We thank both reviewers for their helpful and insightful comments. Below we show how we have amend our manuscript in this revised submission to address the points made, first by Reviewer #1 and then by Reviewer #2.

Review #1

Summary
The authors present an update of the WASP model, using datasets up to the year 2019 of surface temperature, ocean heat content and carbon uptake. They use a time-varying feedback parameter and compare outcomes of climate response and sensitivity on different timescales and using different datasets. They complement this with an analysis of the principal components of their fitted parameters. The model are useful addition to discussions about the information it can be derived from observations and climate sensitivity, and am happy to see an updated version.

We thank Reviewer #1 for their careful reading of the manuscript. We agree with Reviewer#1's finding that the model and method represents a useful addition to the literature, and we are pleased that the reviewer will be happy to see an updated version. Below, we specify how we have updated our revised manuscript to address the points made by the reviewer.

Major points
1. The authors compare without much comment different datasets of global warming and ocean heat uptake. The HadCRUT dataset is incomplete dataset of global temperature, with missing data at the poles, which warm faster than the average. In contrast, Cowtan and Way is an example of a dataset that does have global coverage. I would recommend switching HadCRUT out for another dataset that has taken into account polar warming (for instance NOAAGlobalTemp). Alternatively, wait (one week?) for the new version of HadCRUT, which does account for missing data. Similarly, but probably less important, the authors compared two datasets of ocean heat uptake without comment. According to the IPCC's SROCC report, older estimates of ocean heat uptake have biases that may lead to an underestimate of ocean heat uptake (Bindoff, 2019, p.457). Cheng et al (2017) can be considered superior to the old standard of Levitus (2012).}

We thank the Reviewer for highlighting the importance of the distinction between the different statistical methods are used to generate historical datasets.

Both Reviewer #1 and Reviewer #2 make clear why temperature records with infilling should be preferred over those without infilling. Reviewer #1 also notes that newer estimates of ocean heat uptake (e.g. Cheng et al.) should be preferred over older records with identified biases.

In light of these comments from Reviewer #1 and Reviewer #2, we have highlighted how our findings show that different climate sensitivities arise from these different methods of statistical historical reconstruction. Principally, this revised manuscript has updated to the new HadCRUT5 dataset, which includes statistical infilling, and compares to the HadCRUT5 without statistical infilling dataset. We clearly set out how the HadCRUT5 dataset is our preferred choice for constraining climate sensitivity (Lines 205-210):

"The preferred combination of observational datasets is HadCRUT5 & Cheng et al., as these represent the most up to date methodologies for their respective temperature (Morice et al., 2021) and heat content (Cheng et al., 2017) reconstructions. The other dataset combinations are included to assess the sensitivity of our method to different heat content datasets (HadCRUT5 & NODC) and the sensitivity of our findings to the statistical infilling of missing data (HadCRUT5 (no infill) & Cheng et al.). It is noted that most other temperature datasets now reconstruct similar historic global mean temperature anomalies to HadCRUT5 (e.g. see Morice et al. 2021)."

2. I didn't get an intuitive understanding of how the time varying feedback parameter works. Why is there a difference between equation 4 and 5? It would be nice if some additional details could be included here and a reference to the first paper which you derive this.

The revised manuscript now contains an extensive explanation of the representation of time-varying climate feedbacks in the supplementary material (Supplementary Information Section S3), including a new Supplementary Figure S7 giving an example of climate feedback responses to two idealised step-function scenarios in radiative forcing.

Minor points
Abstract: it might be easier to include the 140 year response time scale, for better comparability with climate models?

We agree that improving comparability with complex climate models will enhance the manuscript, and we have presented 140-year response timescale analysis to be directly comparable with complex model experiments run for this length of time (see revised version Figures 4, 5 and Table 1). The figures quoted in the abstract refer to the 140-year ECS values we calculate to

L61: should multiple be two?

Agreed that greater clarity is required: The general code for the WASP model allows multiple climate feedbacks to act over different response timescales (see Goodwin, 2018 referenced on line 62). Here, we use the Planck feedback (acting over an instantaneous timescale) plus two more feedbacks. Our revised manuscript now states this (Lines 61-64):

"This study considers the instantaneous Planck feedback and two further timescales of climate feedback: a multi-diurnal feedback representing a selection of fast climate processes, such as water vapour and clouds, and a multi-decadal climate feedback representing slower processes, such as the surface warming pattern effect."

L 71: the first word is a typo, right?}

Agreed, this first word is a typo and has been removed.

