Review of the manuscript by P. W. Thorne, J. M. Nicklas, and co-authors
"How well can we quantify when 1.5°C of global warming has been exceeded?"

This paper addresses several aspects of the reconstruction of global mean temperature from the various observational datasets currently available, with particular attention to the different possible interpretations of the Paris Agreement, and in particular of the goal of keeping the global mean temperature increase below 1.5°C. The paper first addresses several technical topics: (i) the definition of global mean temperature, including the GMST/GSAT distinction; (ii) the definition of the reference period representative of the pre-industrial era. A key element of the paper is the evaluation of a wide range of methods for estimating the current level of climate change, 'current level' being understood as an unbiased warming level representative of the present year. This evaluation focuses primarily on realised warming, defined as the temperature that would be/will be observed on average over a 20-year period centred on the year under consideration. From this evaluation, the authors derive a subset of methods showing good properties. The paper then proposes two important methodological developments: (i) the combination of several methods for estimating the current level of warming, which in particular yields better performance (estimation accuracy) than any of the individual methods considered; and (ii) a synthesis of all currently available datasets (reconstructions) of global mean temperature, covering the full range of known sources of uncertainty for this variable, and in particular allowing both the uncertainty estimates provided by existing datasets and the spread among datasets produced by different centres — using different assumptions or data processing choices — to be taken into account. The proposed method further allows datasets that do not cover the entire historical period (i.e. that do not extend back to 1850) — reanalyses in particular — to be incorporated into this synthesis. Finally, the paper provides an estimate of the current (2024) level of both realised warming and human-induced warming since the pre-industrial reference period used, namely 1850-1900.

This paper very clearly presents numerous advances on the question of global mean temperature reconstruction, which are particularly welcome ahead of a new assessment of the available scientific literature through the IPCC's 7th Assessment Report, and in the context of observed warming increasingly approaching the 1.5°C level — a level that is, at this point, symbolically, politically ('policy-relevant'), and potentially even judicially significant. Several papers could easily have been drawn from this substantial body of work. The proposed synthesis of current datasets, bringing together the full range of known uncertainties within a single synthetic ensemble, is in itself a major result, one that is likely to have a wide impact on how global warming and its associated uncertainties are estimated. The evaluation of a very broad range of methods for estimating the current warming level, carried out in the spirit of open science — including the release of the code used and the results obtained — is a considerable and highly impressive undertaking, and a very strong contribution to this field of research. For all these reasons, I recommend that this manuscript be published in ESSD.

However, I have a number of questions, requests for clarification, and remarks on several points. These comments do not detract in any way from the quality and interest of the present manuscript. They concern most sections of the manuscript.

MAJOR COMMENTS

1. Pre-industrial reference (Section 2)

I would like to challenge, to some extent, the argument put forward by the authors for choosing the 1850-1900 period as the reference representative of the pre-industrial period. The authors cite the text of the Paris Agreement, which refers to the pre-industrial period as the reference. They then estimate that the 1850-1900 period is 0.1 [-0.1 to 0.3]°C warmer than the period centred on 1750, a warming that appears to be largely attributable to human activity, whose influence is estimated at 0.1 [0.0 to 0.2]°C. [I have not examined in detail the evidence underlying these estimated temperature differences, and therefore defer to the authors on these figures.]

Taken together, these two elements seem to support fairly clearly the idea that using the 1850-1900 period leads to an underestimation of the current warming level (whether total or human-induced) of about 0.1°C, with associated uncertainties.

The argument put forward by the authors for retaining 1850-1900 as the reference is that using an earlier period would increase uncertainties. This is entirely correct, but two remarks can be made on this point:

- the text of the Paris Agreement does not mention uncertainty as a factor to be taken into account when defining the reference period;
- following the same reasoning, some parties could argue that a period such as 1900-1950 would be preferable to 1850-1900, since its uncertainties are smaller — admittedly at the cost of a larger bias. In other words, if uncertainty is to be taken into account when choosing a pre-industrial reference period, in what sense does the 1850-1900 period represent an optimum in terms of the bias-variance trade-off? (I tend to interpret the authors' argument as a bias-variance trade-off.)

I further note that the method proposed in Section 5 for combining datasets covering different time periods could plausibly be applied back to 1850, in order to combine instrumental datasets starting in 1850 with other sources of information — proxy data in particular — available over a longer period; this suggests that the technical constraint of computing a series over a period longer than 1850-present can in fact be overcome.

In conclusion on this point, I am not necessarily arguing for a different choice than the one proposed by the authors, but as it stands, the argument given for retaining the 1850-1900 reference seems fragile to me. Can it be strengthened, and if not, is 1850-1900 really the best choice?

