Rethinking Machine Learning for Climate Science: A Dataset Perspective

Aditya Grover[1,2]

[1] Department of Computer Science
[2] Institute of the Environment and Sustainability
University of California, Los Angeles

Abstract

The growing availability of data sources is a predominant factor enabling the widespread success of machine learning (ML) systems across a wide range of applications. Typically, training data in such systems constitutes a source of *ground-truth*, such as measurements about a physical object (e.g., natural images) or a human artifact (e.g., natural language). In this position paper, we take a critical look at the validity of this assumption for datasets for climate science. We argue that many such climate datasets are uniquely biased due to the pervasive use of external simulation models (e.g., general circulation models) and proxy variables (e.g., satellite measurements) for imputing and extrapolating in-situ observational data. We discuss opportunities for mitigating the bias in the training and deployment of ML systems using such datasets. Finally, we share views on improving the reliability and accountability of ML systems for climate science applications.

1 Introduction

Large datasets are fueling major advances in the scaling of machine learning (ML) systems for a variety of real-world usecases of relevance to science and society, ranging from creative art and text generation (Ramesh et al. 2021; Brown et al. 2020) to protein folding (Jumper et al. 2021) and drug discovery (Vamathevan et al. 2019). This has led to a growing optimism for the broad field of climate change as well (Rolnick et al. 2022). With advancements in sensory, storage, and network technology, we now have large datasets available for many domains of interest to climate change, such as weather forecasting (Rasp et al. 2020), agriculture and forestry (Zheng et al. 2019), and chemical and materials discovery (Kirklin et al. 2015; Chanussot et al. 2021), among others.

As the first step of any ML pipeline, the choice of a training dataset is critical to the downstream performance of ML systems. Both the quantity and quality of a dataset play an important role, as demonstrated by numerous prior studies (e.g., (Gebru et al. 2021)) that correlate the size, noise, and bias within training datasets with broad and holistic indicators of downstream performance, such as accuracy and fairness. Given the growing enthusiasm in using ML for climate change, it begs the question: are datasets for climate domains aligned with ML pipelines in use today?

In this position paper, we argue that climate science domains can present unique challenges for ML systems given how datasets are collected and generated. In particular, we note that climate datasets used in practice are routinely based on *reanalysis* or *gridding* that combine disparate real and simulated/proxy measurement sources. While such a procedure ensures that datasets have excellent coverage, it leads to a bias that can propagate within standard ML pipelines. This calls for a rethink of both training and deployment of data-centric ML pipelines for climate science, as well as community guidelines for dataset and model release.

The rest of the paper is structured as follows: in Section 2, we briefly review current data collection practices in climate science and the role of ML in improving climate projections. In Section 3, we present opportunities for aligning machine learning with data practices in climate science, as well as community guidelines for improving the transparency and accountability of ML models. Finally, we conclude in Section 4 with a summary and discussion on broader impacts, including implications of this research on domains focusing on climate change mitigation and adaptation, as well as other disciplines within ML.

2 What Makes Climate Data Unique?

Climate modeling is fundamental to understanding the interactions between the atmospheric, oceanic, and land surface process, including anthropogenic interventions. Such models can be used for short-term weather forecasts or long-term projections of the Earth's climate under different interventions. Beyond scientific pursuits, the outputs of these models inform regional and international policy aimed at near- and long-term climate mitigation and adaptation.

Typically, climate models couple our physical understanding with on-ground observations. However, such models can be insufficient for certain downstream usecases due to limited accuracy and/or spatiotemporal resolution. For example, nowcasting requires very short-horizon weather predictions (up to 2 hours ahead) that is greater than the time it takes to spin up numerical weather prediction (NWP) systems (Ravuri et al. 2021). Similarly, many general circulation models (GCM) and earth system models (ESM) that are

used for projecting future climate operate at a 2 degree resolution (200km), which is much lower than typically needed (<0.1 degrees) for effective mitigation planning at a regional level (Fowler, Blenkinsop, and Tebaldi 2007).

In such scenarios, data-driven solutions involving machine learning can play a big role in overcoming the limitations of current climate models. However, the quality of a ML system depends significantly on the availability of high quality datasets. This presents two key challenges. First, historical in-situ observational records for climate variables are irregularly sampled due to uneven access to sensory technology, leading to a geographical bias. Second, for climate change in particular, we require projections of future climate under different interventions (e.g., different fossil fuel usage) — many of these scenarios have never been observed in the past, but are necessary for governments and international organizations to analyze and formulate policies.

