Stability in Climate Change Attribution

Corey Dethier

[Forthcoming in *Philosophy of Sceince*. Please contact at corey.dethier[at]gmail.com before citing or quoting.]

Abstract

Climate change attribution involves measuring the human contribution to warming. In principle, inaccuracies in the characterization of the climate's internal variability could undermine these measurements. Equally in principle, the success of the measurement practice could provide evidence that our assumptions about internal variability are correct. I argue that neither condition obtains: current measurement practices do not provide evidence for the accuracy of our assumptions precisely because they are not as sensitive to inaccuracy in the characterization of internal variability as might be worried. I end by drawing some lessons about "robustness reasoning" more generally.

0 Introduction

The science of climate change attribution involves measuring the human contribution to warming.[1] Like many measurements, attribution relies on establishing a *baseline*, or the state that the system would exhibit absent the phenomenon being measured. In the case of attribution, the baseline is the natural or "internal" variability of the climate system. Internal variability is generally estimated using climate models, but the accuracy of the relevant estimates is hard to confirm. We can't directly measure internal variability, because there is no climate system unaffected by climate change. Indirect measures—especially

[1] "Attribution" is often also used in the context of determining the causes of extreme weather events. This second use won't be my focus in this paper.

those based on historical proxies—require risky extrapolations, particularly because it's expected that the internal variability of the climate is also affected by rising temperatures. At face value, then, internal variability is an impediment to trustworthy measurement of anthropogenic climate change—as philosophers such as Joel Katzav (2013) and Wendy Parker (2010) have explicitly argued.

Parker (2010, 1090–91) suggests that attribution studies could themselves provide evidence for the accuracy of estimates of internal variability: were we to show that these studies are successful, we would have reason to think that their assumptions are (relatively) accurate (see also Katzav 2013, 437). The evidence as of the late 2000s was not sufficient to support this kind of argument, however. Parker draws particular attention to the inconsistency of results within the field; once we move beyond gross qualitative similarities, there is—or was—little that different attribution studies agree on. This paper re-evaluates Parker's conclusion: given that attribution science has changed dramatically in the last decade—and that many of these changes have been focused on internal variability (IPCC 2021, 429–30)—can we *now* say that the results of attribution studies give us reason to think that estimates of internal variability are accurate?

The answer is *no*. While the most recent attribution results are (remarkably) stable, we're not justified in inferring that estimates of internal variability are accurate. The reason why we're not, however, is that research has indicated that we would expect similar results even if the representations of internal variability were (quite) inaccurate. Contra the reasoning laid out above, therefore, the last decade of research shows that internal variability is not quite the impediment it has traditionally been assumed to be. After working through the detailed argument for this conclusion, I discuss the implications for our understanding of stability (or "robustness") more broadly. As we'll see, stability plays many different roles throughout the case study, which might suggest that there are many kinds of robustness. I argue the opposite: there is only one logic of stable results. What matters is simply whether the hypothesis predicts stability and its negation does not.

1 The logic of stable measurement

Consider a classical tube thermometer where the scale—the lines on the outside of the tube that allow us to convert observed height of a column of fluid to temperature—was designed using the ideal gas law. We might say that measurements of temperature using this thermometer are *theory-mediated* in the sense that the measurements can be expected to be accurate only if the

ideal gas law is as well; if the ideal gas law is sufficiently inaccurate, the thermometer should record the wrong temperature. If we have some way of independently measuring temperature, therefore, we can use the behavior of the thermometer to test the accuracy of the theory that it presupposes.

Of course, we are often in situations where we have no way of independently measuring the quantities that interest us and so cannot evaluate the accuracy of our assumptions by simply checking the success of the measurement. But even in these cases we can often check the *stability*—sometimes "robustness" or "agreement"—of the measurement across various permutations of background conditions and assumptions. We can, for instance, design thermometers using different fluids and compare the results. If the assumptions are accurate and the measurements successful, they should be stable. If—in addition—an inaccuracy in the assumptions would lead to unstable results, then observing stability is potentially powerful evidence for the accuracy of the assumptions.

This second condition is not guaranteed: whether a particular variation should be expected to generate instability if the assumptions in question are inaccurate depends on the details. Consider an example from George E. Smith (2014), namely the measurement of the mass of the sun by way of the observed acceleration of another body.[2] Historically, this measurement relied on various elements of Newtonian theory, such as the law of inertia. If the law of inertia is sufficiently inaccurate, then a body can accelerate without being acted on by a force, and thus the magnitude of acceleration is not a good proxy for either the magnitude of the accelerating force or for the magnitude of the gravitational mass generating that force. So: if the law of inertia is inaccurate, then any measurement of the mass of the sun by way of the observed acceleration of another body cannot be expected to be accurate either.

