No Location Left Behind: Introducing the Fairness Assessment for Implicit Representations of Earth Data

Daniel Cai*
Department of Computer Science
Brown University
daniel_cai@brown.edu

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

Abstract

Encoding and predicting physical measurements such as temperature or carbon dioxide is instrumental to many high-stakes challenges – including climate change. Yet, all recent advances solely assess models' performances at a global scale. But while models' predictions are improving *on average over the entire globe*, performances on sub-groups such as islands or coastal areas are left uncharted. To ensure safe deployment of those models, we thus introduce FAIR-EARTH, a fine-grained evaluation suite made of diverse and high-resolution dataset. Our findings are striking–current methods produce highly biased predictions towards specific geospatial locations. The specifics of the biases vary based on the data modality and hyper-parameters of the models. Hence, we hope that FAIR-EARTH will enable future research to design solutions aware of those per-group biases.

1 Introduction

Our planet exhibits phenomena that operate across a wide range of spatial and temporal scales, from local microclimates with hourly fluctuations, to global climate patterns that evolve over decades. Capturing those dynamics is crucial for climate change monitoring and mitigation. Existing methods include observation-based networks on the ground and sea (18), satellite-based remote sensing (21), and more recently, computer-based climate simulations (10). However, all aforementioned approaches suffer from some combination of discretization error, modeling error, data inconsistency, and resource-intensive inference (1). By being able to learn nonparametric models from arbitrarily high-resolution, multi-modal data, AI–and in particular Implicit Neural Representations (INRs) (20)–offer an attractive and efficient alternative. The use of INR to learn the underlying physical dynamics of geospatial data has seen a rapid increase in interest (6). Such methods learn to encode the underlying generative process and representation of the data through an implicit function mapping coordinates to data realisations–through a Deep Neural Network (DNN) (11).

However, all recent INR advancements emerge from improving *average* test performances, i.e., by assessing predictions' quality over the entire Earth. This begs the following question:

How reliable are current state-of-the-art solution when looking at specific sub-groups such as coastal land or islands?

As INRs are being deployed increasingly for climate monitoring, it is of utmost importance to develop specialized evaluation suites to quantify the fairness of current solutions. This of course is practically motivated: for tasks like natural disaster risk assessment where consequences are severe, there is a natural emphasis on improving worst-case performance rather than average-case metrics (15). To ensure that practitioners both in climate and AI research can better assess their model's

*Equal contribution
†Equal contribution

Machine Learning and the Physical Sciences Workshop, NeurIPS 2024.

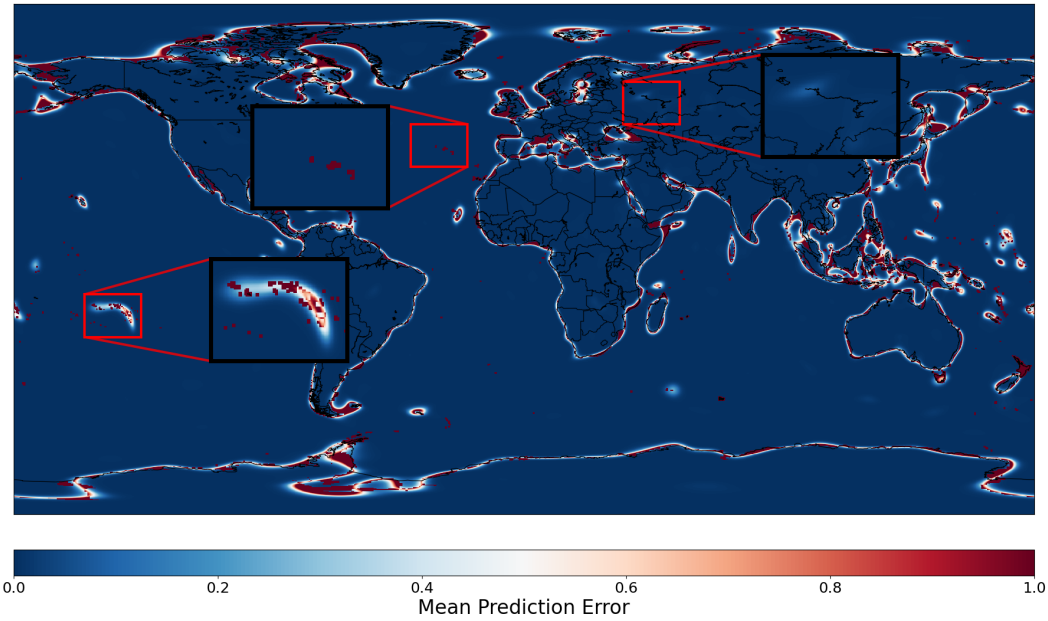


Figure 1: Heatmap of the spatial distribution of approximation error using an Implicit Neural Representation to model land-sea data of the Earth. **Clear bias against islands is observed where the error magnitude is significantly higher**. Further details and plots: Section 4, Table 3.

fairness we propose the first unified and fine-grained evaluation suite, coined FAIR-EARTH. Upon a rigorous evaluation of state-of-the-art INRs against FAIR-EARTH, we obtain the following striking observations:

1. "localized" groups such as coasts and islands are left behind (Fig. 1, Table 3)
2. global performance is *negatively correlated* to island performance (Fig. 2)
3. per-group performance can not be competitive on all sub-groups concurrently (Fig. 10).

