

Reviews and responses

We thank the several reviewers for their substantive and constructive reviews and the Editor for their significant efforts.

We submit here one collated response including the original comments in full and our responses. Original review text is in black and our responses in blue to aid readability hereafter. To further ensure clarity each reviewer comment and our response starts on a new page. Often reviewers had comments which were either expanded upon or linked to those of others and we believe a combined response will aid readers, reviewers and the Editor make better sense of how we intend to respond and revise the manuscript to address the concerns and queries raised.

Requests for immediate guidance from the editor to aid finalisation of a resubmission

The reviews have raised some issues which we would appreciate immediate guidance from the editor upon as we finalise the redrafting:

1. CC3 and Anonymous Reviewer 1 had both suggested that we add accessible summaries at the start of sections (we are assuming at the level of 2, 3, 4, 5 and possibly 6 (although as that is not technical it may be best not to)). We can do so but were unsure of journal practice, whether you saw it as desirable, and if they were to be added would appreciate guidance on expected length and remit.
2. The reviewers have variously alluded either directly or indirectly to the potential usefulness of a roadmap figure being added to the introduction. There is precedent for such a figure in the IGCC papers on indicators by Forster et al that we could use as a starting point. Would adding such a figure in your view be useful?
3. Whether the list of acronyms should be elevated from the supplement to the main text as indirectly alluded to in the group review in CC3.

Opening remarks

We thank all the community reviewers and the 4 formal reviewers for their various perspectives and suggestions which we will use to improve the manuscript. In addition, in the interests of full transparency, we received informally via email a number of suggestions from a range of experts highlighting papers and perspectives. Where possible these have been taken into account but are obviously not part of the formal journal record.

In addition to the formal reviews, to ensure the paper is up to date at point of publication we have updated all analysis in Section 5 to include the calendar year 2025 and an assessment of the latest suite of observational products, their publication status and update protocols. This has led to some changes in the suite of products used for both warming to the modern period and modern period warming and their joining. This will be reflected in tracked changes in the manuscript but, in brief and with major changes italicised and new datasets in addition bolded:

- *HadCRUT 5.1.0.0 analysis update from version 5*
- *HadCRU MLE v1.4 update from version 1.3*
- Kadow et al. one additional year
- *Berkeley Earth high resolution new version of Berkeley Earth added*
- *NOAA GlobalTemp v6.1 update from version 6.0*
- NOAA GlobalTemp v5.1 unchanged
- China-MST3.0-lmax one additional year
- *DCENT_MLE v1.2 update from version 1.1*
- ***DCENT-I new interpolated version of DCENT added***

[Note that DCENT is not used in preference to these interpolated versions]

- COBE-STEMP3 one additional year
- GloSATref unchanged
- GISTEMPv4 one additional year
- ERA5 one additional year
- JRA-3Q one additional year
- ***CMA-GMST new dataset added to list***
- ***Cooper et al. new dataset added to list***

The addition of 2025, unsurprisingly, has also increased the likelihood that we have already exceeded 1.5 and this is reflected also in key numbers and language in the abstract and conclusions.

During the review period some issues in realtime monitoring for some products used arose. These have generally been fixed by the dataset producers but this has served to highlight the criticality of timely monitoring of data quality. Some discussion to this end has been briefly added to the main manuscript and expanded in the Supplement.

Gareth Jones

I found the pre-print interesting. I have some comments that I hope the authors will consider.

We thank Gareth Jones for his thoughtful and constructive review.

L171-176 and L179-195:

"Figure 1: (left) A common misconception is that an increase in the estimate of historical warming (due to scientific progress) brings us closer to a level of warming at which some projected impacts may occur."

I don't think it is fair to call this a "misconception". While future climate impacts are linked to "1.5C" it is natural for people to expect that to mean at 1.5C with respect to a pre-industrial period.

Being explicit by quoting the temperature level relative to the more recent reference period would avoid any "misconceptions".

We have swapped misconception for misunderstanding based upon this feedback and also nuanced the balance of the sentence. We think this is a more factual wording that better reflects our intent. We thank Gareth for pointing out this issue.

Another source of possible confusion is the choice of recent reference periods are used. Allen et al. 2018 (Section 1.2.1.2) mentions different reference periods, but not 1981-2010.

We note this but Allen et al is not referenced here and we prefer to keep the introduction brief noting well-founded concerns regarding length articulated elsewhere in the reviews as a whole. Our discussion later around reference periods chose to prioritise the AR6 WGI approaches on the assumption (backed up by various RSO documented decisions arising in consecutive COPs since 2021) that these represent the basis for current policy decisions in UNFCCC and other fora.

L796-799:

It should be restated here what other temperature dataset is used to enable ERA5 to produce an anomaly relative to 1850-1900 (lines 598-599)

We prefer to provide this clarity in the figure caption by back reference to the section description than to break the text flow and lengthen the passage. This is particularly with a view to well stated concerns by other reviewers over length.

As the warming from 1850-1900 to 1991-2020 comes from other datasets, does ERA5 actually provide anything useful here? At best it just gives the change of temperature since ~2005!

This should be made clear.

This practice is well established and has been in usage by WMO in its annual reporting and others. What ERA5 can add over and above the traditional datasets is both timeliness of updates and a daily resolution of monitoring. Given the now common practice and broad acceptance of such an approach we do not feel that the perception here is justified by evidence and practice. ERA5 goes from the mid-20th Century to present so provides valuable direct information back much further than 2005. Furthermore, such issues are not limited to ERA5 alone. Based on this no changes are proposed to be made here.

L226-229, L364-368, L383-392:

Jones 2020 showed that it is "generally appropriate to use global near-surface air temperature diagnostics to compare simulated historic climate change with observed temperature changes".

Differences in model land coverage, model top ocean level depths, and in absolute model sea-ice coverages make interpreting differences between GMST and GSAT very difficult.

The differences in simulated GMST and GSAT are smaller than other model and observational uncertainties, so should not be overstated.

As noted, observational evidence for MAT warming faster than SSTs is not strong, indeed evidence is there for MAT warming less than SSTs.

It should also be noted that GloSATref (Morice et al., 2025), has a smaller warming trend in MAT than in the SST in HadCRUT5.

As the uncertainties in the observational records are highlighted (Lines 383-392) other uncertainties, such as how models simulate ocean temperatures and sea-ice coverage, should also be highlighted.

Thanks for this feedback. Lines 226-229 are introducing the issue and we don't believe require alteration as such. We have added a reference to Morice et al. at the appropriate juncture in lines 383-392. We in addition now cite Jones (2020) through the addition of a sentence at the end of the summary for this brief section including explicit mention of aspects highlighted above. We feel this strengthens the overall text.

L375-380

Although models do have a "sea surface temperature" diagnostic, this is actually generally deduced from the temperature of the top layer of the ocean. That should be made clearer.

We have clarified this in edits to this segment of the caption.

L1090-1091

This somewhat mischaracterises the Otto 2015 approach. Otto 2015 use a least-squares fit approach to regress global observed temperatures against the output of a "simple two-component impulse–response model".

