Authors response to two reviewers and comments of L.A. Smith

In the following we will respond to the major points given by two reviewers and L.A. Smith. We first give their comments in italic, followed by our response.

Answers to major points of reviewer # 1 (Peter Thorne)

I find the hook to pre-industrial tenuous given that the authors make no attempt to estimate a true pre-industrial based value. They would be better, in my view, to state that they are making an estimate relative to the late 19th Century / early global instrumental record. This would be a fairer reflection of what is actually done and consistent with e.g. IPCC AR5 which deliberately avoided in the published version implying that 1850-1900 constituted pre-industrial as noted in Hawkins et al. Indeed, the final plenary of the WG1 involved a long discussion that I was personally involved in around the topic whereby the parties agreed that pre-industrial was earlier than 1850. It would be unwise, in my view, for the authors to reopen this issue. I note in a couple of places that there are phrases which could imply IPCC used 1850-1900 as pre-industrial, and they did not. Such implications absolutely must be avoided outright in any resubmission.

The discussion linking their work to the pre-industrial era would be far better being given exclusively in the Discussion section and, to my view, the authors should remove allusions to providing an estimate relative to pre-industrial earlier than this. Bottom line: They either should estimate relative to true-pre-industrial or be honest with respect to what they are estimating relative to for the paper to be acceptable. As I see it there is no rigorous attempt to estimate changes since pre-industrial. Rather, there is a rigorous attempt to estimate it since 1880 which in itself is useful and valuable. The authors should be honest in this regard and not oversell their work by claiming it's an estimate relative to pre-industrial when it demonstrably is not.

We agree with the reviewer and will adapt the text in the way he suggests. Our uncertainty and sensitivity analysis is clearly relative to 1880, and not 1850, or relative to the period 1720-1800 (as in Hawkins et al. 2017), or even relative to the period 1400-1800 (as in Schurer et al., July 24, 2017 - Nature Climate Change). The reason we choose for 1880, is (i) data availability and (ii) the increasing uncertainties in GMT estimates for years earlier than 1880. For example, the Hadley Centre estimates the GMT value plus uncertainty in 1900 to be -0.20 [-0.34, -0.06] ºC (95% confidence limits). For 1850 the GMT estimate is -0.37 [-0.59, -0.16] ºC.

We propose to follow the comment to treat the role of 'pre-industrial' solely in the discussion. We propose to add the results of Schurer et al. (2017) who analyze the role of GHGs, solar radiation and volcanic dust from 1401 onwards. They find that GHGs had a significant effect on global warming if the period 1401-1800 is compared to 1850-1900: from 0.02 to 0.20 ºC (5-95% confidence limits). If all forcings are combined (GHG, solar, volcanic) they find 0.09 [0.03 - 0.19] °C.

I also find the Section at the end of the paper alluding to RCPs and end of Century to be out of scope and a distraction. It should either form an integral part of the paper integrated throughout or be dropped. Given journal scope I would lean heavily toward its removal. The year 2100 is not in the past (at least yet)!

Agreed. We propose to remove 'the future' in Section 5.2, including Figure 5.

Finally, given the authors apparent desire to explore uncertainty I find the omission of the JMA observational analysis and the NOAA 20CR product odd. I could see a case for omission of 20CR, but I see no logical case why the JMA analysis should be omitted here as it has the same non-peer-reviewed basis as e.g. the Berkeley global (not land, but global) estimate. JMA uses a fundamentally distinct set of SSTs and so would better span uncertainty that the authors lament in Section 2.1.

We have studied the global series of the Japan Meteorological Agency (JMA). There is a simple practical problem, however: this series starts in 1891. Thus, we miss the important period 1880-1890. Next to that, there are only texts in Japanese which explain how this data product is composed. Finally, studies which show the JMA series are limited. Our choice for five GMT data products as described in Section 2.1 is consistent with recent studies such as Medhaug et al. (2017 - their figure 1a) or Rahmstorf et al. (2017 - their figures 1 and 2). Therefore, we propose not to add trend analyses for JMA to Table 1.