Stability is a different question. If the law of inertia is inaccurate, for example, we might expect that repeated measurements of the mass of the sun by way of the observed acceleration of (say) Mars would be stable. If elliptical orbits of the planets are just a kind of brute fact, to take one way that the law of inertia could be wrong, we should expect that Mars will remain in the same orbit indefinitely, meaning that we should get the same (wrong) result for the mass of the sun across repeated measurements because the acceleration will be the same. By contrast, measuring the mass of the sun by way of the acceleration in generates in *different* bodies should yield unstable results when the law of inertia is inaccurate. If the orbits are just brute facts, it would be a remarkable coincidence for the acceleration of each planet to have the same

[2] My discussion here is the kind of simplification that arises out of cramming a monograph into a few paragraphs, but should do for our purposes.

relationship with (the inverse-square of) its distance from the sun.

As Smith stresses, the story does not end here, because astronomical work on celestial bodies did not end once we had an estimate for the mass of the sun. There were many more masses to be estimated, and the methods adopted in these latter cases depended not just on Newtonian theory but on the earlier results as well: to determine the mass of Jupiter (for example) we look at how much the acceleration of other bodies (such as Mars and Saturn) deviates from the new baseline defined by a model that includes only the effects of the sun's mass. And this means that the stability of measurements of Jupiter's mass provides evidence for the accuracy of the measurement of the sun's mass, which in turn provides further evidence for the accuracy of assumptions—like the law of inertia—that underwrote the latter. Thus do successive measurements serve to (repeatedly) close "loops" of reasoning that constrain our assumptions more and more tightly.

The key lesson of this section: in evaluating the implications of stable measurement results for the theoretical assumptions that those measurements rest on, the crucial question is whether the failure of those assumptions would lead to instability in the measurements. If so, then the observed stability is good evidence for the accuracy of the assumptions. If not, it isn't.

2 Stability and internal variability

In this section, I'll examine the stability of attribution results and discuss the implications for internal variability. In the next, I'll consider the implications for the trustworthiness of the attribution results themselves.

First, though, how do attribution studies actually work? Speaking broadly, attribution studies are regressions: they take the signals of different possible causes of climate change and assign each of them a weight based on how well the resulting combination fits the observed data. Typically—though see below—an estimate for internal variability is employed as a filter on the data before the regression step, with the goal of isolating that part of the data that is in fact the result of warming. The results of an attribution study are the weights—i.e., the weight assigned to the CO_2 signal is what tells us how much CO_2 has contributed to climate change over the relevant period. (For discussion of the details, see Dethier (2022a) or Hammerling et al. (2019).)

It is difficult to assess the stability of attribution results. For one thing, there's no fixed period of evaluation for attribution studies, and there's no principled way of quantitatively comparing results covering (say) 1951-2010 and 1906-2005. Similarly, some attribution studies decouple the effects of aerosols

	1986-2005	1995-2014	2006-2015	2010-2019
Observed	.69 (.52-.82)	.86 (.67-.98)	.94 (.76-1.08)	1.06 (.88-1.21)
Gillett et al.	.63 (.32-.94)	.84 (.63-1.06)	.98 (.74-1.22)	1.11 (.92-1.30)
Haustein et al.	.73 (.58-.82)	.88 (.75-.98)	.98 (.87-1.10)	1.06 (.94-1.22)
Ribes et al.	.65 (.52-.77)	.82 (.69-.94)	.94 (.80-1.08)	1.03 (.89-1.17)

Table 1: The °C change in temperature relative to the period 1850-1900. The first row is the observed change (IPCC 2021, 320). The other rows are estimates for the warming attributable to humans (IPCC 2021, 442).

and greenhouse gases in their analyses, while others don't and combine the two into a single anthropogenic factor. As the estimates for the contribution of greenhouse gases and aerosols are *not* independent, we cannot compare the two studies by (say) summing the two factors together.[3]

Of course, climate scientists are sensitive to the potential value of stable results in attribution. Virtually every study on the subject examines the degree of stability across variation in the models employed in regressing climate data. These examinations do not provide the kind of evidence that we're looking for here. For one thing, attribution is *not* stable across individual models: while analyses based on individual models never provide evidence that climate change isn't attributable to humans, they often fail consistency checks or don't discriminate between anthropogenic and natural causes. For another, there's an important sense in which the results generated by individual models are not properly seen as the *outcome* of a measurement process: individual models are more analogous to individual data points; potentially indicative but not a sufficient basis for analysis (Dethier 2022b). Examining stability across individual models is thus more like examining the statistical properties of a data set to confirm that the method or experiment is behaving as expected—and, indeed, this is how (in)stability across individual models is normally used in the science (see, e.g, Ribes and Terray 2013).

