BAMS
Article

Identifying and Categorizing Bias in AI/ML for Earth Sciences

Amy McGovern, Ann Bostrom, Marie McGraw, Randy J. Chase, David John Gagne II, Imme Ebert-Uphoff, Kate D. Musgrave, and Andrea Schumacher

KEYWORDS:
Atmosphere;
Ocean; Artificial intelligence;
Other artificial intelligence/machine learning

ABSTRACT: Artificial intelligence (AI) can be used to improve performance across a wide range of Earth system prediction tasks. As with any application of AI, it is important for AI to be developed in an ethical and responsible manner to minimize bias and other effects. In this work, we extend our previous work demonstrating how AI can go wrong with weather and climate applications by presenting a categorization of bias for AI in the Earth sciences. This categorization can assist AI developers to identify potential biases that can affect their model throughout the AI development life cycle. We highlight examples from a variety of Earth system prediction tasks of each category of bias.

SIGNIFICANCE STATEMENT: As artificial intelligence (AI) grows in popularity, its methods are being applied to a wide range of Earth system prediction tasks. Although AI can facilitate more accurate prediction at many tasks, it is not without potential pitfalls, especially if the developers are not as familiar with its potential drawbacks. In this paper, we provide a classification system for the types of bias that one is likely to see in applying AI to Earth sciences. Our classification system will assist current and future AI developers to recognize where their AI system or data are biased so they can take steps to alleviate this bias.

https://doi.org/10.1175/BAMS-D-23-0196.1

Corresponding author: Amy McGovern, amcgovern@ou.edu

In final form 2 January 2024

AFFILIATIONS: McGovern—NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES), and School of Computer Science and School of Meteorology, University of Oklahoma, Norman, Oklahoma; **Bostrom**—NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES), and Evans School of Public Policy and Governance, University of Washington, Seattle, Washington; **McGraw, Ebert-Uphoff, and Musgrave**—NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES), and Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado; **Chase**—NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES), and School of Computer Science and School of Meteorology, University of Oklahoma, Norman, Oklahoma, and Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado; **Gagne**—NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES), and National Center for Atmospheric Research, Boulder, Colorado; **Schumacher**—NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES), and Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, and National Center for Atmospheric Research, Boulder, Colorado

Applications of artificial intelligence (AI) and machine learning (ML) in the Earth sciences have grown exponentially over the past few years. We refer to AI/ML more generally as AI throughout the rest of the paper. It is critically important that AI developers create methods in an ethical and responsible manner lest AI be developed and deployed in a manner that could cause harm. In this work, we build on our earlier research (McGovern et al. 2022), which demonstrated multiple ways where AI could go wrong for environmental sciences and Earth science applications. Here we focus specifically on the issue of *bias* as it is one of the key threads throughout much of the recent work on ethical AI (e.g., Peng et al. 2021; McGovern et al. 2022; Balagopalan et al. 2022; Almuzaini et al. 2022; Buolamwini 2023).

Bias is recognized as a key issue that must be addressed in developing ethical and responsible AI for AI in general. It is one of the key issues discussed by the National Institutes of Standards and Technology (NIST) as part of their focus on creating standards for trustworthy AI (Schwartz et al. 2022) and is addressed in the new Executive Order on AI.[1] For Earth sciences applications, it is relatively new to consider bias [see the recent American Geophysical Union AI guidelines (Stall et al. 2023)].

Biased AI models can cause harm in a variety of ways, including affecting people's abilities to obtain a job, have stable housing, and more. For examples of such effects, see O'Neil (2016), Eubanks (2018), Benjamin (2019), and Kantayya (2020). When negatively biased models are deployed and then make the news, they can erode public trust in AI overall. Such models have already been deployed by both private industry and government. Creating and understanding trustworthy AI is a key focus of everyone involved in this work, as all are members of the NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES). Our overall goal with this work is tightly intertwined with our goals of ensuring that AI for the Earth sciences is trustworthy: ensuring that the models being developed and deployed now are as free of harmful bias as possible.

At first glance, bias may not seem to be an issue with AI for the Earth sciences, as compared to AI applications more broadly. Recent work has shown that AI can be successful at applications ranging from meteorology, climate, hydrology, seismology, and more

[1] https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

(e.g., McGovern et al. 2017; Schneider et al. 2017; Reichstein et al. 2019; Bauer et al. 2021; Labe and Barnes 2021; Chantry et al. 2021; Tsai et al. 2021; Zhang et al. 2022; Bi et al. 2023; Lam et al. 2023). Such success stories combined with a false impression that meteorological data are "objective" could lead AI developers to believe that bias is not an issue with Earth science applications. Unfortunately, as we will demonstrate, bias exists in most Earth science data and must be addressed.

As AI is being developed for a wide range of Earth science applications, underlying biases in the data can affect the AI models' performance. Deploying such models could unintentionally exacerbate environmental and climate injustices. For example, consider developing a highly accurate tornado prediction system that relies on the United States national network of weather radars. We demonstrated in McGovern et al. (2022) that many of the counties in the southeastern United States lack good low-level radar coverage. If such an algorithm was developed and deployed without knowledge of the underlying bias in the data, it may unintentionally and incorrectly miss tornadoes in these critical areas.

The main contribution of this work is a new categorization of AI bias focused on the entire life cycle of AI development and application in the Earth sciences. While our categorization is inspired by the one presented by NIST (Schwartz et al. 2022) our focus is on AI for the Earth sciences and how identifying bias can guide developers in AI for Earth science applications. To ensure that the categorizations are useful for AI developers in that area, we provide examples from the Earth sciences domains.

It is not possible for AI developers to mitigate bias until they can identify it. By providing a classification system for AI for the Earth sciences, we enable developers to systematically recognize what the possibilities for bias are in their problem domains. This is a first step toward measuring and mitigating such biases. Stating what makes an AI model "good" and free of bias is a difficult task, similar to stating what is a "good" forecast (Murphy 1993). While some measures of goodness are easy to measure, others are more challenging. This is also true of biases. Some of the categories we provide here will be relatively easy to measure and some are harder. In many cases, mitigating the biases is not straightforward. Due to the in-depth approaches needed to address many of the biases, we will address the mitigation in a future paper.

Bias categories

Our full bias categorization builds on Fig. 2 in Schwartz et al. (2022). We restructure this through a lens focused on human judgment and decision-making, while recognizing the bounded nature of human rationality (Fischhoff and Broomell 2020; Kahneman et al. 1982; Simon 1990). We restructure the NIST categorization into four main categories of bias and focus our framework on the full development and deployment life cycle of AI for the Earth sciences. Our bias categorization is shown in Fig. 1. The four main categories are ordered by the AI development life cycle:

- systemic and structural bias (shown in blue),
- human bias (green),
- data bias (orange), and
- statistical and computational bias (pink).

In the development of an AI model, each type of bias may interact across the categories. Thus, it is critical to understand each in order to develop and deploy AI models in an ethical and responsible manner. We discuss each of the categories and subcategories in Fig. 1 in more detail in the following sections. The colors around the subcategories indicate strong interactions across categories. It is also possible for the types of bias to cascade along the chain of