As a test we estimated linear trends for all five data products shown in Table 1 and additionally the JMA series. It appears that the JMA incremental value for the period 1891-2016 equals the low end of the five data products we apply in our Table 1 (i.e., the incremental value of the HadCRUT4 series). Thus, the incremental value based on the JMA series, does not fall outside the range of values based on HadCRUT4, NASA, NOAA, HadCRUT4 adapted by Cowtan and Way, and BEST.

As for the NOAA 20CR series we have two arguments for not adding it to our study. First, the 20CR series covers the period 1851-2011. Thus, data for the important period 2012-2016 are missing. Second, the series is a combination of modeling (weather prediction models) and data. For our study we prefer to make a distinction between GMT series directly derived from temperature registrations and models, be it GCMs or weather prediction models.

We propose to name these arguments in the first item of the discussion section.

Minor comment #12. Line 202 you should clarify what the implications of ignoring this are or, preferably, perform the extra work necessary for its inclusion. Presumably the impact would be artificially reduced uncertainty ranges? In which case is it really safe to ignore this issue? I'm not entirely convinced and would suggest that extra work leading to its inclusion is instead warranted. Even if it ends up showing no change it would make the piece more robust. As you yourselves state the effect is statistically significant, in which case it really should be included.

Agreed. We propose to show the effect on correcting for this small but significant AR(1) correlation in the innovation series of our Kalman filter model.

Answers to major points of Anonymous Referee #2

This study considers the question of estimating by how much global temperatures have changed since ’pre-industrial’ times, assessing the uncertainty in different trend models and due to different global temperature datasets. The analysis is interesting, though the results are not too surprising. However, I have some major concerns:

1) Framing: the authors emphasize repeatedly that they are estimating changes since a particular baseline and implying that this is what the Paris agreement meant by ’preindustrial’.

This is not the case - the introduction of Hawkins et al. (which the authors cite) discusses this issue at length. In addition, Schurer et al. (2017, NCC) was very recently published, highlighting again that there was likely some additional warming due to anthropogenic factors before 1850. The authors may also like to examine Otto et al. (2015) for an alternative approach to estimating the warming since the 19th century. The text in the discussion on this topic is appropriate however.

Agreed. We propose to treat the topic of 'pre-industrial' more clearly in the discussion section, as we pointed out in our response to Reviewer #1. We will add the references to Schurer et al. (2017) and Otto et al. (2015). Consequently, we will address their findings that GHGs had a significant effect on global warming if the period 1401-1800 is compared to 1850-1900: from 0.02 to 0.20 ºC (5-95% confidence limits). If all forcings are combined (GHG, solar, volcanic) they find 0.09 [0.03 - 0.19] °C.

2) Terminology: some of the phrasing is very confusing when referring to and/or distinguishing between natural *forced* variability (volcanic, solar) and internal *unforced* variability. These terms are sometimes mixed and it’s not always clear what the authors mean. For example, in the abstract (and L86) the authors claim the models are corrected for natural variability, when they mean the forced component, but the introduction uses natural variability to mean both forced and unforced variations. On L133, the authors refer to the ’historicalNat’ runs ’for natural unforced variability’, which is not true - those runs include both natural forced and internal unforced variations as the next sentence correctly states. Variability is also used for the spread or range between different estimates, adding further confusion. The authors should carefully check each use of this type of phrasing and make it far more precise.

Agreed. We propose to check the phrasing of 'natural variability' carefully, this in combination with the terms 'forced' or 'unforced' or both, 'internal variability' and 'spread'.