More promising for present purposes are cross-study comparisons. Unfortunately, these are rare. The only non-cursory example I'm aware of is found in the most recent IPCC report, which offers a like-for-like comparison of three cutting-edge studies: Gillett et al. (2021), Haustein et al. (2017), and Ribes et al. (2021). The results are displayed in table 1 and are stable in at least two important senses. First, they're stable in a sense given by Smith and Seth (2020, 138): the relevant error bounds from the different studies consistently overlap (indeed, the best estimates given by each study fall within the error

[3]There are further practical problems. A systematic review would require re-examining and perhaps re-analyzing data sets that I, at least, have been unable to acquire.

bounds of each of the others). Overlapping error bounds allows each study to be accurate within their stated margin of error; insofar as we have good reason to think that the margins of error are accurate, results that are stable in this respect are much more powerful than results that are not and that thus require us to assume that at least one of the error bounds is inaccurate.

Second, they're stable in a sense outlined by Dethier (2021): they deliver what is effectively the same answer to the major theoretical questions in the area. In attribution, the major theoretical question is whether humans are the primary driver of observed climate change. The estimates for observed climate change delivered by the most recent IPCC report are given in the first row of table 1; they are indistinguishable in both estimate and error bars from the estimates for attributable warming delivered by the different studies. Regardless of which study we choose to rely on, humans are responsible for essentially all of the observed warming since the industrial revolution.

While these results are stable, this stability does not provide good evidence in favor of the accuracy of estimates of internal variability, precisely because we would still expect stability were the estimates inaccurate. It's true that the three studies estimate internal variability in different ways: Haustein et al. (2017) use CMIP5 models, Gillett et al. (2021) CMIP6 models, and Ribes et al. (2021) fit a model that mixes processes with "short" and "long" memories to the difference between empirically observed temperatures and the CMIP6-generated trend line. It's also true that they make use of estimates of internal variability in *dramatically* different ways. Gillett et al. (2021) employ the traditional approach to attribution developed by Hasselmann (1993) and use internal variability as a filter on the data to isolate the signal. Ribes et al. (2021) adopt a Bayesian methodology in which model simulations are used as priors and internal variability plays a role only in the updating step. And so far as I can tell, Haustein et al. (2017) employ internal variability only in the context of generating the uncertainty bands around their best estimate.[4] So there's substantial variation in these three studies with respect to internal variability. Nevertheless, we shouldn't expect this kind of variation to result in unstable estimates for the human contribution to warming because empirical studies such as Imbers et al. (2013, 2014) and Sippel et al. (2021) have shown that attribution results are in fact relatively insensitive to different representations of internal variability.

We'll focus on Sippel et al. (2021), which asks the simple question "what

[4]At time of writing, I have been unable to verify the nature of the methods employed in Haustein et al. (2017) to my satisfaction. If their discussion is indicative, it would represent a dramatic break from the traditional approach.

if internal variability were larger than we thought?" They operationalize this question in two main ways: (1) by simply doubling (the primary empirical orthogonal functions of) internal variability and (2) using estimates for internal variability generated by the highest-variability simulations. They report their results in terms of the *minimum* percentage of 1980-2019 warming attributable to external forcings. Depending on the regression technique employed, doubling internal variability results in best estimates for this minimum percentage between roughly 55% and 80% (Sippel et al. 2021, fig. 6b). Similarly, assuming that internal variability lines up with the top 5% of simulations—note, *not* that it is at $p = .05$ level, but that it falls in the top 5% of the sample—generates a minimum percentage that falls between roughly 45% and 75% (Sippel et al. 2021, fig. 6a). They then combine the two by doubling internal variability *and* assuming that it lines up with the top 5% of simulations. In this extreme situation, *some* methods fail to rule out the hypothesis that external forcings have no effect. The method most similar to that employed by Gillett et al. (2021) still yields a minimum percentage of around 20%, while the preferred methodology of Sippel and coauthors yields one in the low 50s.