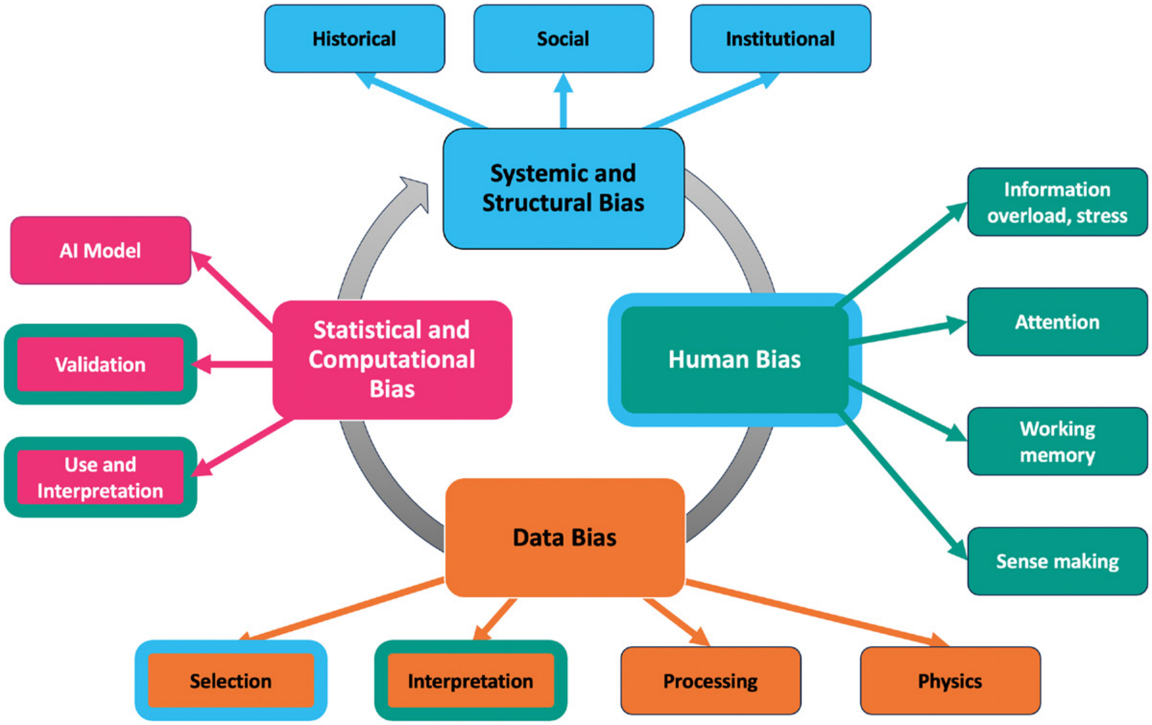

Fig. 1. Full bias categorization for AI for the Earth sciences. Each of the categories interact (shown by the main cycle arrow in the center). Single colors on the subcategories indicate one category, and a second color outlining the circle indicates significant interaction to the category matching that color.

bias, creating a multiplying effect. For example, systemic and structural bias and human bias strongly interact and may cause choices made in the data bias category to thus affect the performance of the model and create additional statistical and computational biases. The gray cycle under the four main categories represents this interaction.

Systemic and structural bias

Systemic and structural biases are biases that are present in the background structure of society or the institution where data may be measured. We put these at the top of Fig. 1 because they are the overarching fabric of the society in which we live and they affect all of our other categories. For in-depth discussions on such biases in society in general (and why we chose the names), see March and Olsen (1984), Friedman and Nissenbaum (1996), Henry (2010), Suresh and Guttag (2021), and Schwartz et al. (2022). Critically for AI, data that are not measured cannot be used to train an AI model. In D'Ignazio and Klein (2020), the authors point out in many ways that "what gets counted counts." Although none of their examples are in weather and climate, their points are still valid for this domain. For example, in places where there are unreliable historical records of temperature, it is challenging to train and validate a climate model on past data. Likewise, if forecasts (human or machine generated) are not archived, there are no data to measure how forecasts have changed over time and there are no data for an AI model to train with or to compare to.

We define three specific subcategories of systemic and structural bias: historical, social, and institutional. We acknowledge that, although we have broken them into three distinct subcategories, biases that fall into systemic and structural bias strongly interact with each other. There are likely additional subcategories that we could have chosen here but we focus on the effects of bias from those three as they relate to Earth science.

Historical bias. As sensors are changing and improving over time, there exist biases in the historical records of such data. Because training AI models requires a large historical record of data, historical bias could skew the AI model predictions to not match current reality across a wide variety of Earth science prediction tasks. The need for large training datasets is especially true for deep learning models.

Some historical biases are already well known. For example, for AI researchers working on climate-related tasks, records such as historical temperature may have existing biases (e.g., Peterson and Vose 1997; Menne et al. 2010). Such biases include spatial and temporal gaps in data or biases due to older measuring instruments that may have been less accurate. Biases are also known to exist in reconstructed paleoclimate data (e.g., Coats et al. 2020) and recent work proposed a framework to assess the quality of such data (Pacchetti et al. 2021). As the climate warms, the statistical distribution is also shifting from historical records, which provides an additional bias. Any of these biases could skew an AI model unless accounted for in the model training. Even the climate change distributional shift can prove problematic given that AI models may be predicting something completely out of scope from their training data.

Although many of the shifts in distributions from climate change are well known, there are less obvious historical biases that can exist in Earth science as well. For example, the uncertainty of historical tropical cyclone (TC) counts differs significantly between the presatellite and postsatellite eras (Vecchi and Knutson 2011). Furthermore, intensity estimates of tropical cyclones are prone to historical biases related to increases in the spatial resolution of satellites and improvements made to aircraft reconnaissance instrumentation (Emanuel 2008).

Often AI developers seeking to train on large datasets will obtain historical data through reanalysis datasets, with ERA-5 (Hersbach et al. 2020) being one of the most popular. While ERA-5 is clearly an outstanding dataset for training AI models (and the authors have used it for much of their work), it is known to have biases (e.g., Yilmaz 2023). The ERA-5 documentation lists some of the limitations.[2] We have observed that AI developers who obtain such data but who are not codeveloping with domain scientists often are unaware of the limitations of the data and will instead assume it provides a singular source of truth, potentially leading to overly confident assessment of AI model performance.

Social bias. This type of bias can be due to reliance on stereotypes or other broadly shared cultural assumptions or practices. At first glance, one might assume that social biases do not apply to Earth sciences applications. Unfortunately, that is not the case. For example, Anbarci et al. (2011) demonstrated that forecast accuracy is improved in locations with higher average household incomes than in locations with lower household incomes. Another example of social bias relevant to AI for the Earth sciences includes gender bias in open-source community tools, which tend to not to support problem solving strategies commonly used by women [for specific examples, see Mendez et al. (2018)]. This can affect the diversity of the AI developers, thus impacting long-term solutions.

Another example of social bias is the use of stereotypes and cultural assumptions made by developers about potential users. This could affect the data that are collected, but it can strongly influence the model that is chosen. For example, probabilistic information is not shared by many weather forecasting organizations partially because they do not believe the general public is sophisticated enough to make use of that information (e.g., Pappenberger et al. 2013), but in practice many people can and do use such information to make sophisticated decisions (Morss et al. 2010; Ripberger et al. 2022). For example, people hedge their risk by changing their daily routine to account for potential weather threats. If developers just assume that end-users do not want or need probabilistic information, they may choose

[2] https://confluence.ecmwf.int/display/CKB/ERA5%3A+data+documentation#ERA5:datadocumentation-Knownissues

an inferior deterministic model. Social science research with diverse user groups is critical to address this bias.

Historical and social biases can overlap. For example, there has been relatively little historical investment in ground-based sensors such as radar or precipitation measurements in the global south (Saltikoff et al. 2019). Such sensors are used in data assimilation and global weather prediction and the lack of sensors leads to a disparity in forecasting between the Northern and Southern hemispheres. This can be seen in the ECMWF performance charts;[3] for example, the "Lead time of anomaly correlation coefficient (ACC) reaching multiple thresholds" shows a significant difference between the hemispheres historically, with the gap narrowing only relatively recently [this is explored in many works, see, for example, Haiden et al. (2021) and Brands et al. (2023)]. This lack of data can lead to an inability to develop accurate models, leading to additional lives lost [for examples, see World Weather Attribution (2023) and Harvey (2023)].

Institutional bias. The final subcategory of systemic and structural bias that we identify is institutional bias. This type of bias stems from the norms within an institution, such as academia, or the weather enterprise, or within an agency or organization. Such norms may come from written rules or unwritten norms and expectations. For example, written rules may specify that certain types of data are not collected or archived. Such rules were likely created well before the advent of AI and the need for large datasets. Many of these rules are historical in that they were created when storage was more expensive. For example, many forecasts were deemed to be of low value after the forecast time had expired and there was not enough storage, so they were simply removed, unless someone saved them into a private repository (e.g., see the Iowa Environmental Mesonet archive[4]). Such data repositories can be very valuable for AI but their lack of availability could lead to biases in training and verification.

Institutional bias also exists within specific groups of people with particular cultures. For example, in the National Weather Service (NWS), which is part of the National Oceanic and Atmospheric Administration (NOAA), tropical cyclone formation declarations have historically been initiated more often during daylight hours (Fig. 2c). One possible reason for this observed trend is that visible satellite imagery, which plays an important role in identifying the existence of a closed low-level circulation (as illustrated in Figs. 2a,b), is not available at night. To the best of our knowledge there is no official rule that states TC formation declarations should wait until daylight hours. However, given the importance of daylight-dependent data sources in the forecasters' process, there appears to be an unofficial "culture" within the National Hurricane Center of declaring TC formation during daylight hours. Since systems can become tropical cyclones at any time of the day or night, this institutional bias should be accounted for or it could impact TC formation research and disaster preparations.

Institutional bias could be especially challenging for people to identify if they are working inside the specific institution or culture. Because the expectations are inherent in that culture, they may not think about the implications when collecting data or building a model, potentially leading to continued bias in the model. This is one of many places where diverse teams can help to address bias issues.