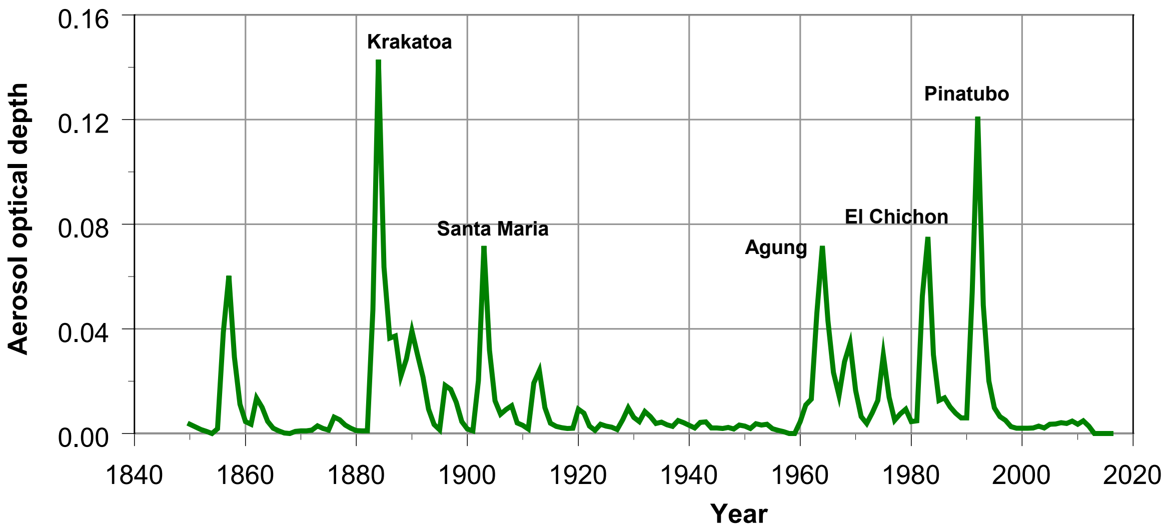
Additionally, we propose to treat the role of natural unforced variability and natural forced variability (i.e., the role of changes in irradiance of the sun and changes in volcanic activity) separately in a second item in the discussion section.

The trend analyses as given in our Table 1 are based on the IPCC definition of climate change (Glossary AR5): anthropogenic forcing combined with decadal to centennial natural variability. However, UNFCCC defines climate change as originating from GHG forcing only. In their philosophy we could argue that the Paris limits of 1.5 and 2.0 C should originate solely from anthropogenic forcing. We propose to quantify this second view on the Paris limits.

To do so we make use of the recent study of Schurer et al. (2017, their figures S2 and S3), and the lower panel of figure 4 in our manuscript. Next to that we estimated the role of volcanos in a time-series setting by extending the Integrated Random Walk (IRW) model. For details we will refer to Visser and Molenaar (1995) and Visser et al. (2015).

It shows that the incremental values shown in Table 1 for the IRW trend are 0.04 ºC degree lower. If estimated in combination with the OLS straight line, i.e. a regression model with one explanatory variable, estimates are 0.02 °C lower than those shown in table 1. This effect, although small, will be due to the Krakatoa eruption in the period 1880-1890.

The indicator for volcanic dust is taken from NASA: aerosol optical depth (AOD). See graph below:



3) GCM analysis: the 106 members used cannot be 'one per model' as there were not that many models in CMIP5. It's not clear what the authors have used here - there must be more than one historical part of the runs for some of the models.

The reviewer addresses a good point. What we meant here is that we used one member per model, **given the use of a specific RCP scenario**. Thus, we have used 42 members for emission scenario RCP4.5, 25 members for emission scenario RCP6.0 and 39 members for emission scenario 8.5, making up a total of 106 members. We propose to clarify this in the text.

There are also 43 piControls on Climate Explorer, and very few are less than 200 years, not only the 20 that the authors have used - why have they not used the others?

Agreed. We have calculated all AR(1) coefficients for all 41 piControl runs, available in the KNMI Climate Explorer. Three of those runs showed a jump or a strong linear trend over the simulation period (varying from 200 to 1000 years). We omitted these. For the remaining 38 runs we have omitted the lowest two AR(1) coefficient estimates (lying around 0.0) and the two highest estimates (lying around 0.75). The remaining range equals the range given in our manuscript: [0.28 - 0.60]. We propose to adapt the text for this finding.