Hasselmann 1997 is an optimal fingerprinting approach which describes the rotation of spatio-temporal climate response patterns to improve the signal-to-noise of the wanted signals in what effectively becomes a regression.

Otto 2015 does not use spatial information and is non-optimal, so can't really be said to build on Hasselmann 1997.

That is not to say it is not a useful approach, but a much better reference to use that Otto 2015 builds on is a paper like Schonwiese and Stahler, 1991.

The ROF approach (L1114) can be said to be built on Hasselmann 1997, that should be reflected.

Thanks for this careful reading and suggested alterations. We have adjusted the text accordingly.

L1109-1110:

No, the KCC method relies heavily on emulators to estimate GHG and natural global mean climate responses.

We have removed 'rather than emulators here' to avoid the confusion.

L1112-1113:

I would disagree with the statement that the KCC method can produce "attributable temperature changes".

The method assumes that global mean temperature observations are interchangeable with the simulated historical forced climate change, "assuming model-truth exchangeability" (Ribes et al., 2021).

This means they start with the assumption that the observed change is attributable to historical forcing changes.

To deduce attributable changes would thus be circular reasoning.

While the method can't be used for attribution, it can be useful to constrain future climate changes based on how well simulated past changes match observed changes, only as long as the assumptions it relies on are stated.

This is probably debatable. Nevertheless the use of the word attributable is not necessary here so to avoid potential for issues we have removed it.

L1120-1125:

It should be noted that any method that just uses global mean temperature can be prone to over-fitting, and it should actually be expected that the net anthropogenic warming would be more "tightly constrained" whilst global temperatures are increasing in such approaches. As noted the separate GHG and other anthropogenic contributions are much more uncertain, so it should raise concerns that the net anthropogenic attribution is over-confident (Jones et al., 2016).

We thank Gareth for raising this important point. While it is definitely true that net anthropogenic warming is more tightly constrained by observations than GHG and other anthropogenic forcings alone, as has been repeatedly pointed out in the attribution literature over the past twenty years, this is somewhat distinct to the point we were trying to make here. Having discussed this internally we feel broadly comfortable that this summary text is a fair reflection of the preceding text and the literature and would prefer to retain as is. Concerns as to any potential overfitting are further mitigated through the benchmarking exercises discussed later in Section 4 and further again in Section 5 by comprehensively spanning observational structural uncertainty in the estimation approach.

L1153-L1159:

The "ROF & KCC" in bottom left should really be separated from each other.

ROF uses coupled climate models for the estimates of responses to different forcing combinations, and for estimating internal variability.

KCC uses coupled climate models for historical and GHG responses, but also simple relationships for some GHG responses, and for natural responses, and an assumed AR process for internal variability.

I don't understand why the GWI method has what appear to be a complex relationship between land, ocean and air, but the far more complicated relationships in coupled climate models is not represented.

We have revised this figure and hopefully the revisions address some, if not all, of the concerns raised here. Any schematic is bound to in some senses oversimplify. The schematic was added because of concerns that to non-specialists the technically dense text may be indecipherable.

We have attempted to alter the figure in a way that may address at least some of the valid concerns raised here.

References:

Allen et al, 2018: Framing and Context. In: Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty, Cambridge University Press

Jones, Stott and Mitchell, 2016, Uncertainties in the attribution of greenhouse gas warming and implications for climate prediction, Journal of Geophysical Research

Jones, 2020, "Apples and Oranges": On comparing simulated historic near-surface temperature changes with observations, Quarterly Journal of the Royal Meteorological Society.

Morice et al., 2025, An observational record of global gridded near-surface air temperature change over land and ocean from 1781, Earth System Science Data.

Ribes et al., 2021, Making climate projections conditional on historical observations, Science Advances

Schonwiese and Stahler, 1991, Multiforced statistical assessments of greenhouse-gas-induced surface air temperature change 1890-1985, Climate Dynamics

Leon Hermanson

We thank Leon Hermanson for the careful review of section 4.5 and the feedback.

I have some small comments to increase the accuracy of the language in this article in section 4.5. Throughout this section (see for example line 1169) there are references to "decadal forecasts" with ESMs. Very few ESMs are used in decadal climate prediction. The vast majority of prediction systems are coupled atmosphere-ocean-sea ice-land models, but have no biology and chemistry (beyond these being prescribed), so to refer to them as ESMs is inaccurate.

Reviewing the section it is really only this point 1 in the enumerated list that is not using ESMs. In points 2 and 3 it truly is principally or exclusively ESMs. We have thus substituted ESM only once in the enumerated bullet 1 to the suggested more accurate descriptor although note that to many the distinction will likely, sadly, be unclear.

Line 1169. "WMO Annual-to-Decadal Climate Prediction multimodel ensemble" not "WMO decadal forecast". If appropriate, feel free to reference the website: www.wmolc-adcp.org

We have amended accordingly and added the link which we agree might be useful for the reader to refer to.

Similarly, line 1200, might be better to use "decadal climate predictions" and use another term than ESM (as long as the new term is applied consistently throughout the document)

We have edited to clarify and for consistency along the lines suggested and ESM no longer appears here.

Line 1172-3: "some of these have high climate sensitivity". There is no evidence the multimodel ensemble is biased high, the situation is better described as for the projections, "cover a range of climate sensitivity in an unstructured way".

We have edited along the lines suggested here.

Line 1179: Atlantic Multi-decadal Variability in models has a timescale of 60-80 years, so unlikely to average out over 10 years. This should perhaps be acknowledged. Papers to cite can be found at the bottom of this page: <https://climatedataguide.ucar.edu/climate-data/atlantic-multi-decadal-oscillation-amv>

Given considerable ambiguity in whether AMV is a result of external forcing, internal physical variability with a defined spectrum or a combination of external forcing exciting an internal variability amplification we prefer not to expand here. The paper cited showed that over the whole historical period at the global average the result held. This does not mean that it holds regionally but as the paper is about global change we prefer to retain as is.

Thanks for writing such a detailed and necessary paper!

Annika Hogner (group comment)

The paper is comprehensive and your extensive assessment of different methods to determine global temperature increase in near-real-time, as well as your proposal for a synthesis methodology, fill a policy-relevant gap.

We thank Annika Högner, Verena Kain and Carl-Friedrich Schleussner for their careful reading of the manuscript and their constructive suggestions.

We suggest a number of larger edits to improve the clarity of the manuscript. In the current form, it takes great effort to read the full paper, and the overarching steps you undertake as well as the relationships between them only emerge once the reader has read more or less all of it. The proposed edits could increase the ease of access to your important work for a general readership, including policy-makers:

A paper of this scope and length would greatly benefit from an index that gives an immediate overview of its different sections and parts.

We would be happy to add such an index but it does not appear to be consistent with the journal style rules. Based upon precedent in Forster et al. we could add a visual summary of the paper structure to the end of the introduction but this would exacerbate length concerns expressed by at least one official reviewer. We would defer to the Editor's guidance on the matter but would be happy to attempt such a visual guide. See Figure 1 in <https://essd.copernicus.org/articles/18/3889/2026/> which we could build from as an initial template.

