The Case of the Mislabeled Axis
On the difference between trustworthiness and accuracy

Corey Dethier

[Preprint; forthcoming in *Philosophy of Science*. Please contact at corey.dethier[at]gmail.com before citing or quoting.]

Abstract

What makes a graph good or bad? I examine a controversial graph of climate model accuracy to motivate a partial answer to this question. The graph has two main problems: visual distance is supposed to measure model accuracy but does not, and the graph is better understood as depicting model accuracy in a possible world than in the actual one. I argue that these two problems are indicative insofar they are not primarily a function of the graph's truth conditions. Generally: the features that make a graph trustworthy or epistemically valuable depart systematically from those that make a graph accurate.

0 Introduction

In testimony before congress in 2015 and 2016, John Christy argued that the consensus view within climate science—that the earth is warming and we're responsible—is false. Unsurprisingly, Christy's testimony was controversial. For one thing, his argument relied on evidence from a single layer of the atmosphere—the mid-troposphere—that has traditionally been the layer that the models struggle the most to accurately model. Christy's critics therefore accused him of cherry-picking the worst data set to focus on. There were also concerns about the data set and Christy's treatment of it; concerns that had

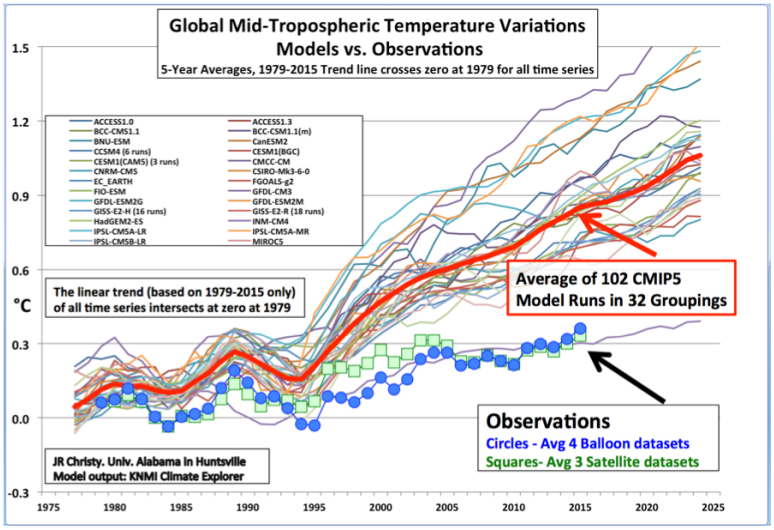


Figure 1: A comparison of model predictions and observations for the mid-troposphere between 1979 and 2015. From Christy (n.d.).

been documented in both the scientific (e.g., Santer et al. 2008, 2009) and philosophical (Lloyd 2012) literature prior to Christy's testimony.

It wasn't just what Christy said that was controversial—it was also the graphs that he presented. One example is given in figure 1. What's notable about the graph is the large gap between the model predictions (the various lines) and the observations (the dots). This gap appears to be evidence that the models are inaccurate. It isn't, however—or, at least, it isn't *good* evidence. As critics have pointed out, the size of the gap is an artefact—a function of Christy's choices about how to graph the information—rather than something that can be found in the data. Indeed, as Gavin Schmidt (2016) shows, you can graph the same data in such a way that the gap entirely disappears.

In this talk, I use Christy's graph to motivate deeper conclusions about the epistemology of visual representations like graphs. I begin by arguing that there are two important problems with treating Christy's graph as evidence for the inaccuracy of contemporary climate models. First, doing so requires us to treat visual distance as a (reliable) measure of accuracy in a context in which it is not plausibly treated as such. Second, doing so misconstrues the nature of the question that Christy's graph actually answers: given his representational choices, his graph doesn't represent how accurate the models *are*, but how accurate they *would be* in a counterfactual context. I then generalize: after offering a sketch graphical truth conditions, I argue that the examples are

representative in that the features that make a graph epistemically trustworthy or epistemically valuable depart systematically from those that make the graph accurate.