Note that these scenarios are not supposed to be realistic: doubling internal variability is essentially an exercise in stress-testing different attribution methods, not a way of evaluating a scenario that might actually obtain. And for our purposes, the main upshot of these studies is that even if our estimates of internal variability are inaccurate, we should expect attribution results to be relatively stable. After all, these studies indicate that we would have to be *very* wrong about internal variability before we started to see major differences in attribution results—before we would expect to see results that were unstable in the senses outlined above. It turns out that attribution isn't *that* sensitive to the details of the representation of internal variability—at least not when we restrict our attention to realistic or relatively probable errors.

As a consequence, we should expect that attribution results will be (relatively) stable across the kinds of differences examined in this section *even if* they rest on inaccurate assumptions about the nature of internal variability. Reasoning that parallels that outlined by Smith in the astronomical context is thus untenable here: we cannot reason from the stability of the measurement of the human contribution to climate change to the accuracy of assumptions about internal variability.

3 Stability and attribution

The conclusion at the end of the last section is a negative one, but there are two elements of the analogy to the astronomical case that remained unexamined. First, we have not considered the implications of stable attribution estimates for the measurements themselves: does stability give us a reason to trust these measurements even if it doesn't give us a reason to trust the assumptions that they rest on? Second, we have not yet considered the possibility of results that depend on estimates for the human contribution to warming in the same way that measurements of the mass of Jupiter depend on estimates for the mass of the sun. If such measurements exist and are stable or independently confirmable, they may provide evidence for the accuracy of attribution results in the same way that the stability of the measurement of Jupiter's mass provides evidence for the accuracy of the measurement of the sun's mass.

Both of these additional elements do in fact provide some evidence in favor of the accuracy of attribution results. Take the stability of attribution results themselves first. The studies discussed in the last section provide us with evidence that even relatively large errors in the estimate of internal variability will not lead to substantively different estimates for the human contribution to climate change. The reasoning here is simple. Previously, we were worried that inaccuracies in the estimate of internal variability would lead to inaccuracies in the estimate of the human contribution to warming. The studies discussed above survey the realistic ways that our estimates about internal variability could be wrong in a relatively thorough way—if the attribution results were inaccurate because of a misrepresentation of internal variability, we would expect to see more instability in these studies. Stability thus gives us reason to think that the attribution results are in fact accurate.

Now consider results that depend on attribution results. At first pass, there is nothing analogous to the case of the mass of Jupiter in climate change attribution. Attribution studies do not proceed in the piecemeal fashion of Newtonian astronomy; prior attribution results are not *directly* implicated in later studies.[5] At least so far as I am aware, no attribution study has presupposed a prior estimate of (e.g.) the anthropogenic contribution to warming in estimating the solar contribution, and we would need stability across varied studies of this sort to motivate an analogy to the astronomical case.

Things are more complicated on closer examination, however. Since the early days of attribution, it has been common for climate scientists to use attri-

[5]They may be implicated indirectly, by serving to guide or shape research. Unfortunately, the epistemic consequences of this kind of influence is outside the scope of the present paper.

bution results to estimate other key climate variables, particular the transient climate response (TCR), which measures how the climate reacts to increases in CO_2 concentration. These estimates presuppose the accuracy of attribution results: roughly, they estimate TCR by dividing their estimate for the contribution of CO_2 to warming by an estimate for the amount of CO_2 that's been emitted since the beginning of the industrial period. Crucially, there are other means of estimating TCR; we can compare the attribution-based estimates to these alternative estimates with the hope of finding the kind of stability that would indicate that our estimates are correct.

IPCC (2021) gives two examples of studies that adopt this method—Schurer et al. (2018) and Ribes et al. (2021)—which report estimates of 1.8°C (9-95% band of 1.2-2.4) and 1.84°C (± .51) respectively. IPCC (2021, ch. 7) surveys a variety of other estimates, such as those delivered by theory (2.0°C [1.6-2.7]), instrumental records (1.9°C [1.3-2.7]), and the observed response to volcanic eruptions (1.9°C [1.5-2.3]). Once again, there is a noteworthy level of agreement to be found among these different estimates.