A list of abbreviations would also be helpful, or alternatively a decrease in the use of abbreviations that do not frequently occur or are not widely established.

We have a list of abbreviations in the Supplement. This could be raised to the main paper but as some reviewers have noted the paper is already very long, and again we are unsure whether it is consistent with journal style rules. We have reviewed the abbreviations once again, updated the table in the Supplement accordingly, and attempted to strengthen the signposting of this resource from the main text. We have also attempted to further reduce the number of acronyms used although note concerns about length which this would exacerbate. If the Editor would prefer the list of abbreviations to appear in the main paper this can, of course, be done but we fear its elevation may break the flow and exacerbate rather than ameliorate length concerns.

Each overarching section would benefit from a section-specific "Figure 1", i.e. a schematic overview figure that summarises what the respective section is assessing and how the different parts relate (e.g. schematic overview of major dimensions of uncertainty in determining whether we have reached 1.5°C; schematic overview of methods intercomparison

and testing; schematic overview of suggested synthesis dataset+method for (quasi) real-time tracking of global warming)

This is an intriguing suggestion but, again, we have been critiqued for already being too long and it would yield a broad range of schematic figures. We will be preparing a range of outreach material with colleagues expert in communications that will accompany the paper. It would be possible in theory to insert such summary figures for each section because we will anyway be preparing such materials but it may take some time and we have concerns around balance. We have requested guidance of the editor in the first instance as to appropriateness given article length and journal policies before making a final determination.

Generally, we noticed that you have structured the paper in a way where you often go into detail first, then provide a summary at the end of a section, rather than starting with a summary of what you will do, then doing it. Reversing the order of this pattern would aid the reader to follow along and more easily provide a higher-level path through the paper if a reader cannot commit to reading the full paper.

Thanks for this suggestion. As you will note this has been elevated to a request for clarification to the editor up front in this response document. In principle we could do this but for some sections we would need to add new material and we are unsure of what balance of summary would be appropriate and how to avoid being somewhat repetitious with the overarching paper abstract. We will follow the guidance provided by the editor here,

We would note that in addition the WMO and UK Met Office have been working on an accessible summary for communication of the results to more generalist audiences once the paper has been published.

We also plan to work with GSCC - an environmental communications group to ensure results are clearly articulated.

Further, it makes a lot of sense to clarify the distinction between realised warming and anthropogenic warming and to untangle how to extract the anthropogenic component as well as what the difficulties are in doing so, as you do in the paper. Given the UNFCCC Definition (Article 1.2) of climate change, however, we consider it unambiguous that under the UNFCCC this refers to anthropogenic warming, which we therefore conclude also determines how the LTTG should be interpreted. Acknowledging other existing definitional ambiguities, we argue not to introduce doubt where the guidance is clear. We do not find your argumentation in Consideration 2 convincing that realised warming is a more relevant metric for impact assessments, as it is commonly global warming levels, referring to anthropogenic warming in ESMs, that are used to assess climate impacts directly. Considering the policy-relevance of any (quasi) real-time assessment of global adherence to the LTTG, we see a strong need for clear and unambiguous guidance accessible to policy-makers and the larger public. Maintaining two options without a very strong reason or evidence in policy guidance will likely cause confusion.

Therefore, we suggest to adapt Consideration 2 in favor of assessing anthropogenic warming, and accordingly remove the pulldown menu option from Figure 28 and show anthropogenic warming, not realised warming, in Figure 28.

This has been the source of much debate within the author team and the above represents one of several views that have been formulated, articulated and robustly debated within the team over the almost two years that we developed the draft. Other, equally strong and reasoned, views have been articulated in formulating this section that it is realised warming that matters. The basic premise here is that hazards are understood in AR6 to be first-order dependent upon the global warming level and are agnostic to the source of that warming. Were the world to warm 3 degrees due to non-anthropogenic sources we would need to adapt to that warming even though it was not due to human activities. From an impacts and adaptation viewpoint it is therefore realised warming that matters.

Ultimately, this is one of several ambiguities where parties to the Paris Agreement could make an explicit determination and this would serve to reduce ambiguity. We prefer, on the basis of attempting to be a community paper and representing the views of the team as a whole, to broadly retain the framing as submitted while recognising the validity of many of the points raised in this comment that many, but far from all, on the team share.

Ultimately, we're providing a tool, not explicit policy guidance or interpretation of a geopolitical text.

Footnote 7 (page 8): This differentiation between reached and exceedance in the AR6 language bears significance. If we recall correctly, it has been introduced as global warming under SSP1-1.9 was not *likely* to exceed 1.5°C. It could be of interest to reflect on the relevance of this differentiation in the context of statements of concurrent levels of warming (i.e. Line 1980) that speak to warming levels being reached. Ultimately, this is of course a question of value judgements rather than science, but there probably should be a difference in confidence implied by using concepts of *reached* (Median estimate?) and *exceedance* of a warming level (which may require a higher confidence?). This might be a relevant differentiation to unpack a bit further and provide a reflection on the appropriate use of language also for the scientific community given the seminal nature of this contribution.

We have reviewed the footnote and text highlighted against the AR6 SPM and underlying chapter and are comfortable that the text faithfully reflects the AR6 findings and intended meanings. Many of the author team were involved in the negotiation of the text in Part B1 of the SPM and we recognise that this is open to different interpretations arising from politically negotiated text. It comes down to whether we refer to the median, or some confidence interval. If we refer to the median, then (given the currently monotonic global warming trend) 'reaching' a level in one year would imply 'exceeding' that level in the year after. Whereas if we use a confidence interval, then we'd also have to say what percentile we mean for a 'reaching' statement, but then exceedance would still follow in the year after for that same percentile. The

difference only arises if one assumes that a 'reaching' statement can only ever mean the median, and that exceeding should use some other percentile. In reality there is likely no enduring (ontological) distinction between reaching and exceeding. We have, for clarity, simplified language about remaining below which may help address some of the issues raised here.

We would further note the exact AR6 text is:

Under the five illustrative scenarios, in the near term (2021–2040), the 1.5°C global warming level is very likely to be exceeded under the very high GHG emissions scenario (SSP5-8.5), likely to be exceeded under the intermediate and high GHG emissions scenarios (SSP2-4.5 and SSP3-7.0), more likely than not to be exceeded under the low GHG emissions scenario (SSP1-2.6) and more likely than not to be reached under the very low GHG emissions scenario (SSP1-1.9).

This is slightly distinct from the implication from the comment.

As for your policy recommendations, the Periodic Review of the long-term global goal under the Convention is established as the process under which the leading question of the paper is assessed and where necessary clarifications in terms of interpretation and methodology have taken place in the past and more of such could take place in the future. It would make sense to reformulate your recommendations accordingly and emphasise how your work can help inform the next periodic review. We also note that some policy recommendations demand definitional clarity from a political process that it may not be equipped to provide. As you also discuss in your Consideration 1, the 2nd Periodic Review has provided clarity that “*the long-term global goal ... is assessed over a period of decades*” (Paragraph 5, decision 21/CP.27). We find your approach very much in accordance with this as you show clearly how different methods align, but it might be worth discussing this policy guidance in terms of your present warming level estimation and your confidence that it aligns with a multi-decadal average.