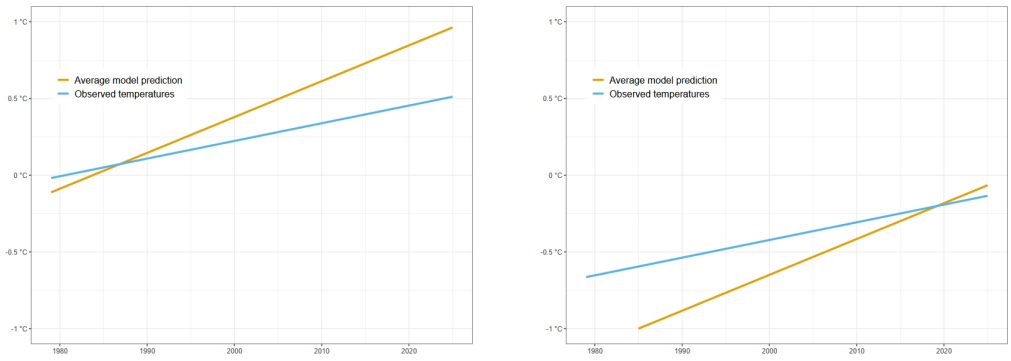
1 Meaningless distances

To understand the first problem with Christy's graph, it's important to recognize that climate model predictions concern relative rather than absolute temperatures. That is, model predictions typically concern changes in temperature or anomalies—deviations from some set baseline. Before we can evaluate the ability of the model to simulate the climate, therefore, we first need to fix a baseline. Here's a typical example. First, scientists run a simulation using the relevant model without introducing external "forcings"—factors like increases in CO_2—until the simulated climate stabilizes. The state of the model at this juncture is taken to represent the pre-industrial period, which usually operationalizes as something like the period 1850-1900. Second, the models are run with the forcing introduced, and the difference between the "forced" temperatures and the stable baseline taken to represent the actual temperature changes observed since the industrial revolution. So, e.g., a 1-degree anomaly relative to the average temperature of an unforced simulation represents a 1-degree anomaly relative to the average temperature recorded between 1850 and 1900.

A similar procedure is required if we are to *visually* compare model outputs to observations. In Christy's graph, the relevant time period is 1979-2015; with ten more years of data, we can extend the comparison to 2025. Do we take 1979 to be the baseline, so that a 1-degree anomaly relative to the average model output for 1979 represents a 1-degree anomaly relative to the average temperature in 1979? Do we take 2025 to be the baseline, so that a 1-degree anomaly relative to the average model output for 2025 represents a 1-degree anomaly relative to the average temperature in 2025? Do we use some other baseline, like the average of the whole period? These choices have dramatic effects on the appearance of the graph, as we can see in figure 2, which compares the first two options using just the trend lines for simplicity.

Notice something about these two graphs—they (appear to) imply very different things about whether the models are getting more accurate or less. The graph on the left appears to say that the models used to be very accurate but they're getting much worse. The one on the right appears to say that the models used to be very inaccurate but they're getting much better.

These are illusions. Or, more precisely, they're artefacts. This data doesn't



(a) 1979 baseline.

(b) 2025 baseline.

Figure 2: Two visual comparisons of model predictions and observations for the mid-troposphere between 1978 and 2025 using different baselines. Observational data from Spencer (2025); model information from Scoccimarro and Gualdi (2014).

tell us whether the models are getting better or worse. The lines on these graphs are just trend lines; what we're seeing are just the differences in the rate of change of the model predictions as opposed to the observations. The same is true in Christy's graph: the gap or space between the models and observations at any particular point is a function of the choice of baseline.

To emphasize the point: the problem here is not that Christy has chosen the wrong baseline (although more on that momentarily). Indeed, the problem is not even that it's hard (or impossible) to evaluate what the visual distance between the two lines means without knowing how wide the error bands are. For notice that introducing error bands into the graph doesn't actually address the problem so long as the choice of baseline is unconstrained: if the distance between the lines is influenced by the choice of baseline, then so too would be the space covered by the error bands. Instead, the problem is simply that so long as the choice of baseline is unconstrained, visual distance is not a reliable measure of model accuracy. Drawing conclusions about the latter from the former is just not warranted.

