TOWARDS DIVERSE AND REPRESENTATIVE GLOBAL PRETRAINING DATASETS FOR REMOTE SENSING FOUNDATION MODELS

Jacob Arndt, Philipe Dias, Abhishek Potnis, Dalton Lunga

Geospatial Science and Human Security Division
Oak Ridge National Laboratory, USA

ABSTRACT

The design of a pretraining dataset is emerging as a critical component for the generality of foundation models. In the remote sensing realm, large volumes of imagery and benchmark datasets exist that can be leveraged to pretrain foundation models, however using this imagery in absence of a well-crafted sampling strategy is inefficient and has the potential to create biased and less generalizable models. Here, we provide a discussion and vision for the curation and assessment of pretraining datasets for remote sensing geospatial foundation models. We highlight the importance of geographic, temporal, and image acquisition diversity and review possible strategies to enable such diversity at global scale. In addition to these characteristics, support for various spatial-temporal pretext tasks within the dataset is also critical. Ultimately, our primary objective is to place emphasis on and draw attention to the data curation stage of the foundation model development pipeline. By doing so, we think it is possible to reduce biases of geospatial foundation models, as well as enable broader generalization to downstream remote sensing tasks and applications.

***Index Terms*—** foundation models, datasets, pretraining, self-supervised learning, unsupervised learning

1. INTRODUCTION

Foundation models (FMs) are large neural networks trained in a self-supervised or unsupervised manner on vast amounts of unlabeled data. Such models are generalizable to a broad range of downstream tasks by adapting the features learned in the pretraining process. They are an emerging paradigm in artificial intelligence and have recently become a focus in the remote sensing and geospatial communities [1, 2, 3].

The data creation and curation process is emerging as an important precursor to pretraining foundation models. In the context of the full foundation model ecosystem and development process, data creation and curation occur before training, adaptation, and deployment [4]. In the curation stage, data is selected and filtered to meet requirements or constraints and ensure quality and relevance. The outcome of this curation step is a pretraining dataset.

The pretraining dataset is highly important for the development of a foundation model. In addition to suitable architectures and learning objectives, vast and diverse pretraining datasets have been a critical component determining the capabilities a model can acquire [4], as it provides the source for learning useful representations that can later be leveraged at the adaptation stage for a variety of downstream tasks. Furthermore, dataset composition is a critical factor in establishing biases and contributing to the overall risks and challenges of developing these models. For geospatial foundation models, these risks and challenges include geographic bias, geographic fidelity, temporal bias, spatial scale, heterogeneity, and generalizability [5]. As such, processes for curating data and the resulting datasets must be well designed, documented, understood, and managed.

Numerous benchmark training datasets exist in the remote sensing community, some of which have been used to pretrain geospatial foundation models [6, 7]. However, many of these datasets are not global and lack geographic, temporal, and image acquisition diversity. To overcome these limitations, new datasets specifically designed for self-supervised learning and pretraining geospatial foundation models have been proposed [8, 9, 10, 3, 11]. Given such variety in the usage of pre-existing datasets and new datasets being developed specific for pretraining, a discussion and review of datasets and dataset curation processes for foundation model pretraining is needed to help the field moving forward in the development of new FMs as well as better understanding their limitations.

Based on a brief review summarized in Section 2 of this manuscript, we identify several limitations of current literature in pretraining datasets for remote sensing foundation models, including: (1) lack of clarity in dataset curation process and the contents of the dataset; (2) general lack of metrics, metadata, and dataset descriptions to inform their usage and determine their diversity and representativeness; (3) some applications, for example glaciology, illegal mining, and deforestation, are not represented in existing labeled datasets or unlabeled datasets leading to question the usefulness of the pretraining data and resulting pretrained model for several downstream applications; (4) lack of consistency in pretraining data and evaluation protocols across works leads to question if the limitations of the model are a result of the pre-

training data, learning scheme, pretext tasks, architecture, or other factors. This ultimately limits our community's ability to compare results and draw conclusions on the state of the art.

