Response to Editor and short comments

Editor comments (E) and author responses (A)

E: I would like to add two additional requests. One apparent simplification is that noise is assumed to be Gaussian. Climate model signals (and observed climate data) has considerable spectral structure (ENSO, PDO, AMOC, etc.). Presumably in these models this is simplified into some modelled split between noise and signal. Could a more realistic noise representation impact the results?

A: While the noise assumption is a little simplistic, it mirrors that made by Cox et al, and is also widely used in the literature. We believe that the heteroscedastic and nonlinear behaviours shown in Fig 5 would be robust to the details of the noise as they arise directly from (a) the reduced damping with large S, and (b) the influence of ocean heat uptake, respectively. The results with CMIP5/6 models also show the relationship between sensitivity and variability to be is weak.

E: It would also be helpful if you could comment if the assumption for the magnitude of the noise (sigma) plays a role in the final results since this appears to be a fixed assumption "generally use the value $\sigma = 0.05$".

A: We have add a comment in the manuscript (p 3 line 26). The value of psi scales with the magnitude of the noise but the results are not very sensitive to it. It was chosen as a reasonable compromise between compatibility with model results and surface temperature observations.

In addition, some least significant digits have changed due to sampling variability following the re-ordering of some of the simulations for the new figure arrangement.

Short comment from Nic Lewis (SC) and author responses (A)

SC: [comments about transient response]

A: p6 l30 We note the stronger relationship to transient response, albeit it is not a primitive parameter in our model.

SC: Page 3 line 20: the model in equations (1) and (2) is not the Winton et al (2010)
model: it is the Held et al (2010) model. The Winton et al model, although similar to the
Held et al. model, has a basic difference in that its efficacy parameter ε applies to total
ocean heat uptake, and is found to vary significantly over time in AOGCMs, whereas in
the Held et al model ε only applies to deep ocean heat uptake, as in equation (1), and
is found to fit AOGCM behaviour with (as here) a constant ε value. The cited Geoffroy
et al (2013a) paper uses the Held model, not the Winton model.
R:

A: done

SC Page 4 Table 1: there is a sign error in the default value for the radiative feedback
parameter, λ. The way this parameter is used in equation (1) implies it is negative, but
Table 1 defines it as 3.7/S not –3.7/S.

A: done

SC: Page 13, line 30: the title of Kass and Raftery's 1995 paper is just "Bayes Factors".

A: done

Short comment from Stephen E. Schwartz

A: No direct changes necessary, though we do mention (p6

l8) that the results are insensitive to detrending or not in the unforced case.

Response to Anonymous Referee #1
Referee comments (R)
Author comments (A)

R:
1 Summary
Annan et al. (2020) examine the utility of using natural variability of global mean surface temperature (or ocean mixed layer temperature) to constrain equilibrium climate sensitivity. They extend recent work on this topic by using a two layer energy balance model in a “perfect model” framework. They find that the strength of variability-based constraints is substantially weaker when climate sensitivity is large, which is true for both simulations with no external forcing and simulations that use estimates of historical external forcing. For moderate climate sensitivity (2.5 K), the uncertainty in ECS is approximately 4 degrees using information from the entire time series and including aerosol forcing uncertainty. For simpler constraints, the uncertainty range is even larger. This work is a useful expansion of recent literature on this topic. The manuscript is clearly written, though I suggest the authors make a minor modification to the organization, expand their discussion in several places, and consider condensing some figures (or adding a summary figure) to help compare results across the various experiments.

2 General Comments
One suggestion to improve the manuscript is to make it easier for the reader to compare across experiments / figures. For example, Figure 1 and 5 are similar and it would be useful to compare all of these results together (perhaps via plotting them on the same axis with color coding in an additional summary figure or grouping or by putting them on a common figure with different panels). It would be similarly useful to intercompare the various

posterior estimates (Fig. 2 and 6 as well as 3, 7, and 8). Perhaps plotting lines corresponding to the 5 - 95% CI and a dot for the most likely value value (which would illustrate the skewness) would help compress the figures (though this may run afoul of the Bayesian framework).

A: Thanks for the suggestions. As per our initial response in the open review, we have combined figs 2 with 6, also 3, 7 and 8 and think the changes are an improvement. Figure captions and legends have been edited appropriately. We draw attention in the text to differences between figs 1 and 5.

R:
There were a few places (noted below) where it would be helpful to more directly compare and discuss this work in the context of other literature. For example, in some places I thought that Cox et al had commented on some issues (e.g., de-trending, two layer models, etc.) and it wasn't immediately clear how to put that work in the context of this manuscript.
I was confused by the factor you used as well as the references for the two layer model used here (see below). It would be helpful to clarify some of this in the revised manuscript.

A: We didn't want to focus too directly on the details of Cox et al and the various comments/responses, as we consider the analysis using the full time series to provide a more compelling result that obviates a detailed investigation of their approach as it puts a strict bound on the potential for variability, however it it is analysed, to provide a constraint. The "garden of forking paths" is a very clear danger when a large number of choices can be made in the analysis procedure, and therefore such choices must be supported by an underlying theoretical basis. However we have added discussion of Kirk-Davidoff and made a number of other changes described in more detail below.

R:

In terms of organization, I thought it would be helpful to include the data (CMIP + HadCRUT) somewhere in the beginning (e.g., a renamed Methods section), rather than introducing with the manuscript's results.

A: Subsection on Additional Data to describe CMIP5/6 and HadCRUT data has been included.

R:
3 Specific Comments
Abstract (line 2) and Page 1 / Line 24: In the abstract I wasn't sure which studies you were referring to that tried to constrain ECS with the trend. This might be an oversimplification of these approaches (the Gregory et al. 2002 and Otto et al. 2013 papers cited), since these publications also considered radiative forcing and ocean heat uptake. I believe more recent work by Jimenez-de-la-Cuesta and Mauritsen (2019; doi:10.1038/s41561-019-0463-y) and Nijsse et al (2020; doi: 10.5194/esd-2019-86) are consistent the abstract language (with caveats that they focus on TCR and a specific time period). I suggest revising this language to reflect the "energy budget constraint" rather than the trend and/or citing these other relevant publications.

A: Citations in the abstract are deprecated by the journal but we have changed the wording a little to the deliberately less specific "trends in observational time series" and expanded the introduction to read "energy balance as constrained by the warming trend in atmospheric and oceanic temperatures "

R:
Abstract / Line 15: "observed. . .observational" consider using "inferred from the detrended observational record"

A: Done as suggested

R:
Page 2 (general comment): Other studies that discuss the

utility of variability in understanding the climate response to external forcing include Langen and Alexeev (2005, doi:10.1029/2005GL024136) and Kirk-Davidoff (2009, doi: 10.5194/acp-9-813-2009).The latter publication seems particularly relevant to the manuscript under review and could be compared to the results here.

A: Kirk-Davidoff citation added, and mentioned again p6 l17

R:
Page 2 / Line 21: Cox et al replied (Cox et al, 2018b) to these comments with some analyses relevant to this manuscript. In it, they discuss issues such as the importance of de-trending, their own two layer experiments, and the effect of historic external forcing. Given the relevance to this manuscript (e.g., they two box model results), some of this information could be presented in the introduction or at least compared to the two layer results shown in this work.

A: Cox 2018b Citation has been added but additional discussion is retained for Section 4. Given their original (2018a) result was justified solely on the single layer unforced model we think it is appropriate to focus initially on this situation.

R:
Section 2: Consider describing the CMIP data and HadCRUT observations here

A: Brief description of HadCRUT and two CMIP ensembles added.

R:
Page 3 / Line 7: In Cox et al (2018b), they did test a two layer model (building off their
one-layer model results)

A: No change needed (Cox 2018b is cited and discussed elsewhere).

R:
Page 3 / Line 20 and Equation 2: I was confused why epsilon did not appear in Eq. 2 or why it wasn't absorbed into gamma in both Eqs. 1 and 2. In quickly looking at the Winton et al (2010) paper: don't they use this term in part because it is a one layer model (line 10 and line 20 seem to imply this was a two layer model, but I realize now this may not have been intended)? Can epsilon be removed here since you explicitly have deep ocean representation? I would appreciate more text justifying / clarifying the purpose of epsilon. It looks like some of what you attribute to Winton et al (2010) should be attributed to Held and Winton (2010, doi: 10.1175/2009JCLI3466.1)?

A: Epsilon does not play a significant role here; we include it primarily because this is the standard version of the model that is widely used to mimic GCM behaviour. We now make this point in the manuscript. It is not quite correct to say that it epsilon can be subsumed into the gamma parameter, as its effect on energy balance is more akin to an additional feedback into space, the strength of which depends on the degree of disequilibrium. As for the references here, while the Held et al reference is actually clearer as to the formulation of the model (see their equations 9 and 10) they did specifically cite Winton et al as the origin of this approach. We now cite both papers (l12).

R:
Page 3 / Line 26: On first read, I thought you had used a value of the average mixed and deep layer depth based on observations. Suggest making this more clear with something like, "We assume a mixed and deep layer depth of 75 and 1000 m, respectively, which are used to calculate the heat capacities (Cm and Cd, respectively) based on ocean coverage of 70% of the planetary surface

area.”

A: Done

R:
Page 3 / Line 27 - 29: It would be useful to provide more information about how you chose your parameter values (and later information about how you get the range of plausible values), citing literature relevant to the selection of these values. It was unclear to me why you didn’t simply use the mean or median from Geoffrey et al. (2013), for example.

A: Aim is not to specifically emulate a particular GCM ensemble but just to cover a reasonable range wherein we believe reality could plausibly lie. Added sentence to explain, "Our aim here is not specifically to replicate or mimic this ensemble but to allow for a reasonable range of parameter values."

R:
Page 5 / Line 11 - 12: This suggests that you checked this using the two-layer model,
consider putting “(not shown)” to indicate that you checked this.

A: done. Yes we did look at this briefly.

R:
Page 5 / Line 19 - 20: Does detrending (using 55 year windows as in Cox et al) alter the analysis? It seems like it would be reasonable to linearly detrend (as done in the Cox et al calculation). In the Cox et al reply, they note that it is still important to de-trend unforced simulations using a two layer model.

A: We did test and detrending the unforced runs made very little difference. We thought it was more in the spirit of the original derivation to not do this in the main analysis, as the stated purpose was to remove the forced trend which we know to be zero in this instance.

R:
Page 5 / Line 23 - 29: This is interesting and useful, though I am not sure how to square this with the results presented in Cox et al. (2018). Could you comment more on this? Is this heteroscedasticity included in their estimate of ECS via linear regression? For example, if you varied S to correspond to the 16 models used in Cox et al. (2018), ran a 150 simulation, and performed linear regression (Ψ versus ECS) would the fit be significantly different from what would be obtained using the 1000-year simulations? Or, in another way, is this issue included in the Cox et al (2018) ECS estimate because the increased variance in high ECS models has the effect of making their linear fit between Ψ and ECS more uncertain (and thus contributes to the uncertainty in GCM Ψ values and, in turn, ECS)? Or is the strength of the relationship in Cox et al fortuitous (as suggested by the importance of what GCM simulations are included in the regression as seen in the Po-Chedley et al comment)? Or perhaps the take home message should be that the observed value of Ψ is uncertain. The Cox et al reply (Extended Data Figure 2, top left) suggests that you would generally expect to get a reasonably strong relationship between Ψ and ECS (even though heteroscedasticity should influence their results, too).