In other contexts, visual distance is a reliable indicator—sometimes even an extremely reliable one. And it could be rendered one here, given (a) a privileged baseline and (b) error bands that indicate how likely the observed temperatures are according to the model.[1] Without those elements, though,

[1]If we treat the chosen baseline as an estimate of some unknown "true" baseline, we could treat any deviation between the chosen and true baseline as a "measurement error"

visual distance is simply not a reliable measure of model accuracy; it's too heavily influenced by factors that have nothing to do with how well the models represent their target. Which means that the graph is not good evidence that the models are (in)accurate. It is not trustworthy or epistemically valuable, at least not with respect to the question that we're interested in asking.

2 A mislabeled axis

The complaint made in the last section slightly misrepresents Christy's approach. Christy's graph doesn't actually use a baseline, at least not in the traditional sense. If you look at the box on the left side of the graph, you will find the sentence "The linear trend (based on 1979-2015 only) of all time series intersects at zero in 1979." You might naturally read this as an empirical claim, something we learn when we find the line of best fit. It's not. It's a methodological stipulation.

What this stipulation means is the following. Rather than picking a year like 1979 and then taking a 1-degree anomaly relative to the average model output for 1979 to represent a 1-degree anomaly relative to the average temperature in 1979, Christy does something that is subtly but importantly different. Described intuitively, his method involves calculating the slope of each line, placing the virtual analogue of his pen at (1979,0), and then drawing the lines starting from there. Less intuitively: he finds the line of best fit and then shifts his data so that the line intersects the x-axis in 1979. Schmidt (2016) notices this, commenting "To my knowledge this is a unique technique and I'm not even clear on how one should label the y-axis." As the title of this section indicates, I think this is an understatement: the result of Christy's method is that his axes are simply mislabeled.

Consider how the choice of baseline affects the meaning of a particular mark on the graph. If the baseline is 1979, placing a marker at the point at (1980, .239) means that in the year 1980, there was a .239°C departure from the temperature in 1979. Importantly, when we fix 1979 as the baseline, we fix both the models and the observations in the same way. In a sense, you can think of a baseline as the "perspective" that we're viewing the information from. When 1979 is the baseline, you are viewing both the data actually recorded and

in the technical sense that phrase has in statistics (see Carroll et al. 2006). Notably, on this approach, the more years go into estimating the baseline, the narrower the error bands, which would obviate the problem raised in the prior paragraph—you couldn't muck around with extreme baselines without a concomitant change to the error bands. Unfortunately, without some privileged "true" baseline, the point is moot.

the models' actual predictions from that perspective. And while there's not a privileged perspective that is inherently preferable to all others, the different perspectives are different views of the same information—namely, what the actual numbers are.

The same is not true with Christy's method. In his graph, placing a point at $(1980, .239)$ means that in 1980 there was a .239°C departure from what is essentially an arbitrary constant—namely, the value taken on by the *trend line* in 1979. Notably, the value taken on by the trend line of the observations in 1979 is not the same as the temperature actually measured in 1979; the two have only a tenuous relationship. The same is true for the value of the trend line of the models. Further, the two trend lines will yield different arbitrary constants, and these two arbitrary constants will have different relationships with the values actually observed and predicted for 1979.

In other words, the dots that represent observations in Christy's graphs do not represent the actual observations; there's no year or even average of years such that we can say that the dot picks out the anomaly or deviation relative to that year. Instead, they represent the temperatures that we would have observed in a possible world where we observed the temperatures picked out by the shifted line of best fit. Similarly, the lines that represent the models do not represent predictions of the actual models. Instead, they represent the temperatures that the models would have predicted in a possible world where we predicted the temperatures picked out by the shifted line of best fit. Crucially, not only are the two shifts not the same, there's no reason to think that the two shifts go together in any sense—it's not like the model predictions he uses are the predictions that we would have gotten had we observed the temperatures that he uses.