Our understanding is that the periodic review is not necessarily binding in the way that e.g. the decision on GWP metrics was. It is also unclear from the available UNFCCC web resources what the status of the Periodic review might be. The main landing page for the periodic review (which we assume to reflect the current status) states:

At its twenty-ninth session (2024), the COP started consideration on the continuation of the periodic review, taking into account experiences from the 2013-2015 periodic review and the second periodic review, as well as from the first global stocktake. The consideration, which will continue at COP 30, will take into account potential overlaps and synergies between the periodic review, the global stocktake and other relevant processes under the Convention and the Paris Agreement, as well as the availability of new information relevant to the periodic review.*

COP30 deferred the discussion to COP31 with little substantial discussion. It would therefore seem unwise to explicitly tie to this process. This work is, in part, intended to be useful to whichever processes, including but not limited to the periodic review, parties might wish to use to take the question forwards.

On the clause from 21/CP.27 para.5, upon further reflection we would suggest that this clause does not provide full clarity, but still leaves a definition of the LTTG underspecified; a discussion on this has been added in section 4.2. In brief, the periodic review was instigated with the explicit condition that it would result in a redefinition or amendment of the LTGG (5/CP.25), and the clause itself is a note, one of the weakest verbs in the UNFCCC lexicon. By contrast, where parties have in the past prescribed metrics, this has been done explicitly, see for example the adoption of GWP-100 in 18/CMA.1, para. 37. Either way, this doesn't weaken a fundamental argument of the paper; a metric for the LTTG remains unspecified, and we believe it would be of use for the community to assess the best options available to us now, to determine the level of warming.

L1865: This section lays out the issues well, but we find the argumentation not fully convincing. Yes, key risks are assessed at lower levels of warming in AR6, but how does this justify a particular interpretation of how to interpret the targets in terms of evolving evidence on historic warming? That's a rather hand-waving connection. If this was ever to change again in the future, would we adjust our interpretation of temperature records also the other way?

Under evolving scientific understanding our estimate would change if estimates of long-term warming changed. We are not clear how we could better explain this differently than presently. We would note that understanding of observed changes continues to evolve (see opening remarks in this response to reviewers) and thus it is all but guaranteed that understanding of historical changes will continue to evolve as new datasets and insights accrue.

We think the more convincing argument one could and should make is that the UNFCCC followed the AR6 approach in its GST decision, and point to relevant paragraphs of the CMA decision that substantiate this (i.e. paragraph 15a CMA5).

The decision given in https://unfccc.int/sites/default/files/resource/1_CMA.5.pdf paragraph 15a restates the IPCC AR6 findings and indeed would imply that UNFCCC in the context of GST1 had regard to the IPCC AR6 findings and evolving best scientific understanding but does not provide a definitive iron-clad basis to decide which interpretation to follow. Furthermore, the UNFCCC language of *noting* is weak in comparison to the far stronger language (*shall*) used where the UNFCCC does formally mandate a metric as discussed in our response above.

We would also recommend to additionally track warming since the Paris Agreement, considering that 2015 warming itself can now easily be assessed with a 20-year centred mean and can, thus, serve as a reference point. Tracking the warming since the Agreement would

provide a very useful addition for evaluations of the appropriateness of mitigation efforts undertaken since the commitment to the LTTG was made.

This suggestion is functionally similar to our worked suggestion in the submitted manuscript to assess additional warming since the period 1981-2010 with a fixed estimate of warming up until that juncture. The question of whether to use warming as understood at the time of the Paris Agreement entering into force or our current understanding of that metric remains. Given that updates to all products occurred over 2015-2024 however it would only be possible to create an estimate since the Paris Agreement came into force using an estimate of the warming to the 20-year period centred on 2015 using modern scientific understanding. It would not be possible to create a similar estimate using 'frozen understanding' of what negotiators in 2015 understood the warming to be at that time. As noted in IPCC AR6 WGI CCB2.3 and Section 3 of our manuscript there were substantive changes to historical period warming estimates thereafter.

Given the broad functional equivalence to the proposed use of 1981-2010 plus a delta to reflect warming up to that period already deployed and concerns over paper length we do not pursue this suggestion further here at this time. We also have concerns that only 11 years have passed since the Paris Agreement and therefore how stable such an estimate might be. However, this does not preclude its use in future monitoring efforts by WMO or others and it is procedurally simple for anyone to do using the data and code provided.

Finally, three minor edits/comments:

page 32 line 762, please introduce NMHS (or just spell it out)

We have added the definition to the list of acronyms given in the supplement

page 49 line 1174ff, the scenario-based projections are not up to date with current historical emissions even when they first come out, e.g. the CMIP7 ScenarioMIP scenarios are harmonised to 2023 historical emissions; may be worth adding a sentence noting this, as additional difference to the initialised forecasts and potential limitation when using CMIP projections to estimate current warming

There is substantive discussion over the forcings issue and their updates elsewhere in the manuscript already and we would like to avoid repeating these points already made about forcings and their updates given concerns articulated by reviewers about length.

page 88 line 2020 "repatriations" we assume should say "reparations"

Yes, you are correct. Thanks for spotting this typo which we have now fixed.

Annika Högner, Verena Kain, Carl-Friedrich Schleussner

Anonymous referee 1

Review of 'How well can we quantify when 1.5°C of global warming has been exceeded?' by Thorne et al. submitted to ESSD.

This paper aims to assess changes in the global surface temperature to answer the question posed in title. For this purpose the authors use best available observations and a wide spread of methodologies many of which provide alternative ways to quantitatively assess the changes in surface temperatures, thus providing multiple lines of evidence. The authors conclude that with respect to preindustrial levels we at the end of 2024 were on the level of +1.40 °C global warming of which 1.34 is human-made.

The paper represents a very timely post-AR6 assessment and goes far beyond consideration provided in different AR6 chapters. Moreover, the paper employs unprecedented number of data and various methodologies for estimating uncertainties. This makes the paper highly important for scientific community and I recommend it to be published subject for minor revisions.

We thank the anonymous reviewer for their summary and their recommendation that the paper be published subject to minor revisions.

Comments

1. The authors give comprehensive argumentation for selecting 1850–1900 as a 'pre-industrial' level used in this study. However, I would suggest making a clear statement (like section highlight) in the beginning of Section 2 before discussing different issues associated with this definition. This will help wide readership to navigate further the material.

This is highly similar to the suggestion in CC3 of summary segments to each section up front. As noted in the responses there we have elevated this as a general query to the editor and will update or otherwise depending upon the guidance we receive back.

2. Try to think whether encapsulating section 3.1 and placing it before section 2, or even the alternating the whole sections 2 and 3 would be a reasonable approach providing a better shape of the paper. To my view these changes to the structure of the manuscript will make it better shaped.