The FAIR-EARTH dataset and the accompanying experiments described in this paper represent our attempt to address these challenges head-on. By providing a common, open-source[3] playground for researchers and practitioners, we aim to accelerate progress in tackling pressing environmental and societal challenges.

2 Background

The (un)fair impact of solely focusing on improving *average* test performance has been observed to be detrimental to sub-group performance within the context of natural image classification tasks (3; 16). However, to the best of our knowledge, no such assessment has been proposed for INR, and in particular for INR on Earth observations. Prior to diving into the fairness assessment in Section 3, we begin with some background on INRs.

Implicit Neural Representations. Throughout our study, we will focus on the latest state-of-the-art INR model developed by (19). The model introduces a novel integration of SPHERICAL HARMONIC location embeddings with SirenNets (20), demonstrating consistently superior performance over earlier location encodings such as GRID and THEORY in various ablation studies (2; 17).

The crux of this procedure lies in decomposition of the underlying signal, e.g., land boundaries, temperature, as continuous signals on the globe $f : (\lambda, \phi) \mapsto \mathbb{R}$. Specifically, for well-behaved functions, e.g., with exponential decay of their eigenvalues, the following decomposition can precisely

[3]The full codebase and dataset will be released upon completion of the reviewer process

Table 1: Per sub-group cross-entropy test loss across various spatial resolution of the dataset. **We observe that the bias of the model in missing "island" persists even for high resolution dataset.**

RESOLUTION	5000	10000	15000	20000	25000	30000
TOTAL	0.16 ± 0.02	0.13 ± 0.03	0.12 ± 0.03	0.11 ± 0.04	0.10 ± 0.04	0.10 ± 0.04
LAND	0.21 ± 0.03	0.15 ± 0.04	0.14 ± 0.05	0.14 ± 0.05	0.12 ± 0.06	0.16 ± 0.06
SEA	0.11 ± 0.02	0.16 ± 0.02	0.09 ± 0.03	0.08 ± 0.03	0.08 ± 0.03	0.80 ± 0.03
ISLAND	2.74 ± 0.03	3.25 ± 0.49	2.85 ± 0.33	2.66 ± 0.26	2.61 ± 0.20	2.51 ± 0.21
COASTLINE	1.06 ± 0.08	1.00 ± 0.06	0.95 ± 0.05	0.90 ± 0.04	0.82 ± 0.04	0.81 ± 0.05

recover the original signal:

$$f(\lambda, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} w_l^m Y_l^m(\lambda, \phi), \tag{1}$$

where Y is the class of spherical harmonic functions as depicted in Fig. 9, w are learnable scalar weights, and l and m are the degrees and orders of the basis functions Y_l^m (in practice, we impose an upper bound on l, effectively capping the representable frequency). Eq. (1) can easily be seen as a "linear" network with weight w. INR then extends that formulation by enabling that weighting of $Y_l^m(\lambda, \phi)$ to be a nonlinear function (DNN) of l and m; we direct the reader to Appendix A.1 for further details and theoretical intuition as to why Eq. (1) may exhibit geospatial biases when dealing with earth observations.

3 The FAIR-EARTH Dataset

We now describe FAIR-EARTH, our evaluation suite enabling localized performance and per-group performance assessment of INRs. The fairness assessment of current SOTA solutions will be provided in the following Section 4. This design of FAIR-EARTH is mostly motivated by two core principles.
High-resolution monitoring. The FAIR-EARTH dataset employs a uniform 0.1° x 0.1° gridding of the globe, yielding a consistent 1800 x 3600 map size for all variables. We leverage two modalities. *Land-ocean data:* Based on (13), this component contains coarse signals like continental landmasses while also providing high-resolution boundaries for fine-grained signals such as islands and coastlines. *Climate data:* Derived from the GRACED2021 (8) and CHELSA (14) datasets, this component provides an assortment of coarse and ultra-fine-grained signals, as well as high resolution along the time dimension. This temporal granularity allows for analysis of both long-term climate trends and short-term weather patterns.
Sub-group monitoring. We propose attributes and metadata including landmass size, coast distance, and population density for each location hence enabling to disentangle the global prediction performance into meaningful subgroups. To define binary thresholds for islands and coastlines, we provide flexibility with adjustable thresholds. This feature allows researchers to fine-tune their analyses based on specific definitions of islands or coastal zones, which can vary depending on the research question or application domain. For the purposes of our analyses, islands are defined as landmasses with size under $30,000$ sq. miles, encapsulating most of the "minor islands" as defined by (7).