A: Their ordinary least squares analysis makes no assumption of (and therefore does not account for) heteroscedasticity, and we believe they must just have been lucky in their set of models (combined to some extent with optimising various choices during their analysis). Our analysis using the exact likelihood renders the entire debate moot. Furthermore, the failure of their approach for CMIP6, and the significantly different relationship exhibited by these models, is additional evidence that their original result was unreliable.

We now also mention the difference between the CMIP ensembles in the conclusions.

R:
Page 5 (Figure 1 discussion): Note that Po-Chedley et al (2018) found that you only recover a strong Ψ - ECS relationship using all of the piControl data and the relationship is weaker with shorter time series. Kirk-Davidoff (2009) also concludes that shorter time series cannot accurately diagnoses climate sensitivity. Of relevance, Nijjse et al (2019, doi: 10.24433/CO.6887733.v1) show that a metric of decadal surface temperature variability (from piControl data) scales with ECS.

A: Po-Chedley and Kirk-Davidoff cited.

R:
Page 6 / Line 31 - 32: It would be useful to provide a reference regarding this point (since the subsequent linguistic calibration was not immediately intuitive).

A: This has been be reworded. It is intended as a literal translation of what P(psi|S) means, however.

R:
Page 6 / Line 32: Should this be “Relative” (where it says “Likelihood values can be
read. . .”?

A: yes, text changed

R:
Figure 1 (and others): Suggest adding a legend with “Single layer (gamma = 0)” and
“Two layer”. In general, it would be helpful to the reader to have a legend on all of the
figures (with the possible exception of Fig. 4) and there appears to be plenty of white
space to do so.

A: Legend added here (and also to other figures).

R:
Figure 2: Suggest “dashed” instead of “dotted.” The blue appears purple on my monitor.
Page 8 / Line 2: You could cite Roe and Baker (2007; doi: 10.1126/science.1144735)
and perhaps others here.

A: Changed to dashed lines

R:
Page 9 / Line 5: I don’t have intuition for what the relative uncertainty should be, but
20% struck me as small (particularly given the later remark that the observational interannual variability looks relatively large compared to the model simulations).

A: Observational variability has a large component of observational uncertainty, so this must provide an upper bound on internal variability.

R:
Page 9 / Line 8: Consider using “purple” instead of “blue” so the reader doesn’t get
confused with the cyan line (at least this appears purple on my screen).

A: We think this colour scheme makes sense, using blue, cyan, magenta and green, with dashed lines representing the results for multiple uncertain parameters. We have added a legend which should help.

R:
Page 11 / Line 26: Does this correspond to any published values of the aerosol forcing
uncertainty? Later, you say that a larger aerosol forcing corresponds to larger ECS,
but this range suggests that you do not, by default, consider larger aerosol forcing - is
that right? If so, why?

A: This scaling considers larger aerosol forcing (alpha >1) just as likely as smaller (alpha <1) and the 95% range is 0-2x the standard value, which we consider to include all reasonable estimates.

R:
Page 11 / Line 33 onwards: Could it also be that the noise term in the two layer models
is too small?

A: This is possible and we've changed the text to mention this.

R:
Page 12 / Line 2: At first I didn't understand what 0.13 represented. Consider re-writing as something like: For each of the three simulations, the RMS differences between model output with no internal variability and observations is 0.13 oC.

A: Done

R:
Page 13 Line 1 (and elsewhere): This record contains more than the 20th century. Consider using a more generic term (like historical) and noting the time period considered. Also consider using "first" instead of "firstly."

A: Changed to historical. Also changed firstly to first.

R:
Page 13 / Line 4: In Figure 5, you also only show two layer solutions, which is different
from Fig. 1 and could be noted here.

A: noted in text at the start of Section 4.1

R:
Page 13 / Line 18: Should some of this CMIP information be included in the Methods Section (perhaps revised to

“Data and Methods”)? What RCP scenario was used to extend the historical time series. This would be a good place to cite Taylor (2011, doi: 10.1175/BAMS-D-11-00094.1). There is also some language that is suggested for acknowledging ESGF and modeling groups (https://pcmdi.llnl.gov/mips/cmip5/citation.html). It would also be helpful to say what variable you are using (tas?).

A: "Additional data" section 2.3 has been included as suggested.

R:
Page 13 / Line 25 - 27 / Figure 5: It is very useful seeing many CMIP5 realizations on this plot. This is a nice illustration of one of your key points (the range of Ψ values can be quite large for a given model).

A: Thank you

R:
4 Grammatical / Other Comments
Page 1 / Line 23: Should this be *mid* 19th century to the early *21st* century
Page 2 / Line 1: focussed -> focused
Page 2 / Line 2: Suggest: “and this topic” -> “which”
Page 2 / Line 31: Remove “to” in “from to”
Page 3 / Line 7: Suggest changing “equilibrium sensitivity” to “equilibrium climate sensitivity”
Page 6 / Line 16: Insert: “so *we* perform. . .”
Page 9 / Line 25: I was not familiar with “i.i.d.” - suggest writing this out.

A: All done as suggested.

Response to Anonymous Referee #2

A:
Thank you for your helpful comments.

R:

General comments:
Annan et al study whether the variability can constrain sensitivity in an idealized twobox model setting. They find it works well for lower values of sensitivity although loses power for higher values (around 5K). They also report that using the forced response in addition can constrain estimates further. It is well written and seems to be technically correct. I have a few questions listed below and some minor suggestions for ease of comparison with previous work.

R:
Specific comments:
Ranges are reported in the 5-95% interval. It would be great to see the 33%
to 66% for comparison with the IPCC ranges particularly in the abstract, as was
done in Cox et al (2018).

A: We have added some extra ranges to the values for the results with the full time series (our most optimistic scenario) to help with comparisons. (IPCC likely range is 17-83%, which we assume is what the reviewer meant). End of Section 4.2

R:
Section 3.1, p.6, lines 25-30. Is the increasing uncertainty in estimated sensitivity due to the larger timescales in the higher sensitivity models relative to the length of the time series which is fixed at 150 years? If climate sensitivity scales proportionally with timescale then the 1K sensitivity has 10 times greater effective sampling for a finite length record. Is this what's going on here?

A: Section 3.1 Yes in part this is the case. We cannot accurately infer the intrinsic time scale when the time series is not long compared to it.

R:

Section 4, p. 11. There are other deterministic forcing factors uniform in amplitude and phase across all the CMIP GCMs (and the real Earth system) not present in the two-box model simulations. These deterministic forcing factors could further separate and discriminate the model sensitivities in addition to the IPCC annual forcing timeseries. Examples of such factors are the diurnal and seasonal cycles of solar insolation. Even though these forcings (and responses) are averaged over when using annual GMSAT, there might still be traces of it in the responses (at least in the more complex and nonlinear CMIP model responses and the real world) and this could act to further help reveal the sensitivity. This is not considered in the ideal model scenario and would be interesting to test in the two-box.

A: Possibly, but we don't think such a simple model can address this with much precision. Weak results relating to the annual cycle were obtained by: Knutti, R., Meehl, G., Allen, M. R., & Stainforth, D. A. (2006). Constraining climate sensitivity from the seasonal cycle in surface temperature, Journal of Climate 19(17), 4224–4233.

R:
section 4, p. 12. Worth noting that there are differences between forcing from well mixed GHGs and aerosol forcings. Well mixed GHGs tend to act uniformly over the globe while aerosol forcing is quite geographical. You're not going to be able to compare the effects of aerosol forcing with well mixed GHG forcing in a box model, at least in a simple way.

A: Sect 4 yes we now note this at the top of p14 (and also cite Aldrin et al as an example). Hemispheric models take advantage of this point.

R:
Technical corrections:
line 25: The usual approach to integrating stochastic

differential equations is the Euler-Maruyama method to simulate the correct variance on the random variable. If the timestep is 1 unit (as it appears here) it shouldn't make any difference but is worth noting for the reader.

A: We selected the noise sampling deliberately to achieve the desired variance in temperatures.

R:
Table 1: Would be good to get all the parameters in the same units as Geoffroy et al (2013a) for direct comparison (particularly Cm and Cd in W yr /m2 /K). I'm aware these are not standard SI units, but just like kWhrs, they are much more convenient to calculate with. It would also be good to list the resulting timescale ranges τf and τs for the same reason.

A: Table 1 We have added the capacities in these alternative units.

What could we learn about climate sensitivity from variability in the surface temperature record?

James Douglas Annan[1], Julia Catherine Hargreaves[1], Thorsten Mauritsen[2], and Bjorn Stevens[3]

[1]Blue Skies Research Ltd, The Old Chapel, Albert Hill, Settle, BD24 9HE, UK
[2]Department of Meteorology, Stockholm University, Stockholm, Sweden
[3]Max Planck Institute for Meteorology, Hamburg, Germany

Correspondence to: jdannan@blueskiesresearch.org.uk

Abstract.

We examine what can be learnt about climate sensitivity from variability in the surface air temperature record over the instrumental period, from around 1880 to the present. While many previous studies have used ~~the trend in the~~ trends in observational time series to constrain equilibrium climate sensitivity, it has also been argued that temporal variability may also be a powerful constraint. We explore this question in the context of a simple widely used energy balance model of the climate system. We consider two recently-proposed summary measures of variability and also show how the full information content can be optimally used in this idealised scenario. We find that the constraint provided by variability is inherently skewed and its power is inversely related to the sensitivity itself, discriminating most strongly between low sensitivity values and weakening substantially for higher values. It is only when the sensitivity is very low that the variability can provide a tight constraint. Our investigations take the form of "perfect model" experiments, in which we make the optimistic assumption that the model is structurally perfect and all uncertainties (including the true parameter values and nature of internal variability noise) are correctly characterised. Therefore the results might be interpreted as a best case scenario for what we can learn from variability, rather than a realistic estimate of this. In these experiments, we find that for a moderate sensitivity of 2.5°C, a 150 year time series of pure internal variability will typically support an estimate with a 5–95% range of around 5°C (e.g. 1.9–6.8°C). Total variability including that due to the forced response, as ~~observed in~~ inferred from the detrended observational record, can provide a stronger constraint with an equivalent 5–95% posterior range of around 4°C (eg ~~1.7–5.6~~1.8–6.0°C) even when uncertainty in aerosol forcing is considered. Using a statistical summary of variability based on autocorrelation and the magnitude of residuals after detrending proves somewhat less powerful as a constraint than the full time series in both situations. Our results support the analysis of variability as a potentially useful tool in helping to constrain equilibrium climate sensitivity, but suggest caution in the interpretation of precise results.

