Can Attribution Science Close the Loop?

Corey Dethier

0 Plan

In this talk, I'll defend two claims:

1. Uncertainty about internal variability doesn't undermine our ability to measure the human contribution to climate change.
2. There is only one "logic" of stability / robustness.

The plan:

1. Background: why (uncertainty about) internal variability is a problem.
2. An analogy to measurements in astronomy.
3. One way that the stability of results in attribution science fails to solve this problem.
4. One way that it does.
5. Why there is only one logic of stability.

1 Background

To measure the **human contribution to climate change**, we need to know the state the climate system would have exhibited absent human actions.

> **Essentially**: the human contribution = observed change − change absent human actions.

A widely-acknowledged problem is that there is substantial uncertainty about how the system would behave absent human actions (IPCC 2021, 429–30; Katzav 2013; Parker 2010):

1. The climate is extremely complex, exhibiting variability across time and location.
2. The natural or "internal" variability (IV) of the system cannot be derived from first principles.
3. It also can't be directly observed – there's no global climate unaffected by human actions.
4. Furthermore, IV depends on factors that are affected by climate change.

In short: there's substantial uncertainty about IV.

Which means: erroneous assumptions about IV are a serious potential source of error in measuring the human contribution to climate change.

Parker (2010) suggests a potential remedy:

> *if* different measurements of the human contribution were stable, *then* we *might* be able to offer a "robustness reasoning"-style argument for the accuracy of our assumptions about IV.

Problem: as of 2010, measurements of the human contribution to warming were not stable.

> That's changed.

	1986-2005	1995-2014	2006-2015	2010-2019
Gillett et al. (2021)	.63 (.32-.94)	.84 (.63-1.06)	.98 (.74-1.22)	1.11 (.92-1.30)
Haustein et al. (2017)	.73 (.58-.82)	.88 (.75-.98)	.98 (.87-1.10)	1.06 (.94-1.22)
Ribes et al. (2021)	.65 (.52-.77)	.82 (.69-.94)	.94 (.80-1.08)	1.03 (.89-1.17)

The table above gives three distinct measurements of the °C human contribution to warming, where 0°C represents the average temperature in period 1850-1900 (IPCC 2021, 442).

These results are **stable** in at least two senses:

1. The error bars overlap (Smith and Seth 2020).
2. They deliver (approximately) the same answer to the major theoretical question of interest – namely: is climate change primarily driven by human actions? (Dethier 2021)

Question: does this stability actually provide evidence that our assumptions about IV are correct?

2 An analogy

Our **question** has analogues in other settings such as (e.g.) Newtonian astronomy.

To measure the gravitational contribution of a body to another body's motion, we need to know the behavior the system would have exhibited absent the first body.

> **E.g.**: The gravitational effect of the sun = the observed motion − the motion absent the sun.

We face the same problems here as in the climate case:

1. There's no guarantee that celestial objects obey the same first principles as terrestrial ones.
2. There's no solar system unaffected by the sun's gravity for us to observe.

As Smith (2014) documents, **stability** played a key role in solving this problem.

Imagine that our assumptions about the behavior the system would have exhibited absent the sun are wrong – e.g., the principle of inertia is highly inaccurate.

Since any individual measurement depends on those assumptions, individual measurements can't reveal this error. But (in)stability across multiple measurements can.

> **Consider**: if the principle of inertia is accurate, for any planet x, $m_{sun} \approx G^{-1} a_x r^2_{sun-x}$
>
> **Or**: we should get the same value for m_{sun} regardless of which planet's values we plug in.
>
> **But**: if the principle of inertia is inaccurate, $m_{sun} \not\approx G^{-1} a_x r^2_{sun-x}$

Which means: if our assumptions are accurate, stability across measurements using different planets is **predictable**. If our assumptions are inaccurate, we expect instability: the a_x and r^2_{sun-x} differ wildly with x; it would be surprising if they happened to have the same ratio.

So: stability of measurements *can* confirm the assumptions those measurements rely on.

Iterating: the stability of m_{sun} confirms our assumptions about inertia; the stability of $m_{jupiter}$ confirms both our assumptions about inertia and our assumptions about m_{sun}, etc. Why?

> For any planet x, if m_{sun} is accurate, $m_{jupiter}$ should be approximately proportional to the difference between the observed acceleration and the acceleration given just the sun's gravity times the square of the distance between that planet and Jupiter.
>
> And thus stability in $m_{jupiter}$ is predictable if m_{sun} is accurate.
>
> If m_{sun} is (highly) inaccurate, we should see instability instead.