4 Observed (Un)fairness of Existing Solutions

We now present a unified fairness assessment of current SOTA INRs. We recognize that climate data are often confounded by outside factors, e.g., cloud cover or population based crowd-sourcing (12). As our goal is to highlight the presence of structural biases in existing INRs, we will focus our assessments on the land-ocean dataset which is noiseless. Details on training procedures can be found in Appendix A.2.

Global Performance is not Representative of Local Performance We begin by examining the trends in training and evaluation. In particular, we leverage the metadata in FAIR-EARTH to compare algorithm performance over local signals to global signals. We frame land and sea data as global signals, with the localized counterpart being island data.

Based on preliminary correlation analysis, as expected, our results (Appendix A.5 and Fig. 2) indicate an almost-linear relationship between land and sea data. However, we notice a striking

moderate negative trend between land loss and island loss ($R^2 = 0.59$). When stratified along training resolution, this trend becomes stronger, and suggests that when optimized for total loss, state-of-the art INRs suffer a tradeoff between global and local performance. As a proxy test for this

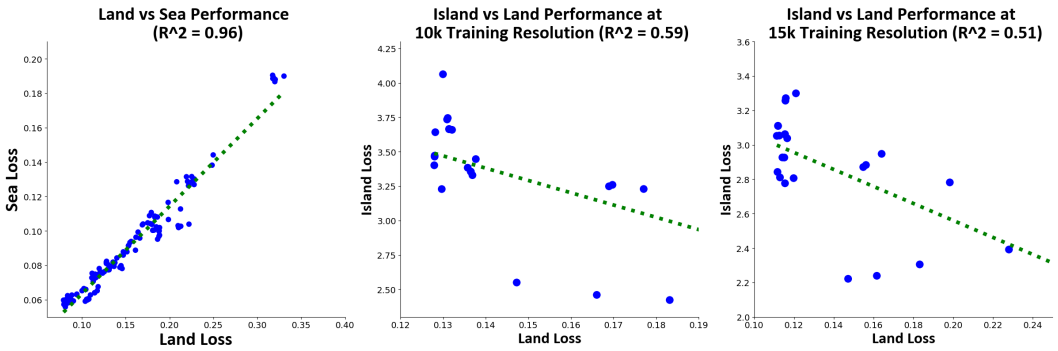


Figure 2: Correlation Analysis Between Local and Global Signals. Stratification along training resolution reveals a stronger negative trend between island and land performance.

hypothesis, we select the best-performing models for each sub-group (global, island, and coastline), and examine their performance on other subgroups (Fig. 10). As expected, there exists a sharp degradation in island performance for top global models, and an even sharper degradation in global performance for top island models. As a result, regardless of stratification, INRs of the form given by Eq. (1) seem incapable of *competitive* performance across all subgroups at once. Finally, coastline performance indicates a nuanced behavior of spherical harmonics for location encoding. While coastlines represent a similarly fine signal to islands, total loss is actually *positively* correlated with coastline loss, suggesting that Eq. (1) is capable of representing signals that are *fine*, but not *localized*.

Aliasing as a Result of Overfitting As referenced earlier, Eq. (1) is global in nature. The implications of this are evident in Fig. 8, where the smoothness of land and sea signals are compromised in an effort to fit to islands. This results in the aliasing and error spikes within landmasses observed in Fig. 1, as the algorithm overfits towards islands. In short, in the effort to precisely fit the islands, Eq. (1) is forced to learn an *unnatural* representation of the land and sea masses. As we could not identify a clear set of hyperparameters able to maintain competitive performance in both subgroups simultaneously, this suggests current SOTA solutions relying on INR modeling require further development to reach truly equitable predictions.

Biases Across Multiple Modalities and Subgroups While the land-sea binary classification task reveals natural biases against fine and localized areas, we emphasize that one of FAIR-EARTH's main strengths is in its ability to easily quantify subgroup disparities across multiple modalities. In particular, we perform similar stratified evaluation against FAIR-EARTH's benchmarks for environmental signals. For the surface temperature dataset, which exhibits similar sharp variations across recognizable boundaries, we note similar trends. Namely, FAIR-EARTH reveals systematic patterns in representation quality: regions with sharper variations, particularly near the coast, show significantly higher average representation loss ($\text{MSE}_{\text{Land}} = 0.87$, $\text{MSE}_{\text{Coast}} = 0.101$) compared to regions with smoother variations ($\text{MSE}_{\text{Sea}} = 0.43$, $\text{MSE}_{\text{Island}} = 0.49$) (Table 5).

Moreover, this analysis extends naturally to downstream biases through FAIR-EARTH's rich metadata. At the country level, we observe that representation challenges at the feature level manifest as systematic performance disparities. For instance, SPHERICAL HARMONIC and THEORY encodings particularly struggle with to demarcate the Spain's fine Mediterranean coastline, while all studied encodings show degraded performance in coastal countries due to sharp temperature gradients at land-sea boundaries (6).