1 Introduction

2 ~~Introduction~~

For many years, researchers have analysed the warming of the climate system as observed in the modern instrumental temperature record (spanning the ~~late~~ mid 19th to early ~~20th~~ 21st century), in order to understand the response of the climate system to external forcing. For the most part, the focus has been on the long-term ~~warming trend~~energy balance as constrained by the warming trend in atmospheric and oceanic temperatures (e.g. Gregory et al., 2002; Otto et al., 2013; Lewis and Curry, 2015). However, some research has ~~focussed~~ focused more specifically on the temporal variability exhibited in ~~this~~ the surface air temperature record (Schwartz, 2007; Cox et al., 2018a), ~~and this topic~~ which is the focus of this paper.

Schwartz (2007) argued on the basis of a simple zero-dimensional energy balance model that an analysis based on the fluctuation-dissipation theorem (Einstein, 1905) could be used to directly diagnose the sensitivity of the Earth's climate system S — here conventionally defined as the equilibrium surface air temperature response to a doubling of the atmospheric CO_2 concentration — from variability in the observed record of annually and globally averaged surface air temperature observations over the observational record. While we do not wish to repeat the arguments here, we will note that several researchers disputed this analysis, demonstrating *inter alia* that this method did not reliably diagnose the sensitivity of climate models, and also arguing why it could not be expected to do so, given their complexity ~~(Foster et al., 2008; Knutti et al., 2008)~~ (Foster et al., 2008; Knutti et al., 2008; Kirk-Davidoff, 2009). Perhaps as a consequence of these arguments, this line of research was largely ignored for the subsequent decade.

More recently however, Cox et al. (2018a) reopened this question with an analysis based on an emergent constraint approach. That is, rather than following the directly diagnostic approach of Schwartz (2007), they instead observed that a quasi-linear relationship existed across an ensemble of CMIP5 models (Taylor et al., 2012), between the sensitivities of these models, and their interannual temperature variabilities as summarised in a statistic which they denoted Ψ. It has been cogently argued that an emergent constraint should only be taken seriously if supported by some theoretical basis (Caldwell et al., 2014), and Cox et al. (2018a) did indeed present an analysis — again based on simple zero-dimensional energy balance modelling — which qualitatively underpinned this linear relationship. Using the value of Ψ obtained from observations of surface air temperature, together with the empirical relationship between Ψ and S they had derived from the climate models, they produced a best estimate of the equilibrium ~~sensitivity of the climate system of~~ climate sensitivity of 2.8°C with a likely (66% probability) range of 2.2–3.4°C, a substantially tighter range than most previous research. However, questions have also been raised about this result ~~(Brown et al., 2018; Rypdal et al., 2018; Po-Chedley et al., 2018)~~ (Brown et al., 2018; Rypdal et al., 2018; Po-Chedley et al., 2018; Cox et al., 2018b).

In this paper, we explore the question of to what extent temporal variability in the globally and annually averaged temperature record can be used to constrain equilibrium climate sensitivity. We consider both the internal variability of the climate system itself, and also the total variability including deviation from a linear trend due to the forced response. Our investigations are performed in the paradigm of a simple idealised modelling framework, using a two-layer energy balance model which has been widely used to simulate the climate system and which generalises and improves on the performance of the zero-dimensional

model. As part of our investigations, we examine the relationship between the Ψ statistic and the equilibrium sensitivity in the model. We also show how the full time series of variability can be used to constrain climate sensitivity, under a variety of idealised scenarios. Our results are based on "perfect model" experiments and therefore may be more readily interpreted as a best case scenario for what we can learn from variability, rather than a realistic estimate of this.

In the next section, we present the two-layer energy balance model and briefly outline the experimental methods used in this paper. We ~~firstly~~ first focus on internal variability, that is to say, the temporal variability arising entirely from ~~to~~ internal dynamics of the climate system in the absence of forcing. We evaluate the power of the Ψ statistic in constraining equilibrium sensitivity, and also consider the more general question of what could in principle be learnt from the full time series. We then consider variability over the period of the observational record (primarily the 20th century, but with some extension into the 19th and 21st centuries). This includes forced variability due to temporal changes in both natural and anthropogenic forcings as well as the internal variability of the climate system. Throughout the paper, the term variability refers simply to all temporal variation in the annually-averaged temperature time series after any linear trend is removed.

2 Methods

2.1 Model

The basic underpinning of previous work is energy balance modelling of the climate system, from which it is anticipated that interannual variability may be informative regarding the equilibrium sensitivity. While previous research was based on analysis of the simplest possible zero-dimensional single layer planetary energy balance, there is evidence that the behaviour of the climate system over the ~~20th century~~ historical period is poorly modelled by such a system (e.g. Rypdal and Rypdal, 2014). Therefore, we use here a slightly more complex two-layer model based on ~~Winton et al. (2010)~~Winton et al. (2010); Held et al. (2010) . This model has been shown to reasonably replicate the transient behaviour of the CMIP5 ensemble of complex climate models (Geoffroy et al., 2013b, a). The model is defined by the two equations:

$$C_m \frac{dT_m}{dt} = F^t + \lambda T_m - \epsilon\gamma(T_m - T_d) + C_m\delta^t \tag{1}$$

$$C_d \frac{dT_d}{dt} = \gamma(T_m - T_d) \tag{2}$$

This is a two-layer globally-averaged energy balance model which simulates the mixed (T_m) and deep (T_d) ocean temperature anomalies in the presence of time-varying forcing F^t. $\lambda = -F_{2\times}/S$ is the radiative feedback parameter where S is the equilibrium sensitivity and $F_{2\times}$ is the forcing due to a doubling of the atmospheric CO_2 concentration. C_m and C_d are the heat capacities of the mixed-layer and deep ocean respectively and γ represents the ocean heat transfer parameter. The parameter ϵ was introduced by Winton et al. (2010) to represent the deep-ocean heat uptake efficacy, and while it is not important for our analysis, we include it for consistency with the broader literature. In a slight modification to Winton et al. (2010), we add a

Parameter	Default Value (Prior)	Description
S	3.0 (U[0,10])	Equilibrium climate sensitivity ($^\circ C$)
λ	~~$3.7/S$~~ $3.7/S$	Radiative feedback ($Wm^{-2}\,^\circ C^{-1}$)
γ	0.7 (N(0.7,0.2^2))	Deep ocean heat uptake parameter ($Wm^{-2}\,^\circ C^{-1}$)
ϵ	1.3 (N(1.3,0.3^2))	Deep ocean heat uptake efficacy
$F_{2\times}$	3.7	Forcing of $2\times CO_2$ (Wm^{-2})
D_m	75	Depth of mixed layer (m)
C_m	$4.2\times10^6\times0.7\times D_m$	Heat capacity of mixed layer ($Jm^{-2}\,^\circ C^{-1}$)
C_m	7.0	Heat capacity of mixed layer ($W\,yr\,m^{-2}\,^\circ C^{-1}$)
D_d	1000	Depth of deep ocean (m)
C_d	$4.2\times10^6\times0.7\times D_d$	Heat capacity of deep ocean ($Jm^{-2}\,^\circ C^{-1}$)
C_d	93	Heat capacity of deep ocean ($W\,yr\,m^{-2}\,^\circ C^{-1}$)
σ	0.05 (N(0.05,0.01^2))	Gaussian noise parameter ($^\circ C$)

Table 1. Adjustable parameters and default values

noise term δ^t to the first equation to represent the internal variability of the system as was originally introduced in a single layer energy balance climate model by Hasselmann (1976). Here δ^t is sampled on an annual (ie, time step) basis from a Gaussian $N(0,\sigma)$ where we generally use the value $\sigma = 0.05$ which generates deviations of order 0.05°C on an annual basis~~.~~, reasonably compatible with both GCM results and observations of the climate system. Our conclusions are not sensitive to this choice. The mixed layer temperature T_m is considered synonymous with the globally averaged surface temperature. The equations are solved via the simple Euler method with a one year time step.

The values of the various adjustable parameters are listed in Table 1. ~~The depths of~~ We assume depths of 75m and 1000m for the mixed and deep ocean layers respectively which are used to calculate the heat capacities C_m and C_d respectively based on ocean coverage of 70% of the planetary surface area. The default values for adjustable parameters are given in Table 1 and the values used here lie close to the mean of those obtained by fitting the model to CMIP5 simulations by Geoffroy et al. (2013a). Our aim here is not specifically to replicate or mimic this ensemble but to allow for a reasonable range of parameter values. If we set $\gamma = 0$ and ignore the deep ocean then we recover the single layer model of Hasselmann (1976) which was used by both Schwartz (2007) and Cox et al. (2018a) in their theoretical analyses.

2.2 Bayesian estimation

Our investigations are performed within the paradigm of Bayesian estimation. In general, the Bayesian approach provides us with a way to estimate a set of unknown parameters Θ from a set of observations O via Bayes' Theorem,

$$P(\Theta|O) = P(O|\Theta)P(\Theta)/P(O). \tag{3}$$

Here $P(\Theta|O)$ is the posterior probability distribution of Θ conditioned on a set of observations O, $P(O|\Theta)$ is the likelihood function that indicates the probability of obtaining observations O for any particular set of parameters Θ, which in this paper will always contain S and may include other parameters. $P(\Theta)$ is a prior distribution for the parameters Θ, and $P(O)$ is the probability of the observations which is required as a normalising constant in the calculation of the posterior probability distribution.

Formally, the value of the observations is fully summarised by the likelihood function $P(O|\Theta)$, but we primarily present our results as posterior pdfs in order to provide an easily interpreted output which can be directly compared to previously published results. We therefore use a uniform prior in S as this is typically the implicit assumption in emergent constraint analyses (Williamson et al., 2019). This choice results in the posterior being visually equivalent to the likelihood even though their interpretation is somewhat different. In some experiments, we will consider that only the sensitivity is unknown, but in others we will consider a wider range of parametric uncertainties. The priors that we use for all uncertain parameters are shown in Table 1.

2.3 Additional data

While this study primarily focusses on the behaviour of the simple energy balance model, we also use and present some data from external sources. In order to perform simulations of the historical period, we force our climate model with annual time series for the major forcing factors based on IPCC (Annex II: Climate System Scenario Tables 2013). Our two-layer model with a one-year time step (and Euler method of numerical integration) reacts rather too strongly to short-term spikes in forcing and thus we scale the volcanic forcing to 70% of the nominal value in order to give more realistic simulations. We show some outputs of the model together with surface air temperature observations from HadCRUT (Morice et al., 2012) as a purely visual indication of the model's performance. We do not use these real temperature observations in any of our analyses, however.

For comparison with our simple model results, we also present some results calculated from historical simulations performed by climate models in the CMIP5 (Taylor et al., 2012) and CMIP6 (Eyring et al., 2016) ensembles. For CMIP5, we use results from 23 models obtained from the Climate Explorer website (https://climexp.knmi.nl/). Where multiple simulations were performed with a single model, we show all results (amounting to 89 model runs in total) and these vary substantially due solely to the sample of internal variability in each simulation. Output from CMIP6 models was provided to the authors by Martin Stolpe. Due to the highly variable size of initial condition ensembles in this set of simulations, we limited use to at most 5 simulations from each model, resulting in a sample of 117 simulations from 31 models.