To reiterate the lessons from above, insofar as our aim is using the stability of TCR estimates as evidence for the accuracy of not just the attribution-based estimates of TCR but the attribution results themselves, we need to show that we would expect less stability were attribution results inaccurate. Schurer et al. (2018, 8658) gives us some reason expect less stability: examining the variability in TCR results when the attribution step relies on a single model rather than the multi-model mean indicates that different values for the human contribution to warming will result in moderately different estimates for TCR. These results are less definitive than we might like, however. They indicate that we should expect somewhat worse agreement between attribution-based estimates for TCR and other lines of evidence were the assumptions of the former inaccurate, but both the strength of this expectation and how much worse remain up in the air.

Still, when we survey all of the evidence considered in this section, the picture is fairly clear: estimates of the human contribution to climate change are accurate to within their relevant confidence bounds. There are (of course) ways that we could be wrong and ways that this evidence could be misleading. But these results are not so sensitive to estimates of internal variability that it's reasonable to distrust them on that basis. On the contrary, the evidence available indicates that our estimates for internal variability could be quite inaccurate without seriously undermining attribution results.

4 The many faces of robustness

My focus in this paper has been stability. I could easily have spoken of "robustness" instead, which is largely used as a cognate term. To parrot I.J. Good (1983), however, there are more views about the nature of robustness than there are philosophers who work on the subject. Indeed, many philosophers think that there is more than one kind of robustness—a position most famously argued for by Woodward (2006).

One of the lessons of our case study in attribution science is that this view, at least so flatly stated, is wrong.[6] There is only kind of robustness or—better—there is only one logic of stable results. What Woodward and others get right is that the details matter: not every case is similar, and the mere existence of stability tells us nothing. On the contrary, stable results confirm a hypothesis when (and to the extent that) the hypothesis predicts stability and its negation predicts the opposite.

Consider again the examples of stability that we examined above.

- Measurements of the sun's mass by way of its effects on other bodies. Newtonian theory predicts stability; were the theory (sufficiently) inaccurate, we'd expect instability. Stability confirms the theory.
- Measurements of Jupiter's mass by way of its effects on other bodies. We expect stability if the measurement for the sun's mass is accurate, instability if it's not. Stability confirms the assumption.
- Measurements of the human contribution to warming that use internal variability in different ways. We expect stability regardless of whether estimates of internal variability are accurate. Stability doesn't confirm.
- Measurements of the human contribution to warming using different values for internal variability. If the measurements are in error *because the estimate for internal variability is inaccurate*, we'd expect instability; if the mis-estimation of internal variability is not a problem, we'd expect stability. Stability confirms that the measurements are not inaccurate due to internal variability.
- Measurements of TCR by way of methods that both do and don't rely on attribution results. If the attribution results are accurate, we'd expect stability; if they're not, we *might* see instability. Inconclusive, but somewhat positive.

And three we didn't explicitly discuss:

[6] I'm unconvinced that Woodward (2006) should actually be interpreted so flatly, but much of the subsequent literature seems to read him this way.

- Measurements of Jupiter's mass by way of its effects on other bodies. Newtonian theory predicts stability; were the theory (sufficiently) inaccurate, we'd expect instability. Stability confirms the theory.
- Measurements of TCR by way of methods that both do and don't rely on attribution results. If the TCR results are accurate, we'd expect stability; if they're not, ???. Inconclusive.
- Measurements of TCR by way of methods that both do and don't rely on attribution results. If estimates of internal variability are accurate, we'd expect stability; if they're not, we'd still expect stability. Stability doesn't confirm.

The reasoning for each of these latter conclusions follows from one of the earlier examples.

This list illustrates *some*, but certainly not all, of the many faces of robustness. As we can see, stable results may or may not confirm. Indeed, a particular set of stable results may be confirmatory with respect to one question, but not confirmatory with respect to another. And while we've treated confirmation as a simple binary, the real world is much more complicated: stability across measurements of Jupiter's mass may offer powerful evidence in favor of the truth of Newtonian theory but only weak evidence in favor of the accuracy of the estimate of the sun's mass. *Nevertheless*, in each of these cases the logic is exactly the same. The question is always whether (and to what extent) the hypothesis predicts stability and its negation predicts the opposite. The point generalizes beyond these cases as well. Indeed, it's a simple consequence of basic Bayesian principles identified by Myrvold (1996): where there is stability to be found, what matters confirmation-wise is just what the different hypotheses have to say about it.

5 Conclusion

In this paper, I've argued for three distinct conclusions. First, that the stability of attribution results does not provide evidence for the accuracy of our assumptions about internal variability. Second, that the same stability does provide evidence for the accuracy of those results—a conclusion at least partly supported by the stability of TCR results. Third, that while stability has many different implications, it has only one logic.