For these reasons, I think the best way to view the situation is to see Christy's graph as answering a fundamentally different question than the question of how accurate the models actually are. That is, even if we ignore the problems of the last section, Christy's graph only gives us direct insight into how accurate the models are in an alternative possible world where we observed and predicted different temperatures than were observed and predicted in the actual world. And while in principle we *can* work backwards, no one can do that in their head, let alone reasonably reconstruct what the graph would look like if we depicted the actual temperatures instead.

Why, given the above concern, describe Christy's axis (or axes) as mislabeled? Because it is at minimum misleading and arguably simply incorrect to call the y-axis "temperatures," let alone temperatures in °C, without indicating that these are possible-world temperatures or temperatures shifted by an arbitrary amount. What Christy has done is not different in principle

from shifting every model prediction up by 100°C and observation down by the same amount, and while the box on the graph announces his method, that doesn't render the values that the graph actually depicts any more worthy of being labeled temperatures in °C.

In any case, the takeaway from this section is another reason for thinking that Christy's graph is not trustworthy or epistemically valuable with respect to the question of how accurate the models are. It looks like it answers that question, but on close inspection it actually answers a different one.

3 Graphical accuracy

What kinds of general conclusions can we draw from the two problems identified above? In the rest of this paper, I'll argue that the foregoing examples are representative in at least one respect: in our discussion of the problems that undercut the trustworthiness or epistemic value of Christy's graph, at no point did we discuss whether the graph was "accurate" or "true." That is, the features that make a graph trustworthy or epistemically valuable systematically depart from those that make it accurate. But to see that, we'll need more of an account of both aspects. This section addresses the question of what makes a graph accurate.[2] The next considers what makes one trustworthy.

At least on first pass, the most basic distinction in graphing is between *marks* and *frames*. A mark is an object like a dot, line, or shape. A frame consists of elements like axes, tick marks, grid lines, etc. A graph is meaningful—think *well-formed*—when a frame is properly applied to (a collection of) marks. And while both marks and frames are necessary for a graph to be meaningful, these two elements play very different roles in determining the meaning of a graph.

At this level of abstraction, the syntax of a graph is much like the syntax of a language: the mark-frame distinction mirrors the subject-predicate distinction. On closer inspection, however, the analogy starts to break down. In particular, a frame is not like an individual predicate or relation term, that we apply to a fixed number of objects to form a single proposition. Instead, when we apply a frame to a set of marks, it's more like repeatedly applying the same predicate or relation to each of the marks individually, generating a single proposition each time. The truth conditions of the graph as a whole

[2]Essentially, I'll be offering an account of graphical truth conditions. Previous philosophical work on this subject includes Dethier (2025), Irving (2011), Kulvicki (2010), and Perini (2005, 278–79), though all touch on the questions addressed here only in passing. For a more thorough treatment than I offer here, see Dethier (manuscript).