Human bias

Human bias is the second category of bias in our diagram (Fig. 1) because it also directly affects the later two categories yet it is itself influenced by the systemic and structural biases. We show this interaction with the blue circle around the entire human bias category (see Fig. 1).

[3] https://charts.ecmwf.int/

[4] https://mesonet.agron.iastate.edu/archive/

In a complex world, it is procedurally rational to seek satisfactory solutions (i.e., satisfice) rather than optimize (Simon 1990). Yet satisficing processes like pattern recognition and heuristic search can introduce bias. Our categorization of human biases relies on prior efforts to understand the ways in which mental shortcuts (heuristics) and mental models can result in biased judgments and decisions (Fischhoff and Broomell 2020; Kahneman et al. 1982), and efforts to catalog known biases (Arkes 1991; Benson and Manoogian 2016; Benson 2017).

We highlight four categories of human bias that represent ways in which information processing can lead to bias. There are many additional categories and ways of categorizing of human bias as presented in the references above; we chose these four overarching categories as most representative of biases likely to cause issues for AI developers for the Earth systems.

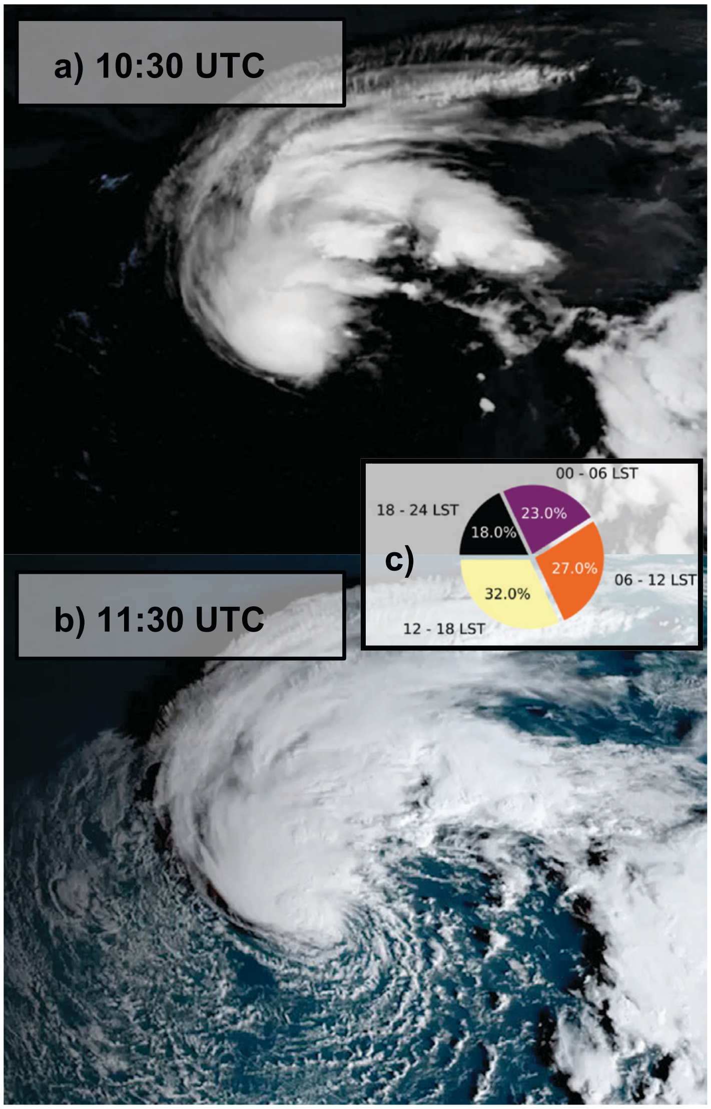

Fig. 2. CIRA Geocolor (Micke 2018) images of the formation of Hurricane Martin (2022). (a) Martin prior to tropical storm designation (1030 UTC 1 Nov 2022). (b) Martin after local sunrise (1130 UTC 1 Nov 2022). (c) Numbers of tropical cyclone formations declared by the National Hurricane Center at various local times from 2000 to 2022, as reported in the HURDAT2 dataset (Landsea and Franklin 2013).

Information overload. Developers and end-users of AI models both face an exponentially increasing complexity of data, which can lead to information overload. AI developers have an increasing amount of data to choose from in training the model. Identifying the best data source is a challenging task, even aside from all of the biases related to these choices that are discussed below (e.g., see selection and processing biases in the next section). Examples of such data sources include new sensors being launched and the increasing amount of data being shared online by government meteorological agencies.

Information overload can also come from increasing professional and societal pressures, such as expectations to keep up with the increasing volume of research (Bornmann et al. 2021) and data governance issues (Nelson and Office of Science and Technology Policy 2022). Institutions are also creating increasingly complex rules and expectations around AI, especially as AI is becoming more visible for weather and climate applications.

As with all of our subcategories, this one can also interact with other subcategories. Specifically, stress from sources other than information overload can also further constrain and bias a human developer's information processing abilities. When the human's ability to process information is constrained, it promotes quick decision strategies (Arkes 1991), which may not be the best choices for creating an unbiased AI model.

Attention. Attention determines what information goes into working memory and what information is filtered out. This directly interacts with information overload, as overload can affect the ability to attend to different pieces of information. For example, *confirmation bias* may result from the combined effects of information overload, sense-making processes, and selective attention.

Attention can be driven by exposure to information, physical format or context (e.g., motion and/or color; see Wolfe 2021), prior beliefs or mental models, personal motivations, and social norms. Attention can affect all parts of AI model development and deployment, including the interpretation and use of AI models by end-users. An example of this is attending to a single aspect of data quality, such as the time period the data cover, and paying less attention to spatial extent, representativeness of places experienced by specific population groups, or other aspects of data quality.

Working memory. Working memory has limited capacity and affects decisions by constraining what is considered and how it is considered at a given time (Baddeley et al. 2020). This can affect the development of the AI models as well as the deployment and use of them. For example, weather forecasters work with multiple sources of guidance in preparing forecasts, which they evaluate critically in their work. Increasing update frequencies and incorporating uncertainties in AI guidance may, at least initially, tax working memory, making it a challenge to track, synthesize, and critically evaluate the guidance. This could bias use of AI guidance toward more familiar or simpler inputs that tax working memory less, even beyond conscious biases. Demuth et al. (2020) states, "When [forecasters] cannot easily understand the workings of a probabilistic product or evaluate its accuracy, this reduces their trust in information and their willingness to use it."

Sense making. Humans have an inherent need to make sense of data, which can bias our judgments and choices. An example of this is our tendency to see patterns even in sparse data (Tversky and Kahneman 1971). As with the other types of human bias, this can affect the entire life cycle of development from data selection through model validation and interpretation. For example, developers may interpret graphical presentations of model verification statistics through preexisting graph schemas, i.e., the types of graphs with which they are familiar (Bancilhon et al. 2023).

Human biases can emerge in both individual and group judgments and decisions (Jones and Roelofsma 2000). Groups can enhance biases when there is social projection as in the case of false consensus (Mullen and Hu 1988; Fischhoff and Johnson 1996), and through discussion, which can produce groupthink (Tajfel 1982), group polarization, and group escalation of commitment. These sources of bias interact. For example, a developer may face many sources of information, some of which conflict, and be drawn to examine the information that is more salient or accessible but potentially less relevant, thereby inadvertently introducing bias into a model.

Both individuals and groups can also mediate biases, for example by considering the opposite hypothesis of why a judgment or decision might be wrong or by bringing attention to alternative viewpoints. Changing the decision environment for data selection and interpretation or for making AI modeling choices can also help reduce biases (Larrick 2004).

Data bias

Data bias is crucial to understand and address, as the data chosen for AI model training and validation will directly affect the bias of the final AI model. This is chosen as the third category in Fig. 1 since the choices of data occur before the model training itself. If the underlying data are biased from existing historical environmental or climate injustices, it is unlikely that the AI model will be able to address the injustices, and instead will likely perpetuate them.

We break data bias into four subcategories, arranged mostly in the order in which they occur in the AI life cycle. Both systemic and structural biases and human biases affect the data biases and all of the subcategories that we propose. We discuss this in more detail with the specific biases but it is important to understand that none of our proposed bias categories exist independently of one another. Understanding the full cycle of bias is critical to ensuring an AI model is as free of bias as possible.