3 Unforced (internal) variability

3.1 Using scalar measures of variability to estimate S

Schwartz (2007) and Cox et al. (2018a) both summarised the variability in the temperature record with a scalar measure that they argued (based on simple energy balance modelling) should be informative regarding the sensitivity. Schwartz (2007) sum-

marised the variability via the characteristic decorrelation time constant $\tau = \tau(\Delta t) = -\Delta t / \ln(\rho_{\Delta t})$ where Δt is an adjustable lag time and $\rho_{\Delta t}$ is the autocorrelation coefficient of the temperature time series at a time lag of Δt. The method of selecting Δt and therefore the estimation of τ was not presented in an entirely objective algorithmic form, but for the simple one-layer climate model that was considered, the expected value of τ calculated from a long unforced time series is independent of lag. Cox et al. (2018a) argued that the function $\Psi = \sigma_T / \sqrt{-\ln \rho_1}$ should be linearly related to the equilibrium sensitivity. In this function, ρ_1 is the lag-1 autocorrelation of the time series of annual mean surface temperatures, and σ_T is the magnitude of interannual variability of these temperatures. Ψ and τ are closely related and co-vary very similarly over a wide range of sensitivity when other model parameters are held fixed (not shown). Henceforth in this section we focus solely on Ψ as it is more precisely defined and has been recently discussed in some detail (Williamson et al., 2019). However very similar results are also obtained when equivalent experiments are performed using τ.

We now present some investigations into the relationship between Ψ and S in unforced simulations of the two-layer model introduced in Section 2. We perform a multifactorial experiment in which 1000-member ensembles of simulations are integrated for both 150 and 1000 year duration, over a range of S from 0 to 10C, and with γ set to either the default value of 0.7 or alternatively set to 0 in which case we recover the single-layer version of the model. All other model parameters are held fixed at standard values in these experiments. Since there is no forced trend in these experiments, we do not include any explicit detrending step in ~~the analyses.~~ these analyses. However the results are insensitive to detrending.

Figure 1 shows the results obtained when Ψ is calculated from the time series of annual mean surface temperatures produced by these simulations. For 150-year simulations using the single layer model, there is a strong linear relationship between the mean value of Ψ obtained, and the sensitivity of the model, just as Cox et al. (2018a) argued. However, Cox et al. (2018a) did not consider sampling variability, that is to say, the precision with which this expected value of Ψ might be estimated from a finite time series. As our results show, there is substantial uncertainty in the value of Ψ obtained by individual runs, and there is also strong heteroscedasticity, that is to say, the variance of each ensemble of Ψ values increases markedly with sensitivity. This variation arises from the sequence of noise terms which generate the internal variability in each simulation of the model and is therefore an intrinsic aspect of the theoretical framework relating Ψ to S. For these unforced simulations, it seems quite possible for a model with its sensitivity set to a value of 5°C or even higher to generate a time series which has a modest value for Ψ of say 0.1, even though the expected value of Ψ from such model simulations would be much higher. Similarly to the results shown by Kirk-Davidoff (2009), an accurate diagnosis of Ψ could in principle be made with a sufficiently long time series of internal variability, but the sampling uncertainty only decreases with the square root of the duration of the time series (as expected from the Central Limit Theorem), so this is unlikely to be of use in practical applications with real data.

When we consider the two-layer model using the standard parameter value of $\gamma = 0.7$ then the situation is a little different. In this case the relationship between sensitivity and Ψ is flatter and more curved, with the expected value of Ψ changing slowly for $S > 4$°C. The underlying explanation for this is quite simple. Any small perturbation to the surface temperature is damped on the annual time scale by a relaxation factor which varies in proportion to $\epsilon\gamma - \lambda$, and $\epsilon\gamma$ is equal to 0.91 for standard parameter values. Therefore, when S is large, changes in $\lambda = -3.7/S$ have relatively little impact on the total damping and thus both the magnitude and autocorrelation of variability are relatively insensitive to further increases in S. Williamson et al. (2019)

also presented a theoretical analysis of this two-layer model in which they argued that the response of Ψ was close to linear across the GCM parameter range, and our result confirms this for sensitivity values from around 2 to 4 or even 5°C. However, the gradual curvature for larger values results in a saturation of the response of Ψ to increases in S and this, together with the increasing sampling uncertainty, has consequences that will be shown in subsequent experiments. In fact the relationship between Ψ and the transient climate response (i.e. the warming observed at the time of CO_2 doubling under a 1% per annum increase) is more close to linear, than the relationship between Ψ and S. Thus our work does not challenge the underlying analysis that they presented, but augments it with additional details.

We now directly consider the question of how useful an observed value Ψ^o can be as a constraint on the equilibrium climate sensitivity through Bayesian estimation. It is not trivial to directly calculate the exact value of the likelihood $P(\Psi^o|\Theta)$ for a given observed value Ψ^o, as Ψ is itself a random variable arising from the stochastic model and thus depends on the sequence of random perturbations that were generated during the numerical integration of the model. Therefore, we use here the technique known as Approximate Bayesian Computation (Diggle and Gratton, 1984; Beaumont et al., 2002). This is a rejection-based sampling technique in which samples are drawn from the prior distribution, used to generate a simulated temperature time series, and rejected if the value of Ψ calculated from this time series does not lie within a small tolerance of the observed value. The set of accepted samples then approximately samples the desired posterior. We have no observations of long periods of unforced climate variability with the real climate system, so we perform a number of synthetic tests in which different hypothetical values for Ψ^o are tested.

Our experiments take the form of a 'perfect model' scenario, where the model is assumed to be a perfect representation of the system under consideration, with no structural imperfections. Our uncertainties here are due solely to unknown parameter values and internal variability noise. In these experiments, we assume that Ψ for the true system is calculated from a 150 year temperature time series of the unforced system, without any observational error whatsoever. The results of four experiments — using values of Ψ^o which range from 0.05 to 0.2 in regular increments — are shown in Figure 2. There is not necessarily an immediate correspondence between these synthetic values and the observationally-derived value that Cox et al. (2018a) calculated, as we are using unforced model simulations here. Nevertheless, the results are qualitatively interesting. With other model parameters set to the default values, the four values of Ψ^o used here correspond to the expected value generated by 150 year integrations with sensitivities of approximately 1, 2.5, 5 and 10°C respectively. The figure shows that in this experimental scenario, Ψ can only provide a tight constraint in the first case where the sensitivity is very low. In this case, the 5–95% range of the posterior is an impressively narrow ~~0.7~~0.6–1.7°C. For the case $\Psi^o = 0.1$, the equivalent probability interval is 1.8–8~~.0~~.1°C and for higher values of Ψ the posterior is very flat indeed with just the very lowest values of S excluded. Similar results are obtained when equivalent values of τ are used as observational constraints.

Strictly, when considering the strength of the constraint ~~, it is~~ obtained from the variability, we should focus on the likelihood $P(\Psi|S)$ ~~that we should be considering,~~ rather than the posterior pdf $P(S|\Psi)$, since the latter depends also on the prior ~~.~~ ~~Likelihood values~~ which is in principle independent of the observations. The likelihood for different values of S, which tells us the relative probability of any particular sensitivity value generating the observation, can be directly read off from Figure 2 as the height of the appropriate density curve at ~~specific sensitivities, rather than the integral under it~~the specific sensitivity value.

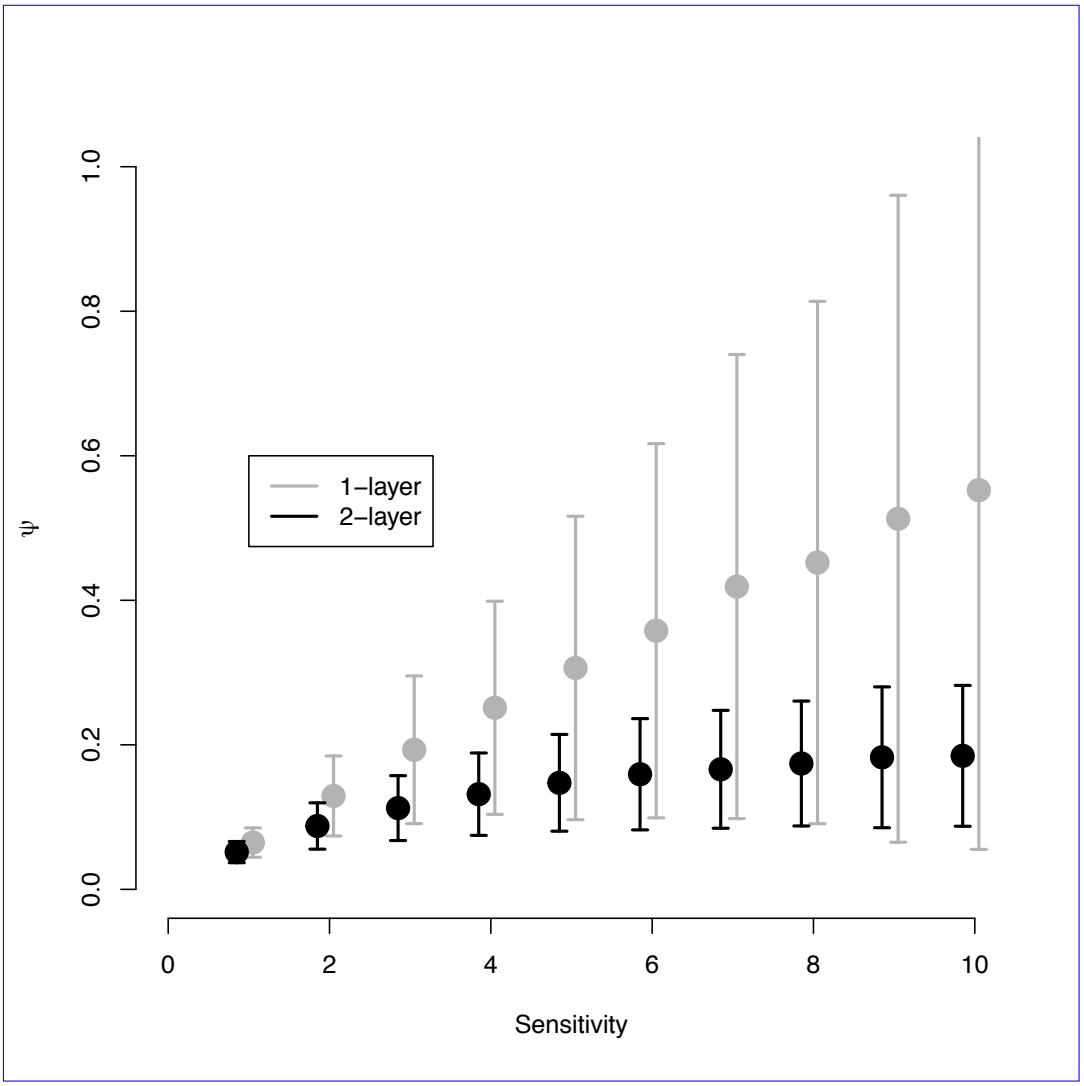

Figure 1. Ψ estimated from 150y time series for 1- and 2-layer model. Grey results are for single-layer model and black results for two-layer model. Large dots show means of 1000 simulations, with error bars indicating ±2sd ranges for each ensemble. Results are calculated at each integer value of sensitivity and offset slightly for visibility.