5 Conclusions

Our experiments and analyses leverage the FAIR-EARTH dataset to provide a nuanced re-examination of state-of-the-art algorithmic performance in Earth system modeling. The high-resolution nature of the FAIR-EARTH dataset has revealed fine-grained patterns and phenomena that were previously undetectable. Our findings indicate that the SOTA algorithm's performance is more sensitive to

data characteristics than previously understood, suggesting that robust and equitable evaluation datasets like FAIR-EARTH are necessary to consider a wider range of data scenarios when assessing algorithmic performance. FAIR-EARTH will be available as open-source for all practitioners and research. While our current study was limited to the land-sea dataset, FAIR-EARTH also provides multiple other learnable signals; we encourage users to reproduce and iterate on existing results.

References

[1] Myles R Allen, JA Kettleborough, and DA Stainforth. Model error in weather and climate forecasting. In *ECMWF Predictability of Weather and Climate Seminar*, pages 279–304. European Centre for Medium Range Weather Forecasts, Reading, UK, 2002.

[2] Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification, 2019. URL `https://arxiv.org/abs/1906.05272`.

[3] Randall Balestriero, Leon Bottou, and Yann LeCun. The effects of regularization and data augmentation are class dependent, 2022. URL `https://arxiv.org/abs/2204.03632`.

[4] H. E. Beck, E. F. Wood, M. Pan, C. K. Fisher, D. G. Miralles, A. I. J. M. van Dijk, T. R. McVicar, and R. F. Adler. Mswep v2 global 3-hourly 0.1° precipitation: Methodology and quantitative assessment. *Bulletin of the American Meteorological Society*, 100:473–500, 2019. doi: 10.1175/BAMS-D-17-0138.1.

[5] Center for International Earth Science Information Network - CIESIN - Columbia University. Gridded population of the world, version 4 (gpwv4): Population density adjusted to match 2015 revision un wpp country totals, revision 11. Technical report, NASA Socioeconomic Data and Applications Center (SEDAC), Palisades, New York, 2018. Accessed 29 August, 2024.

[6] Elijah Cole, Grant Van Horn, Christian Lange, Alexander Shepard, Patrick Leary, Pietro Perona, Scott Loarie, and Oisin Mac Aodha. Spatial implicit neural representations for global-scale species mapping. In *International Conference on Machine Learning*, pages 6320–6342. PMLR, 2023.

[7] C Depraetere, AL Dahl, and G Baldacchino. A world of islands. an island studies reader. 2007.

[8] X. Dou, J. Hong, P. Ciais, et al. Near-real-time global gridded daily co2 emissions 2021. *Scientific Data*, 10:69, 2023. doi: 10.1038/s41597-023-01963-0.

[9] Phil PG Dyke and PP Dyke. *An introduction to Laplace transforms and Fourier series*. Springer, 2001.

[10] Nicholas Geneva and Dallas Foster. Nvidia earth2studio, 2024. URL `https://github.com/NVIDIA/earth2studio`.

[11] Michael Hillier, Florian Wellmann, Eric A de Kemp, Boyan Brodaric, Ernst Schetselaar, and Karine Bédard. Geoinr 1.0: an implicit neural network approach to three-dimensional geological modelling. *Geoscientific Model Development*, 16(23):6987–7012, 2023.

[12] Thomas Hovestadt and Piotr Nowicki. Process and measurement errors of population size: their mutual effects on precision and bias of estimates for demographic parameters. *Biodiversity and Conservation*, 17:3417–3429, 2008.

[13] G. Huffman, D. Bolvin, D. Braithwaite, K. Hsu, R. Joyce, and P. Xie. Integrated multi-satellite retrievals for gpm (imerg), version 4.4. Technical report, NASA's Precipitation Processing Center, 2014. URL `ftp://arthurhou.pps.eosdis.nasa.gov/gpmdata/`. Accessed 29 August, 2024.

[14] D. N. Karger, O. Conrad, J. Böhner, T. Kawohl, H. Kreft, R. W. Soria-Auza, N. E. Zimmermann, P. Linder, and M. Kessler. Climatologies at high resolution for the earth land surface areas. *Scientific Data*, 4:170122, 2017. doi: 10.1038/sdata.2017.122.

[15] Luke Kemp, Chi Xu, Joanna Depledge, Kristie L Ebi, Goodwin Gibbins, Timothy A Kohler, Johan Rockström, Marten Scheffer, Hans Joachim Schellnhuber, Will Steffen, and Timothy M Lenton. Climate endgame: Exploring catastrophic climate change scenarios. *Proceedings of the National Academy of Sciences*, 119(34):e2108146119, 2022. doi: 10.1073/pnas.2108146119. PMID: 35914185; PMCID: PMC9407216.