Thanks for the suggestion and encouraging us to reconsider structure and narrative continuity here. We tried various ways of reorganising this text offline and reading for narrative flow. Splitting the differences in metrics of global surface temperature from the discussion of observed changes we don't feel helps the narrative overall. Placing the definition of pre-industrial after section 3 might help the early narrative but would yield a more jarring disconnect with section 4. There probably isn't a perfect way to do this so we

prefer on balance to retain the present structure which requires fewer overall changes being made. We have tried where possible without adding undue length to strengthen the overall narrative, however, in particular to address specific comments raised by the reviewers.

3. Fig 2 and associated text. Over ocean you show the number of months with data over the ocean. However, the number of months may not necessarily reflect accurately the sampling problem as for the regions with relatively high intramonthly / synoptic / subsynoptic variability the number of reports is equally important. I understand that in section 3.2.2 the issue of sampling will be discussed, however already here it would be useful to provide information about the completeness of the records in terms e.g. of the number of months with such and such number of obs (or another metric can be used). Anyway, the reader has to be informed how many months with e.g. 1-2 VOS report and how many with 10-30, especially in the regions with highly variable surface conditions.

We agree that sampling aspects need to be made clearer here in the caption and have done so. We may add to the supplement a figure of observation counts over time and reference it from the text. The relevant authors are developing and discussing this presently.

4. GSAT vs GMST (SST vs MAT), lines 381+. First it would be useful to state that conclusions here are valid for long-term change estimation, not to implicitly provoke using the conclusions for the analysis of short-term variability, especially on the local/regional scales. For this purpose it is necessary to highlight clearly somewhere in sections that it is obvious that SST and MAT are physically different variables and their consideration for estimation of trends. AR6 and before does not imply the possibilities for extensions across scales. Maybe also the problem of e.g. formal definition of surface fluxes could be mentioned if SST anomalies are used for quantifying SAT over oceans.

We have added explicit text around local and high frequency differences to attempt to address this concern. Given the length of the paper we would rather not discuss issues around fluxes which would be a (or perhaps more correctly at least one other very long!) paper in its own right.

5. Sections 3.2.2 and 3.3.3. It would be good to articulate that besides the infilling algorithms for the data void areas there is a problem of sampling uncertainties associated with insufficient number of samples even for formally non-gappy cells (I also mentioned this issue above). Definitely, this problem is less sensitive for temperatures compared to e.g. winds or flux products (see, e.g. <https://doi.org/10.1175/JCLI4010.1>, <https://doi.org/10.1175/JCLI4008.1>) but anyway,

this should be mentioned. In some works of Liz Kent these issues are covered and uncertainty estimates are quantified. It would be also good to mention that sampling uncertainties (not infilling fully unsampled regions) might be regionally important but probably less important for global temperature uncertainty estimates.

We have added some text to the start of 3.2.3 here to try to make this point clearer to the reader without adding undue length. We have also had to adjust the text to incorporate the Cooper et al analysis and the new DCENT infilling.

6. Section 4. I would suggest splitting this section into 3. Sections 4.1 and 4.2 can constitute one section, section 4.3 could stay as another separate section and the remaining sub-sections could constitute another separate section.

This is really a question of presentational style. At the moment the Section 4 is structured as a large continuous section because all aspects feed to a greater or lesser extent into the Section 4.7 summary. Splitting into these smaller pieces would be problematic for that reason, although not insurmountable. On balance whilst recognising that section 4 is very long we prefer to keep it as one section with well signposted structure.

7. Figure 16 is hard to handle. It has to be either decomposed into separate figs or/and simplified. In the left column there too many symbols whose meaning can hardly be handled. I am not sure, that the right column is really necessary as it stands now. Intercolumn links are hard to assess. Maybe 2 figs + associated Table could serve better.

This figure also was commented on by Gareth Jones. We are redrafting for clarity and for decision as to retention or otherwise in the redraft. This redraft is taking into account this feedback.

8. To my understanding fig 20 and table 2 have large overlap in the contents. I think, Table (with minor extensions) could serve better for the paper.

This is the first of several such paired Table / Figure presentations of results in this subsection. We agree that there is a degree of overlap but the overlap is far from 1-to-1. Also we are wary that the diverse readers of the journal will prefer figure or tabular presentations. Choosing one over the other feels like it might actively harm the ability of the broad journal readership to understand and digest these key results. We therefore prefer to

retain both the table and the figure presentation whilst absolutely recognising the valid concern that this means a degree of information redundancy here.

9. Table 5 can definitely be moved to the Supplementary material.

We have removed Table 5 and will rely in the resubmission instead on Table S4 in the supplement which already extended Table 5

10. What is the reason for using B-W colors in Figure 25? Color would serve effectively here.

Figure 25 represents equiprobable solutions drawn from the large ensemble (whilst recognising that this is not a truly random draw - see response to Aurelien Ribes). The choice of black and white is deliberate to convey this equiprobability. For those who wanted to investigate individual members the timeseries are made available. We will clarify in the caption.

Anonymous Referee 2

This timely paper thoroughly reviews data and methodologies to assess global mean temperatures and long-term warming estimates. The authors propose a practical and scientifically sound methodology to combine observations and long-term warming estimates for policy-relevant monitoring of present-day warming. They also highlight knowledge gaps and the need for policy clarification, offering concrete solutions.

Hereafter come some minor suggestions for improvement.

We thank the anonymous reviewer for their summary and their recommendation that the paper be published subject to minor revisions.

Please mention early on in the main text that a list of acronyms is available in Supplement (Table S1) and possibly add the following entries: EBM; ECS; OHCA; TCR

We have added the requested terms. The Table S1 is pointed to at the end of the introduction when signposting overall paper structure. There isn't really anywhere obvious that this could be introduced earlier in the text in a coherent way. We therefore retain the referencing as is while noting the request from the community review to elevate this material to the main text. We seek guidance from the editor as noted in the prior response and will update accordingly based upon the feedback received.

Section 3.2, although well structured, is difficult to follow. A table of content could help. Particularly §3.2.3 calls on methods that are not fully named.

This is well noted but in many cases the long names of datasets etc add little value to the analysis and in the interests of respecting concerns from others about length there is a limit to what we can do. If the journal editor suggests we add summaries per section then we will try to use the summary to help better guide the reader here.

In Fig.17 It is not clear what metric was used to assess how close the UKCP ensemble members were "close to the HadCRUT5 observations' previous 10-year mean"? The mean of the previous 10 years or the year 2024.

It is a combination of both but we are unsure how to make this clearer than is already the case.

Footnote S1 should appear in the main text rather than in Supplementary material.

We have copied the footnote also to appear in section 4.6.1 but see value in retaining it also in the supplement for completeness and reader understanding

L542: EOF is not defined in main text (only in Table S1)

We have added the definition to the text

L515: EOT not defined at this point. Refer to section 3.2.3.

We have edited accordingly and also defined EOTs in the footnote

L761: NMHS not defined in text nor in Table S1. Supposedly National Meteorological and Hydrological Services

We have now defined in the Table S1

Fig. 12: yellow and purple boxes are not explained in the legend.