In this paper, we build upon these observations and describe: (1) a vision for summarizing existing and future pretraining datasets; and, (2) a vision for developing a globally diverse and representative pretraining dataset. As a path forward, we promote the following key aspects for a pretraining dataset: (1) geographic diversity in landcover, biome, climate, population density, built environment, and socio-economics; (2) temporal diversity to support applications where temporal invariance may or may not be desired; (3) image acquisition parameter diversity such as ground sample distance, sensor, and viewing geometry; (4) support for different pretext tasks (spatial context, geolocation, temporal sequences); (5) support for numerous downstream tasks and applications (e.g. ocean, land, wetland, snow and ice, forest monitoring, agriculture, urban and built environment, disaster assessment); (6) support for different dataset sizes to promote different use cases and data-centric studies related to pretraining geospatial foundation models; and (7) detailed documentation of the dataset following best practices established in [12] along with metadata such as spatial resolution, sensor, geospatial coordinates, and other relevant information for each sample in the image dataset. As noted in [6], properties of diversity, richness, and scalability are highly relevant for creating benchmark datasets, and these same principles should be applied to geospatial foundation model pretraining datasets.

2. RELATED WORK

Pre-existing Benchmark Datasets for Pretraining. Numerous annotated remote sensing datasets have been gathered over the years for benchmarking tasks such as image classification [7, 6], object detection [13, 14], and segmentation [15, 16]. Several existing studies in self-supervised or unsupervised pretraining have approached building a pretraining dataset by combining many of these existing datasets and ignoring their labels [1, 17, 18, 19, 20].

In addition to combining various benchmark datasets together, it is also common to incorporate additional unlabeled imagery into the pretraining dataset [21, 2]. Two popular staples in many of these pretraining datasets are the million-AID [6] and the functional map of the world (fMoW) [7] datasets. Many works (e.g., SatMAE [2], ScaleMAE [17] and RingMo [1], [3]) use such imagery to train transformer-based architectures using a Masked Autoencoder (MAE) scheme for model pretraining, where the model is tasked to reconstruct pixels of masked image patches given only the remaining visible patches. However, there is a lack of works comparing the capabilities of models pretrained using different datasets but under similar training configurations (i.e., model architecture, pretraining scheme and training schedule), which compromises establishing a consensus on the best methodology for merging these datasets or which datasets are better than others for pretraining.

Purely Unlabeled Datasets for Pretraining. Previous works for developing datasets from large archives of unlabeled remote sensing imagery has also recently come into focus. Most of these works focus on building in diversity to the dataset by stratifying the study area using existing ancillary geospatial data (coordinates, landcover, populated places, climate) and leveraging sampling methods for obtaining geographic, temporal, and image acquisition diversity [8, 21, 10, 3, 11, 9]. For example, in SeCo [8] locations of populated places along with temporal information are used in a sampling strategy to build a pretraining dataset from Sentinel-2 tiles.

Datasets to Support Pretext Tasks. In addition to MAE-schemes, contrastive learning schemes leveraging additional information have also been explored in the literature. For such purposes, building pretext tasks into the dataset is also common practice. GASSL [19] leverage the images' geographic coordinates to aid learning in a constrastive-fashion. Tile2Vec [22] takes advantage of ideas of spatial autocorrelation by sampling neighboring image patches and distant neighbors in a triplet sampling and learning scheme, while GASSL[19], SeCo [8], and SatMAE [2] leverage multiple temporal views of the same location to build temporal diversity in the dataset to develop models capable of being sensitive or insensitive to temporal change. Moreover, ScaleMAE [17] utilize an image's ground sample distance with a novel positional encoding module to train a model that is more capable of being adapted to images with a wide range of ground sample distances.

Table 1 provides a non-exhaustive list of examples of datasets previously used in foundation model pretraining and self-supervised model training studies.

3. A VISION FOR SUMMARIZING PRETRAINING DATASETS

Developing Dataset Metrics. Due to the large number of existing datasets and the never-ending interest in creating new ones, a method for summarizing datasets must be developed. A first goal is to formulate a suite of metrics for large remote sensing foundation model pretraining datasets that describe the geographic, temporal, and image acquisition diversity and representativeness of the dataset. While difficult due to the vast number of potential remote sensing applications, it could also be beneficial to quantify a dataset's capability to support any number of downstream tasks. One possible metric for diversity to consider is entropy, which was used by [21] to indicate improvement in their GeoPile dataset's content relative to a Sentinel-2 dataset.