[16] Polina Kirichenko, Mark Ibrahim, Randall Balestriero, Diane Bouchacourt, Ramakrishna Vedantam, Hamed Firooz, and Andrew Gordon Wilson. Understanding the detrimental class-level effects of data augmentation, 2023. URL https://arxiv.org/abs/2401.01764.

[17] Gengchen Mai, Yao Xuan, Wenyun Zuo, Yutong He, Jiaming Song, Stefano Ermon, Krzysztof Janowicz, and Ni Lao. Sphere2vec: A general-purpose location representation learning over a spherical surface for large-scale geospatial predictions, 2023. URL https://arxiv.org/abs/2306.17624.

[18] M.A. Palecki, J.H. Lawrimore, R.D. Leeper, J.E. Bell, S. Embler, and N. Casey. U.s. climate reference network products, 2013. Accessed 29 August, 2024.

[19] Marc Rußwurm, Konstantin Klemmer, Esther Rolf, Robin Zbinden, and Devis Tuia. Geographic location encoding with spherical harmonics and sinusoidal representation networks, 2024. URL https://arxiv.org/abs/2310.06743.

[20] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473, 2020.

[21] Soroosh Sorooshian, Kuolin Hsu, Dan Braithwaite, Hamed Ashouri, and NOAA CDR Program. Noaa climate data record (cdr) of precipitation estimation from remotely sensed information using artificial neural networks (persiann-cdr), version 1 revision 1, 2014. Accessed 29 August, 2024.

A Appendix

A.1 Details on Spherical Harmonics Encoding

In particular, the sinusoidal nature of the location embedding introduces some interesting challenges as Fourier representations, and by extension, the spherical harmonic representations, are known to be particularly suited for stationary signals (9). This phenomenon arises from a fundamental property of these representations: to accurately capture highly localized signals, an extensive set of basis functions is required. However, this translates to several practical issues in the context of geospatial data. First, there is a tangible cap to basis size, due to numerical and tractability issues. Second, with this limited basis size, the embedding is limited in its ability to represent highly localized and fine-grained features such as islands, peninsulas, or intricate coastlines.

A.2 Training details

Our findings are the results of training over 150 model variations to adequately measure the effects of hyper-parameters such as embedding size, training resolution, and weight decay. To generate training data, we following the sampling procedure in (19), and similarly sample a validation set of size $0.2 \times$ *Num. Training Samples*. Finally, our evaluation is conducted over the entire 1800×3600 grid, which consists of a uniform 2D grid with $0.1^\circ \times 0.1^\circ$ resolution in longitude and latitude.

Training leverages the closed-form spherical harmonic generating function by (19). For faster training, (19) also provides analytic equations up to a certain embedding size, but these need to be re-calculated when using larger bases. On a simple 8-core CPU machine, the full training suite takes roughly 48 hours to complete.

A.3 Dataset Info

Table 2: FAIR-Earth Components

Category	**Description**	**Misc.**	**Source**
Land-Sea	Binary and continuous data, with data on percent water surface coverage for every grid point.	Additional metadata available for island and coastline labeling.	(13)
Population	Population distribution data for regions across the Earth, based on primary sources and interpolation.	Errors in Egypt and Greenland are smoothed via nearest-neighbor interpolation.	(5)
Precipitation	Global precipitation patterns and measurements.	Includes time-slice data from the year 2018 with monthly resolution.	(4)
Temperature	Temperature data and trends across different regions.	Includes time-slice data from the year 2018 with monthly resolution.	(14)
Carbon Dioxide Emissions	Data on CO2 emissions from various sources globally.	–	(8)