We do not see any yellow and purple boxes in Figure 12 so are unclear what the concern is here.

L948: for consistency, cite IPCC 5th assessment report as IPCC AR5

Changed accordingly

L1011 SLCF first defined at L1042

We have moved the definition up accordingly

L1038 Acronym EBM is not defined at this point, supposedly Energy Balance Model, could be defined already at L650: “Chan et al. (2023) used a more-detailed energy balance model to compare SSTs with their coastal LSAT” or at section 4.4.2 Energy balance models with a Kalman filter.

Defined now at the earlier point at L650

Footnote 26: "ocean heat content change ~~change~~ between 1971 and the present day"

Thanks. Typo corrected

L1087: It is not clear to which indicator this sentence refers.

We have tried to clarify here that the discussion is of human induced warming as an indicator

L1127 FAIR should be FaIR

Thanks. Typo corrected

L1128 misplaced parenthesis and subject appears twice

We are unclear what is meant here. To us this reads as intended

Fig. 15 Legend reads "Finite Amplitude Impulse Response" for FaIR, which appears nowhere in the text (only in Table S1). Maybe add to section 4.4.3.

Added to Section 4.4.3 as suggested

L1239 and Fig 18 lower panel: text and caption mention 4 most skillful methods, but only three shown in plot, since HadCRUT5 central estimate is the "ground truth". Also, the three methods shown here do not correspond to those with the highest log-likelihood in table S3. The caption could mention that the lower panel shows the best performing methods for realised warming with a log-likelihood > 2 . Then it would work with Table 2.

Thanks for noticing this and the constructive suggested adjustments which will be incorporated

Fig. 19 The caption mentions 1990, which is not shown in the plot. Likewise, the 8 years prior to 1995 with at least one day $> 1^{\circ}\text{C}$ and the 20-year median shown. Maybe move these two sentences to the main text, rather than the caption.

The information specified here is supplementary to the plot, included simply to reassure the reader that nothing of much interest happens outside the range of the plot, or in the case of the median that it can be a bit different from the arithmetic mean. These are minor points that would be distracting in the body of the text, and it is considered preferable for them to be placed in the figure caption.

L1268–1290 Description of Figure 20: it would be easier to follow if the comments were in the same order as the groups from top to bottom.

This is an intuitively appealing suggestion. However, we have played around with ordering here and feel that text coming in the figure appearance order leads to a feeling of text jumping around between topics which makes for a harder read so we would rather keep the text order as is here.

Table 2 caption wrongly states that `GWI_tot_CGWL` is highlighted in green, whereas the table reports `FAIR_nonat_unB` as having the lowest RMSE. `GWI_tot_CGWL` has however the largest log-likelihood. Please update caption to match data shown.

We had to update the table as apparently colours are not allowed in Tables and have updated the caption holistically to cater for this. In doing so we will have accounted for this useful point raised.

Figure 21 caption incoherent. It mentions ensemble members 11 but actually shows data from ensemble member 20. (I could find the figure for ensemble member 11 in the github repo (https://github.com/jnickla1/Thorne_15/blob/main/fut_ssp_spaghetti11.png)). Justification of choice doesn't match the data. 20 seems to be on the lower end, 11 closer to the mean. Please re-write caption to be coherent.

Thanks for pointing this out. We have comprehensively redrafted for clarity.

Table 3 caption wrongly states that GWI_tot_CGWL is highlighted in green.

See response to similar point on Table 2. Corrected.

L1528-1531: The sentence “we found that both the PIVW and CSCM approaches were more robust to different future scenarios (especially SSP1-2.6) if the weights were trained on 90 years of error (2000-2090) with respect to the centred 20-year running mean, rather than just 65 years of error (2025-2090).” doesn't seem to be in line with results shown at Table 5.

Table 5 has been removed per an earlier review request by the other anonymous reviewer and we will rely on the supplemental table. We will verify that this result is robust in revisions.

L1573 Last sentence is not clear. Please rephrase.

We have redrafted for clarity as requested here.

Aurelien Ribes

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.

We thank Aurelien for his comprehensive, careful and constructive review of the manuscript and for highlighting in this summary the multiple potential applications of the findings.

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 $0.1 [-0.1 \text{ to } 0.3]^{\circ}\text{C}$ warmer than the period centred on 1750, a warming that appears to be largely attributable to human activity, whose influence is estimated at $0.1 [0.0 \text{ to } 0.2]^{\circ}\text{C}$. [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?

You are correct that relying on reduced uncertainty during 1850-1900 compared to earlier periods to justify its use as a pre-industrial baseline is fundamentally a value judgement. You are also correct that our Section 5 approach of merging estimates across multiple time periods could be extended to account for pre-1850 warming, such as by using estimates from PAGES2K and GloSAT.

The results comparing 1850-1900 with 1750 are those from the AR6 assessment (cited in the text), which draw on a number of different lines of evidence (e.g. early instrumental, proxies, models), a similar approach to that suggested above. Such approaches have also been applied at national level to assess warming since 1850-1900 in countries with insufficient instrumental data from that period to use instrumental data alone (e.g. Grose et al. 2023 for Australia).

We have changed the text in the end of the last paragraph of Section 2 and spun out as its own paragraph to instead emphasize that the 1850-1900 baseline is a convenient working definition that we do not alter for our paper, but which could be revisited in the future. We have stressed the historical precedent of this working definition across successive IPCC cycles, the decision in AR6, and the fact that the Paris Agreement negotiations were predicated upon such a definition. We have also stressed explicitly that this is a normative judgement call rather than a rigorous scientific choice per se.

Grose MR, Boschat G, Trewin B, Round V, Ashcroft L, King AD, Narsey S, Hawkins E. (2023) Australian climate warming: observed change from 1850 and global temperature targets. *Journal of Southern Hemisphere Earth Systems Science* 73, 30–43. <https://doi.org/10.1071/ES22018>

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

I.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.

We have modified this text to clarify that a metric of realised warming requires some level of smoothing to satisfy the LTTG interpretation. We have also added a footnote to explicitly recognise that different forcings / internal variability can yield spatially distinguished signals for the same nominal warming level. This is a more general issue that pervades, however, multiple approaches to analyses such as the use of GWLs in WGI and burning embers in WGII of IPCC amongst many others and is a whole (set of) paper(s) in its own right. We hope our minor edits give a flavour of this to sufficiently address the well stated concern here.

It should also be noted that ENSO-related variability in GSAT, while significant on annual timescales, is much less significant on decadal and longer timescales - for example, Trewin (2022) found that the 10-year mean of GSAT did not vary by more than 0.03 °C from the 30-year mean with the same midpoint for any 10-year period since 1970.

- 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.

We agree here but believe this to be well covered later in Section 4 with robustness assessed against the NorESM1 ensemble precisely to account for this. Given well stated

length concerns elsewhere we made alterations at the start of Section 4 and hope the reader can easily enough infer this.

- “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.