Selection bias. The first step in training an AI model is to identify the data available and necessary for training. Selection bias is particularly affected by the systemic and structural biases that may exist in the environment. If the data do not even exist, it is impossible for AI developers to choose the data to train a model. Though we chose to color the interactions in Fig. 1 for selection bias only by systemic and structural biases, human biases can also strongly influence the choice of data for training, as it involves active choices on the part of the developer.

Sometimes AI developers want to choose all of the available data and let the AI model identify what is critically important. While some AI models can handle such large datasets, it is likely that this choice will create training data with strong correlations across the data, which can impede AI model learning and performance. Understanding the characteristics of the data are especially important for applying post-model interpretation techniques (Flora et al. 2024).

As an example in weather and climate, historical records of temperature are limited in both space and time and are often reconstructed from data where there could be additional biases in place. Figure 3 shows an example of selection bias for the task of rapid

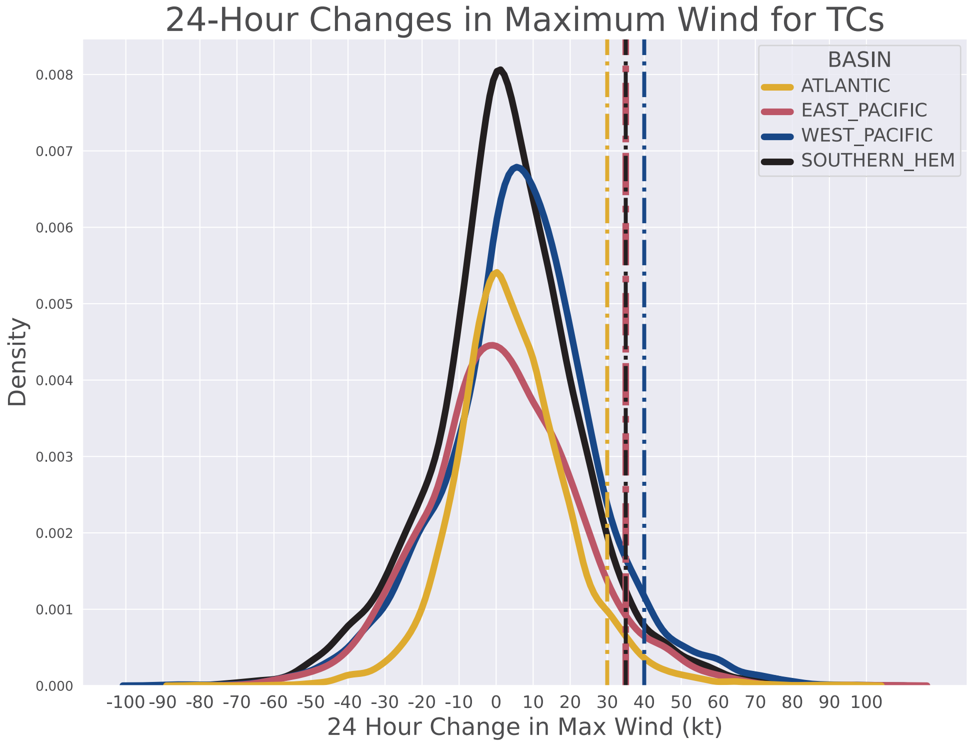

Fig. 3. Distribution of 24-h changes in maximum wind speed for tropical cyclones from 2005 to 2021, as identified by the International Best Track Archive for Climate Stewardship (IBTrACS), and separated by ocean basin. Dashed lines indicate the 95th percentile of maximum wind speed for each basin, which often serves as the threshold for defining rapid intensification (Kaplan and DeMaria 2003).

intensification of tropical cyclones. In this case, data differs by basin and a model that fails to account for this, e.g., by selecting Atlantic basin data for training, would not perform well globally. As an example in the broader Earth sciences, consider the task of detecting and predicting landslides. Developing an AI system to improve such predictions could save countless lives but the data are not available for many places where landslides are most likely to occur (Casagli et al. 2023).

Interpretation bias. Once the data have been chosen, the next step is to identify what type of interpretation the AI developer is putting on the data. Data in its raw form is rarely AI ready and there typically is a human layer of interpretation on the data to help prepare it for AI training. This interpretation is at the conceptual level and not at the implementation level, since the actual processing of the data would fit into the next subcategory (processing bias). However, this conceptual level is important as it influences what is actually implemented and processed.

To help make this more concrete, we provide several examples drawn from the Earth sciences. One example is processing satellite data. There are usually multiple channels available, and interpretation bias could lead to the choice of channel that poorly informs the modeling task. A second example is data available at the census scale. Here the developer must choose the level of aggregation, such as zip code or even more fine-grained criteria. This choice can have considerable consequences on the use of the data (e.g., Kenny et al. 2021; Lang and Pearson-Merkowitz 2022) A third example comes from rain gauge data. If the data are fine-grained enough, such as the 5-min data provided by mesonets (e.g., McPherson et al. 2007), the choice of how to aggregate that data to match coarser-grained data such as hourly radar estimated rainfall, could create biases in the training data.

Processing bias. Once the data are selected and the proposed interpretation is ready, they must be processed before use in AI training. This processing can both introduce or adjust for known biases. The processing step is often intertwined with the interpretation step, yet we separate them for clarity. The interpretation step focuses on the conceptual level of how data will be aggregated or combined while the processing step focuses on the implementation. It is possible that by choosing one interpretation or one method of processing data over another, that a bias toward one solution is either introduced or corrected.

For example, if data are subsampled, a skew toward a specific outcome could be intentionally or unintentionally produced. Subsampling, upsampling, and data augmentation are commonly used strategies in AI for addressing skewed datasets as well as datasets with missing data. Skewed data often arise in rare-event prediction tasks in weather and climate. Since ML models typically struggle to learn effective general models with highly skewed data, sampling approaches are a very common strategy to address such data. Sampling from a 99%/1% split (example: tornadoes, aircraft turbulence, and many more rare but impactful phenomena) to a more equitable 50%/50% split may create a model that can predict the rare class but it may also significantly overpredict the rare class. Sometimes AI developers will also subsample the testing data, thus reporting nonrepresentative results if the model were to be deployed.

If a dataset has missing data, synthetic data can be created and used to provide more information. For tropical cyclones, radar data are only available if the storm is within the coverage area; synthetic radar can be used to fill in some gaps in coverage. Similarly, microwave sensors on satellites (critical for observing precipitation and cloud structure in tropical cyclones) have a low temporal sampling rate. Synthetic microwave data can be generated at a higher temporal resolution, similar to that of geostationary satellites. Care must be taken when creating such synthetic data so that bias is not introduced into the dataset. For example, if synthetic radar data are used to fill in gaps in global coverage but it was trained only in one location such as the United States, it will not generalize well over the full globe.

In other cases, datasets can be smaller than is needed for training AI methods. In these cases, data augmentation strategies such as rotating or translating images (Lagerquist et al. 2020) or adding realistic noise patterns to synthetic images (Schreck et al. 2022, 2023) can provide useful synthetic new data. However, AI developers must be careful when applying the standard data augmentations techniques from computer vision as they could introduce additional biases by creating non-physically-realistic data. For example, a standard image flip of meteorological data changes the physical meaning of the data (flow could be reversed, Coriolis force might be opposite what is required for the hemisphere, etc.).

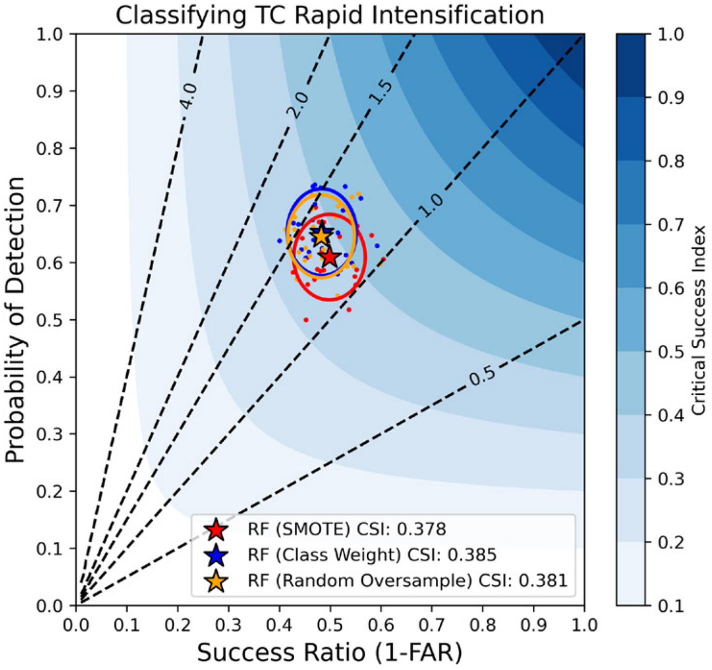

Fig. 4. Classifying tropical cyclone rapid intensification (RI) events from 2005 to 2021 using different random forest sampling strategies (class weighting, random oversampling, and SMOTE). The classification task was redone 25 times for each model using a bootstrapping approach. The stars indicate the mean values for each model across the 25 samples, and the ellipses show the 5th–95th percentile ranges for POD and SR.