References

Dethier, Corey (2021). Climate Models and the Irrelevance of Chaos. *Philosophy of Science* 88.5: 997–1007. DOI: 10.1086/714705.

— (2022a). Calibrating Statistical Tools: Improving the Measure of Humanity's Influence on the Climate. *Studies in the History and Philosophy of Science* 94: 158–66. DOI: 10.1016/j.shpsa.2022.06.010.

— (2022b). When is an Ensemble Like a Sample? 'Model-Based' Inferences in Climate Modeling. *Synthese* 200.52: 1–20. DOI: 10.1007/s11229-022-03477-5.

Gillett, Nathan P. et al. (2021). Constraining Human Contributions to Observed Warming since the Pre-industrial Period. *Nature Climate Change* 11: 207–12. DOI: 10.1038/s41558-020-00965-9.

Good, Irving J. (1983). *Good Thinking: The Foundations of Probability and Its Applications*. Minneapolis: University of Minnesota Press.

Hammerling, Dorit et al. (2019). Climate Change Detection and Attribution. In: *Handbook of Environmental and Ecological Statistics*. Ed. by Alan Gelfand et al. Boca Raton: Chapman and Hall: 789–817.

Hasselmann, Klaus (1993). Optimal Fingerprints for the Detection of Time-dependent Climate Change. *Journal of Climate* 6.10: 1957–71. DOI: 10.1175/1520-0442(1993)006<1957:OFFTDO>2.0.CO;2.

Haustein, Karsten et al. (2017). A real-time Global Warming Index. *Scientific Reports* 7 (15417): 1–6. DOI: 10.1038/s41598-017-14828-5.

Imbers, Jara et al. (2013). Testing the Robustness of the Anthropogenic Climate Change Detection Statements using Different Empirical Models. *Journal of Geophysical Research: Atmospheres* 118.8: 3192–99. DOI: 10.1002/jgrd.50296.

— (2014). Sensitivity of Climate Change Detection and Attribution to the Characterization of Internal Climate Variability. *Journal of Climate* 27.10: 3477–91. DOI: 10.1175/JCLI-D-12-00622.1.

IPCC (2021). *Climate Change 2021: The Physical Science Basis*. Ed. by Valérie Masson-Delmotte et al. Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge: Cambridge University Press.

Katzav, Joel (2013). Severe Testing of Climate Change Hypotheses. *Studies in History and Philosophy of Science Part B* 44.4: 433–41. DOI: 10.1016/j.shpsb.2013.09.003.

Myrvold, Wayne (1996). Bayesianism and Diverse Evidence: A Reply to Andrew Wayne. *Philosophy of Science* 63.4: 661–65. DOI: 10.1086/289983.

Parker, Wendy S. (2010). Comparative Process Tracing and Climate Change Fingerprints. *Philosophy of Science* 77.5: 1083–95. DOI: 10.1086/656814.

Ribes, Aurélien and Laurent Terray (2013). Application of Regularised Optimal Fingerprinting to Attribution. Part II: Application to Global Near-surface Temperature. *Climate Dynamics* 41.11-12: 2837–53. DOI: 10.1007/s00382-013-1735-7.

Ribes, Aurélien et al. (2021). Making Climate Projections Conditional on Historical Observations. *Science Advances* 7.4: 1–9. DOI: 10.1126/sciadv.abc0671.

Schurer, Andrew P. et al. (2018). Estimating the Transient Climate Response from Observed Warming. *Journal of Climate* 31.20: 8645–63. DOI: doi.org/10.1175/JCLI-D-17-0717.1.

Sippel, Sebastian et al. (2021). Robust Detection of Forced Warming in the Presence of Potentially Large Climate Variability. *Science Advances* 7.43: 1–17. DOI: 10.1126/sciadv.abh4429.

Smith, George E. (2014). Closing the Loop: Testing Newtonian Gravity, Then and Now. In: *Newton and Empiricism*. Ed. by Zvi Beiner and Eric Schliesser. Oxford: Oxford University Press: 262–351.

Smith, George E. and Raghav Seth (2020). *Brownian Motion and Molecular Reality: A Study in Theory-Mediated Measurement*. Oxford: Oxford University Press.

Woodward, James (2006). Some Varieties of Robustness. *Journal of Economic Methodology* 13.2: 219–40. DOI: 10.1080/13501780600733376.