A.4 Extended Figures

A.4.1 Dataset Figures

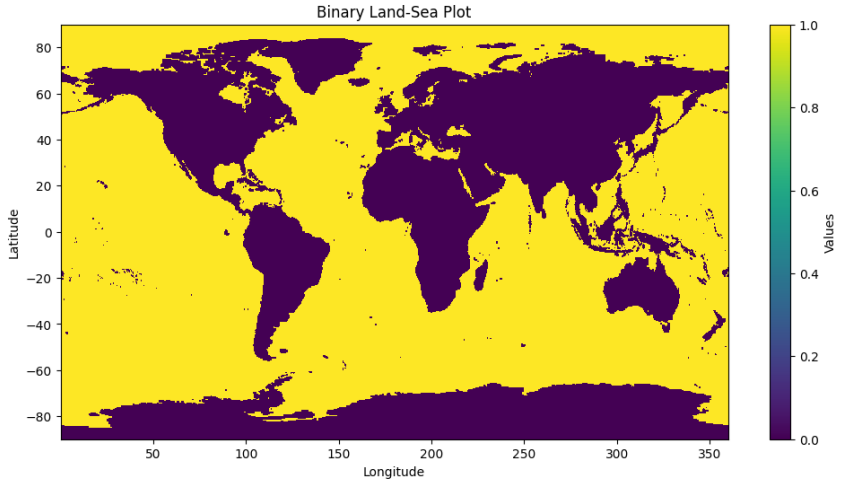


Figure 3: Binary Land-Sea

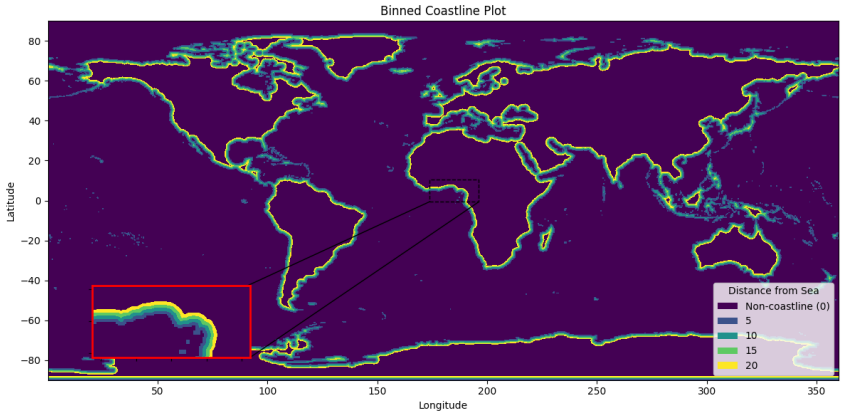


Figure 4: Coastline

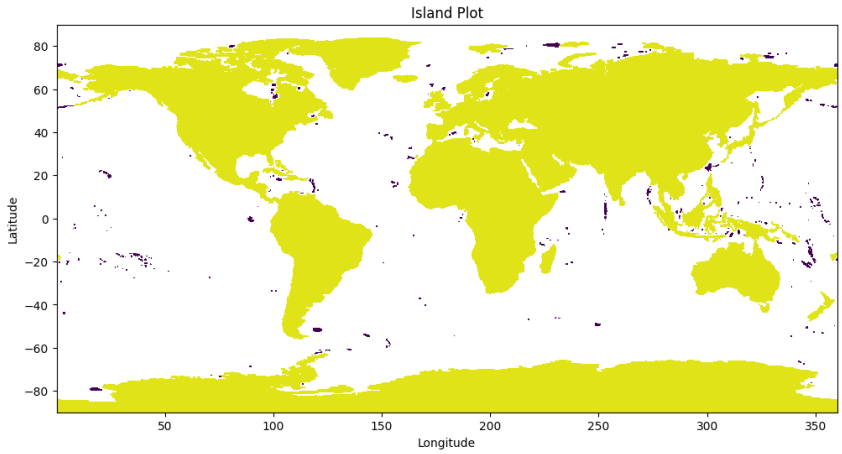


Figure 5: Islands (in Dark Purple)

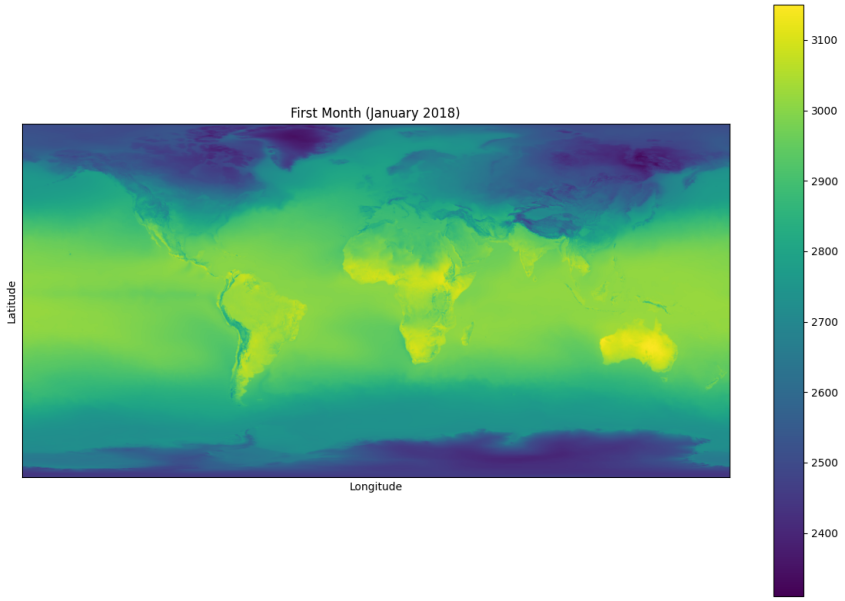


Figure 6: Air Surface Temperature Plot (Jan. 2018)