In the experiment performed with $\Psi^o = 0.1$, the maximum likelihood value is achieved at a value of ~~$S = 2.4$~~$S = 2.5$°C, and the likelihood drops by a factor of 10 at both S=~~1.4~~1.3°C and S=~~7.6~~7.9°C. Kass and Raftery (1995) suggest that a likelihood ratio of 10 or more between two competing hypotheses could be taken to represent "strong" evidence in favour of one over the other, so if we adopt this linguistic calibration we could say that the observation of $\Psi^o = 0.1$ represents strong evidence in favour of ~~$S = 2.4$~~$S = 2.5$°C versus all values outside of the range ~~1.4~~1.3–~~7.6~~7.9°C (but conversely, does not represent strong evidence to discriminate between any pair of values inside that range). It is somewhat coincidental that this range seems quite

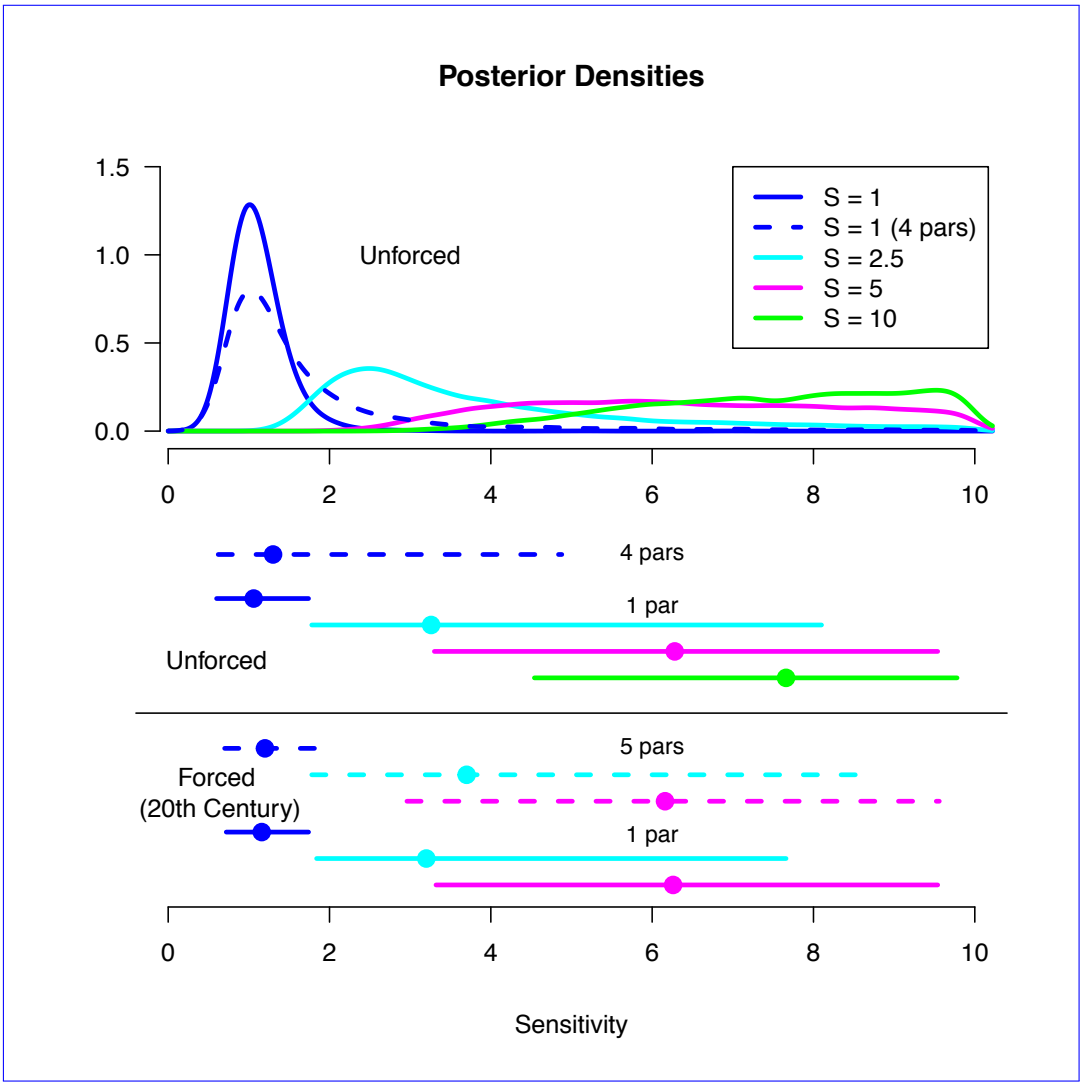

Figure 2. Posterior estimates of sensitivity ~~based on~~ inferred by using observations of Ψ ~~estimated from 150y time series with unforced~~ to constrain parameters in the 2-layer model. ~~Four~~ Top panel: four solid-line pdfs in blue, cyan, magenta and green represent estimates based on 150y unforced simulations, assuming observations of $\Psi^o = 0.05$, 0.1, 0.15, and 0.2 respectively, where only S is uncertain with uniform prior. ~~Dotted~~ Dashed blue line represents posterior estimate for $\Psi^o = 0.05$ with additional ~~modelling~~ parametric uncertainties as described in ~~main text~~Section 3.2. Horizontal lines and dots in "Unforced" central panel indicate 5–95% ranges and median respectively of these experiments. Horizontal lines and dots labelled as "Forced (20th Century)" are similar results based on forced simulations of historical period as described in Section 4.1. Solid lines: only S is uncertain. Dashed lines: multiple uncertain parameters.

similar to the 5–95% range of the posterior pdf as the philosophical interpretation of the ranges is rather different. There is a strong skew in this range, which extends much further towards higher values of S than lower ones, compared to the maximum

likelihood estimate. We stress that this skew is a fundamental property of the physical model and is not due to the Bayesian analysis paradigm.

3.2 Additional uncertainties

The ~~experiments summarised in~~ pdfs plotted in the top panel of Figure 2 assume that all model parameters other than S are known with certainty. In reality, we have significant uncertainty as to what values we should assign to several other parameters. We consider just three of these: the ocean heat uptake parameter γ, the efficacy or pattern effect parameter ϵ and the internal noise parameter σ. Geoffroy et al. (2013a) fitted the two-layer model to various GCM outputs in order to estimate parameter values including γ and ϵ and based on these results we use as priors for these parameters the distributions $N(0.7, 0.2)$ and $N(1.3, 0.3)$ respectively which generously encapsulate their results. Geoffroy et al. (2013a) did not consider internal variability and thus we do not have such a solid basis for a prior in σ and assume a comparable relative uncertainty of 20%, i.e. a prior of $N(0.05, 0.01)$. When we repeat the previous experiments but include these additional parametric uncertainties, then for the experiment where we use $\Psi^o = 0.05$ as a constraint, the posterior for S widens substantially from the previous spread of ~~0.7~~0.6–1.7°C, to ~~0.7~~0.6–~~4.8~~4.9°C as also shown as the dashed blue line in Figure 2. The largest factor generating this substantial increase in uncertainty is due to the uncertainty in σ. The equivalent posteriors using the larger observational values for Ψ^o also broaden somewhat but this is less visible in the results as they are of course always constrained by the prior range.

3.3 Using the full time series

Although the results in Figure 2 show that an observation of Ψ taken from a short unforced simulation cannot tightly constrain equilibrium sensitivity in this model (except perhaps in the most exceptional of circumstances), it could still be hoped that a more precise constraint could possibly be gleaned by a more advanced analysis that uses some different diagnostic of the time series. In this section, we show how the total information of the time series can be used. By doing this, we create the most optimistic possible scenario for using internal variability to constrain equilibrium sensitivity of this simple climate model.

This approach requires us to calculate the likelihood for the full set of observations, $P(O|\Theta)$ where here $O = T_m^i, i = 1 \ldots N$ is the full time series of annual surface temperature anomalies. Once the model parameters and forcing are prescribed, the time series of surface temperature anomalies is uniquely determined by the series of random noise perturbations δ^i. Thus, in the absence of observational error, we can invert this calculation to calculate (up to machine precision) the sequence of annual random noise perturbations δ^i, $i = 1 \ldots N$ that are required in order to replicate any given observed temperature time series. This is why we selected a model time step ~~of one year: with one observation per year, we can precisely invert the model integration to calculate one uncertain noise input per year and thereby reproduce the full model integration to within machine precision~~as long as one year, as it results in the number of observations being as large the number of noise terms making this exact inversion possible. The probability of the model (with a particular set of parameters) generating the observed sequence is exactly the probability of the required noise sequence being sampled. This value is readily calculated, since the joint density of these ~~i.i.d. generated~~ independent and identically distributed δ^i is simply the product of their individual densities. This unusual approach which we do not believe has been previously implemented in this context is possible here as we are assuming zero

observational uncertainty. With this exact likelihood calculation, the Bayesian estimation process is straightforward. In the case where only sensitivity is considered uncertain, it can be performed by direct numerical integration, sampling the sensitivity on a fine regular grid and calculating the likelihood (and therefore posterior) directly at these values.

It is worth emphasising that this calculation represents an absolute best case scenario for using the time series of temperature anomalies as a constraint. There can be no diagnostic or statistical summary of the observations that provides more information than the full set of observations themselves contain. Thus, we cannot hope to obtain a better constraint by some alternative analysis of the temperature time series.

Figure 3 shows results obtained from this approach, in the case where only S is considered uncertain. To aid visual comparability with Figure 2, the y-axis scale is fixed at the same value despite cutting the peaks of some pdfs. It is not possible to define what a "typical" noise sequence might look like and therefore we plot 20 replications with different randomly generated instances of internal variability for each sensitivity value tested. ~~We can see~~ It seems that the results ~~are generally~~ may be a little more precise than was obtained using Ψ alone (as shown by the pdfs generally having higher peak densities), though this depends on the specific sample of internal variability that was obtained. It is still only in the case of the lowest sensitivity value of 1°C that we reliably obtain a tight constraint. With the true sensitivity of 2.5°C, the posterior 5–95% range, averaged over the samples, is 1.9~~–6.8~~–7.0°C, ~~a little~~ marginally narrower than the 1.8–8.1°C range obtained previously when an equivalent Ψ value of 0.1 ~~is~~ was used. When additional parametric uncertainties are considered in this unforced scenario, the constraints again weaken, though not to quite such an extent as in Section 3.2 when only Ψ is used as a constraint. We do not show these results here.

