We appreciate the useful comments by both reviewers on the manuscript *"A lower and more constrained estimate of climate sensitivity using updated observations and detailed radiative forcing time series"*. Below follows our responses to the comments by the reviewers, as well as descriptions of how the manuscript has been modified. The original reviewer's comments are in italics.

Anonymous Referee #1

General Comments

Using an energy balance model, the authors estimate the equilibrium climate sensitivity (ECS) using updated observations of global temperature and ocean heat uptake change. The method is based – and extended here – on previous papers of Aldrin et al. 2012 and Skeie and co-workers. As the authors note, there are five main differences between this study and Aldrin et al. 2012 (end of section 2.1.). As main result of their study, the authors find a lower and more constrained estimate of ECS in their model. A broad background into the general topics of the paper (constraining climate sensitivity, observations of ocean heat uptake) are given. However, it is suggested to focus on the new additions presented here (for example their treatment of long-term internal variability).

Their results and approach is interesting and their lower values of ECS is in line with results from previous study (e.g. Otto et al. 2013; Lewis et al. 2013). Since the method and approach of using reduced complexity models to estimate ECS has been extensively used over the last ten years, it is suggested that the authors strengthen the focus of the paper on their results and not so much on general background, and in particular on the physical interpretation of their lower climate sensitivity estimate. The main results of their study are presented in the paragraph starting "The drastic reduction in uncertainty [...] with ten more years of data may be surprising. We believe there are two main reasons for this". It is suggested that this important finding is discussed in more detail. In particular, the limitations of the climate model in reproducing the observed record and its relation to low climate sensitivity values should be discussed. In that sense, it is also suggested that already the Abstract highlights that this is a model based result and the limitations and advantages of both the climate model and the method employed here could be more highlighted.

Response: The reviewer suggest to focus more on the new addition in this paper compared to Aldrin et al. (2012). To highlight the difference between this paper and Aldrin et al. (2012), we have moved the text on p 791 line 3, to the beginning of the Method section. We have also throughout the text focused more on the results and physical interpretation of the results and less on the general background, as commented elsewhere in this response to the reviewers.

The limitations of the climate model in reproducing the observed record is discussed and clarified. This is partly related to the long term internal variability, which have got more focus in the revised version of the manuscript. These issues are also raised in other review comments and answered in details there (e.g. see response to point 3 of referee #2).

The reviewers suggest that "the two main reasons […]" is discussed in more detail. In the section "Uncertainties in Radiative Forcing" we have tried to focus more on the RF development last 10 years (see also response on the comment by Referee #2 p796 top). In section "The role of ocean heat content" we have rewritten the first paragraph:

"Including observations from the last decade had a large effect on the ECS estimate. The OHC time series is extended by about 20% and previous studies (Tomassini et al., 2007; Urban and Keller, 2009; Aldrin et al., 2012) have shown that OHC data have the potential to constrain the ECS estimate. In Appendix E we show that when adding only near surface temperatures between 2000 and 2010 (excluding the OHC data for this decade, Fig. E1g) the ECS estimate is only slightly narrower than using data up to 2000 (Fig. E1c), highlighting the information provided by the OHC data. "

We have included the following text to highlight advantages and limitations of the method at the end of the abstract:

"If we do not explicitly account for long-term internal variability, the 90% C.I. is 40% narrower than in the main analysis and the mean ECS becomes slightly lower, which demonstrates that the uncertainty in ECS may be severely underestimated if the method is too simple. In addition to the uncertainties represented through the estimated probability density functions, there may be uncertainties due to limitations in the treatment of the temporal development in RF and structural uncertainties in the EBM."

In that regard, the structure of the Results section could be particularly strengthened. In its current state, the main analysis presents the results and their discussion at the same time. For example, section 3.1. presents the estimates of ECS, but also includes a discussion of their relation to previous results and to the likely range presented in the IPCC AR4. It is suggested that the Results section presents the results in a short, concise way and the discussion of the results is shifted to the Discussion section.

Response: We agree that the structure of the results section could be improved. We have moved the text on p794 line 1 to 12 to the discussion to a new subsection "*Inter hemispheric differences*". The discussion regarding mid-century warming and the sensitivity test using HadCRUT4 are moved to a new subsection in the discussion "*Surface temperature observations*". We have also rewritten the paragraph regarding the fitted temperature and OHC (Fig. 3). We think these changes significantly improve the readability of the section.