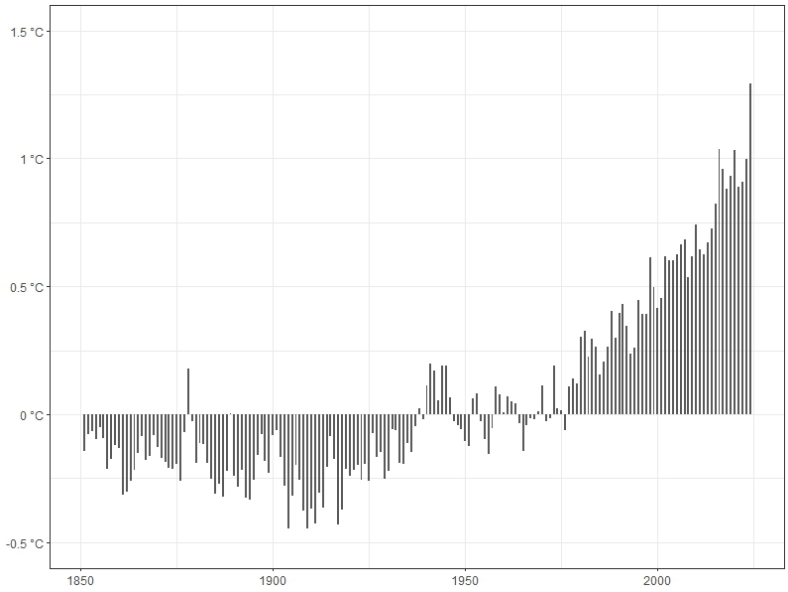


Figure 3: Mean global temperatures between 1851 and 2024, with a baseline of the 20th Century average. Data from National Centers for Environmental Information (2024).

are then a function of the truth conditions of these individual messages. In a slogan, graphs are more like a code than a cypher: each mark on the graph is a meaningful element of its own.

A basic account of the truth conditions of a graph—the semantic content, in other words—follows naturally from this slogan. Take figure 3, for example. Both the predicate-like placement within the frame and the graphical marks have meanings. A mark placed at the location (1979, .12) represents both the year 1979 and the temperature .12°C. In turn, the frame of figure 3 defines a relation—the mean global temperature being y above the 20th century average in year x. Putting the two together yields the proposition that these two objects stand in that relation. So: the semantic content of each mark-in-a-frame is determined compositionally by the meaning of the mark and the meaning of the properties and relations defined by the frame. The content of the graph is then, at minimum, the conjunction of the contents of each mark-in-the-frame.

This basic account leaves many important questions unanswered. In particular, while it's clear that propositions like

(1979) In 1979, the mean global temperature was 0.12°C above the 20th century average.

are part of the semantic content of figure 3, it's less clear whether either of the

following are:

(NEG) In 1979, the mean global temperature was *not* 0.90°C above the 20th century average.

(DIF) The mean global temperature was 1.43°C higher in 2024 than it was in 1851.

Say that the set of propositions corresponding to the individual marks on the graph—propositions like (1979)—is the graph's *primary content*. Both (NEG) and (DIF) are in some sense consequences of the primary content. But they are consequences in an importantly different sense.

We can already see that a single graph (typically) has many propositions as its content; in this respect, graphs are more like models than sentences. It thus strikes me as natural to say that the semantic content of the graph is just any proposition that the graph satisfies in something like the familiar logical sense. Importantly, that means that (NEG) is "part of" what figure 3 says: by failing to place a mark at $(1979, .90)$, figure 3 fails to satisfy the sentence "In 1979, the mean global temperature was 0.90°C above the 20th century average." Since (NEG) is the negation of that sentence, figure 3 thus satisfies (NEG). That's not something we infer from the graph, properly-speaking, because it's already included in the content. That's just what it *means* to not place a mark at $(1979, .90)$.

By contrast, (DIF) is not a proposition that figure 3 satisfies in the same sense. To derive (DIF) from figure 3, we need additional information. Specifically, we need information about the nature of the objects represented by the x and y axes. (DIF) is a consequence of figure 3 only because the y axis represents something with the structure of a quantity, namely mean global temperature. If instead our graph depicted (e.g.) the day with the highest average temperature, it would make no sense to subtract the y-value of the leftmost mark from that of the rightmost mark. The move from figure 3 to (DIF) thus looks more like an inference properly-speaking: we reason our way to (DIF) by way of combining the representation with background information about what it represents. I'm thus inclined to say that (NEG) is part of the content of figure 3 whereas (DIF) is not.

The argument just given is barely a sketch—at minimum, we would need to say more about what "satisfies" means in this context. For present purposes, what's important is the connection (or lack thereof) between the account of graphical content just outlined and the problems identified in the prior section. At best, only the second of our two problems concerns features that would make the graph inaccurate. On the account offered, the first problem—that the visual distance between the models and observations is being used

to support claims about the accuracy of the models when it has no bearing on that question—is a problem with how the graph is being used, or perhaps with what it implies. A graph that was totally unimpeachable with respect to accuracy could still exhibit this problem.