L 83: halocarbons is not capitalised}

Agreed, halocarbons now a appear uncapitalized throughout (Line 98 of main manuscript, and also see Supplementary Information).

L92: I thought all the data used was after 1850. Why do you need volcanic aerosols before that date?}

Agreed that an explanation helps clarify this. The default setting for WASP model simulations is to start in the year 1765 for RCP scenarios or 1700 for ssp scenarios, with sources of radiative forcing defined with non-zero values from some date after the model start date. Since the temperature in year 1850 (and just afterwards) is affected by the volcanic aerosol (and all other) sources of radiative forcing just prior to 1850, we keep the with default WASP model configuration.

This is now explained (Lines 92-95):

"The WASP model starts simulations at year 1700 by default (e.g. Goodwin, 2018), with different sources of radiative forcing defined from some time after that date. While the observational constraints used in this study start in year 1850, the model state in 1850 is affected by radiative forcing received prior to that date. Therefore, this study imposes radiative forcing on the WASP model prior to 1850."

L111: should the j be an i?}

Agreed, the sentence now states (Line 124-125):

"for each of the *i* sources of radiative forcing".

L118: why not use the default definition of TCR of a 20 year average?

Agreed, we have adopted the default definition of a 20-year average for the TCR when calculating using the WASP simulations in a revised manuscript. (e.g. Lines 130-132):

"Here, the transient climate response, TCR, is calculated as the 20-year average warming centred at the year of CO_2 doubling for a scenario with a 1 per cent per year rise in CO_2 and no other forcing (hereafter: 1pctCO_2 scenario)."

L240. This section or the discussion can do with more context. Why is this interesting?(I think it is, but I needed some brain racking!)

Agreed. The key point are that this type of analysis may ultimately help reduce uncertainty in the ECS and TCR by simplifying what the sources of uncertainty are. Also, conducting this type of statistical analysis on efficient model ensembles may ultimately help analyse complex model ensembles. The additional context is now provided for this section on Lines 272-277:

"The observational records provide constraints on the parameters of the posterior ensembles that manifest not only as posterior distributions for these parameters but also as relationships between them, as well as between model parameters and key model outputs of interest (such as ECS(t)). While the correlation structure of the 25 parameters' joint posterior distribution is generally quite complex, some key structures emerge that indicate how ECS and TCR uncertainties might be reduced. This method of analysing variation, and simplifying the degrees of freedom of variation, in large data-constrained efficient model ensembles may ultimately help explore parameter space in more complex Earth system models."

L344: Figs 2 -> Fig 2

Agreed, this has been changed.

Review #2 This study uses a series of observations and a relatively simply climate model with explicit parameters to try to constrain climate sensitivity (ECS) and transient response (TCR) to CO2 doublings. The model includes feedbacks on two timescales which leads to larger ECS than what would be the case if feedback is assumed constant. Overall, I find the paper is fairly clear and fills a niche in the literature, nevertheless, I did not notice some room for improvements. Therefore I recommend only to accept this study for publications after major revisions have been undertaken.

We thank Reviewer #2 for their careful reading of the manuscript and insightful comments. We are pleased the reviewer finds our manuscript clear and to fill a niche in the literature. Below, we identify how have improved the manuscript during revision in light of the points raised by the reviewer.

Major points

I am worried that the authors are overconfident in the ability to constrain slow feedback based on historical warming. Slow feedbacks are known to evolve continuously from years to centuries (e.g. Rugenstein et al. 2020), but in this study they are limited to acting over timescales of a few decades. It is in this conjunction, where in historical warming happened for the most part over a *period of~50 years (since the 1960-70s), that I am concerned as to whether sufficient signal is available to constrain the slow feedback. At the very least the*

We agree that the method used in the manuscript does not constrain the 'slow feedback' consisting of all feedbacks evolving over timescales from years to centuries. Rather, our manuscript uses the historical record to get a constraint on the multidecadal feedback (timescale 25-40 years). To avoid confusion, we have amended the terminology used in the manuscript: the term that was called lambda_slow in the previous version is, in this revised version, now renamed lambda_multidecadal (e.g. lines 83, 123, 163, 221, …)

Also, when lambda_multidecadal is introduced we now explicitly state that this manuscript does not attempt to constrain feedbacks acting over timescales slower than multidecadal (Lines 86-88):

"Note that slow climate feedbacks with timescales longer than multi-decadal are not explored here, since the historical records of temperature and heat content changes do not extend long enough to offer a reliable constraint on processes acting on such long timescales."

