Synthetic Data before "Data Driven" Science: Data and Representation in American and Indian Meteorology

Anya Martin

The ascension of "data-driven" science in the early 2000s, and of the machine learning paradigm in the last few years, has led to increased attention on the diversity (or lack thereof) of foundational datasets. Basic datasets like Imagenet and the Pima Indian Diabetes Dataset (PIDD), which have served as standard tests for image classification and data mining algorithms, are deeply biased: Imagenet contains "a range of derogatory and offensive categories in the "person" subtree" (Prabhu & Birhane, 2020; Raji et al., 2021), and the 1980 Pima Diabetes Dataset was made longitudinal under and through regimes of mass immobilization and "biosocial suffering" (Radin, 2017). These works overwhelmingly focus on the representation of biometric, image, or language data; I aim to trace historical and current practices of representative and synthetic *meteorological data*. We see "representation" function not on the level of individual people but on the level of observation densities, climate events, and weather models' ability (or lack thereof) to track crucial phenomena like ocean-air interactions, peatland fires, and monsoons.

While "big data" and "data-driven science" appeared around the early 2000s, massive datasets have existed since the 1900s, not coincidentally around the birth of statistics. This took two major forms, and the first was census data, which has progressively evolved into our current biometric analysis schemes. The second was environmental and specifically meteorological data. As Tsing (2011) notes, colonial scientists relied on a *transcendent and nonsocial* Nature that "could travel across cultures and across empires," a thing that could be studied outside of people and as moving outside human action. By ignoring e.x. Indonesian botanists and speaking "to Nature itself," it was possible to alienate and appropriate existing sciences into a study of Nature outside the human. This "global Nature" continues to animate modern meteorology, which has proven capable of assimilating a staggering range of different air/ocean/forest/cryosphere models, but which continues to place social/economic data as an awkward post-hoc addition.

"Weather data" consists of observational inputs to all these different models, but just as crucially consists of *reanalysis data*, the perfectly-spaced and "gridded" data produced as the outputs of global climate models. Foundational datasets like ERA-5, produced by the European Center for Mid-Range Weather Forecasting, have served as a "ground truth" for the field, and they have also been used as "ground truth" data for the new machine learning models which have increasingly become relevant. We see that meteorology has a fifty-year-old "big data" paradigm (global Nature), a modeling scheme (global climate models, GCMs),

and a synthetic data paradigm (reanalysis data). Calls for effective representation in meteorological data and modeling have worked through these terms, which can be loosely classified into three categories:

- Appeals for "physics representation" to make a global climate model more effectively represent atmospheric dynamics like monsoons.
- Appeals for "data coverage," arguing that more/better data can improve prediction on a large scale.
- Appeals for a paradigm shift, arguing that existing methods cannot effectively serve "real world" aims of providing services to citizens.

First, appeals for representation can demand attention to *specific atmospheric interactions*. This is most clearly seen in the India Meteorological Department's (IMD's) 2010s efforts to "couple" air and ocean models. Prior to around 2010, global models of the air and ocean updated one another loosely in a way which made it difficult to model complicated and large-scale air-ocean interactions, which made GCMs very bad at predicting monsoon dynamics. Papers in this period (Chaudhari et al., 2013; Pokhrel et al., 2012; Saha et al., 2014) repeatedly invoke "representation" in terms of how well models could track air-ocean dynamics; "ill-representation of sea ice" and "improper representation of teleconnection in terms of remote forcing via Walker circulation," among others, were linked to general biases in air-ocean dynamics. These "representation" and "ill-representation" decisions are profoundly political; effective representation is both a regional call and a global call to make transcendent "global Nature" models actually describe basic Indian Ocean dynamics. Because these models are then used to generate synthetic reanalysis data, effective "representation" then appears not just in theory but in the implantation of these dynamics into future synthetic datasets.

So one method to increase "representation" is to change the dynamics which generate synthetic data. Another less "high theory" method is to advocate increased coverage of actual observations. Observations are unevenly distributed across national lines and within nations; most mesonets are national efforts, and most data collection takes place at airports, seaports, major government offices, and universities. This means that initiatives to improve data density often interweave with rural/urban advocacy – buoy programs like OMNI-RAMA are justified based on their use by fishers, and rural stations are justified by their critical role for farmers. Because meteorological stations require a low density of people (and a lack of nearby buildings) to achieve good error bounds, meteorological "big data" often operates at the nadir of big biometric data schemes, and so different methods must be employed.

Finally, we see AI increasingly come to bear as a method for addressing existing met data problems. This can come in the form of assimilating new and more "representative" datasets, or it can imply a new focus on "impact" which makes the data more useful for constituents (Kooshki Forooshani et al., 2024). As is hopefully clear by now, the "big data" schemes common to AI/ML have different synthetic data schemes and different paradigms of inclusion than the "global nature" of meteorology, and so scientists dealing with the crippling problems of physics-based modeling may turn to AI as a method of denoising/debiasing data, assimilating more "equitable" data, or as a way to reject the meteorological paradigm altogether, choosing to replace one set of bias problems with another.

References

Chaudhari, H. S., Pokhrel, S., Saha, S. K., Dhakate, A., Yadav, R. K., Salunke, K., Mahapatra, S., Sabeerali, C. T., & Rao, S. A. (2013). Model biases in long coupled runs of NCEP CFS in the context of Indian summer monsoon. *International Journal of Climatology*, *33*(5), 1057–1069. https://doi.org/10.1002/joc.3489

Kooshki Forooshani, M., van den Homberg, M., Kalimeri, K., Kaltenbrunner, A., Mejova, Y., Milano, L., Ndirangu, P., Paolotti, D., Teklesadik, A., & Turner, M. L. (2024). Towards a global impact-based forecasting model for tropical cyclones. *Natural Hazards and Earth System Sciences*, *24*(1), 309–329. https://doi.org/10.5194/nhess-24-309-2024

Pokhrel, S., Chaudhari, H. S., Saha, S. K., Dhakate, A., Yadav, R. K., Salunke, K., Mahapatra, S., & Rao, S. A. (2012). ENSO, IOD and Indian Summer Monsoon in NCEP climate forecast system. *Climate Dynamics*, *39*(9), 2143–2165. https://doi.org/10.1007/s00382-012-1349-5

Prabhu, V. U., & Birhane, A. (2020). *Large image datasets: A pyrrhic win for computer vision?* (arXiv:2006.16923). arXiv. https://doi.org/10.48550/arXiv.2006.16923

Radin, J. (2017). "Digital Natives": How Medical and Indigenous Histories Matter for Big Data. *Osiris*, *32*(1), 43–64. https://doi.org/10.1086/693853

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). *AI and the Everything in the Whole Wide World Benchmark* (arXiv:2111.15366). arXiv. https://arxiv.org/abs/2111.15366

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y.-T., Chuang, H., Iredell, M., Ek, M., Meng, J., Yang, R., Mendez, M. P., Dool, H. van den, Zhang, Q., Wang, W., Chen, M., & Becker, E. (2014). The NCEP Climate Forecast System Version 2. *Journal of Climate*, *27*(6), 2185–2208. https://doi.org/10.1175/JCLI-D-12-00823.1

Tsing, A. L. (2011). *Friction: An Ethnography of Global Connection*. Princeton University Press.