Also, in section 3.2, the authors could use the AR(1) value from each model's own control run to fit a spline to the historical run of that same model, rather than assume the same across every model. Also, how has the correction for natural forcings been applied (L250)? Has the mean across the historicalNat runs been subtracted from each historical run? If so, this is inconsistent as the response to volcanic eruptions varies significantly across models.

In our revision we propose to give values for smoothing by splines with φ=0.28 and φ=0.60, similar to shown in our figure 3. Period: 1861-2016. This gives a small change in the upper panel of our figure 4. The spread is for both smoothing options identical ± 0.50 C (2σ). The mean value of all 106 increments is 1.15 for the smoothing option with φ=0.28 and 1.00 for φ=0.60.

Answers to major points of L.A. Smith

Visser et al (2017) provide an interesting and insightful discussion of signal detection in global mean temperature (GMT), focusing on the 1.5 degree target of the Paris Agreement of 2015. This paper could be made more informative by further consideration of three topics: (1) clarifying what is meant by "signal" and by "noise", and more specifically how (whether) natural variability can be "corrected for" in an evolving nonlinear system, (2) implications of using CMIP5 models, given that those models display a wide range of values for today's GMT, and (c) a cleaner definition of how one would detect failure to stay "well below" a temperature target, or to exceed it. These points are expanded upon below.

Specific Comments

"Natural variability" is said to be a dominant source of uncertainty which has been "corrected for" (24). Although discussions of a climate signal coming "out of the noise" are common, the notions underlying the distinction between signal and noise in the climate context is unclear; it is not the traditional distinction of observational noise superimposed on a imprecisely measured but well-defined signal. Superposition can only be assumed in nonlinear systems given purely observational noise that has no impact on the system: natural variability, internal variability and the like alter the dynamics, and thus the "signal" itself, if such a separation exists (Smith (2001,2002)). A more appropriate conceptualization in nonlinear systems is found in consideration of an ensemble of systems each subject to a common driving and independent realizations of the relevant noise. In this case, the ensemble median would provide a well-defined signal while the distribution about it would capture the effects of noise processes. This view is of limited utility in climate science, where there is only one realization (the Earth): particular realizations need not reflect the (unobservable, non-empirical) "signal"; indeed they can diverge arbitrarily far from it. So in no sense can one expect "the" signal to emerge from the noise, given observations of a single realization. While vague appeals to something somewhat reminiscent of an adiabatic change in thermodynamics may be voiced, clear clarification of the meaning of signal and noise in the climate context would be of value.

In short: it would be useful to clarify how "natural variability" and "internal variability" might be isolated in the case of a complicated, nonlinear, evolving planetary system.

How are we to make sense of the traditional notions of "signal" and "noise" given that the "noise" is not mere observational noise but actually a component of the system dynamics, and given that in nonlinear systems we cannot appeal to a principle of superposition of solutions (Smith, 2002).

The modeling of climate data by stochastic climate models have been described in Mudelsee (2014, sections 2.5.1 and 2.6). He describes the suitability of climate modeling with AR(1) processes (and the more general ARIMA models as well) to describe the persistence in data.

The reviewer is right that correlated noise is not the same as climate variability arising from nonlinear systems. However, statistical modeling has proven fruitful in a wide field of ecological modeling. To stick to the modeling of global mean temperatures, we refer to our review of (statistical) trend analyses in the peer-reviewed literature in the Supplementary Material section of our manuscript (table S.1). Furthermore, Visser et al. (2015) show in their table 1 that researchers in the field of sea level rise apply 30 trend methods for quantifying "the signal" in sea level data, all with different mathematical formulations.

Note: we do not use trend models for **prediction**. Next to that, projections up to the year 2100 will be removed.