Thus, there appears to be the potential for internal variability, as represented by the full temperature time series, to provide a slightly better constraint than that obtained by a summary statistic alone, but the improvement is marginal and even our optimal calculation which uses the exact likelihood of the full time series cannot accurately diagnose equilibrium sensitivity except when the true value is very low. These results again show a skew similar to that obtained when Ψ was used as the constraint in Sections 3.1 and 3.2. Thus this non-Gaussian likelihood is again an inherent property of the physical model and not an artefact of the analysis. We mention again that these calculations are made under the three optimistic assumptions that (a) the model is perfect and we have exact knowledge of all other model parameters, (b) we know the forcing to be zero over this time period, and (c) there is no observational error.

4 Forced variability

While the theoretical underpinning of Cox et al. (2018a) was originally based on the properties of unforced internal variability, Cox et al. (2018b) acknowledged that their approach may have benefited from some signature of forced variability entering into their calculations. In order to calculate their Ψ statistic, they applied a windowed detrending method in order to focus on variability of both model simulations and observations of the ~~20th century~~historical period. However, the window length of 55 years that they used was justified primarily in empirical terms and cannot remove shorter-term variations in forced response.

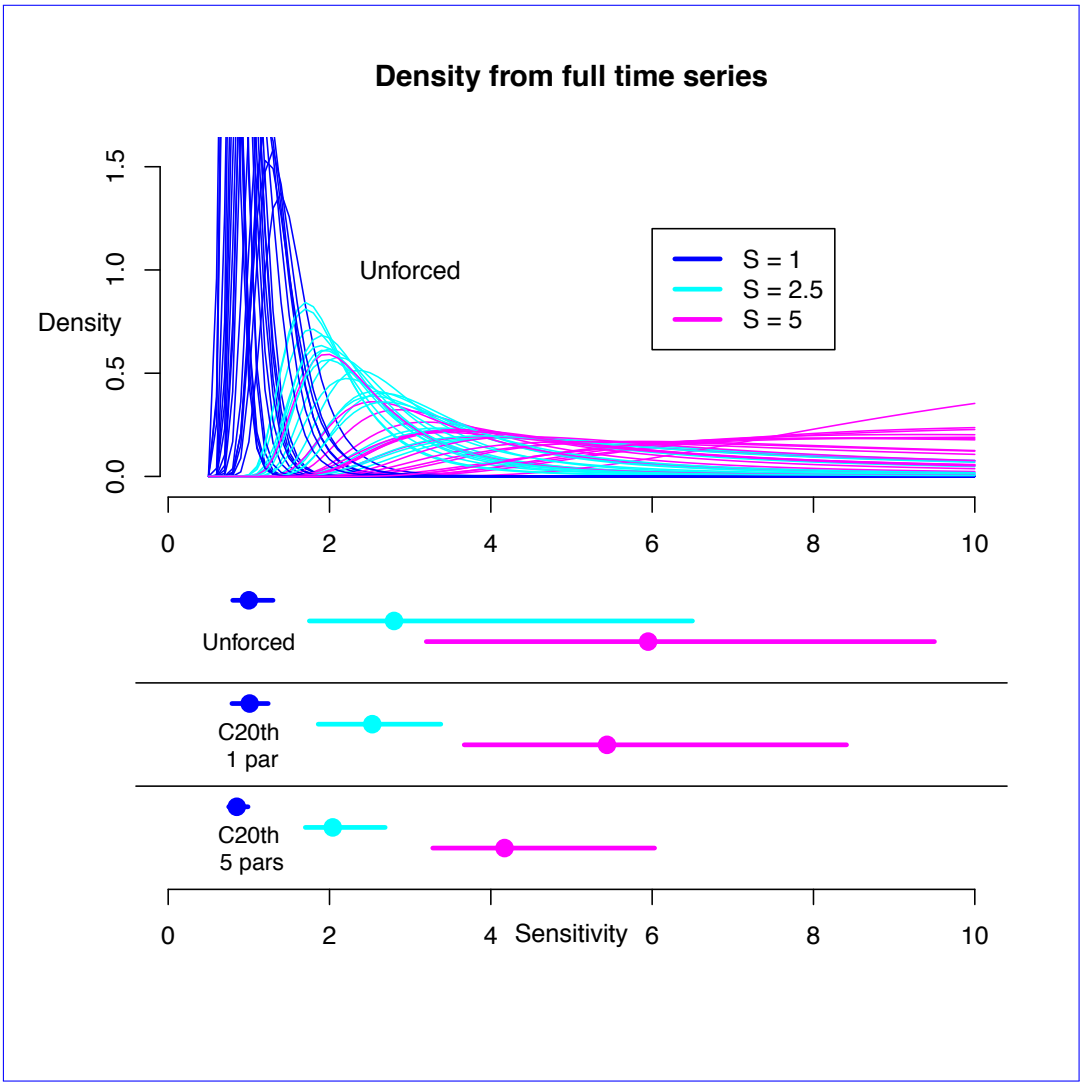

Figure 3. Posterior estimates for the climate sensitivity ~~using~~ from Bayesian estimation using the full ~~150 year~~ time series of annual mean surface temperatures. Main plot: Results from 150y unforced simulations as discussed in Section 3.3. 20 replicates are performed for each true sensitivity of 1, 2.5, 5°C as indicated by the colour blue, cyan and magenta respectively. Horizontal lines and dots immediately below top panel show means of the 5–95% range and median of each set of results. Horizontal lines labelled "C20th" show analogous results using simulations of historical period, with only S uncertain or with 5 uncertain parameters as discussed in Section 4.2

In this section, we perform a series of analyses based on ~~20th century~~ historical forced simulations, in order to investigate more fully the potential for such forced effects to improve the constraint. We force the climate model with annual time series for the major forcing factors based on IPCC (Annex II: Climate System Scenario Tables 2013). Our two-layer model with a one-year time step (and Euler method of numerical integration) reacts rather too strongly to short-term spikes in forcing

and thus we scale the volcanic forcing to 70% of the nominal value in order to give more realistic simulations. In some of the following experiments, we consider aerosol forcing as a source of uncertainty in addition to that arising from the internal parameters of the model. This uncertainty is implemented via a scaling factor denoted by α which is uncertain but constant in time, applied to the original aerosol forcing time series. In these cases, our prior distribution for α is ~~$N(1,0.5)$~~$N(1,0.5^2)$.

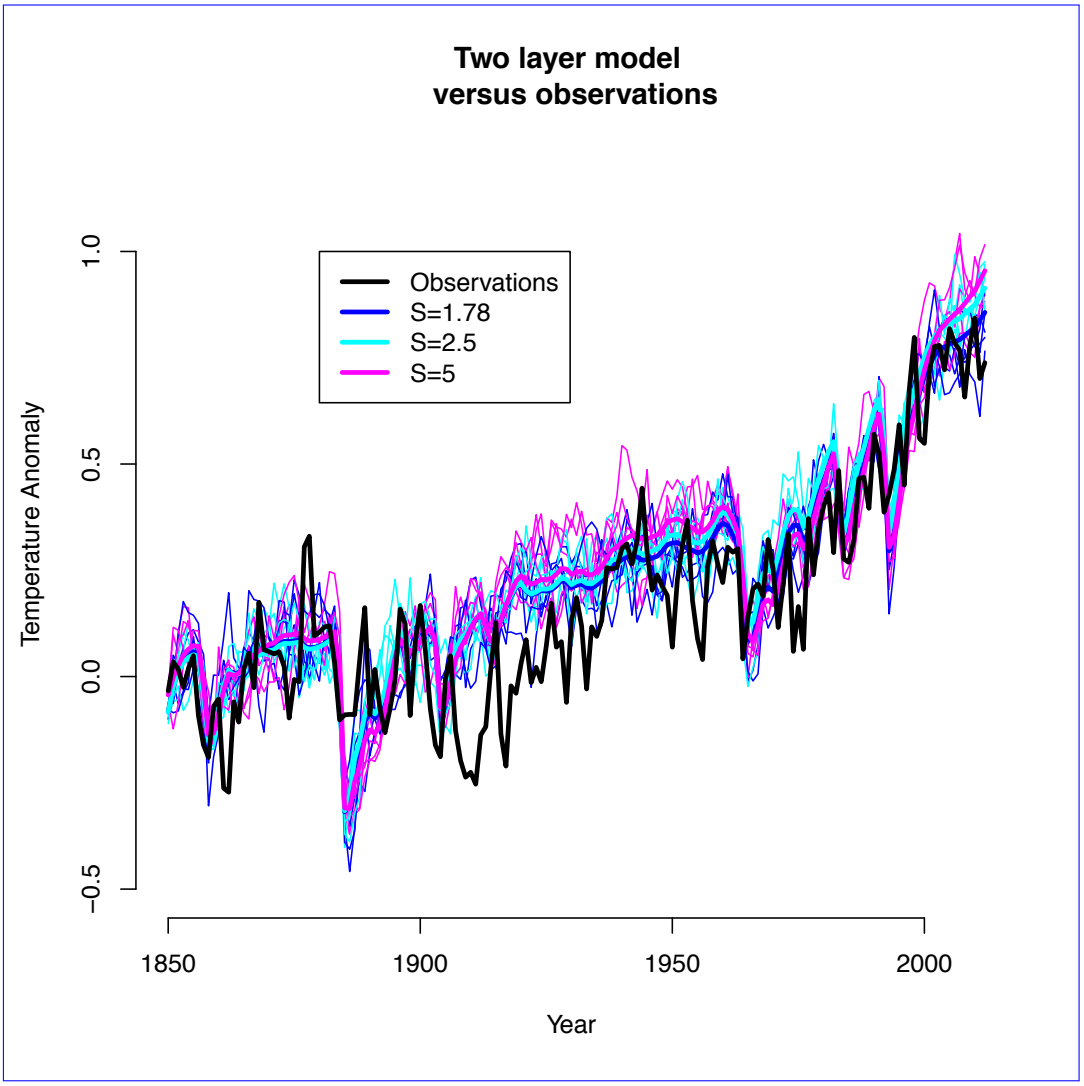

Figure 4. Simulations of instrumental period with 2-layer model. Thick lines are forced response excluding internal variability, thin lines are 5 replicates of each parameter set including internal variability. Blue lines: $S = 1.78$°C, $\gamma = 0.7 W m^{-2}\ ^{\circ}C^{-1}$, $\epsilon = 1.3$. Cyan lines: $S = 2.5$°C, $\gamma = 1.0 W m^{-2}\ ^{\circ}C^{-1}$, $\epsilon = 1.7$. Magenta lines: $S = 5$°C, $\gamma = 1.0 W m^{-2}\ ^{\circ}C^{-1}$, $\epsilon = 1.7$, $\alpha = 1.7$. Black line is HadCRUT data.