2. Choice of realised warming (20-yr average) as the primary variable of interest / evaluation target (Section 4)

The paper uses, as its main reference for defining a current warming level, the 20-year running mean of global mean surface air temperature (GSAT), which by construction includes the response to anthropogenic forcing, to natural forcing, as well as a share of internal variability, all averaged over 20 years. This quantity is defined as 'realised warming.'

a. Potential drawbacks

The use of this variable, and more generally the inclusion of a share of internal variability in the definition of the GWL, seems to me to raise the following questions:

- Differentiated impacts

l.879-882: the authors write: "*The counter-argument to use of a human-induced warming metric is that climate impacts and adaptation requirements depend on the actual sustained warming irrespective of its underlying cause(s). If the LTTG was motivated primarily to avoid specific impacts, then the realised temperature change is the relevant metric.*"

This could be challenged. As an example, the leading mode of variability influencing GSAT is ENSO. 0.1°C of global warming due to an enhanced greenhouse effect would produce very different impacts from 0.1°C of warming due to ENSO, in terms of warming patterns, which extreme events are favoured or disfavoured, and so on. In some regions, 0.1°C of ENSO-induced warming could even result in cooler temperatures, fewer extreme heatwaves, etc. In this respect, disentangling internal variability from mean warming could be valuable for assessing the climate impacts of climate change. To a lesser extent, the same reasoning could support setting aside the response to natural forcing as well — ultimately pointing toward a focus on human-induced warming.

- Causality

This is a non-causal variable, in the sense that its value at a given time t depends on both the past (before t) and the future (after t). This is well illustrated by the discussion of volcanic eruptions.

- "Long-term"

The term "Long-Term Temperature Goal (LTTG)" is often invoked to support the idea that the 1.5°C level refers to "a long-term average," but as the authors themselves point out, the term "long-term" does not actually appear in the text of the Paris Agreement, and stems instead from a later interpretation.

- Arbitrary time window

If the intent is indeed to include internal variability, then the choice of an appropriate averaging period — here, 20 years — seems fairly arbitrary. Using a 20-year average implicitly suggests that 20-year-averaged internal variability is 'appropriate,' whereas a 10-year average or a single-year value would be considered inappropriate variability in this respect.

- Historical perspective

From a historical point of view, this choice might look somewhat like a reversal. 20-year averages have traditionally been used in model projections to remove most of internal variability, thereby approximating the forced response. This is indeed a very simple (and thus convenient) way of removing internal variability — even if imperfect — particularly in the context of climate projections, where no abrupt changes in natural forcing occur. Treating this as the best available reference for the forced response in the observed historical climate seems more debatable. In particular, other, arguably better, estimates of the forced response could be derived instead. Several of the methods reviewed here now provide year-by-year estimates of 'mean temperature,' where 'mean temperature' is understood as the mathematical expectation of possible GSAT values across multiple realizations of internal variability.

* As an alternative, focusing on the total forced response seems attractive for several reasons (note: focusing on human-induced warming offers the same advantages):

Interpretation

The forced response can be interpreted as the average — in the sense of a mathematical expectation — over a large number of realisations of internal variability. The notion of an average is thus preserved, but taken over an ensemble of possible realisations rather than over a time period.

Single-year estimates

Various methods now provide estimates of the forced response (and hence of expected GSAT) for the current year, thereby avoiding any arbitrary choice of averaging window such as 20 years.

Causality

The forced response at time t is determined by external forcings applied before time t, and is therefore a causal quantity, unaffected by anything that might happen after time t.

* In light of the above, could the authors add some discussion of this issue? I am fully aware that they have carried out a very substantial body of work and analysis, and that redoing the entire analysis with a different evaluation target is probably not feasible. I am also aware that making a definitive choice between realised warming (including both forced response and internal variability), total forced warming, and human-induced warming is probably beyond the scope of this paper. However, adding some discussion of these potential caveats could help prevent an overly restrictive use of the results (methods here are evaluated against realised warming), avoid the impression that the Paris goals must necessarily be understood in terms of 20-year-average warming (a point the text already makes, but the evaluation itself is biased toward realised warming), and could also open the door to further comparisons using different evaluation targets or metrics.

On this topic, a clear position from the scientific community on the most appropriate variable for interpreting the Paris Agreement would certainly be welcome, but again, I assume this question lies beyond the scope of this particular paper — and it is entirely possible that reaching a consensus on this point within the scientific community would be difficult.

3. Evaluation of methods for estimating the current warming level

A substantial part of Section 4 concerns the evaluation of the various methods (in fact, a remarkably large and diverse set of methods) for estimating the CGWL. I have the following comments on the proposed evaluation approach:

a. Evaluation period

For various reasons, including limitations on the applicability of certain methods, not all methods are evaluated over the same time period. Some are applied over the full 1850-2025 period, others over a shorter period. This may contribute to differences in performance. Furthermore, using data from, say, the pre-1950 or even pre-1970 period is likely to pose difficulties for certain methods that no longer really arise in the present period. These difficulties include: the extrapolation of very weak or only emerging trends in the early 20th century; and the relative disagreement between HadCRUT5 observations and CMIP models over this period (at least for methods that use information from these models in one way or another), which may be linked to a shortcoming in the observations themselves.