In Earth sciences, rare or extreme events are often of particular interest. Effectively representing rare and extreme events in training datasets can be difficult. In Fig. 4, we provide an example of how common choices for addressing missing data (which fall into processing bias) affect model performance. Here we show how the results for a classification task of identifying a rare event (tropical cyclone rapid intensification) differ based on sampling strategy. The random forest models are otherwise identically constructed, with the only differences being choice of sampling algorithm (or the use of class weighting). We see in Fig. 4 that while each model has a similar overall critical success index (CSI), the synthetic minority oversampling technique (SMOTE) model more effectively reduces the overprediction bias (i.e., the SMOTE model has fewer false alarms), but at the expense of a lower probability of detection (POD) compared to the other two models.

Physics bias. Unique to Earth sciences are potential biases introduced by the laws of physics, which can limit data availability. Typically such limitations do not exist in traditional AI datasets, where one may be training an AI image recognition system from photographs. As a weather example, Fig. 1 of McGovern et al. (2022) highlighted the regions of the southeastern United States that had better and poorer areas of radar coverage. This coverage is limited by the laws of physics, in that radar beams are straight lines and the curvature of the Earth limits how far away they can sense phenomena near the surface.

We provide a satellite-based example in Fig. 5, which shows an example of Hurricane Dorian. Although a tropical cyclone is present continuously in time and space, and is constantly evolving, the data are available at nonregular temporal intervals. This is entirely due to the laws of physics as data can only be observed when the satellite passes over the tropical cyclone.

Additional examples of biases introduced by the laws of physics include satellite parallax and radar sampling issues. Parallax stems from the angle and height at which phenomena are

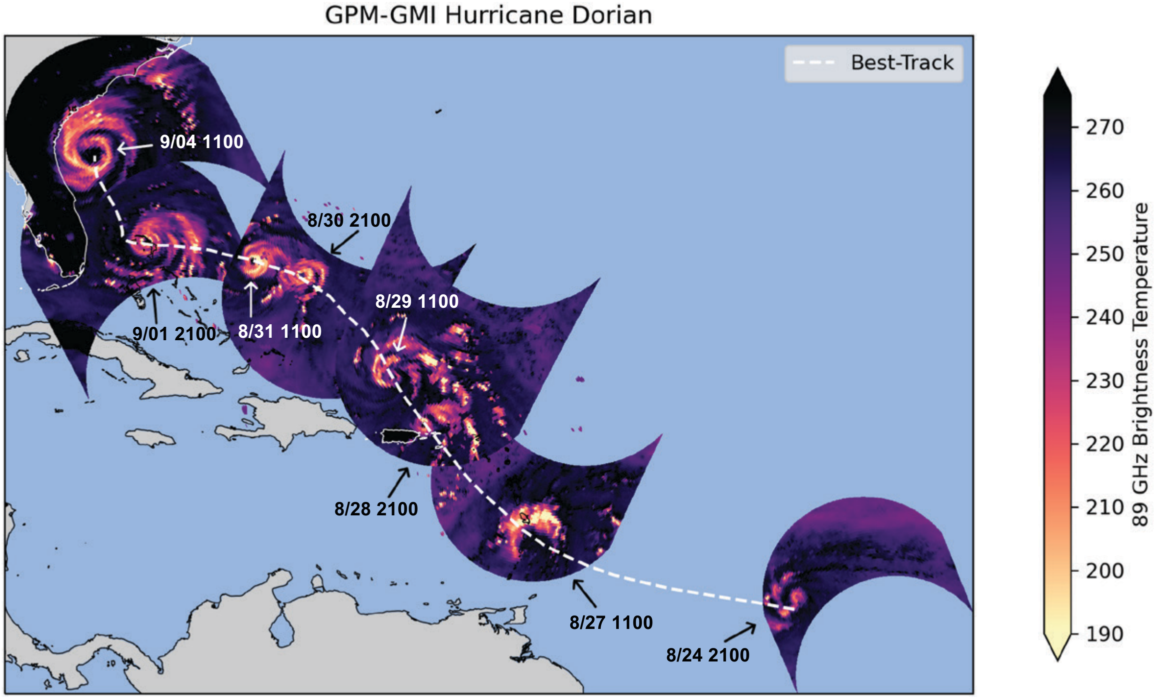

Fig. 5. Global Precipitation Measurement (GPM) mission Microwave Imager (GMI) observations of Hurricane Dorian. The images shown are 89-GHz brightness temperatures (horizontal polarization). Times of observations are annotated on the plot.

observed from a satellite. Although parallax can be corrected, it relies on an accurate cloud height retrieval which is imperfect. Similarly, radar beam effects (e.g., lower resolution at further range; height above ground increases with range) results in different radars observing different portions of potentially the same phenomena. The correction for either of these effects is challenging and will lead to a bias in most AI applications.

Physics bias can overlap with processing bias discussed above. While synthetic data could be introduced to address the missing data, one must be careful to introduce synthetic data that are free of bias itself, or the problem will continue to propagate. This overlaps with the discussion of data augmentation techniques.

Statistical and computational bias

The final category of bias that we identify in Fig. 1 focuses on the AI model itself. Although data bias can be the underlying cause of AI model bias, we identify three categories of bias within the statistical and computational steps of training the AI model. As before, we order these roughly in the order in which they occur in the AI development life cycle. As with other subcategories, these can interact with each other.

AI model bias. AI model bias can take two forms. The first is that the AI developer must choose the model(s) that they are going to apply to the task at hand. This choice can be influenced by human bias. For example, an AI developer may be biased against deep learning due to a perceived lack of interpretability, preferring more “traditional” AI models such as decision trees, when it could be that deep learning would produce better results on the task (or vice versa).

The second form of AI model bias comes from the AI model itself. One of the key issues in applying AI to Earth science tasks is that the AI models lack understanding of the laws of physics. While there is work on developing physics-based AI models (e.g., Lapuschkin et al. 2019; Kashinath et al. 2021), this is still in its infancy and most AI models are unconstrained by the true physics of a phenomenon, and can learn idiosyncrasies of the data. AI models that are purely data driven, such as Bi et al. (2023) and Lam et al. (2023), are successful in many situations but may struggle especially in situations outside of the training data distributions. A physics-based model such as a numerical weather prediction model would still be able to provide realistic answers in situations such as unprecedented heat waves.

Validation bias. Validation bias can also take multiple forms, all of which interact with the developer's initial choice of model and with human biases. First, model validation data and metrics are chosen by AI developers, thus interacting with human biases. An AI developer may cherry-pick the case studies to highlight some aspect of a model or may choose a metric that looks more favorable in certain situations. For example, a developer could choose accuracy instead of a more appropriate skill score when predicting a rare event. Choosing the right metric to measure the goodness of a weather forecast is challenging (Murphy 1993) and the same approach applies across the Earth sciences.

In addition to choosing the validation scores or case studies, the human researcher may also choose to validate the AI method using explainable AI (XAI) methods. However, these approaches have biases that an AI developer must account for (Mamalakis et al. 2023). Likewise, the results of XAI are often subject to human confirmation bias (Burnett 2020), where the developer may confirm their existing expectations of what the model learned and dismiss other parts as "noise" when they in fact can significantly affect the outcome of model deployment.

Use and interpretation bias. Once an AI model is trained, model developers must choose how to use, interpret, and potentially deploy the model. These choices intersect strongly with human biases. For example, the choice of metric and/or case studies for model evaluation may come from recent experience, a form of *recency bias* that humans are subject to. The metric may be chosen from a paper the developer most recently read but may not be the metric that best measures performance for this model's deployment. Likewise, a case study may be chosen from a recent high-impact phenomena, while neglecting additional use cases that should be studied before deployment.

Early deployment of nontrustworthy AI models can lead to significant downstream consequences. This was highlighted in our earlier paper (McGovern et al. 2022), where we discussed a model that predicted earthquake risk that was deployed and used too early. Another financial example comes from the insurance industry, which has lately been focused on revamping their risk models due to the changing climate. If a model is deployed without adequate adaptation to our changing risks and climates, risks may be underestimated and thus people may not be covered in cases of major disasters. Consequences of deploying and trusting a model that is biased could even include lives lost from overreliance on an underperforming model. This bias also interacts strongly with the validation bias.