Figure 4 presents 18 simulations from the model, consisting of 5 instances of internal variability for each of three different parameter sets which were chosen to give reasonable agreement with observational data, and additionally one simulation for each of these parameter sets in which internal variability was not included in order to show the pure forced response. These simulations are shown merely to indicate the typical behaviour of the model under ~~20th century forcing~~ historical forcing estimates and are not directly relevant to our analyses. Note that the observations of the real climate system which are also plotted here include observational error (estimated to be roughly $\pm$ 0.05°C at the 1 standard deviation level) whereas the model output is presented as an exact global temperature. Thus it is to be expected that the model results are somewhat smoother and less variable than the observations~~. The~~ , although it may also be the case that the two layer model has insufficient variability. For each of these three simulations without internal variability, the RMS differences between model output and observations is ~~identical to two significant figures for the three simulations without internal variability, at~~ 0.13°C.

When we hold other parameters at default levels, best agreement between model and data (defined here simply by RMS difference between the two time series) is achieved for a rather low sensitivity of 1.78°C. If the γ and ϵ parameters are increased slightly above their defaults then we can achieve an equally good simulation (again as measured by RMS difference) with a higher value for sensitivity of 2.5°C. If, additionally, aerosol forcing is also increased above the default value, then a higher sensitivity still of 5°C achieves an equally good match to the observed temperature time series ~~. These~~ at the global scale. While there are hemispheric differences in these forcings that may provide some additional information (Aldrin et al., 2012) , these simulations help to illustrate why it has been so challenging to effectively constrain equilibrium sensitivity from the long-term observed warming.

4.1 Using Ψ

In order to assess what we can learn about sensitivity from the variability of ~~20th century~~ historical temperature observations, we ~~firstly~~ first consider the utility of Ψ calculated from ~~20th century simulations~~ simulations over this period. Cox et al. (2018a) proposed that the effect of forcing over this interval could be effectively removed by a process of windowed detrending. Figure 5 shows results analogous to ~~those of~~ the two-layer simulations in Figure 1, but using forced simulations of the ~~20th century~~ historical period rather than unforced control simulations, and therefore with Ψ calculated via the windowed detrending method of Cox et al. (2018a). One very minor discrepancy with the calculations presented in that paper is that our simulations only extend to 2012 (this being the limit of the forcing time series that we are using) and therefore we omit the last 4 years of the time period that they used. This does not significantly affect any of our results. We do not consider one-layer simulations here as this version of the model is known to provide poor simulations of historical changes (Rypdal and Rypdal, 2014).

Grey dots and error bars indicate results obtained when only S is considered uncertain. These results are qualitatively similar to those obtained by the unforced two-layer simulations in Figure 1, in that they exhibit a nonlinear and heteroscedastic relationship that levels off for large values of S. The values of Ψ obtained ~~for any specific~~ at each value of S are ~~somewhat larger than those of Figure 1~~however somewhat larger in the forced experiments, which supports the claims of Brown et al. (2018) and Po-Chedley et al. (2018) that the windowed detrending has not been wholly effective in eliminating all influence of forcing. ~~The overall nature of the relationship is broadly similar, however.~~ As mentioned previously, ~~results from GCMs~~ analysis of

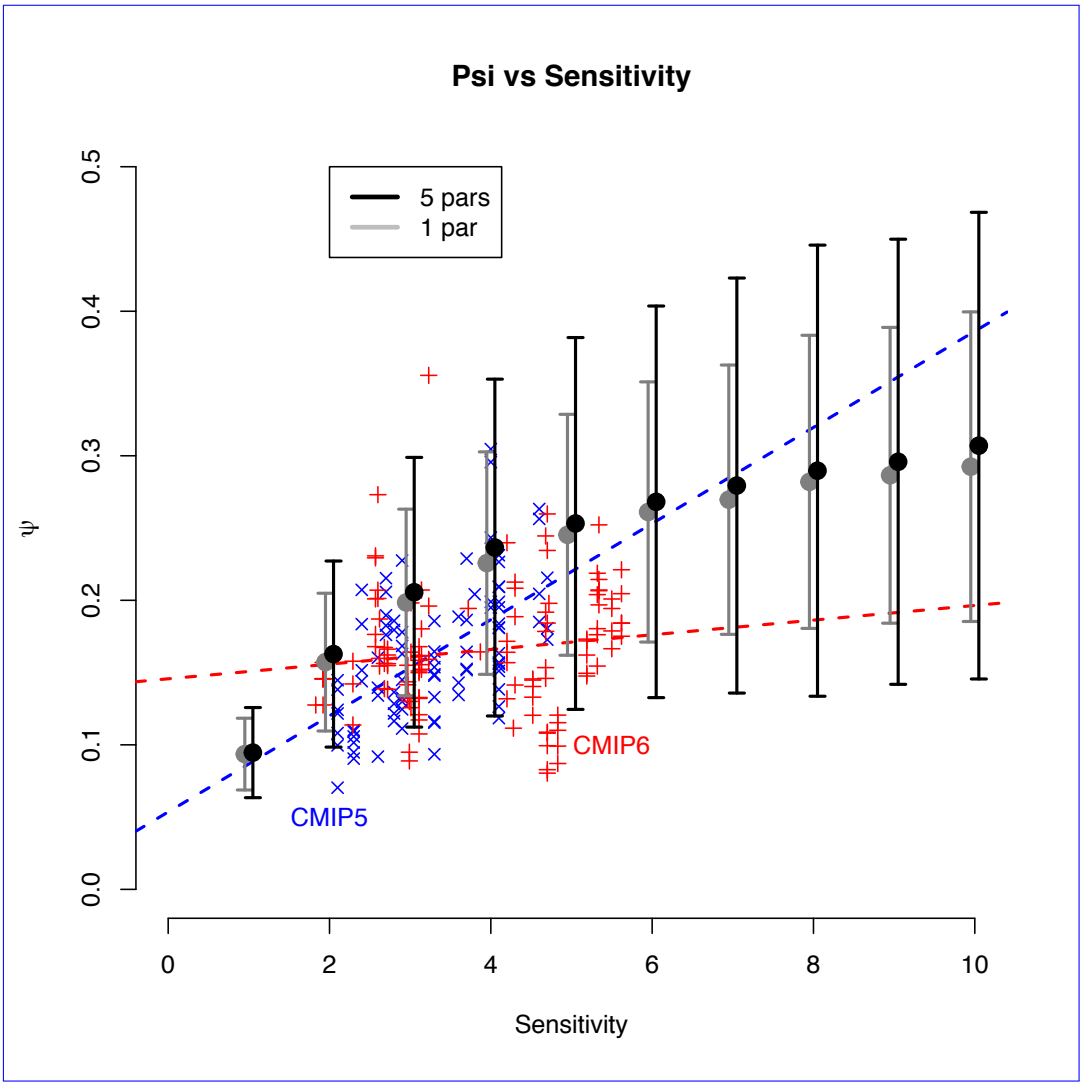

Figure 5. Ψ estimated from ~~20th century~~ historical simulations from 2-layer model. Grey results are based on simulations where only S is considered uncertain. Black results additionally account for uncertainty in γ, ϵ, σ and α. Large dots show means of 1000 simulations, with error bars indicating $\pm$2sd ranges for each ensemble. ~~Crosses~~ Blue and red crosses indicate results generated by CMIP5 and CMIP6 models respectively, together with the best-fit regression lines as dashed lines, in matching colours.

GCM outputs indicate significant uncertainty in other model parameters and therefore we have performed additional ensembles of simulations which account for uncertainty both in model parameters and aerosol forcing. These uncertain parameters, and their priors, are the same as in Section 3.2, with the addition here that we also consider uncertainty in the aerosol forcing through the scaling parameter α. The combined effect of these uncertainties has little systematic effect on the mean estimate

of Ψ but slightly increases the ensemble spread. Interestingly, in contrast to our earlier experiments, no single factor appears to have a dominant effect here.

~~Black~~ Blue and red crosses also shown on this figure show results obtained from the CMIP5 ~~ensemble. We obtained historical simulations performed by 23 CMIP5 models from the Climate Explorer website (). Where multiple simulations were performed with a single model, we show all results (amounting to 89 model runs in total) and these vary substantially due solely to the sample of internal variability in each simulation. The CMIP5~~ and CMIP6 ensembles. The CMIP models appear to generate slightly lower values of Ψ than the two-layer model does with the same sensitivity, although the ~~two sets of~~ results seem broadly compatible. Reasons for the difference may include biases in parameters of the two-layer model, structural limitations, or differences in the forcings used. Although we are not replicating the emergent constraint approach here, we do note that there is a significant correlation in the CMIP5 ensemble results shown here between their sensitivities and Ψ values. However, the relationship seen here is is weaker than that obtained by Cox et al. (2018a) for a different (though overlapping) set of models, and explains a lower proportion of the variance. This remains true whether we perform the regression with S as predictor, as suggested by our Figure, or when using Ψ as predictor as in Cox et al. (2018a). For CMIP6, the correlation is insignificant.

~~Posterior estimates of sensitivity based on Ψ estimated from 20th century simulations with 2-layer model. Three solid-line pdfs in blue, cyan and magenta represent estimates based on Ψ^o = 0.1, 0.18, and 0.25 respectively, where only S is uncertain with uniform prior. Dashed lines represent equivalent posteriors with additional modelling uncertainties as described in main text.~~

~~Figure ??~~ Figure 2 shows results generated when various hypothetical values for Ψ^o are used to constrain model parameters for ~~20th century~~ historical simulations. As before, we test sensitivity values of 1, 2.5 and 5°C which here correspond to values for Ψ^o of 0.1, 0.18 ~~,~~ and 0.25 respectively. As in the previous experiments, the smallest value of Ψ^o generates ~~an impressively~~ a tight constraint with a 5–95% range of ~~0.8~~0.7–1.7°C when only S is considered uncertain. This grows to 1.8–7.~~5~~.2°C for $\Psi^o = 0.18$ and the larger value of $\Psi^o = 0.25$ provide very little constraint. When the additional parametric and forcing uncertainties are considered, the tightest range corresponding to $\Psi^o = 0.1$ grows a little to ~~0.8–1.9~~0.7–2.0°C and the $\Psi^o = 0.18$ case spreads to 1.8–8.8°C.

When we use the observational value of $\Psi^o = 0.13$ (calculated from HadCRU data) and include multiple parametric uncertainties, the 17–83% posterior range is 1.3–2.~~6~~.7°C and the 2.5–97.5% range is ~~1.0~~0.9–5.~~5~~.1°C. These ranges are somewhat larger than the equivalent results presented by Cox et al. (2018a) which were 2.2–3.4°C and 1.6–4.0°C respectively, despite the highly optimistic perfect model scenario considered here.

4.2 Using the full time series

Finally, we repeat the approach of Section 3.3, and use the full information of the time series, by the same method of inverting the model to diagnose the internal variability noise that is required to generate the observed temperature time series. Since we are interested solely in variability, we only consider the temperature residuals after detrending. We use a simple linear detrending over the period 1880-2012, which will leave a signature primarily due both to volcanic events and also the contrasting

temporal evolution of negative aerosol forcing and positive greenhouse gas forcing, which both generally increase throughout this period but which exhibit different multidecadal patterns.