Article 4.1 of the Paris Agreement text refers to Article 2 as the long-term temperature goal, so this language does appear in the Paris Agreement. However, it remains true that long-term in this instance does not refer to a long-term time average. We have added this in the introduction as a footnote on the first definition to clarify. A further discussion of the 1/CP.27 note which has been used to support use of a multi-decadal mean is now included in section 4.2, particularly around the extent to which this interpretation is warranted.

Overall, while the UNFCCC language does not define that a multi-decadal average should be used, it also doesn’t say that it shouldn’t; one of the fundamental ambiguities motivating the assessment of existing methods in this paper. The choice to use the 20-year mean as a baseline in our paper is therefore framed as common practice/interpretation against which to compare methods, and not as something originating from the UNFCCC or Paris Agreement. It is predicated upon its usage in IPCC AR6 WGI in the SPM and the underlying chapters and its reference in subsequent UNFCCC deliberations and decisions.

- 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.

The text at the end of 4.2 has been modified to emphasize that the decision to align with the 20-year period is done for reasons of lineage, policy relevance, convenience and simplicity and follows the approach used in AR6. The existing text already emphasizes that we are not intending to make normative judgements. There is a long history of using e.g. 30-year averages to denote climatology that stretches back across a century or more of practice by WMO and national meteorological services as well as climatologists. This extends to recent literature which we cite. We could expand considerably here but are minded to avoid adding even more length and exacerbate the concerns of reviewers.

- 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.

We would agree that there are fundamentally more optimal ways to estimate e.g. the true forced response but this is not the ambition in the analysis here. Rather the ambition is to estimate robustly what the estimate of a metric (the 20-year centred average) will be when looking back ‘wise after the fact’ in the future without having to wait. For realised warming this includes the non-forced components. A brief overview of the forced response has been added to the discussion of Consideration 2 in section 4.2, and a pointer added that outputs from methods calculating both anthropogenic and total forced warming are included in the quantitative comparison in the paper.

* 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.

This is indeed true but we think well covered by existing text.

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.

We include the assessment of several such estimates in our analysis, only a subset of which end up passing all the tests in subsequent subsections. Thus while an appealing characteristic it does not follow that this is either a necessary or a sufficient condition to

yield a robust estimate of the 20-year moving average of realised warming that will survive the test of time when looked back upon and evaluated retrospectively.

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 .

The forced response, without any smoothing applied, would not be a workable definition aligned with the LTTG interpretation due to volcanic events. We have included some brief text on using the forced response as an option to the text in 4.2.

* 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.

We have tried, without lengthening the paper unduly, to better caveat aspects raised in the above in several places in section 4 and the discussion section. In principle, given the availability of code etc, someone could select a different target metric and repeat the analysis. We would stress that the final paragraph of the discussion is a necessary caveat that this constitutes a snapshot assessment and that a long-term governance mechanism is required to reflect updating scientific understanding in all aspects.

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.

This is correct - it is beyond the scope. Getting an absolute consensus on such a vexed topic would seem implausible. Perhaps more importantly though, even if, by some miracle, the entire scientific community could be somehow shepherded to a strong consensus, what would actually matter is if policy makers were to agree. As we go on to note later in the paper the easiest way forwards (although perhaps equally implausible) would be for parties to the Paris Agreement to agree that for the purposes of tracking adherence to the LTTG it was metric X calculated using approach Y with assumptions A, B and C that applies.

Otherwise, rightly, scientists will continue to question, analyse, develop etc. as new data, tools and insights accrue.

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).

Thanks for raising this issue. Indeed, use of differing period of records, forced by methodological vagaries including in some cases a need for ancillary data, is somewhat limiting here. We could quantify over the common period of record for all approaches but:

1. This would be very limiting given that the common period is rather short and the trend quasi-monotonic in nature over this period representing a very weak test.
2. In reality many of the limitations found in methods that extend to the early period would become readily apparent in later more exacting tests performed even if a slightly larger set of methods were to survive the first cut undertaken in 4.6.1.
3. Repeating the analysis and subsequent analyses steps as required with a different set of surviving methods would require space on the supercomputer resource that was already heavily borrowed in performing the initial calculations.

On the basis of the above while appealing we believe that rerunning would not materially affect the final selection and, given that the work lacks explicit funding support is not, sadly, practically feasible.

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. $\text{RMSE} < 0.06^{\circ}\text{C}$ — 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.

We would agree, and already state explicitly at the start of 4.6, that the twenty-year running mean is an imperfect target for human-induced warming for the reasons stated. However, it is hard to know what a ‘better’ target might constitute that wouldn’t require substantial assumptions as to how to identify and remove external natural forcing and internal variability (noting that many statistical methods outlined in Section 4.3 attempt to do something akin to this and would introduce circularity) and accounting for any co-linearities between them. The fear might be that doing this introduces biases as large or perhaps even larger than the imperfections arising from using the straight 20-year mean. We note in this regard some substantial degree of spread between the intuitively appealing methods in 2.4.3 that try to do precisely this.

The issue would be acute if we had experienced a very large volcanic eruption (a Samalas-1257 for example) that might have departed global temperatures by several tenths of a degree. In a sense we are fortunate to have had a relatively volcanically quiescent 20th Century and early 21st century and estimates of the departure between long-term warming and underlying anthropogenic warming are generally minor (order 0.1°C on decadal timescales) which arguably enables the equivalency. We propose to strengthen the opening paragraph on 4.6 to make more explicit these issues and that human-induced warming estimates will, naturally, on average show greater departures against this benchmark.

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)

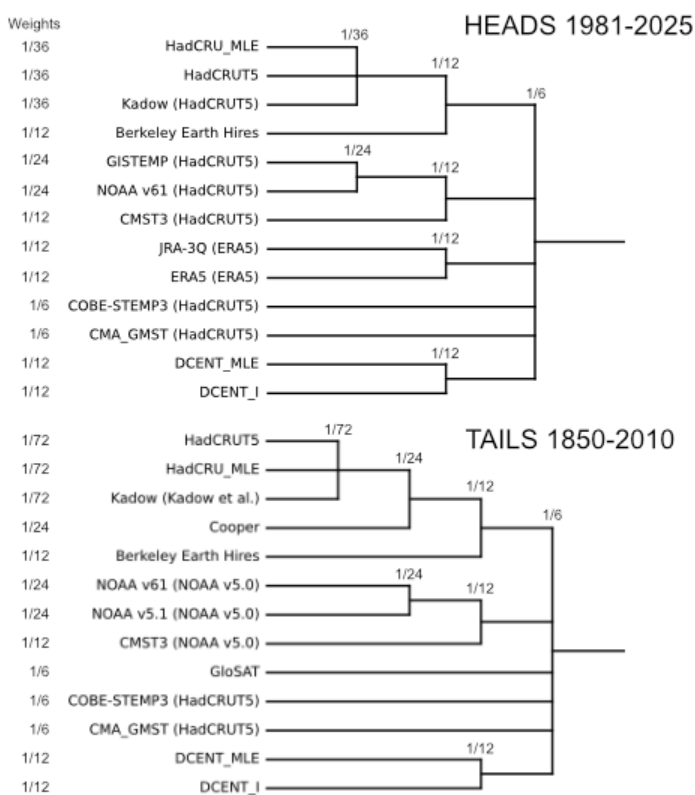
The manuscript explains that 100% probability is assigned at the root of the tree, making it possible, in principle, for the reader to calculate the weight actually attributed to each product. However, to make this easier, it would be helpful to provide the weights used for each product somewhere — either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

Thank you for your suggestion. We have added a figure with weights to the supplement.