Applying Metrics and Review Methodology. Following this, a second goal will be to apply the developed method and metrics to existing datasets used during foundation model

Name	Datasets and Data Sources
Benchmark	
RingMo [1]	MillionAID, DOTA, LoveDA, FAIR1M, DIOR, iSAID
ScaleMAE [17]	fMoW
BillionScale [18]	MillionAID
RVSA [20]	MillionAID
GASSL [19]	fMoW
Custom	
SeCo [8]	Sentinel-2 tiles
Satlas Pretrain [9]	NAIP tiles, Sentinel-2 tiles
WorldStrat [10]	Airbus SPOT 6/7 tiles, Sentinel-2 tiles
Prithvi [3]	Landsat-Sentinel HLS tiles
SkySense [11]	WorldView-2,3, Sentinel-1,2
Mix of Benchmark and Custom	
GeoPile [21]	NAIP tiles, RSD46-WHU, MLRSNet, RESISC45, PatternNet
SatMAE [2]	fMoW, Sentinel-2 tiles

Table 1. Examples of remote sensing pretraining datasets and data sources used in the literature. The *Custom* methods also define sampling strategies for data curation.

pretraining. By doing so we hope to quantify the utility of a pretraining dataset before a model has been trained. Several questions that we would like to answer through this review and analysis are: Where are these samples located geographically? What dates and times were they collected? What are the acquisition parameters? How diverse in terms of image content and geography is this data? Are the datasets missing certain content to enable key downstream applications? Is this data geographically referenced? Is the dataset curation process documented and reproducible? Does the dataset have a datasheet [12]?

4. A VISION FOR REPRODUCIBLE, DIVERSE, AND REPRESENTATIVE GLOBAL PRETRAINING DATASETS

Geographic, Temporal, and Image Acquisition Diversity. To achieve geographic diversity, one could uniformly sample across the entire globe. However, doing so would yield a dataset with numerous imbalances. Populated areas would be underrepresented, certain ecosystems (e.g. tundra) would be underrepresented while others would dominate. One solution to such a problem is to leverage existing geospatial datasets as a guide for sampling. Datasets such as terrestial ecoregions, climate zones, landcover, and population density should be among some of the core inputs to be considered to ensure diversity. The work of [10] on developing WorldStrat, for example, used urban density, landuse, and underrepresented points of interest (e.g. informal human settlements) to guide dataset curation and sampling.

Use of ancillary datasets for sample site selection is also common in land cover classification data products. ESA WorldCover v200 [23], for example, incorporates biome and realm features from a 2017 ecoregions data product [24], climate data from TerraClimate, digital elevation models, and the global human settlement layer. Dynamic world [25], a near real-time landcover data product, also uses the ecoregions layer as a reference during training sample selection. In addition, they partition the globe into three hemispheres to further diversify their samples.

Fundamental to a remote sensing foundation model pretraining dataset is diversity in image acquisition parameters such as sensor, off-nadir angle, ground sample distance, scan direction, sun azimuth, sun elevation, and satellite azimuth. This information is especially relevant to very high resolution satellite imagery where these collection parameters have significant impact on overall image appearance and the objects within those images. Also relevant are image acquisition dates and times, which provide the essential information for incorporating temporal diversity into the dataset.

Support for Pretext Tasks Pretext tasks refer to objectives constructed from unlabeled data that a model is trained to solve. For example, in masked image modeling the pretext task is for the model to reconstruct or predict the original image after portions of the image have been masked or erased. Studies in self-supervised and unsupervised learning for remote sensing data have suggested the importance of spatial and temporal pretext tasks. The works of SatMAE [2], SeCo [8], and GASSL [19], for example, have all indicated temporal views can improve pretraining. Spatial pretext tasks such as those highlighted by [19] and [22] indicate improved pretraining as well. Given these successes, we emphasize the need to support both spatial and temporal pretext tasks within a remote sensing pretraining dataset. Including all possible image metadata (e.g. ground sample distance, off-nadir angle, sensor, coordinates, date and time) is one way to enable further pretraining objectives in the future. This aspect also helps establish documentation for the data.

4.1. Preliminary Dataset Development

In this section we present preliminary work in developing a remote sensing foundation model pretraining dataset that meets the key requirements highlighted in the introduction of this paper. We create two global spatial point datasets of different resolutions that we use during the sampling stage. The first point dataset has 0.05 degree spatial resolution while the second is higher resolution at 0.01 degrees and includes only points at locations where population density is greater than 1000 in the reference ORNL Landscan dataset [26]. The higher resolution point dataset is needed to better capture the diversity of built environment and populated areas, whereas