The calculation is similar to that of Section 3.3, but there are some minor details which are worth mentioning. Although the detrending is performed over the interval 1880-2012, we initialise the simulations in 1850 to allow for a spin-up. In contrast to Section 3.3 where detrending was not performed, knowledge of the residuals after detrending does not actually enable an exact reconstruction of the internal variability of the model simulation, as any random trend in this internal variability will have been removed by detrending. In fact a whole family of different model simulations will be aliased onto the same residuals. Therefore our inversion only truly calculates the noise perturbations after the removal of the component that generates any linear trend. This is not a significant problem for the likelihood calculation as the effect of this aliasing is minor and its dependence on model parameters is negligible.

The results of multiple replicates are shown in ~~Figures ?? and ??~~Figure 3, in which we consider ~~firstly~~ first the case where only S is uncertain, and then our larger set of parametric uncertainties. In the case where only S is unknown, the full time series of detrended residuals provides strong evidence on S which can as a result be tightly constrained except perhaps when it takes a high value such as 5°C. For $S = 2.5$°C, the posterior 5–95% range is typically under 2C in width, with the average of our samples being 1.9–3.5°C. When multiple uncertainties are considered, the constraint is however markedly weakened. For a true sensitivity of $S = 2.5$°C, the posterior 5 – 95% range is typically ~~around~~ over 4 degrees, at ~~1.7~~ 1.6 – ~~5.6°C~~6.0°C, with a "likely" 17–83% range of 2.2–4.1°C. As in the unforced experiments, these optimal constraints ~~are clearly~~ appear somewhat narrower than can be obtained by using the Ψ statistic, but are not necessarily tight enough to be compelling in themselves.

~~Posterior estimates for the climate sensitivity using Bayesian estimation using the full 1880–2012 observational time series of annual mean surface temperature anomalies after detrending. 20 replicates are performed for each true sensitivity of 1, 2.5, 5°C as indicated by colours blue, cyan and magenta respectively. Only S is considered uncertain.~~

5 Conclusions

~~As Figure ?? but for multiple parametric uncertainties.~~

6 ~~Conclusions~~

We have explored the potential for using interannual temperature variability in estimating equilibrium sensitivity. While — as Williamson et al. (2019) argued — there is generally a quasi-linear relationship between S and the expected value of $\Psi = \sigma_T / \sqrt{-\ln \rho_1}$ over a reasonable range of S in the simple energy balance model, this relationship saturates for higher S and furthermore, sampling variability is significant and highly heteroscedastic. These properties undermine the theoretical basis for the linear regression emergent constraint approach which was presented by Cox et al. (2018a), as the ordinary least squares regression method relies on a linear relationship with homoscedastic errors. The behaviour of the model instead results in an inherently skewed likelihood $P(\Psi|S)$ with a long tail to high values for S. Thus, while Ψ statistic can indeed be informative on

S, the constraint it provides based on internal variability in the case of unforced simulations is rather limited. Furthermore, the CMIP5 and CMIP6 ensembles exhibit quite different relationships in the regression framework, suggesting a lack of robustness of the original analysis. We have shown how it is possible in principle to extract the full information from time series of annual temperatures, by calculating the exact likelihood for the complete set of these observations. However, even in this scenario of a perfect model with a few well-characterised parametric uncertainties and no observational uncertainty on the temperature time series, the constraint on sensitivity is seriously limited by the variability inherent to the model. It is only in the case where the true value of the sensitivity is very low that such an approach can generate a tight constraint. For example, if the true sensitivity takes a moderate value of 2.5°C, then we could only expect to generate a constraint with a typical 5–95% range of around 1.9–6.8°C. As was the case when using Ψ, estimates generated from the full time series are rarely close to symmetric and instead are typically skewed with a long tail to high values. This skew is an inherent property of the physical model that defines the likelihood and not an artefact of our analysis methods.

Forced variability, such as that occurring during the instrumental period, does provide additional information in our experiments, and therefore we could in theory hope to calculate a narrower posterior range, with a typical width of around 4°C (e.g. ~~1.7–5.6~~1.8–6.0°C) when the true sensitivity is 2.5°C. It must however be emphasised that these calculations rely on very optimistic assumptions and therefore represent a best case that is unlikely to be realised in reality. Nevertheless, our results do suggest that variability can inform on the sensitivity and may generate a useful constraint in addition to that arising from the longer-term observed trend.

6 Acknowledgements

TM acknowledges funding from the European Research Council (ERC) (Grant agreement No.770765) under the European Union's Horizon 2020 research and innovation program (Grant agreement No.820829).

References

M. Aldrin, M. Holden, P. Guttorp, R. B. Skeie, G. Myhre, and T. K. Berntsen. Bayesian estimation of climate sensitivity based on a simple climate model fitted to observations of hemispheric temperatures and global ocean heat content. Environmetrics, 23(3):253–271, 2012. doi:10.1002/env.2140. URL http://dx.doi.org/10.1002/env.2140.

Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate Bayesian computation in population genetics. Genetics, 162(4): 2025–2035, 2002.

Patrick T Brown, Martin B Stolpe, and Ken Caldeira. Assumptions for emergent constraints. Nature, 563(7729):E1, 2018.

Peter M Caldwell, Christopher S Bretherton, Mark D Zelinka, Stephen A Klein, Benjamin D Santer, and Benjamin M Sanderson. Statistical significance of climate sensitivity predictors obtained by data mining. Geophysical Research Letters, 41(5):1803–1808, 2014.

Peter M Cox, Chris Huntingford, and Mark S Williamson. Emergent constraint on equilibrium climate sensitivity from global temperature variability. Nature Publishing Group, 553(7688):319–322, January 2018a.

Peter M Cox, Mark S Williamson, Femke JMM Nijsse, and Chris Huntingford. Cox et al. reply. Nature, 563(7729):E10, 2018b.

Peter J Diggle and Richard J Gratton. Monte Carlo methods of inference for implicit statistical models. Journal of the Royal Statistical Society: Series B (Methodological), 46(2):193–212, 1984.

Albert Einstein. Über die von der molekularkinetischen theorie der wärme geforderte bewegung von in ruhenden flüssigkeiten suspendierten teilchen. Annalen der physik, 322(8):549–560, 1905.

Veronika Eyring, Sandrine Bony, Gerald A. Meehl, Catherine A. Senior, Bjorn Stevens, Ronald J. Stouffer, and Karl E. Taylor. Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization. Geoscientific Model Development, 9 (5):1937–1958, May 2016. doi:10.5194/gmd-9-1937-2016. URL https://doi.org/10.5194/gmd-9-1937-2016.

G. Foster, J. D. Annan, G. A. Schmidt, and M. E. Mann. Comment on "Heat capacity, time constant, and sensitivity of Earth's climate system" by S. E. Schwartz. Journal of Geophysical Research, 113(D15):D15102–5, August 2008.

O Geoffroy, D Saint-Martin, G Bellon, A Voldoire, DJL Olivié, and S Tytéca. Transient climate response in a two-layer energy-balance model. Part II: Representation of the efficacy of deep-ocean heat uptake and validation for CMIP5 AOGCMs. Journal of Climate, 26(6): 1859–1876, 2013a.

Olivier Geoffroy, David Saint-Martin, Dirk JL Olivié, Aurore Voldoire, Gilles Bellon, and Sophie Tytéca. Transient climate response in a two-layer energy-balance model. Part I: Analytical solution and parameter calibration using CMIP5 AOGCM experiments. Journal of Climate, 26(6):1841–1857, 2013b.

J. M. Gregory, R. J. Stouffer, S. C. B. Raper, P. A. Stott, and N. A. Rayner. An observationally based estimate of the climate sensitivity. Journal of Climate, 15(22):3117–3121, 2002.

K. Hasselmann. Stochastic climate models. Tellus, 28(6):473–485, 1976.

Isaac M Held, Michael Winton, Ken Takahashi, Thomas Delworth, Fanrong Zeng, and Geoffrey K Vallis. Probing the fast and slow components of global warming by returning abruptly to preindustrial forcing. Journal of Climate, 23(9):2418–2427, 2010.

IPCC. Annex II: Climate System Scenario Tables . In T.F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P.M. Midgley, editors, Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel o book section AII, page 1395–1446. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013. ISBN ISBN 978-1-107-66182-0. doi:10.1017/CBO9781107415324.030. URL www.climatechange2013.org.

R.E. Kass and A.E. Raftery. Bayes factors. Journal of the American Statistical Association, 90(773–795), 1995.

DB Kirk-Davidoff. On the diagnosis of climate sensitivity using observations of fluctuations. Atmospheric Chemistry & Physics, 9(3), 2009.

Reto Knutti, Stefan Krähenmann, David J Frame, and Myles R Allen. Comment on "Heat capacity, time constant, and sensitivity of Earth's climate system" by S. E. Schwartz. Journal of Geophysical Research, 113(D15):C05019–6, August 2008.

Nicholas Lewis and Judith A Curry. The implications for climate sensitivity of ar5 forcing and heat uptake estimates. Climate Dynamics, 45 (3-4):1009–1023, 2015.

Colin P. Morice, John J. Kennedy, Nick A. Rayner, and Phil D. Jones. Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. Journal of Geophysical Research: Atmospheres, 117(D8), April 2012. doi:10.1029/2011jd017187. URL https://doi.org/10.1029/2011jd017187.

A. Otto, F. E. L. Otto, O. Boucher, J. Church, G. Hegerl, P. M. Forster, N. P. Gillett, J. Gregory, G. C. Johnson, R. Knutti, N. Lewis, U. Lohmann, J. Marotzke, G. Myhre, D. Shindell, B. Stevens, and M. R. Allen. Energy budget constraints on climate response. Nature Geosci., 6:415–416, 2013. doi:10.1038/ngeo1836.

Stephen Po-Chedley, Cristian Proistosescu, Kyle C Armour, and Benjamin D Santer. Climate constraint reflects forced signal. Nature, 563 (7729):E6, 2018.

Martin Rypdal and Kristoffer Rypdal. Long-memory effects in linear response models of earth's temperature and implications for future global warming. Journal of Climate, 27(14):5240–5258, 2014.

Martin Rypdal, Hege-Beate Fredriksen, Kristoffer Rypdal, and Rebekka J Steene. Emergent constraints on climate sensitivity. Nature, 563 (7729):E4, 2018.

S. E. Schwartz. Heat capacity, time constant, and sensitivity of Earth's climate system. Journal of Geophysical Research, 112(D24): D24S05–12, November 2007.

Karl E Taylor, Ronald J Stouffer, and Gerald A Meehl. An overview of CMIP5 and the experiment design. Bulletin of the American Meteorological Society, 93(4):485–498, 2012.

Mark S Williamson, Peter M Cox, and Femke JMM Nijsse. Theoretical foundations of emergent constraints: relationships between climate sensitivity and global temperature variability in conceptual models. Dynamics and Statistics of the Climate System, 3(1), 2019. doi:10.1093/climsys/dzy006.

Michael Winton, Ken Takahashi, and Isaac M Held. Importance of ocean heat uptake efficacy to transient climate change. Journal of Climate, 23(9):2333–2344, 2010.