Together, the above challenges necessitate the use of alternate data sources, such as *reanalysis datasets* and *gridded datasets*. Reanalysis datasets combine historical observations with the outputs of climate models, whereas gridded datasets rely on statistical tools for imputing missing values or proxy measurements made via satellites. In both cases, the goal is to generate high volume and high coverage datasets for training ML systems. Several such datasets are in use today, such as CHIRPS (Funk et al. 2015), a gridded dataset for high-resolution rainfall combining satellite measurements with in-situ observations, and ERA-5 (Muñoz-Sabater et al. 2021), a reanalysis dataset maintained by the European Centre for Medium-Range Weather Forecasts. These datasets are updated daily and contain historical observations spanning many decades, providing excellent spatiotemporal coverage at the expense of their respective model bias. As a concrete example, consider data for soil moisture available from the ERA reanalysis dataset. Soil moisture is an important climate variable for projecting the agriculture viability of any land area. For validation on real measurements, ERA5 uses in-situ soil measurement data from 14 sites — 4 in North America, 6 in Europe, 1 in Australia, and 2 in Africa, reflecting a highly biased distribution with respect to global demographics and completely omitting some continents.

3 Roadmap for Climate ML Pipelines

In the previous section, we motivated the use of reanalysis and gridded datasets for training ML models, and the inherent bias they encode. How should we train ML systems on such climate datasets? The status quo, as adopted in several papers (e.g., Oses et al. (2020); Baño-Medina, Manzanas, and Gutiérrez (2020)), is to treat the reanalysis dataset as ground-truth. However, this ignores the context in which the dataset was generated and is likely to propagate or even potentially amplify the bias in the dataset. While there is no simple solution, we believe that ML pipelines that explicitly account for this additional context can be far more effective for downstream applications. In this regard, we outline our position on exciting directions for improving the training and deployment of ML pipelines for climate science.

3.1 Training

Model selection. While training benefits immensely from the use of high coverage (but biased) datasets, we can consider alternate strategies for model selection (e.g., via the use of validation datasets). In areas for which we have in-situ observations, we can monitor the model's performance directly on such data for the held-out years, sidestepping any bias due to the use of gridded or reanalysis tools. Also, note that since model selection is less data-hungry than training the model itself, this strategy can also be potentially applied for underserved regions with few in-situ measurements.

Unsupervised learning and domain adaptation. In the last few years, there have been several advances in large scale unsupervised representation learning, including both contrastive and generative approaches (Murphy 2022). While in-situ measurements of climate variables are hard to obtain for arbitrary targets, we can obtain high-quality feature descriptors for unsupervised pretraining.

Alternatively, a closely related problem is that of unsupervised domain adaptation, where we need to transfer ML models trained on one domain to a related domain (with zero or few labels). Various techniques have been developed to enable such a transfer, such as the use of domain randomization (Tobin et al. 2017) for control tasks. In the climate context, we can consider the gridded/reanalysis datasets as the source domain and consider transferring ML models trained on such datasets to points in the target domain of interest.

3.2 Deployment

Uncertainty quantification. Well-calibrated uncertainty estimates can play a key role in reliably communicating the predictions of ML systems trained on gridded and reanalysis datasets and downstream users relying on theses predictions. In principle, one could use any gridded or reanalysis dataset for training a ML model. However, as one might expect, different datasets differ in their imputation strategies and hence, the predictions of ML models trained on these datasets would also differ. Consequently, we can treat these models as an ensemble (Lakshminarayanan, Pritzel, and Blundell 2017) and use the distribution of predictions for each of the ML models as a measure of uncertainty due to the imputation strategy.

Datasheets and model cards. While the need for documenting datasets and models is well-recognized in both the ML and climate communities, the standards and terminologies vary significantly. As we see more real-world deployments, it is important to expand the scope of existing protocols, such as datasheets (Gebru et al. 2021) and model cards (Mitchell et al. 2019) in the ML community, to better document key details relating to the gridded and reanalyzed datasets, such as the details on the auxiliary climate models and data sources used for dataset creation, the distribution of in-situ measurement sites, and any known limitations of the imputation strategy. We believe including such details can significantly improve the transparency and interpretability of ML systems, as well as aid in reproducibility — a growing area of concern for ML in scientific applications (Kapoor and Narayanan 2022).