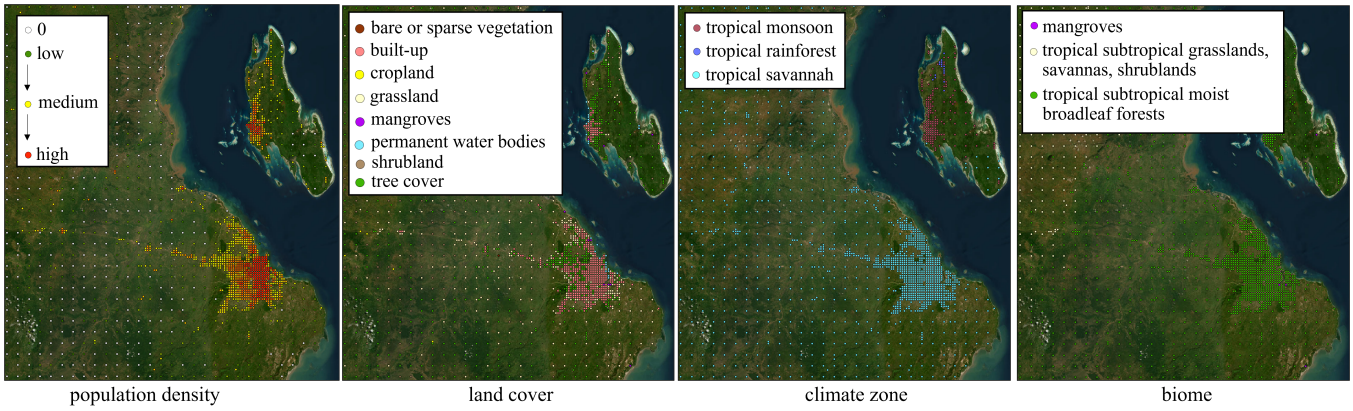

Fig. 1. Example of the sample point grid and attributes considered for *geographic* diversity sampling.

the lower resolution dataset is primarily used to capture non built-up characteristics. We merge these two point grids into a single point dataset and conflate it with the ancillary geospatial layers to obtain biome [24], biogeographic realm [24], climate zone [27], population density [26], and landcover [23] type for each point. Figure 1 provides an illustration of the final spatial point dataset over a small area of interest and several of the attributes used for diversification sampling.

Image acquisition parameter diversity will be achieved by including images from Maxar's WorldView-2 and -3, Planet's Planetscope constellation, and ESA's Sentinel-2. Together, a single location will feature data collected at a variety of ground sample distances of approximately 0.5 meters, 3 meters, and 10 meters. For consistency, all images include Blue, Green, Red, and Near-Infrared channels. We plan to further diversify on an image's viewing geometry attributes such as off-nadir angle, azimuth, scan direction. To achieve temporal diversity and support for temporal pretext tasks, we plan to sample each location at 8 different dates, with each collect spaced approximately 3 months apart.

5. CONCLUSION

Development and usage of foundation models in remote sensing is growing rapidly. The pretraining dataset is an important component in the development process of these models. This preliminary review and analysis of pretraining datasets illustrates that there is not a standard protocol or curation process for developing these datasets. Many existing studies use different pretraining datasets and methodology to assemble datasets.

The volume of existing pretraining datasets and their variable usage in existing research, makes it difficult to understand the contributions that pretraining data make to a foundation model's ability to adapt to downstream tasks. A careful review, summary, and development of metrics for pretraining data should be developed to help us better understand the qualities of these data and their impacts on foundation models.

Several key elements to consider in developing a pretraining dataset are geographic diversity, temporal diversity, image acquisition diversity, support for spatial-temporal pretext tasks, support for different downstream tasks, support of different dataset sizes, and clear documentation. Many of these themes are present to varying degrees in existing pretraining datasets, but through our review we see opportunities to fill existing gaps in developing large, diverse, and representative datasets for remote sensing foundation models.

Our immediate next steps are to further develop our pretraining dataset. Following this, we would like to train a foundation model on this dataset and assess the trained model on downstream tasks including building extraction, road extraction, land use and land cover classification, and change detection of damaged buildings.

6. ACKNOWLEDGEMENTS

We acknowledge that this manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

7. REFERENCES

[1] Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang,

Hao Chang, Qinglin He, Guang Yang, Ruiping Wang, Jiwen Lu, and Kun Fu, "Ringmo: A remote sensing foundation model with masked image modeling," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–22, 2023.

[2] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon, "Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery," *Advances in Neural Information Processing Systems*, vol. 35, pp. 197–211, 2022.

[3] Johannes Jakubik, Sujit Roy, CE Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, et al., "Foundation models for generalist geospatial artificial intelligence," *arXiv preprint arXiv:2310.18660*, 2023.

[4] Rishi Bommasani et al., "On the opportunities and risks of foundation models," *CoRR*, vol. abs/2108.07258, 2021.

[5] Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, et al., "On the opportunities and challenges of foundation models for geospatial artificial intelligence," *arXiv preprint arXiv:2304.06798*, 2023.