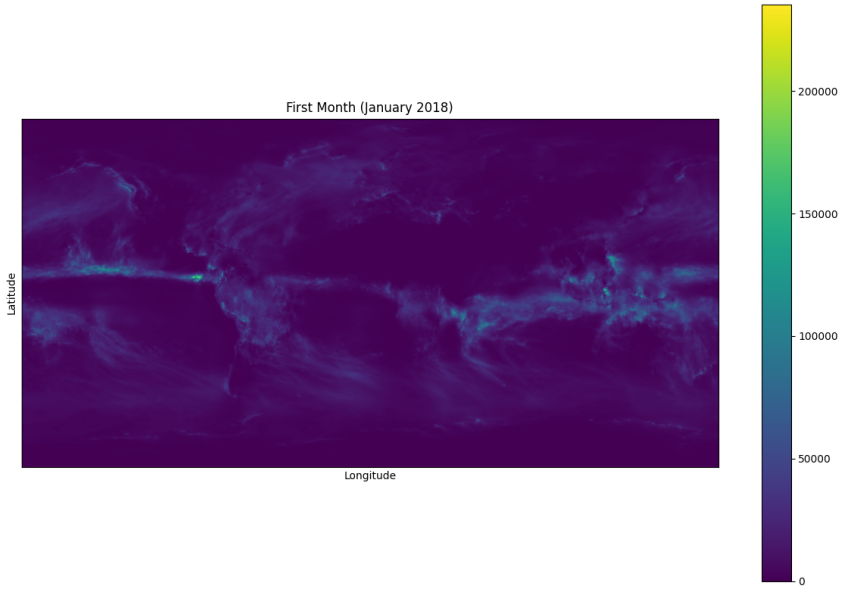


Figure 7: Cumulative Precipitation Plot (Jan. 2018)

A.4.2 Misc. Figures

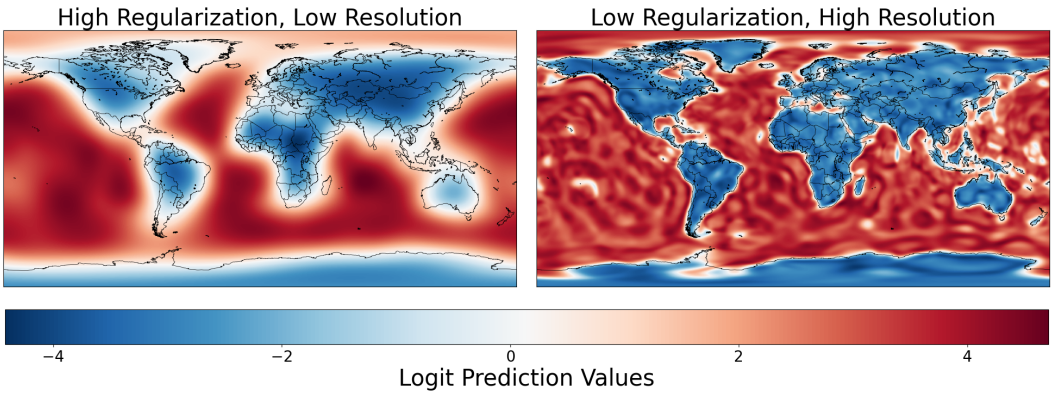


Figure 8: Model Behavior at Different Resolutions and Regularizations. Smaller models fail to capture any fine or local signals. Larger models poorly reconcile local signals with existing global signals.

Missed Islands

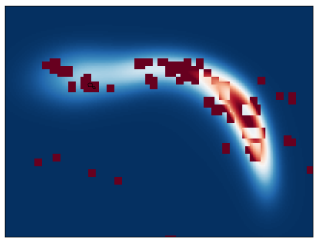
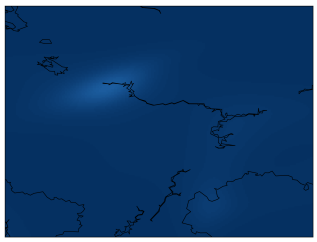
Blending of High-Frequency Islands

Aliasing in the Middle of Landmasses


Table 3: Zoomed-In Inset Plots

RESOLUTION	5000	10000	15000	20000	25000	30000
TOTAL	0.16 ± 0.02	0.13 ± 0.03	0.12 ± 0.03	0.11 ± 0.04	0.10 ± 0.04	0.10 ± 0.04
LAND	0.21 ± 0.03	0.15 ± 0.04	0.14 ± 0.05	0.14 ± 0.05	0.12 ± 0.06	0.16 ± 0.06
SEA	0.11 ± 0.02	0.16 ± 0.02	0.09 ± 0.03	0.08 ± 0.03	0.08 ± 0.03	0.80 ± 0.03
ISLAND	2.74 ± 0.03	3.25 ± 0.49	2.85 ± 0.33	2.66 ± 0.26	2.61 ± 0.20	2.51 ± 0.21
COASTLINE	1.06 ± 0.08	1.00 ± 0.06	0.95 ± 0.05	0.90 ± 0.04	0.82 ± 0.04	0.81 ± 0.05

Table 4: Per sub-group cross-entropy test loss across various spatial resolution of the dataset. **We observe that the bias of the model in missing "island" and "coastline" persists even for high resolution dataset, as improvement plateaus.**