It is also worth noting that the statistics community and the physical science community often hold very different notions of what a trend is: for the first, it is a statistically consistent combination of two well-defined models (the trend model and the noise model), while for the second it is merely a systematic, often obvious drift. Statisticians require, and quantify, consistency between these two components, and reject identification of a trend if that consistency is lacking. Physical scientists often require the observations to look trendy, and the ability to reject simple statistical models given the data, when those models are known by construction not to admit a trend. The second bar is much lower.

The claim that modelling groups "have not been very successful in tuning to the observed trend" (299) suggests some knowledge as to how large the spread would be in the absence of each group knowing the observed trend (aiming for the same target). It has been argued elsewhere that knowledge of such spread would be very useful to have if, perhaps, impractical to obtain.

Visser et al (2017) state that "mean progression derived from GCM-based GMTs appear to lie within the range of the trend-dataset combinations" (311). It would be interesting to see the variations among individual CMIP5 simulations (not the mean over them, but their distribution). The IPCC AR5 reports that variations in the global mean temperature of today's CMIP5 GCMs have a range exceeding 2.5 degrees (see right side axis labels of Figure 9-08 of Flato et al (2013)); what are the implications of our best models showing a range of GMT almost twice the 1.5 degree target? Physical and biological processes are driven by actual temperature, not anomalies. Given the current (limited) level of realism in these models, and the fact there is a great deal more in them than their basis in physical understanding, the authors might wish to reconsider calling today's GCMs "fully physics-based" (86).

The upper panel of figure 4 shows in part what the reviewer asks for. We will discuss the implication of the wide range of incremental values at the end of the discussion section. Here, we will argue that GCM simulations are less suited for tracking the signal in GMTs due to their wide range. Another argument will be that GCM simulations in CMIP5 are up to date up to the year 2005. Estimates for the period 2006-2016 are less reliable.

We propose to add the important comment of the reviewer that GCMs give a wide range of estimates for the global temperature over the period 1961-1990. Not as anomalies but in **absolute temperatures**. Indeed, figure 9.8 of the AR5 WGI report (2013, page 768) shows a range from 12.6 ºC to 15.3 ºC, based on 36 models. This range is almost the double of the 1.5 ºC limit.. Also see figure 1 upper panel in Hawkins and Sutton 2016 BAMS 963-980.

Finally, we propose to remove the expression that GCMs are 'fully physics based'. That is, indeed, not true.

Lastly: what precisely does it mean to hold GMT "well below" (14) some temperature threshold? How would we know if we had missed this target? Can this be phased with sufficient precision to allow, say, an insurance contract or legal wager to hinge onits occurrence? Issues include the duration for which the threshold is exceeded (An instant? A month? A year? A decade?) and how to deal with the imprecision in measuring the global mean temperature, even today. In practice, simply setting the target as an absolute value of GMT, inspired by the agreed 1.5 change, would prove more straightforward both scientifically and legally, even if not politically or diplomatically.

Good point. However, we propose to remove Section 5.2 where we extend the historical analysis to the year 2100. Therefore, this important comment is not directly applicable to our revised text.

Additional changes to a revised manuscript

Due to the comments given by the reviewers we propose to restyle the discussion section to improve readability. The following items are addressed (in this order):

- Data products, trend models and GCM simulations. Would the results presented here, change if (i) other data products would have been chosen (such as JMA), (ii) other trend models should have been chosen, or (iii) other GCM simulations would have been chosen (such as blended simulations)?
- Correcting for forced and unforced natural variability. This is a new item which addresses the question if we should filter a GMT series for short term natural variability, complying with the IPCC definition of climate change, or that we should filter GMTs for all natural forcings, complying with the definition of climate change of UNFCCC.
- Choosing a pre-industrial baseline. Addresses the consequences of results presented by Hawkins et al. (2017) and Schurer et al. (2017).
- Policy recommendation. This is a the slightly extended text taken from Section 5.2.

Furthermore, we remove section 5.2.