[6] Yang Long, Gui-Song Xia, Shengyang Li, Wen Yang, Michael Ying Yang, Xiao Xiang Zhu, Liangpei Zhang, and Deren Li, "On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid," *IEEE Journal of selected topics in applied earth observations and remote sensing*, vol. 14, pp. 4205–4230, 2021.

[7] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee, "Functional map of the world," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6172–6180.

[8] Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9414–9423.

[9] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi, "Satlaspretrain: A large-scale dataset for remote sensing image understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 16772–16782.

[10] Julien Cornebise, Ivan Oršolić, and Freddie Kalaitzis, "Open high-resolution satellite imagery: The worldstrat dataset –with application to super-resolution," in *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022, vol. 35, pp. 25979–25991.

[11] Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al., "Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery," *arXiv preprint arXiv:2312.10115*, 2023.

[12] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021.

[13] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3974–3983.

[14] Xian Sun, Peijin Wang, Zhiyuan Yan, Feng Xu, Ruiping Wang, Wenhui Diao, Jin Chen, Jihao Li, Yingchao Feng, Tao Xu, et al., "Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 184, pp. 116–130, 2022.

[15] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai, "isaid: A large-scale dataset for instance segmentation in aerial images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 28–37.

[16] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong, "Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation," *arXiv preprint arXiv:2110.08733*, 2021.

[17] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell, "Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4088–4099.

[18] Keumgang Cha, Junghoon Seo, and Taekyung Lee, "A billion-scale foundation model for remote sensing images," 2023.

[19] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon, "Geography-aware self-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10181–10190.

[20] Di Wang, Qiming Zhang, Yufei Xu, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang, "Advancing plain vision transformer toward remote sensing foundation model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[21] Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen, "Towards geospatial foundation models via continual pretraining," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16806–16816.

[22] Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon, "Tile2vec: Unsupervised representation learning for spatially distributed data," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 3967–3974, Jul. 2019.

[23] Daniele Zanaga, Ruben Van De Kerchove, Dirk Daems, Wanda De Keersmaecker, Carsten Brockmann, Grit Kirches, Jan Wevers, Oliver Cartus, Maurizio Santoro, Steffen Fritz, Myroslava Lesiv, Martin Herold, Nandin-Erdene Tsendbazar, Panpan Xu, Fabrizio Ramoino, and Olivier Arino, "Esa worldcover 10 m 2021 v200," Oct. 2022.

[24] Eric Dinerstein, David Olson, Anup Joshi, Carly Vynne, Neil D. Burgess, Eric Wikramanayake, Nathan Hahn, Suzanne Palminteri, Prashant Hedao, Reed Noss, Matt Hansen, Harvey Locke, Erle C Ellis, Benjamin Jones, Charles Victor Barber, Randy Hayes, Cyril Kormos, Vance Martin, Eileen Crist, Wes Sechrest, Lori Price, Jonathan E. M. Baillie, Don Weeden, Kierán Suckling, Crystal Davis, Nigel Sizer, Rebecca Moore, David Thau, Tanya Birch, Peter Potapov, Svetlana Turubanova, Alexandra Tyukavina, Nadia de Souza, Lilian Pintea, José C. Brito, Othman A. Llewellyn, Anthony G. Miller, Annette Patzelt, Shahina A. Ghazanfar, Jonathan Timberlake, Heinz Klöser, Yara Shennan-Farpón, Roeland Kindt, Jens-Peter Barnekow Lillesø, Paulo van Breugel, Lars Graudal, Maianna Voge, Khalaf F. Al-Shammari, and Muhammad Saleem, "An Ecoregion-Based Approach to Protecting Half the Terrestrial Realm," *BioScience*, vol. 67, no. 6, pp. 534–545, 04 2017.

[25] Christopher F Brown, Steven P Brumby, Brookie Guzder-Williams, Tanya Birch, Samantha Brooks Hyde, Joseph Mazzariello, Wanda Czerwinski, Valerie J Pasquarella, Robert Haertel, Simon Ilyushchenko, et al., "Dynamic world, near real-time global 10 m land use land cover mapping," *Scientific Data*, vol. 9, no. 1, pp. 251, 2022.

[26] Budhendra Bhaduri, Edward Bright, Phillip Coleman, and Jerome Dobson, "Landscan," *Geoinformatics*, vol. 5, no. 2, pp. 34–37, 2002.

[27] Hylke E Beck, Niklaus E Zimmermann, Tim R McVicar, Noemi Vergopolan, Alexis Berg, and Eric F Wood, "Present and future köppen-geiger climate classification maps at 1-km resolution," *Scientific data*, vol. 5, no. 1, pp. 1–12, 2018.