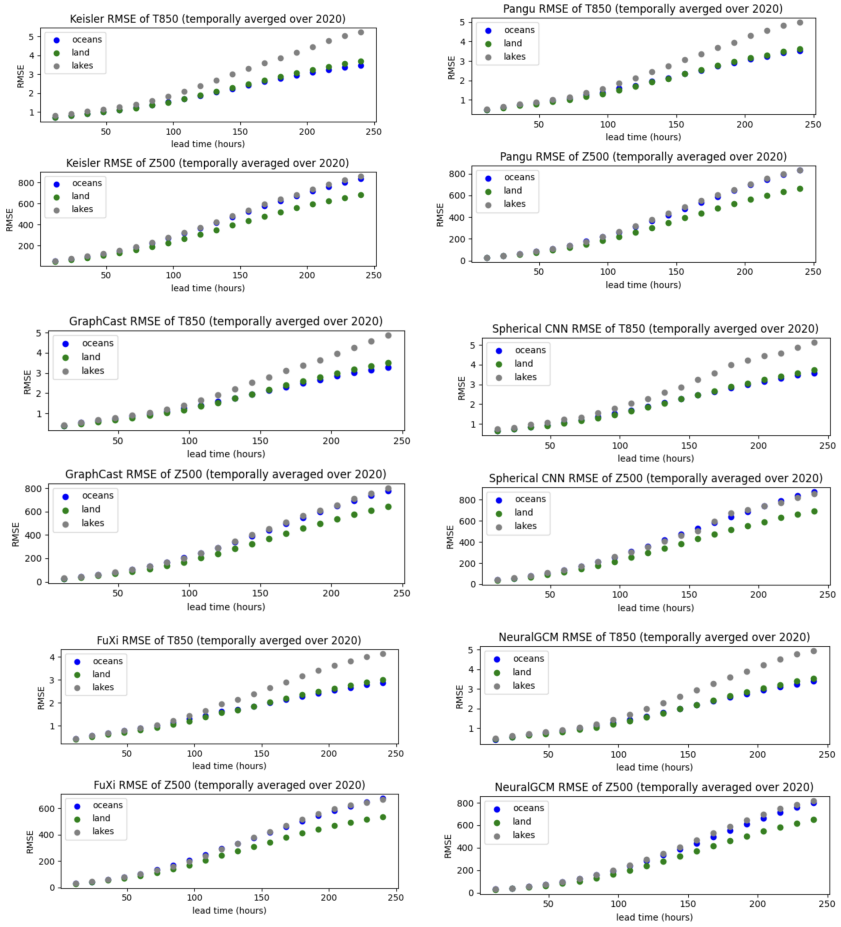

Figure 4: Results of preliminary experiments with landcover as an attribute.

Figure 5: Area-weighted RMSE by subregion by model.