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.

We agree this is useful information. Given that this is highly technical in nature and concerns over length we will better document this in the supplement including adding a figure with explicit weights used as shown below.



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.

You are correct that there are differences in potential dataset quality, such as DCENT having improved SST corrections relative to most other SST products, but there is no benchmark against which absolute correctness can be evaluated, so the assessment will always be somewhat subjective. Also, care must be taken to not assume a single dataset is better everywhere and at all times when improvements are only supported for limited periods or areas. Assessment also tends to focus on the best estimate and not on quantification of uncertainty. Accounting for dataset quality in consensus estimates is something that could be revisited in the future. However, as noted, it is difficult to establish a consensus approach to measuring or accounting for dataset quality, including when setting up an approach intended to accommodate future dataset changes or additions. We have added some additional text on this issue in subsection 5.1 at the end of step 3.

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?

On page 29 of supplement section 3, we mentioned this issue and that the linear correlation coefficient across the two time periods is between -0.1 and -0.2. This negative correlation tends to occur due to our baselining of anomalies during 1981-2010; ensemble members with larger slopes over this period will tend to have more negative anomalies during 1981-1995 and more positive anomalies during 1996-2010. Therefore, our final estimates of temperature change slightly underestimate uncertainty across the join. We have now added a brief mention of this issue in subsection 5.1 at the end of step 3.

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?

A direct comparison with the Craigmile and Guttorp reconstruction would not be very informative as their reconstruction is now based on many outdated datasets, such as the old JMA dataset. During the development of the paper we considered alternative synthesis methods, including improved versions of the Craigmile and Guttorp approach. We did not include an updated version of Craigmile and Guttorp as that would entail a non-trivial amount of work and our paper is already long.

Instead, we have added an additional paragraph comparing our approach with Craigmile and Guttorp in subsection 5.1 at the end of step 3. We mention two potential weaknesses of Craigmile and Guttorp: lack of preserving temporal correlations; and potentially placing too much weight on datasets with narrow uncertainties, noting that narrower uncertainties often reflect less comprehensive uncertainty estimates rather than truly better constrained estimates. We also mention the multi-modality weakness of our approach.

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 $+0.1^{\circ}\text{C}$ 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.

Thank you for bringing up this issue. We recognize multi-modality as a potential weakness of our method. We have added text mentioning this, including the cause of multimodality in the early 1900s, in subsection 5.1 at the end of step 3. This will be a perennial problem when new datasets appear that account for a newly discovered apparent data issue and while existing groups evaluate and, if necessary, update.

Aurélien Ribes

Simon Tett

Review of “How well can we quantify when 1.5o of global warming has been exceeded”. Thorne et al This is an interesting paper which contains a lot of science. Broadly it aims to determine when 1.5C of warming has been exceeded framing the question as probabilistic and using calibrated uncertainty language. My main problem with the paper is the excessive length which precludes a detailed review. Broadly the paper is a combination of a long literature review, an approach to determine how much temperatures have warmed since pre-industrial and if we have exceeded 1.5C warming (or other limits). The problem with the later is it requires some estimate of what the next 10 years of global temperature might be. As the limits are defined through 20-year centred averages. My recommendation is to reject the paper as it is. 100+ pages is excessive.

I think the paper should be split into three:

- 1) A literature review – though I recommend the authors be more critical than is normally the case.
- 2) A paper focused on how much warming since pre-industrial there has been to date.
- 3) A paper focused on exceedance of a limit defined through a long term average requiring future projections.

In doing this I recommend the author team are more focused. The goal should not be 3 40 page papers but 3 20 page papers. For example, resist the calls for community action within the body of the paper. If those are needed, they should go in the conclusions.

We thank Simon Tett for his review and we acknowledge his criticisms, but due to the closeness of global temperature to the LTTG (as illustrated by the recent UNEP overshoot report) and the urgent policy interest in this question, we prefer to continue with the paper in its current format rather than starting afresh with three new papers, which is likely to substantially delay publication.

Some reviewers have requested shortening paper length while others have requested additions. Obviously, we cannot simultaneously satisfy all of these requests.

The authors have at various points discussed length and splitting the paper into somehow linked parts, including suggestions along the lines he has given here. The general feeling has been that what was lacking - and the very reason for the initial genesis of the paper - was a holistic assessment of all aspects of the problem in one place. There have been papers looking at specific subsets of the problem (one reason for the very long reference list) but nothing holistic. We have tried to keep the paper as short as possible while covering the very broad range of topics sufficiently holistically. We would hope that readers may be able to dip in and out in their topical area but see value to the more dedicated reader in a holistic snapshot review that

enables the interconnectedness of the various issues to be explored and serves as a comprehensive snapshot review of the state of the science at the current point in time.

Other thoughts

1) I thought the debate about pre-industrial being 1850-1899 had been settled in the last IPCC report.

The last IPCC report, as in prior IPCC reports, used 1850-1900 as a proxy for true pre-industrial. Between two CCBs across Chapter 1 and Chapter 2 (led by authors of this paper) the topic was broadly explored but AR6 was simply a snapshot in time and IPCC does not have the authority to 'settle' such an issue. On the other hand, as we note, parties to the Paris Agreement could settle what pre-industrial constituted (period, metric etc.) for the purposes of verifying adherence to the LTTG. Otherwise scientific understanding will continue to evolve and successive snapshots of this will continue to be reflected in successive IPCC WG reports which might choose to vary this definition.

2) Given this the paper(s) should focus on changes relative to this period. No need to include other baselines.

We would disagree with this interpretation. As we show extensively, if we use a more modern base period with an assumed warming from 1850-1900 to that point in time, then our uncertainty in the warming estimate shrinks enormously. Very little of the uncertainty arises in the well observed modern period. From a policy perspective this matters. The use of modern baselines also helps to untangle issues over legal interpretation as we touch on in the paper.

It is for reasons precisely such as this that we see the need to keep the paper together as a single, albeit very large, analysis rather than split into smaller pieces whereby such interlinkages and implications of choices would be obscured.

We also note that Aurelien Ribes urged us to consider choices of even earlier baselines. It is clear that the baseline question remains an active area of consideration.

3) I think the current presentation of the third part is quite weak with it being, I think, hidden in section 5.

Section 5 only makes sense in light of the issues and findings that have been reviewed and analysed in the precursor sections so its placement is unavoidable. We have reviewed all aspects of structure to try and avoid any perception that it is hidden.

4) How sensitive are results to changes in definition of family tree?

This sensitivity is shown in the supplement. The bottom-line is not very sensitive except for if scientifically naive and uninformed choices are made. We will try to signpost the results in the supplement more prominently while having regard to your and others' concerns over paper length.

Simon Tett