4 Broader Impacts

This position paper calls for a careful reflection on the use of datasets for ML applications in climate science. We argued that while current reanalysis and gridded datasets might seem to have global coverage at high spatiotemporal bandwidths, these datasets are in fact reflective of the geographic and socioeconomic disparities in access to sensory technology (e.g., satellites, weather balloons). Quantifying and mitigating this bias without compromising on overall accuracy is an open challenge for the ML community. Our work highlights a select group of directions in this regard grounded in metrics concerning accuracy, reliability, and reproducibility.

While the use of gridded and reanalysis datasets is common practice in the climate science community, we also expect similar challenges in other fields relevant to climate change, and ML more broadly. For example, efforts to use ML for computational chemistry are fundamentally bottlnecked by the domain gap in computational simulation softwares and real experimental data. Even more so, with the advent and rapid proliferation of deep generative models, we are likely to find future ML systems trained on mixtures of real and synthetic data, and thus leading to a natural cross-pollination of tools and techniques.

References

Baño-Medina, J.; Manzanas, R.; and Gutiérrez, J. M. 2020. Configuration and intercomparison of deep learning neural models for statistical downscaling. *Geoscientific Model Development*, 13(4): 2109–2124.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *NeurIPS*, 33: 1877–1901.

Chanussot, L.; Das, A.; Goyal, S.; Lavril, T.; Shuaibi, M.; Riviere, M.; Tran, K.; Heras-Domingo, J.; Ho, C.; Hu, W.; et al. 2021. Open catalyst 2020 (OC20) dataset and community challenges. *ACS Catalysis*, 11(10): 6059–6072.

Fowler, H. J.; Blenkinsop, S.; and Tebaldi, C. 2007. Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 27(12): 1547–1578.

Funk, C.; Peterson, P.; Landsfeld, M.; Pedreros, D.; Verdin, J.; Shukla, S.; Husak, G.; Rowland, J.; Harrison, L.; Hoell, A.; et al. 2015. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific data*, 2(1): 1–21.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873): 583–589.

Kapoor, S.; and Narayanan, A. 2022. Leakage and the Reproducibility Crisis in ML-based Science.

Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; and Wolverton, C. 2015. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Computational Materials*, 1(1): 1–15.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *NeurIPS*, 30.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, 220–229.

Muñoz-Sabater, J.; Dutra, E.; Agustí-Panareda, A.; Albergel, C.; Arduini, G.; et al. 2021. ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, 13(9): 4349–4383.

Murphy, K. P. 2022. *Probabilistic machine learning: an introduction*. MIT press.

Oses, N.; Azpiroz, I.; Marchi, S.; Guidotti, D.; Quartulli, M.; and G. Olaizola, I. 2020. Analysis of copernicus' era5 climate reanalysis data as a replacement for weather station temperature measurements in machine learning models for olive phenology phase prediction. *Sensors*, 20(21): 6381.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

Rasp, S.; Dueben, P. D.; Scher, S.; Weyn, J. A.; Mouatadid, S.; and Thuerey, N. 2020. WeatherBench: a benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11): e2020MS002203.

Ravuri, S.; Lenc, K.; Willson, M.; Kangin, D.; Lam, R.; Mirowski, P.; Fitzsimons, M.; Athanassiadou, M.; Kashem, S.; Madge, S.; et al. 2021. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878): 672–677.

Rolnick, D.; Donti, P. L.; Kaack, L. H.; Kochanski, K.; Lacoste, A.; Sankaran, K.; Ross, A. S.; Milojevic-Dupont, N.; Jaques, N.; Waldman-Brown, A.; et al. 2022. Tackling climate change with machine learning. *ACM Computing Surveys (CSUR)*, 55(2): 1–96.

Tobin, J.; Fong, R.; Ray, A.; Schneider, J.; Zaremba, W.; and Abbeel, P. 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, 23–30. IEEE.

Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; et al. 2019. Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6): 463–477.

Zheng, Y.-Y.; Kong, J.-L.; Jin, X.-B.; Wang, X.-Y.; Su, T.-L.; and Zuo, M. 2019. CropDeep: the crop vision dataset for deep-learning-based classification and detection in precision agriculture. *Sensors*, 19(5): 1058.