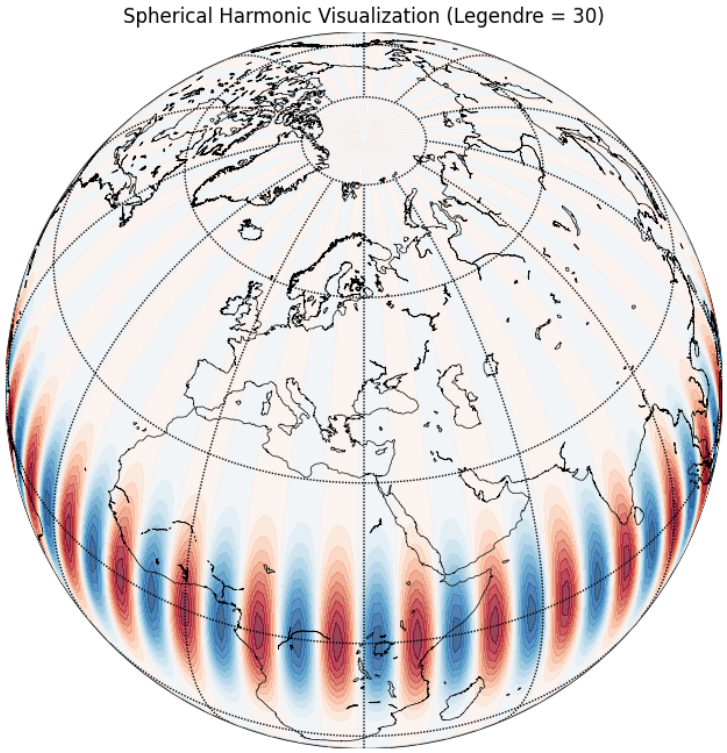


Figure 9: Spherical Harmonics (Eq. (1)) visualized on the Globe.

A.5 Extended Experimental Results

Subgroup	**SH**	**THEORY**	**SPHEREC+**
Land	0.076	0.217	5.867
Sea	0.028	0.063	2.152
Island	0.041	0.047	2.101
Coast	0.083	0.249	5.835

Table 5: Surface temperature regression subgroup losses for various encodings. Coast (underlined) consistently exhibits greater losses.

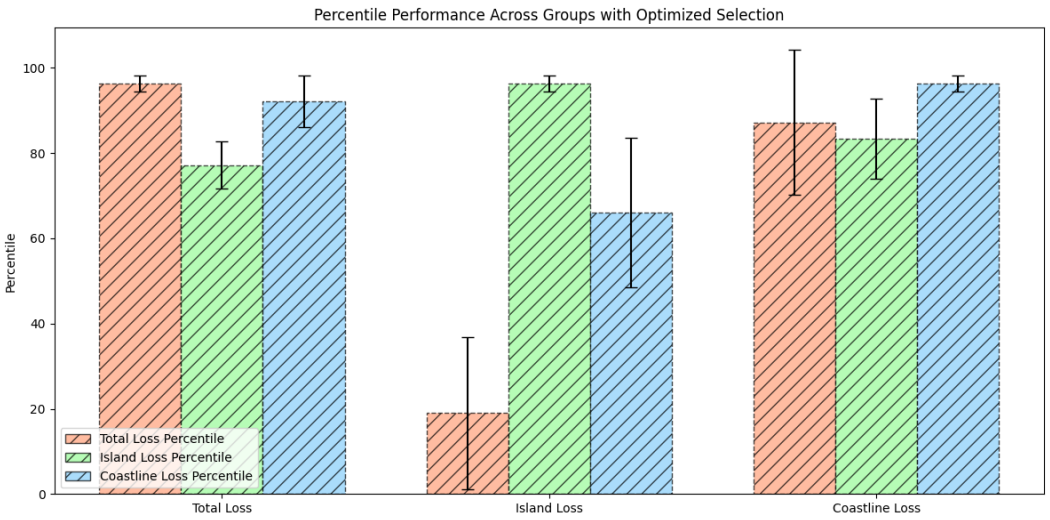


Figure 10: Performance of Top Sub-Group Models on Different Subgroups. Higher percentile indicates better performance. No models show consistent competitive performance along all sub-groups.

Encoding	Land-Sea Classification		Surface Temperature Regression	
	Best Country (Value)	**Worst Country** (Value)	**Best Country** (Value)	**Worst Country** (Value)
SPHERICAL HARMONIC	Honduras (0.001)	Spain (0.795)	Georgia (0.001)	Panama (0.333)
THEORY	Kyrgyzstan (0.001)	Spain (1.114)	Sierra Leone (0.002)	Vietnam (0.561)
SPHEREC+	Austria (0.418)	Chile (1.257)	Romania (0.143)	Greenland (10.425)

Table 6: Country discrepancies via FAIR-EARTH. Losses are respective to each dataset, and only countries with over 100 sampled points are included to mitigate noise. All encoding-dataset combinations exhibit a wide disparity in country-level performance.