Preliminary upshot: a sufficient condition on confirmation of h via stability in x:

1. If h is true, we should expect stability in x.
2. If h is false, we should expect instability in x.

3 Internal variability

Reformulation of **question**: do both the conditions given at the end of the last section hold?

> Plausibly, if our assumptions about IV are accurate, we should expect stability.
>
> Should we expect instability if they're inaccurate?

In favor of a "yes" answer: the studies use different methods that incorporate IV in different ways.

In favor of a "no" answer: empirical studies like Sippel et al. (2021) indicate that measurements of the human contribution to warming aren't very sensitive to IV.

> Roughly: how much does the *lower bound* of the human contribution depend on IV?
>
> In the context of a method akin to Gillett et al. (2021), they find that throwing out the bottom 95% of IV simulations has essentially no effect on the lower bound.
>
> Throwing out the bottom 95% and then doubling the result shifts it by $\sim 10\%$.

Upshot: we'd have to be *very* wrong about IV before such an error would generate noticeable instability in measurements of the human contribution.

Or: the second condition fails.

Which means: we can't argue from stability to the accuracy of our assumptions about IV.

4 The human contribution to warming

The argument suggested by Parker is unsuccessful. But stability can still help here.

Consider again Sippel et al's results. What they show is that we would have to be *very* wrong about IV for it to cause a series error in our measurement of the human contribution.

> Let h be "an error in IV is causing an error in the measure of the human contribution."
>
> If h is true, we should expect instability in Sippel et al's results: the most probable way for h to be true is for a small error in IV to cause large errors in the measurement.
>
> If h is false, we should expect stability: extant measurements are relatively stable but rely on different estimates for IV; so either h is true (contra the assumption) or we should see stability.

There's more stability-related evidence to be had, though it's less definitive.

> Just as measurements of $m_{jupiter}$ depend on measurements of m_{sun}, there are measurements that depend on the human contribution to warming.
>
> E.g., one way of estimating TCR relies on an estimate of the human contribution.
>
> Using this method, Schurer et al. (2018) and Ribes et al. (2021) report estimates of 1.8°C (± .6) and 1.84°C (± .51) respectively.
>
> Qualitatively, these results agree with other estimates that rely on first principles (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]).
>
> To my knowledge, however, there's been no investigation of how sensitive this agreement is to the accuracy of the measurement of the human contribution – meaning that we don't know whether our second condition is true.

Upshot: while stability doesn't show that our estimates of IV are accurate, it does indicate that errors in those estimates are unlikely to be the cause of errors in the measurement of the human contribution to warming.

5 The logic of stability

In this talk, I've reviewed five cases of reasoning from stability:

1. 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.
2. 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.
3. 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.
4. 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.
5. 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.

Three examples we didn't explicitly discuss:

6. 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.
7. 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.
8. 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.)

These examples exhibit dramatic variety; as Woodward (2006) points out, cases of "robustness reasoning" are not all alike.

But: they do all obey the same basic logic: stability in x confirms h if:

1. If h is true, we should expect stability in x.
2. If h is false, we should expect instability in x.

That is: it doesn't matter whether (e.g.) x is a modeling result or a measurement outcome (Dethier 2024). The reasoning involved is exactly the same.

All of this can be captured in a basic Bayesian formalism owed to Myrvold (1996).

Let X be the relationship between the joint likelihood and the product of the individual likelihoods:

$$X = \frac{L(e_1, e_2; h)}{L(e_1; h)L(e_2; h)}$$

X is then equal to the ratio between a measure of the likelihood of stability on h and the same measure of the likelihood of stability on $\neg h$:

$$X = \frac{P(e_1 \& e_2|h)}{P(e_1|h)P(e_2|h)} \times \frac{P(e_1|\neg h)P(e_2|\neg h)}{P(e_1 \& e_2|\neg h)}$$

Which are just quantified analogues of our two conditions.

References

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

— (2024). The Unity of Robustness: Why Agreement Across Model Reports is Just as Valuable as Agreement Among Experiments. *Erkenntnis* 89: 2733–52. DOI: 10.1007/s10670-022-00649-0.

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.

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

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, Saïd Qasmi, and Nathan P. Gillett (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.