"That said, it will probably attract attention that the authors claim to be able to constrain slow feedbacks as amplifying slow warming. Here, however, the prior assumption appears to by a uniform distribution from -3 to +2 Wm-2K-1, i.e. skewed to negative values, and thus assumed a priori to be amplifying. I would like to have the authors choose a prior that is symmetric about zero for lambda_slow."

Agreed, this revised manuscript now uses a prior for lambda_multidecadal that is centred on zero, and so does not assume a priori that lambda_multidecadal is either amplifying or damping. Our new prior for lambda_multidecadal is a uniform distribution from -3Wm-2K-1 to +3Wm-2K-1, thus with an equal chance of being damping as being amplifying (See Figure 2b and Supplementary Table S1).

Also, the revised manuscript now explicitly states how there are regions of our posterior parameter space with amplifying lambda_multidecadal and there are also regions of our posterior parameter space with damping lambda_multidecadal, so it is clear to the reader that our revised manuscript is consistent with both amplifying or damping multidecadal climate feedbacks (Lines 242-247):

"The posterior distributions for fast and multi-decadal climate feedback strengths are bimodal in the HadCRUT5 & Cheng et al. and HadCRUT5 & NODC ensembles (Fig. b,c, red and grey), corresponding to one observation consistent region with weaker amplifying fast feedback ($\lambda_{fast}^{equil} \sim -0.6$ Wm^{-2}) and strong amplifying multidecadal feedback ($\lambda_{multidecadal}^{equil} \sim -1.7$Wm^{-2}) , and another observation consistent region with very strong amplifying fast feedback ($\lambda_{fast}^{equil} \sim -2.2$Wm^{-2}) and damping multidecadal feedback ($\lambda_{multidecadal}^{equil} \sim +1$Wm^{-2}) (Fig. 2d, shown for the HadCRUT5 & Cheng et al. ensemble), noting that the sign convention used implies amplifying feedback from negative λ."

The difference in slow feedback between the two temperature datasets is interesting. However, the explanation provided that they differ mostly be Cowtan and Way having more warming in the recent years seem insufficient. If one plots the difference over the entire record, and not just since 1960, then you realise that mostly the difference arises around the year 1900, and after 1910 the correction is remarkably stable (attached). It would seem that it should be possible to figure from where in the time series the signal that constrains slow feedback comes from?

Agreed, with the revised datasets (HadCRUT5 replacing HadCRUT4) this finding is removed from the study.

The treatment of constraining data is also troublesome. 1) There is no particular reason to use HadCRUT without infilling. HadCRUT is only available where observations were conducted, and so has a low bias as the unobserved high latitude regions, where there is warming amplification according to climate models, are not included. Cowtan and Way infilled datasets, including that of HadCRUT but also based on other datasets such as COBE. I would suggest referring to them as 'HadCRUT in-filled', rather than 'Cowtan and Way'

Reviewers #1 and #2 have both highlighted valid reasons for preferring particular datasets due to their methodologies. In a revised manuscript we will also highlight these reasons for preferring the estimates of climate sensitivity from infilled records of temperature anomaly.

2) I am not sure why the authors include HadSST3.1 as a separate constraint, this data is already part of HadCRUT.

We now explain the reasons why HadSST4 (in the new revised version) is used as an additional constraint to HadCRUT5 in the Supplementary Information, Supplementary Section S4:

"The WASP model contains one input parameter for the ratio of global sea-surface warming to global mean surface warming at equilibrium (Ratio 1 or r1 in Supplementary Table S1) and another for the ratio of global whole-ocean warming to global sea-surface warming at equilibrium (Ratio 2 or r2 in Supplementary Table S1). It is these input parameters that require the use of a separate observational constraint for sea surface temperatures (HadSST4 in Supplementary Table S2) and an observational constraint for ocean carbon uptake (The Global Carbon Budget in Supplementary Table S2) to be applied. As the r1 parameter is varied between prior ensemble members, simulated global mean surface warming and sea surface temperature warming vary differently (relative to each other) across the ensemble. Therefore, the observational constraints for both global surface temperature and sea surface temperature are required to help constrain the posterior values of r1 within the posterior ensembles. As the r2 parameter is varied between prior ensemble members, the relative ocean uptakes of heat and carbon are varied across the ensemble members. Therefore, observational constraints for both ocean heat and carbon uptake are required to constrain the values of r2 in the posterior ensembles."

3) I am worried about including ocean carbon content as a constraint, atmospheric CO2 is prescribed so all this does is to help constrain the exchange rates which are apparently shared with heat transfer. It is, however, well-known that the physical processes of ocean heat- and carbon uptake are different. I suggest removing this constraint.