Discussion and future work

Our goal with this paper is to create a categorization system for AI biases in the Earth sciences that will help AI developers recognize what types of bias they could encounter while creating new AI models. Creating a systematic approach for AI developers to identify biases is the first step toward measuring and mitigating biases. While the scope of the paper does not extend to mitigation strategies, this is a topic of current work by the authors. Such strategies are sufficiently complex that they warrant additional publications to cover the topics in sufficient details.

Understanding the types of bias that can appear throughout the AI life cycle is also critical to creating models that will be used to address environmental and climate justice issues as well as climate mitigation. As the climate is changing and high-impact phenomena change their distributions (IPCC 2022), it is important that we recognize the ways in which an AI model may unintentionally miss high-impact events or perpetuate environmental injustices due to systemic and structural biases (Fig. 1). In addition, an AI model that is unbiased and makes reliable and trustworthy predictions can help to address climate mitigation and adaptation. For example, in the weather domain, AI could be used to create synthetic radar for countries that do not have a full radar network, thus facilitating improved predictions of high-impact events such as floods, droughts, and severe storms (Veillette et al. 2018; Lagerquist et al. 2020; Hilburn et al. 2020). Another example is AI being used at a subseasonal scale to guide agricultural and water decisions (Sun and Scanlon 2019; White et al. 2022) or to address food security issues, which span the range from crop diversity to intelligent robots to help with sustainable agriculture practices, to drought prediction and minimizing food waste.[5]

Trustworthiness is a key focus of research in AI models, even included in the latest Executive Order on AI (The White House 2023). A key piece of trustworthiness is ensuring that AI models are as free of bias as possible. When an AI model is deployed, we would like it to be trusted because it is proving to be useful and is not creating any unexpected biases in the predictions. If the AI model developers are not sufficiently familiar with the domain they are working in nor aware of the potential biases of the data and the AI models, they may put unwarranted trust into the model (Jacovi et al. 2021), which can have significant downstream implications.

Future work on this topic will include focusing on the measurement and mitigation of risk. Our goal is to adapt the recent NIST AI Risk Management Framework (Tabassi 2023) to focus on guidance for developers of AI in the Earth sciences. Additional research is needed to help developers of AI in the Earth sciences identify and debias their work, perhaps in the form of guidelines complementary to those produced by NIST for AI risk management (Tabassi 2023). A promising approach is to build on metadata approaches such as RealML (Smith et al. 2022), datasheets (Pushkarna et al. 2022), and model cards.

Recent work such as Ball (2023) has discussed how AI is approaching a critical threshold of reproducibility. We want to strongly echo one piece of their advice: interdisciplinary and diverse teams are key to the eventual success of an AI model. By bringing together diverse teams with different viewpoints, it is much more likely that AI biases will be identified quickly and addressed before a model is deployed. The culture of developing AI models for any discipline needs to shift to one where all aspects of the ML system are documented and shared with both developers and users, which will help to create stronger, impactful, and less biased AI models.

Acknowledgments. This material is based upon work supported by the National Science Foundation under Grant ICER-2019758. This material is also based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement 1852977.

Data availability statement. Imagery from Fig. 2 is from Cooperative Institute for Research in the Atmosphere at Colorado State and the National Oceanic and Atmospheric Administration (CSU/CIRA and NOAA). Specifically, we leveraged the CIRA satellite library at https://satlib.cira.colostate.edu/. Figures 3 and 4 were both generated using the developmental dataset of the Statistical Hurricane Intensity Prediction Model (SHIPS) (DeMaria and Kaplan 1994). The dataset we used is publicly available via CIRA at https://rammb2.cira.colostate.edu/research/tropical-cyclones/ships/. GPM GMI data in Fig. 5 are available at https://doi.org/10.5067/GPM/GMI/GPM/1B/07.

[5] https://www.weforum.org/agenda/2022/04/ai-can-create-a-resilient-food-system-from-the-lab-to-the-field-to-the-dinner-table/

References

Almuzaini, A. A., C. A. Bhatt, D. M. Pennock, and V. K. Singh, 2022: ABCinML: Anticipatory bias correction in machine learning applications. *FAccT'22: 2022 ACM Conf. on Fairness, Accountability, and Transparency*, Seoul, South Korea, Association for Computing Machinery, 1552–1560, https://doi.org/10.1145/3531146.3533211.

Anbarci, N., J. Boyd, E. Floehr, J. Lee, and J. J. Song, 2011: Population and income sensitivity of private and public weather forecasting. *Reg. Sci. Urban Econ.*, **41**, 124–133, https://doi.org/10.1016/j.regsciurbeco.2010.11.001.

Arkes, H., 1991: Costs and benefits of judgment errors: Implications for debiasing. *Psychol. Bull.*, **110**, 486–498, https://doi.org/10.1037/0033-2909.110.3.486.

Baddeley, A., G. Hitch, and R. Allen, 2020: A multicomponent model of working memory. *Working Memory: The State of the Science*, Oxford University, 10–43.

Balagopalan, A., H. Zhang, K. Hamidieh, T. Hartvigsen, F. Rudzicz, and M. Ghassemi, 2022: The road to explainability is paved with bias: Measuring the fairness of explanations. *FAccT'22: 2022 ACM Conf. on Fairness, Accountability, and Transparency*, Seoul, South Korea, Association for Computing Machinery, 1194–1206, https://doi.org/10.1145/3531146.3533179.

Ball, P., 2023: Is AI leading to a reproducibility crisis in science? *Nature*, **624**, 22–25, https://doi.org/10.1038/d41586-023-03817-6.

Bancilhon, M., L. Padilla, and A. Ottley, 2023: Improving evaluation using visualization decision-making models: A practical guide. *Visualization Psychology*, Springer, 85–107.

Bauer, P., P. D. Dueben, T. Hoefler, T. Quintino, T. C. Schulthess, and N. P. Wedi, 2021: The digital revolution of Earth-System Science. *Nat. Comput. Sci.*, **1**, 104–113, https://doi.org/10.1038/s43588-021-00023-0.

Benjamin, R., 2019: *Race after Technology: Abolitionist Tools for the New Jim Code.* Polity Press, 172 pp.

Benson, B., 2017: Cognitive bias cheat sheet, simplified. Medium, 8 January, https://medium.com/thinking-is-hard/4-conundrums-of-intelligence-2ab78d90740f.

Benson, B., and J. Manoogian, 2016: Cognitive bias cheat sheet: An organized list of cognitive biases because thinking is hard. Medium, 1 September, https://betterhumans.pub/cognitive-bias-cheat-sheet-55a472476b18.

Bi, K., L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, 2023: Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, **619**, 533–538, https://doi.org/10.1038/s41586-023-06185-3.

Bornmann, L., R. Haunschild, and R. Mutz, 2021: Growth rates of modern science: A latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanit. Soc. Sci. Commun.*, **8**, 224, https://doi.org/10.1057/s41599-021-00903-w.

Brands, S., J. A. Fernández-Granja, J. Bedia, A. Casanueva, and J. Fernández, 2023: A global climate model performance atlas for the Southern Hemisphere extratropics based on regional atmospheric circulation patterns. *Geophys. Res. Lett.*, **50**, e2023GL103531, https://doi.org/10.1029/2023GL103531.

Buolamwini, J., 2023: *Unmasking AI: My Mission to Protect What Is Human in a World of Machines.* Random House Publishing Group, 336 pp.

Burnett, M., 2020: Explaining AI: Fairly? Well? *Proc. 25th Int. Conf. on Intelligent User Interfaces*, Cagliari, Italy, Association for Computing Machinery, 1–2, htps://doi.org/10.1145/3377325.3380623.

Casagli, N., E. Intrieri, V. Tofani, G. Gigli, and F. Raspini, 2023: Landslide detection, monitoring and prediction with remote-sensing techniques. *Nat. Rev. Earth Environ.*, **4**, 51–64, https://doi.org/10.1038/s43017-022-00373-x.

Chantry, M., S. Hatfield, P. Dueben, I. Polichtchouk, and T. Palmer, 2021: Machine learning emulation of gravity wave drag in numerical weather forecasting. *J. Adv. Model. Earth Syst.*, **13**, e2021MS002477, https://doi.org/10.1029/2021MS002477.

Coats, S., J. E. Smerdon, S. Stevenson, J. T. Fasullo, B. Otto-Bliesner, and T. R. Ault, 2020: Paleoclimate constraints on the spatiotemporal character of past and future droughts. *J. Climate*, **33**, 9883–9903, https://doi.org/10.1175/JCLI-D-20-0004.1.