For these reasons, it would be worth checking whether an evaluation carried out over a common period for all methods, and restricted to a more recent period (in which climate change is clearly well established), leads to the same conclusions in terms of method performance (and relative performance).

b. Target for human-induced warming

For attribution methods aiming to estimate human-induced warming, the 20-year running mean does not constitute a suitable target, since it represents total warming rather than the quantity these methods are actually trying to estimate. This may be a cause for concern, since the level of precision sought here — e.g. RMSE < 0.06°C — is comparable in magnitude to the warming that would be induced by the mere absence of a major volcanic eruption over the past few decades.

4. Synthesis of multiple datasets (Section 5)

I greatly appreciated the proposed method for synthesising multiple available datasets into a single product, along with the associated ensemble designed to represent the total uncertainty — meaning both the uncertainty sampled by the available uncertainty estimates and the 'external' uncertainty arising from the spread among the various datasets. The split into two sub-periods, the proposed weighting scheme accounting for the (strong) dependence among datasets, and the generation of a representative ensemble (via clustering) are particularly nice features. The final result, suggesting that the resulting synthesis product provides both a best estimate and an uncertainty range consistent with previous assessments, is also quite convincing. That said, I feel that some aspects of the method used in this section would benefit from a more detailed description.

a. Weights (e.g. Fig. 24)

The manuscript explains that 100% probability is assigned at the root of the tree, making it possible, in principle, for the reader to calculate the weight actually attributed to each product. However, to make this easier, it would be helpful to provide the weights used for each product somewhere — either directly in Fig. 24 or in a separate table.

b. Period 1850-2010

My current understanding is that Fig. 24 describes the dependence structure (and hence the weights) used for datasets available over the second period, i.e. from 1981 to the present. The manuscript suggests that a similar weighting approach is used over the first period, i.e. 1850-2010. If this is correct, I would suggest also describing the dependence tree for products covering this period, and specifying the values of the final weights used.

c. Performance

The weighting scheme, while attractive, does not take into account the performance of the datasets considered in any way, and this choice does not appear to be discussed. Taking SSTs as an example, my current understanding of the recently developed DCENT dataset is that it applies improved corrections for inhomogeneities in the SST record, which are now well understood and well documented. The resulting SSTs do not reproduce some of the inconsistencies reported for previous datasets (e.g. the early-20th-century cold period). Altogether, this suggests that this product may offer greater homogeneity than others, which here could be regarded as higher performance, potentially warranting a higher weight in the final product. At some point, with respect to potential future updates of the synthesis product (for which the authors put forward very specific suggestions), strong evidence that one dataset performs better than others at correcting SST measurement bias could lead to other datasets being excluded from the synthesis for failing to meet the state of the art in homogenisation, or being assigned a reduced weight. I am not necessarily calling for any change to the proposed weighting scheme here, but I think the authors should at least discuss this performance issue, and possibly acknowledge that it could lead to substantial changes in the proposed weighting in the future, depending on new available evidence.

d. Splicing in 1996

My understanding of the splicing step between the two periods, in 1996, is that members before 1995 and after 1996 are drawn completely independently of one another. This independent draw seems to me to carry a risk of artificially inflating the variance (depending on the autocorrelation structure of the errors). Could the authors verify, or demonstrate, that this splicing step does not introduce an undesirable increase in variance?

For example, if this procedure were applied to an ensemble currently available, such as HadCRUT5, by introducing an (artificial) split in 1995 and then re-splicing members drawn independently before and after that date, would the variance of the original ensemble be altered?

e. Comparison with other syntheses

The paper mentions that previous studies have attempted to produce probabilistic syntheses of the various available datasets, e.g. Craigmile and Guttorp (2021). Is there any insight or comment on how the synthesis proposed here compares with these previous attempts at such a synthesis?

f. Multi-modality

The fact that only a very small sample of datasets is currently available might induce an undesirable or unrealistic structure in the uncertainty ensemble, such as multi-modality in the distribution of temperature at a given date. In practice, some degree of multi-modality does appear to arise — e.g. in Fig. 25a, around 1910. This might result from one dataset making different choices than the others in correcting inhomogeneities over this period. This feature of the uncertainty ensemble could be questioned, as it is difficult to justify why, for example, the temperature anomaly in 1910 could be +0.1°C or around -0.2°C, but not 0°C. Could the authors discuss this potential multi-modality issue? Possible solutions could involve explicitly assuming a normal error distribution across datasets (while preserving the same variance, so as not to increase the overall variance of the final product). I nonetheless acknowledge that this is likely not a top-priority issue.

Aurélien Ribes