We now explain the use of separate ocean heat and carbon uptake constraints to generate the posterior ensembles (Supplementary Section S4):

""The WASP model contains one input parameter for the ratio of global sea-surface warming to global mean surface warming at equilibrium (Ratio 1 or r1 in Supplementary Table S1) and another for the ratio of global whole-ocean warming to global sea-surface warming at equilibrium (Ratio 2 or r2 in Supplementary Table S1). It is these input parameters that require the use of a separate observational constraint for sea surface temperatures (HadSST4 in Supplementary Table S2) and an observational constraint for ocean carbon uptake (The Global Carbon Budget in Supplementary Table S2) to be applied. As the r1 parameter is varied between prior ensemble members, simulated global mean surface warming and sea surface temperature warming vary differently (relative to each other) across the ensemble. Therefore, the observational constraints for both global surface temperature and sea surface temperature are required to help constrain the posterior values of r1 within the posterior ensembles. As the r2 parameter is varied between prior ensemble members, the relative ocean uptakes of heat and carbon are varied across the ensemble members. Therefore, observational constraints for both ocean heat and carbon uptake are required to constrain the values of r2 in the posterior ensembles."

Minor suggestions

29, Please mention here the sign convention. It seems the authors use a positive sign for the Planck feedback, which is a negative stabilising feedback, and negative signs for the positive feedbacks in the climate system (water vapor, surface albedo). Most readers will be confused over this, although I realise many British authors apply this convention.

Agreed that there are two sign conventions in use in the literature for climate feedback. We adopt the sign convention of positive lambda is physically meaningful. We now initially state this sign convention on Line 91:

"The sign convention adopted has positive overall λ_{eff}, such that negative λ_{fast} and $\lambda_{multidecadal}$ are amplifying."

and mention again where it impacts interpreting the results on line 248:

"… noting that the sign convention used implies amplifying feedback from negative λ."

47, Tokarska et al. (2020) only did TCR, not ECS. ECS was constrained based on recent warming by Bengtsson and Schwartz (2013), Jimenez-de-la-Cuesta and Mauritsen (2019) and Nijsse et al. (2020).

We thank the reviewer for these recommendations, we cite Nijsse et al. 2020 in the revised manuscript (Lines 58, 72).

However, the Tokarska, Hegerl, Schurer, Forster and Marvel "Observational constraints on the effective climate sensitivity from the historical period" (2020) study in ERL does indeed constrain ECS, and so we have kept the reference to Tokarska et al. (2020) as a citation for ECS. Perhaps the reviewer was thinking of the Tokarska, Stolpe, et al (2020) study in Science Advances that does only consider TCR. However, it is the Tokarska, Hegerl et al.(2020) study that constrains ECS that is cited in our manuscript (Lines 578-579)

53, perhaps delete 'at any given time or timescale'

Agreed, this is deleted in a revised manuscript.

58, perhaps worthwhile mentioning those studies that are relying on these models, and why the authors of this study believe their method makes avoiding GCMs for estimating time-dependence is possible? See also major points.

Agreed. We provide an example of a study that requires complex model output to explore ECS is Nijsse et al (2020), although any study that uses an emergent constraint on ECS or TCR could be used as an example (Lines 57-59):

"Our estimates of ECS and TCR are independent of simulated warming responses in complex climate models (in contrast to estimates utilising complex model output via emergent constraints, e.g. Nijsse et al., 2020)."

The way our method is able to constrain ECS and TCR utilising time-varying climate feedbacks is now explored in the main text (Lines 61-72) and Supplementary Information (Supplementary Section S3, equations S4-S10; Supplementary Figure S7).

71, 'Quisque' is not a word in my vocabulary. According to wikipedia it is a pre-historic herring."

Thank you, this word was a typo and does not appear in this revised version.

81, perhaps nit-picking, but surface albedo feedback, at least that associated with seaice, is not as fast as water vapor, see for instance Tietsche et al. (2011) that find a 1-2year timescale.

Agreed, it is true that the surface sea-ice albedo component of the fast feedback does strictly have a timescale longer than the residence timescale of water vapour in the atmosphere. We now state that in our revised manuscript (Lines 88-90):

> "Also, the snow and ice albedo feedback has a timescale longer than the atmospheric water vapour residence timescale, but is included in λ_{fast} here as the timescale snow and sea-ice responds significantly faster than multi-decadal timescales."

89-90, It would be useful to display the used forcing in a figure, for example to show priors and posteriors of for example aerosol forcing, equivalent to Figure 2.