DeMaria, M., and J. Kaplan, 1994: A Statistical Hurricane Intensity Prediction Scheme (ships) for the Atlantic basin. *Wea. Forecasting*, **9**, 209–220, https://doi.org/10.1175/1520-0434(1994)009<0209:ASHIPS>2.0.CO;2.

Demuth, J. L., and Coauthors, 2020: Recommendations for developing useful and usable convection-allowing model ensemble information for NWS forecasters. *Wea. Forecasting*, **35**, 1381–1406, https://doi.org/10.1175/WAF-D-19-0108.1.

D'Ignazio, C., and L. F. Klein, 2020: *Data Feminism.* MIT Press, 328 pp.

Emanuel, K., 2008: The hurricane–climate connection. *Bull. Amer. Meteor. Soc.*, **89**, ES10–ES20, https://doi.org/10.1175/BAMS-89-5-Emanuel.

Eubanks, V., 2018: *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* St. Martin's Press, Inc., 260 pp.

Fischhoff, B., and S. Johnson, 1996: Organizational decision making. *The Possibility of Distributed Decision Making*, Cambridge University Press, 216–237.

——, and S. B. Broomell, 2020: Judgment and decision making. *Annu. Rev. Psychol.*, **71**, 331–355, https://doi.org/10.1146/annurev-psych-010419-050747.

Flora, M. L., C. K. Potvin, A. McGovern, and S. Handler, 2024: A machine learning explainability tutorial for atmospheric sciences. *Artif. Intell. Earth Syst.*, **3**, e230018, https://doi.org/10.1175/AIES-D-23-0018.1.

Friedman, B., and H. Nissenbaum, 1996: Bias in computer systems. *ACM Trans. Inf. Syst.*, **14**, 330–347, https://doi.org/10.1145/230538.230561.

Haiden, T., M. Janousek, F. Vitart, Z. B. Bouallegue, L. Ferranti, F. Prates, and D. Richardson, 2021: Evaluation of ECMWF forecasts, including the 2021 upgrade. ECMWF Tech. Memo. 884, 54 pp., https://doi.org/10.21957/90pgicjk4.

Harvey, C., 2023: Weather warning inequity: Lack of data collection stations imperils vulnerable people. *Scientific American*, 5 July, https://www.scientificamerican.com/article/weather-warning-inequity-lack-of-data-collection-stations-imperils-vulnerable-people/.

Henry, P. J., 2010: Institutional bias. *The Sage Handbook of Prejudice, Stereotyping and Discrimination*, J. F. Dovidio et al., Eds., SAGE Publications Ltd, 426–440, https://doi.org/10.4135/9781446200919.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, https://doi.org/10.1002/qj.3803.

Hilburn, K. A., I. Ebert-Uphoff, and S. D. Miller, 2020: Development and interpretation of a neural-network-based synthetic radar reflectivity estimator using GOES-R satellite observations. *J. Appl. Meteor. Climatol.*, **60**, 3–21, https://doi.org/10.1175/JAMC-D-20-0084.1.

IPCC, 2022: *Climate Change 2022: Impacts, Adaptation and Vulnerability.* H.-O. Pörtner and D. Belling, Eds., Cambridge University Press, 3056 pp.

Jacovi, A., A. Marasović, T. Miller, and Y. Goldberg, 2021: Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. *FAccT'21: Proc. 2021 ACM Conf. on Fairness, Accountability, and Transparency*, Online, Association for Computing Machinery, 624–635, https://doi.org/10.1145/3593013.3593986.

Jones, P. E., and P. H. Roelofsma, 2000: The potential for social contextual and group biases in team decision-making: Biases, conditions and psychological mechanisms. *Ergonomics*, **43**, 1129–1152, https://doi.org/10.1080/00140130050084914.

Kahneman, D., P. Slovic, and A. Tversky, 1982: *Judgment under Uncertainty: Heuristics and Biases.* Cambridge University Press, 55 pp.

Kantayya, S., 2020: *Coded Bias*. 7th Empire Media, https://www.7thempiremedia.com/films-codedbias.

Kaplan, J., and M. DeMaria, 2003: Large-scale characteristics of rapidly intensifying tropical cyclones in the North Atlantic basin. *Wea. Forecasting*, **18**, 1093–1108, https://doi.org/10.1175/1520-0434(2003)018<1093:LCORIT>2.0.CO;2.

Kashinath, K., and Coauthors, 2021: Physics-informed machine learning: Case studies for weather and climate modelling. *Philos. Trans. Roy. Soc.*, **A379**, 20200093, https://doi.org/10.1098/rsta.2020.0093.

Kenny, C. T., S. Kuriwaki, C. McCartan, E. T. R. Rosenman, T. Simko, and K. Imai, 2021: The use of differential privacy for census data and its impact on redistricting: The case of the 2020 U.S. Census. *Sci. Adv.*, **7**, eabk3283, https://doi.org/10.1126/sciadv.abk3283.

Labe, Z. M., and E. A. Barnes, 2021: Detecting climate signals using explainable AI with single-forcing large ensembles. *J. Adv. Model. Earth Syst.*, **13**, e2021MS002464, https://doi.org/10.1029/2021MS002464.

Lagerquist, R., A. McGovern, C. Homeyer, D. Gagne, and T. Smith, 2020: Deep learning on three-dimensional multiscale data for next-hour tornado prediction. *Mon. Wea. Rev.*, **148**, 2837–2861, https://doi.org/10.1175/MWR-D-19-0372.1.

Lam, R., and Coauthors, 2023: Learning skillful medium-range global weather forecasting. *Science*, **382**, 1416–1421, https://doi.org/10.1126/science.adi2336.

Landsea, C., and J. Franklin, 2013: Atlantic hurricane database uncertainty and presentation of a new database format. *Mon. Wea. Rev.*, **141**, 3576–3592, https://doi.org/10.1175/MWR-D-12-00254.1.

Lang, C., and S. Pearson-Merkowitz, 2022: Aggregate data yield biased estimates of voter preferences. *J. Environ. Econ. Manage.*, **111**, 102604, https://doi.org/10.1016/j.jeem.2021.102604.

Lapuschkin, S., S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, 2019: Unmasking Clever Hans predictors and assessing what machines really learn. *Nat. Commun.*, **10**, 1096, https://doi.org/10.1038/s41467-019-08987-4.

Larrick, R. P., 2004: Debiasing. *Blackwell Handbook of Judgment and Decision Making*, John Wiley & Sons, Ltd, 316–338, https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470752937.ch16.

Mamalakis, A., E. A. Barnes, and I. Ebert-Uphoff, 2023: Carefully choose the baseline: Lessons learned from applying XAI attribution methods for regression tasks in geoscience. *Artif. Intell. Earth Syst.*, **2**, e220058, https://doi.org/10.1175/AIES-D-22-0058.1.

March, J. G., and J. P. Olsen, 1984: The new institutionalism: Organizational factors in political life. *Amer. Political Sci. Rev.*, **78**, 734–749, https://doi.org/10.2307/1961840.

McGovern, A., K. Elmore, D. Gagne, S. Haupt, C. Karstens, R. Lagerquist, T. Smith, and J. Williams, 2017: Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bull. Amer. Meteor. Soc.*, **98**, 2073–2090, https://doi.org/10.1175/BAMS-D-16-0123.1.

——, I. Ebert-Uphoff, D. J. Gagne, and A. Bostrom, 2022: Why we need to focus on developing ethical, responsible, and trustworthy artificial intelligence approaches for environmental science. *Environ. Data Sci.*, **1**, e6, https://doi.org/10.1017/eds.2022.5.

McPherson, R. A., and Coauthors, 2007: Statewide monitoring of the mesoscale environment: A technical update on the Oklahoma Mesonet. *J. Atmos. Oceanic Technol.*, **24**, 301–321, https://doi.org/10.1175/JTECH1976.1.

Mendez, C., and Coauthors, 2018: Open source barriers to entry, revisited: A sociotechnical perspective. *ICSE'18: Proc. 40th Int. Conf. on Software Engineering*, Gothenburg, Sweden, Association for Computing Machinery, 1004–1015, https://doi.org/10.1145/3180155.3180241.

Menne, M. J., C. N. Williams Jr., and M. A. Palecki, 2010: On the reliability of the U.S. surface temperature record. *J. Geophys. Res.*, **115**, D11108, https://doi.org/10.1029/2009JD013094.

Micke, K., 2018: Every pixel of GOES-17 imagery at your fingertips. *Bull. Amer. Meteor. Soc.*, **99**, 2217–2219, https://doi.org/10.1175/BAMS-D-17-0272.1.