The connection between graphical truth conditions and the second problem is a little bit muddier. If I'm right, and the axis is genuinely mislabeled, then (on the account just offered) Christy's graph is not just misleading but incorrect or inaccurate. Christy might respond that his procedures are indicated in the graph and that temperatures in a possible world are still temperatures. While this response is unconvincing qua defense of the graph itself, it is more convincing when read as an argument for the claim that the graph does accurately portray its target—it's just that the target is not what any reasonable person would expect. (To be clear: *more* convincing is not the same as convincing *tout court*.) In an important sense, therefore, whether or not we take the second problem to affect the truth or accuracy of Christy's graph is irrelevant to the epistemic criticism. The graph is not trustworthy or valuable as a source of information *either way*.

4 Graphical epistemology

The conclusion of the last section will likely be unsurprising to contemporary philosophers of science. After all, it's a truism of the literature on model evaluation that the ability of models to provide us with worthwhile or accurate information doesn't depend on the truth of those models (see, e.g., Parker 2020). Similarly, recent work on the cognitive science of perception suggest that our perceptual system doesn't just trade in misinformation, that misinformation is often central to the system's functioning as a reliable source of knowledge (see Azzouni 2024). The interesting question is not *whether* the trustworthiness or epistemic value of graphs come apart from their truth conditions, but *why* they do and what kind of departures we should thus expect.

To sketch an answer to this second question, it's helpful to take brief detour into the philosophical literature on diagrammatic proofs in mathematics. One of the main conclusions that has come out of this literature is that diagrams tend to exhibit "free rides": viewers can simply see certain consequences of the represented facts when looking at a diagram in a way that they can't when looking at a linguistic representation of the same phenomena.[3] Take figure 3: the viewer can "just see" that temperatures are going up by looking at the

[3]The term "free rides" originates with Barwise and Shimojima (1995). For a more recent survey, see Giardino (2020).

graphs in a way that you really can't by looking at the data or a linguistic rendition thereof—a set of sentences of the form "In year x, the mean global temperature was y°C above the 20th century average."

What the viewer sees by looking at a graph need not be part of the semantic content—the truth-conditions—of the graph. That temperatures are going up is not a part of the content of figure 3; that the models and observations are far apart or diverging is not part of the content of figure 1. At the same time, it's important to recognize that while viewers see things that are not part of the content of the graph, they also fail to see things that are part of the content. I feel safe in claiming that no one, even on close examination of figure 3, will come away from the graph with the information that in 1979 the mean global temperature was .12°C above the 20th century average, let alone with information about the temperature recorded in every single year. That information is part of the truth conditions of the graph, not just on the account presented here but on any plausible account—on my reading, every prior paper on the subject (see note 2) takes it as given that what I'm calling the "primary content" is part of the truth conditions of a graph. So in this respect, viewers are failing to see things that are part of the graph's content.

In other words, when it comes to graphs, there's no such thing as *free* rides: to buy the extra content that we see when viewing a graph, we have to give up on seeing information that is part the graph's truth conditions. Indeed, I would hazard that, all else equal, the more substantively what's seen when looking graph goes beyond its semantic content, the more it will fall short of it as well.

Whether my hazard is true or not, the general lesson is the same: because the features that make a graph trustworthy or epistemically valuable depend on what we see when looking at the graph, and because what see systematically departs from the truth conditions of the graph, the features that make a graph trustworthy systematically depart from those that make it accurate. To determine exactly how these two categories depart, we need empirical information about how people actually read graphs—fortunately, there's a lot out there (see, e.g., Glazer 2011). Unfortunately, however, we'll have to leave examining that literature to another time.

5 Conclusion

In this paper, I've argued that the features that make a graph trustworthy systematically depart from those that make it accurate. In fact, I've offered an explanation of this fact: what makes the graph accurate is a function of

the precise placement of marks in the frame; what makes a graph trustworthy depends on what the viewer sees when they look at those marks. These are not at all the same thing. To motivate this conclusion, I examined a controversial graph from the climate scientist John Christy, arguing that it exhibits two important problems: visual distance is supposed to measure model accuracy but does not, and the graph is better understood as depicting model accuracy in a possible world than in the actual one. On my account these problems are indicative; the hope for developing a more detailed account of graphical truth conditions is that it might allow those who make graphs to more easily identify and categorize such problems.