Thank you for highlighting this. In our revised manuscript we include a figure showing the prior and posterior distributions of recent radiative forcing (Figure 3), with comparisons to the IPCC estimate and a range of CMIP6 models analysed by Smith et al (2020).

161, However, very strongly cooling aerosols would result in mid-century cooling because of the different evolutions of aerosol and greenhouse gas forcing (e.g. Stevens2015, Bellouin et al. 2019). Supposedly the bayesian method applied automatically filters out these values, which is why I would like to see the posterior distribution of aerosol forcing.

We agree that very strongly cooling aerosols would likely result in mid-century cooling. We also agree that our Bayesian approach filters out combinations of greenhouse and aerosol radiative forcing sensitivities that give rise to historic warming trends that are not consistent with observations. In this revised manuscript we have included the prior and posterior distributions of recent aerosol radiative forcing in a new figure (Figure 3).

"204, Here, I suggest to again remind the reader of the sign convention"

Agreed, although this particular section of the text is removed we do remind the reader of the sign convention (e.g. Line 92 and Line 249_:

> "…noting that the sign convention used implies amplifying feedback from negative λ."

218, why not use a doubling of CO2? This is how ECS is defined.

We agree that a CO2 doubling would also work here. We have chosen to use a 4xCO2 perturbation, in line with one of the standard idealised scenarios for CMIP-class models. Note that WASP does not (yet) have a state-dependence on lambda and so the results of a 2xCO2 experiment would be equivalent (but with a slightly lower signal to noise ratio, where the noise is driven by the imposed interannual variability in Earth energy balance in WASP).

224, by 90 do the authors mean 5-95?

Agreed that this was unclear, we now specify the percentile intervals as well as the ranges (e.g. Line 263:

> "… varying from 2.1 °C (1.6 to 2.5 °C at 90% range from 5th to 95th percentiles)"

248-250, or perhaps a better constraint on total lambda, say based on paleoclimates?

Agreed – but this section has been re-written due to the differing results now that we have updated to HadCRUT5 temperature reconstructions.

270, this section added no new information that had not already been provided. I suggest removing it.
277, yet these components only explain 1/3 of the total variance?

We agree our manuscript that lacked clarity on the benefits and motivation behind the principle component and stepwise regression sections (4.2.2 and 4.2.3 respectively). We now explain why the PC analysis and stepwise regression is conducted (Lines 274-279):

"The observational records provide constraints on the parameters of the posterior ensembles that manifest not only as posterior distributions for these parameters but also as relationships between them, as well as between model parameters and key model outputs of interest (such as ECS(t)). While the correlation structure of the 25 parameters' joint posterior distribution is generally quite complex, some key structures emerge that indicate how ECS and TCR uncertainties might be reduced. This method of analysing variation, and simplifying the degrees of freedom of variation, in large data-constrained efficient model ensembles may ultimately help explore parameter space in more complex Earth system models."

With the updated results (moving to HadCRUT5 as the temperature constraint), the first 5 Principle Components now explain 60% of the variance in the posterior model dataset.

353, this statement requires there are no slow feedbacks acting on timescales from decades to millennia. I recommend to remove this statement, or strongly caveat.

Agreed, we now state how slow feedbacks acting on many timescales from multi-decadal to millennia may affect how our estimate comnpares to those from palaeoclimate studies (Lines 376-379):

"…but note that additional slow feedbacks not considered here, acting from many decades to millennia, may affect how our estimates are comparable to estimates of climate sensitivity from the palaeo-record where any longer, (e.g. Rohling et al., 2012; 2018)."

"368, not 'multiple' but 'two distinct' timescales."

Agreed we now state that we consider the instantaneous Planck feedback and two further timescales of feedback (now Lines 62-65):

"This study considers the instantaneous Planck feedback and two further timescales of climate feedback: a multi-diurnal feedback representing a selection of fast climate processes, such as water vapour and clouds, and a multi-decadal climate feedback representing slower processes, such as the surface warming pattern effect."

"370, IPCC 'likely' means 66 percent probability or better."

Agreed, this is clarified (Lines 394-395):

"…IPCC ECS likely (66% chance or better) …"

371, for Sherwood et al. (2020) probably the number referred to is 17-83 percent.

Agreed, this is clarified in this revised manuscript (Lines 392-393):

> "…the recent Sherwood et al. (2020) Bayesian review has a narrower baseline 17^{th} -83^{rd} percentile (66%) range of 2.6 to 3.6 K."