Morss, R. E., J. K. Lazo, and J. L. Demuth, 2010: Examining the use of weather forecasts in decision scenarios: Results from a US survey with implications for uncertainty communication. *Meteor. Appl.*, **17**, 149–162, https://doi.org/10.1002/met.196.

Mullen, B., and L. Hu, 1988: Social projection as a function of cognitive mechanisms: Two meta-analytic integrations. *Br. J. Soc. Psychol.*, **27**, 333–356, https://doi.org/10.1111/j.2044-8309.1988.tb00836.x.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293, https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

Nelson, A., and Office of Science and Technology Policy, 2022: Ensuring free, immediate, and equitable access to federally funded research. Executive Office of the President of the United States Tech. Rep., 8 pp., https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf.

O'Neil, C., 2016: *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown Publishing Group, 272 pp.

Pacchetti, M. B., S. Dessai, S. Bradley, and D. A. Stainforth, 2021: Assessing the quality of regional climate information. *Bull. Amer. Meteor. Soc.*, **102**, E476–E491, https://doi.org/10.1175/BAMS-D-20-0008.1.

Pappenberger, F., E. Stephens, J. Thielen, P. Salamon, D. Demeritt, S. J. van Andel, F. Wetterhall, and L. Alfieri, 2013: Visualizing probabilistic flood forecast information: Expert preferences and perceptions of best practice in uncertainty communication. *Hydrol. Processes*, **27**, 132–146, https://doi.org/10.1002/hyp.9253.

Peng, K., A. Mathur, and A. Narayanan, 2021: Mitigating dataset harms requires stewardship: Lessons from 1000 papers. arXiv, 2108.02922v2, https://doi.org/10.48550/arXiv.2108.02922.

Peterson, T. C., and R. S. Vose, 1997: An overview of the Global Historical Climatology Network temperature database. *Bull. Amer. Meteor. Soc.*, **78**, 2837–2850, https://doi.org/10.1175/1520-0477(1997)078<2837:AOOTGH>2.0.CO;2.

Pushkarna, M., A. Zaldivar, and O. Kjartansson, 2022: Data cards: Purposeful and transparent dataset documentation for responsible AI. *FAccT'22: 2022 ACM Conf. on Fairness, Accountability, and Transparency*, Seoul, South Korea, Association for Computing Machinery, 1776–1826, https://doi.org/10.1145/3531146.3533231.

Reichstein, M., G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat, 2019: Deep learning and process understanding for data-driven Earth system Science. *Nature*, **566**, 195–204, https://doi.org/10.1038/s41586-019-0912-1.

Ripberger, J., A. Bell, A. Fox, A. Forney, W. Livingston, C. Gaddie, C. Silva, and H. Jenkins-Smith, 2022: Communicating probability information in weather forecasts: Findings and recommendations from a living systematic review of the research literature. *Wea. Climate Soc.*, **14**, 481–498, https://doi.org/10.1175/WCAS-D-21-0034.1.

Saltikoff, E., and Coauthors, 2019: An overview of using weather radar for climatological studies: Successes, challenges, and potential. *Bull. Amer. Meteor. Soc.*, **100**, 1739–1752, https://doi.org/10.1175/BAMS-D-18-0166.1.

Schneider, T., S. Lan, A. Stuart, and J. Teixeira, 2017: Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations. *Geophys. Res. Lett.*, **44**, 12 396–12 417, https://doi.org/10.1002/2017GL076101.

Schreck, J. S., G. Gantos, M. Hayman, A. Bansemer, and D. J. Gagne, 2022: Neural network processing of holographic images. *Atmos. Meas. Tech.*, **15**, 5793–5819, https://doi.org/10.5194/amt-15-5793-2022.

——, M. Hayman, G. Gantos, A. Bansemer, and D. J. Gagne, 2023: Mimicking non-ideal instrument behavior for hologram processing using neural style translation. arXiv, 2301.02757v1, https://doi.org/10.48550/arXiv.2301.02757.

Schwartz, R., A. Vassilev, K. Greene, L. Perine, A. Burt, and P. Hall, 2022: Towards a standard for identifying and managing bias in artificial intelligence. NIST Tech. Rep. 1270, 86 pp., https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf.

Simon, H. A., 1990: Invariants of human behavior. *Annu. Rev. Psychol.*, **41** (1), 1–20, https://doi.org/10.1146/annurev.ps.41.020190.000245.

Smith, J. J., S. Amershi, S. Barocas, H. Wallach, and J. Wortman Vaughan, 2022: REAL ML: Recognizing, exploring, and articulating limitations of machine learning research. *FAccT'22: 2022 ACM Conf. on Fairness, Accountability, and Transparency*, Seoul, South Korea, Association for Computing Machinery, 587–597, https://doi.org/10.1145/3531146.3533122.

Stall, S., and Coauthors, 2023: Ethical and responsible use of AI/ML in the earth, space, and environmental sciences. *ESS Open Archive*, https://doi.org/10.22541/essoar.168132856.66485758/v1.

Sun, A. Y., and B. R. Scanlon, 2019: How can big data and machine learning benefit environment and water management: A survey of methods, applications, and future directions. *Environ. Res. Lett.*, **14**, 073001, https://doi.org/10.1088/1748-9326/ab1b7d.

Suresh, H., and J. Guttag, 2021: A framework for understanding sources of harm throughout the machine learning life cycle. *EAAMO'21: Proc. First ACM Conf. on Equity and Access in Algorithms, Mechanisms, and Optimization*, Online, Association for Computing Machinery, 9 pp., https://doi.org/10.1145/3465416.3483305.

Tabassi, E., 2023: Artificial intelligence risk management framework (AI RMF 1.0). Tech. Rep. NIST AI 100-1, 48 pp., https://doi.org/10.6028/NIST.AI.100-1.

Tajfel, H., 1982: Social psychology of intergroup relations. *Annu. Rev. Psychol.*, **33** (1), 1–39, https://doi.org/10.1146/annurev.ps.33.020182.000245.

The White House, 2023: Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. Tech. Rep. 2023-24283, E.O. 14110 of Oct 30, 2023, 88 FR 75191, Executive Office of the President, 36 pp., https://www.federalregister.gov/d/2023-24283.

Tsai, W.-P., D. Feng, M. Pan, H. Beck, K. Lawson, Y. Yang, J. Liu, and C. Shen, 2021: From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. *Nat. Commun.*, **12**, 5988, https://doi.org/10.1038/s41467-021-26107-z.

Tversky, A., and D. Kahneman, 1971: Belief in the law of small numbers. *Psychol. Bull.*, **76**, 105–110, https://doi.org/10.1037/h0031322.

Vecchi, G. A., and T. R. Knutson, 2011: Estimating annual numbers of Atlantic hurricanes missing from the HURDAT database (1878–1965) using ship track density. *J. Climate*, **24**, 1736–1746, https://doi.org/10.1175/2010JCLI3810.1.

Veillette, M. S., E. P. Hassey, C. J. Mattioli, H. Iskenderian, and P. M. Lamey, 2018: Creating synthetic radar imagery using convolutional neural networks. *J. Atmos. Oceanic Technol.*, **35**, 2323–2338, https://doi.org/10.1175/JTECH-D-18-0010.1.

White, C. J., and Coauthors, 2022: Advances in the application and utility of subseasonal-to-seasonal predictions. *Bull. Amer. Meteor. Soc.*, **103**, E1448–E1472, https://doi.org/10.1175/BAMS-D-20-0224.1.

Wolfe, J. M., 2021: Guided Search 6.0: An updated model of visual search. *Psychon. Bull. Rev.*, **28**, 1060–1092, https://doi.org/10.3758/s13423-020-01859-9.

World Weather Attribution, 2023: Limited data prevent assessment of role of climate change in deadly floods affecting highly vulnerable communities around Lake Kivu. 29 June, https://www.worldweatherattribution.org/limited-data-prevent-assessment-of-role-of-climate-change-in-deadly-floods-affecting-highly-vulnerable-communities-around-lake-kivu/.

Yilmaz, M., 2023: Accuracy assessment of temperature trends from ERA5 and ERA5-Land. *Sci. Total Environ.*, **856**, 159182, https://doi.org/10.1016/j.scitotenv.2022.159182.

Zhang, Q., W. Zhang, X. Wu, J. Zhang, W. Kuang, and X. Si, 2022: Deep learning for efficient microseismic location using source migration-based imaging. *J. Geophys. Res. Solid Earth*, **127**, e2021JB022649, https://doi.org/10.1029/2021JB022649.