References

Azzouni, Jody (2024). Blur and Knowledge from Falsehood: Neural Network Science and Neurophysiology Meets Epistemology. *Journal of Neurophilosophy* 3.2: 281–97. DOI: 10.5281/zenodo.14272677.

Barwise, Jon and Atsushi Shimojima (1995). Surrogate Reasoning. *Cognitive Studies* 2.4: 7–27. DOI: 10.11225/jcss.2.4_7.

Carroll, Raymond J. et al. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. 2nd ed. Boca Raton: Chapman & Hall/CRC.

Christy, John R. (n.d.). Testimony of John R. Christy. *U.S. House Committee on Science, Space & Technology* (): 1–23.

Dethier, Corey (2025). How do you Assert a Graph? Towards an account of depictions in scientific testimony. *Noûs* 59.3: 678–99. DOI: 10.1111/nous.12529.

— (manuscript). Graphical Truths.

Giardino, Valeria (2020). Diagrammatic Proofs in Mathematics: (Almost) 20 Years of Research. In: *Handbook of the History and Philosophy of Mathematical Practice*. Ed. by Bharath Sriraman. Cham: Springer: 1–23.

Glazer, Nirit (2011). Challenges with Graph Interpretation: A Review of the Literature. *Studies in Science Education* 47.2: 183–210. DOI: 10.1080/03057267.2011.605307.

Irving, Zachary C. (2011). Style, but Substance: An Epistemology of Visual versus Numerical Representation in Scientific Practice. *Philosophy of Science* 78.5: 774–87. DOI: 10.1086/662567.

Kulvicki, John (2010). Knowing with Images: Medium and Message. *Philosophy of Science* 77.2: 295–313. DOI: 10.1086/651321.

Lloyd, Elisabeth A. (2012). The Role of "Complex" Empiricism in the Debates about Satellite Data and Climate Models. *Studies in History and Philosophy of Science Part A* 43.2: 390–401. DOI: 10.1016/j.shpsa.2012.02.001.

National Centers for Environmental Information (2024). *Climate at a Glance: Global Time Series*. URL: https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/global/time-series (visited on 06/13/2024).

Parker, Wendy S. (2020). Model Evaluation: An Adequacy-for-Purpose View. *Philosophy of Science* 87.3: 457–77. DOI: 10.1086/708691.

Perini, Laura (2005). The Truth in Pictures. *Philosophy of Science* 72.1: 262–85. DOI: 10.1086/426852.

Santer, Benjamin D. et al. (2008). Consistency of Modeled and Observed Temperature Trends in the Tropical Troposphere. *International Journal of Climatology* 28: 1703–22. DOI: 10.1002/joc.1756.

Santer, Benjamin D. et al. (2009). Incorporating Model Quality Information in Climate Change Detection and Attribution Studies. *Proceedings of the National Academy of Sciences* 106.35: 14778–83. DOI: 10.1073/pnas.0901736106.

Schmidt, Gavin A. (2016). *Comparing Models to the Satellite Datasets*. URL: https://www.realclimate.org/index.php/archives/2016/05/comparing-models-to-the-satellite-datasets/ (visited on 04/02/2023).

Scoccimarro, Enrico and Silviio Gualdi (2014). CMCC-CM model output prepared for CMIP5 rcp45. *World Data Center for Climate (WDCC) at DKRZ*. DOI: 10.1594/WDCC/CMIP5.CMC0r4.

Spencer, Roy (2025). AMSU/MSU Midtroposphere Day/Month Temperature Anomalies and Annual Cycle V6 [Data set]. *NASA Global Hydrometeorology Resource Center Distributed Active Archive Center*. DOI: 10.5067/GHRC/AMSU-A/DATA403. URL: https://www.earthdata.nasa.gov/data/catalog/ghrc-daac-msutmt-6.