We would like to still compare the main results to other studies and the IPCC range in the results section, to put it in perspective. We have rewritten the text where our estimated ECS are compared to other studies. We have included the AR5 range and reference to the other recent studies:

"The ECS posteriori mean is 1.8˚C (Fig. 2a) which is below the lower limit of the likely range (>66% probability) for the ECS of 2 to 4.5˚C in IPCC AR4 (Meehl et al., 2007), but within the AR5 likely range of 1.5 to 4.5 ˚C (Collins et al., 2013). The 90% C.I. of the posterior ECS is 0.9 to 3.2˚C, and the heavy upper tail often seen in estimates of ECS is less pronounced. The probability of ECS being larger than the upper limit of the IPCC likely range of 4.5˚C is 0.014. Tomassini et al. (2007), who used observational data up to 2003, found a probability of 0.16 for the ECS > 4.5˚C. We have used observational data up through 2010, seven more years than Tomassini et al. (2007). Recently, there are studies including observations for the last decade (Lewis, 2013; Otto et al., 2013) and using data from the last glacial maximum (Schmittner et al., 2011) that also find climate sensitivity in the lower range of IPCC."

The discussion could also be shortened and more focused. For example, it gives a broad overview of ocean heat content and for example instrumental issues associated with measuring ocean heat uptake (section 4.2). This is clearly important and has to be mentioned, but the focus of the paper is in using these observations as constraints for the model.

Response: We have rewritten the OHC discussion and focused more on the results, e.g. linked the issues regarding measurements to our results and the observational errors. The comparison between Levitus, CISRO data and Lyman is removed, due to the suggestion by the referee to focus more on our results. The Lyman data has also been updated with data and instrumental bias correction (Johnson et al., 2013) and shown to be closer to the Levitus and the CISRO estimates.

“The fitted OHC from the model is compared to the three historical estimates in Fig. 3. The fitted OHC is a smooth curve compared to the observations but with dips related to volcanoes as is also seen in the observations, and two of the three OHC data sets used in this study show a flattening of the OHC since 2004. There is good agreement with the long-term trend in OHC between the model and the observations. Except for responses to volcanic eruptions there is low correlation between the shorter term variability in the 3 observational datasets (as opposed to the near surface temperature data). As noted in section 3.1 one of the observation curves lies outside the 90% C.I. in the 1950s. The reported and estimated standard errors for this dataset increase back in time (Figs. B3 and B4) and in the 1950s the standard errors are larger than the difference between our estimate and the data. The further back in time the poorer is the spatial coverage of the observations (e.g. Fig. 1 in Abraham et al., 2013). In early 2000s the Argo floats were launched (http://www.argo.ucsd.edu/), which significantly improved the spatial coverage of ocean observations. Prior to the launch of Argo the main data used were collected from expendable bathythermographs (XBT) which have systematic data errors (Gouretski and Koltermann, 2007), but also the Argo data have known biases that needs to be corrected (Abraham et al., 2013). Lyman et al. (2010) found that XBT bias correction was the main source of uncertainty in the warming trend from 1993 to 2008. The differences among the three dataset are larger than between the surface temperature data series, and a further effort in estimation of historical OHC data and its uncertainty is needed. (Fig. 3d)”

The discussion regarding energy budged is shortened: “Since added energy to the climate system is almost exclusively stored as heat in the oceans, a non-zero global radiative imbalance is approximately equal to the rate of change in OHC. The inferred planetary energy imbalance from Hansen et al. (2011) was 0.58 ± 0.15 W m^{-2} during the 6-year period 2005–2010 assuming a stronger aerosol RF (−1.6 ± 0.3 W m^{-2}) and a larger climate sensitivity (3 ± 1.0˚C) than the posterior means in this study. We find a similar planetary imbalance of 0.46 ± 0.16 W m^{-2} over the same time period.”

Some of the text in the Result section is moved to the discussion, and new subsections are introduced.

The discussion section is now separated in 7 subsections:

1. Uncertainties in Radiative Forcing
2. Surface temperature observations (new, text moved from result section)
3. The role of ocean heat content
4. Multidecadal oscillations

5. Interhemispheric differences (new, text from the removed subsection “Hemispheric difference in feedbacks” and from the result section)
6. Comparisons with results from a similar approach

We think these changes have made the discussion more focused.

Also in that regard, it is suggested that the general use of references is reconsidered in the sense that mentioning of previous work and results of the references is more focused on the context and content of this paper. For example, the authors often use the expression: Author et al. (xxxx) investigated / Author et al. found. Albeit important, distilling the essence of the references and how they are specifically related to the paper might enhance the readability of the paper. The references are very up to date and their extent is impressive.

Response: We have taken this comment into consideration when revising the paper.

The authors present a very interesting sensitivity study of their results. The results of the sensitivity study could be presented in special section dedicated to this and not as currently combined with the results of the main analysis. This would make it easier for the reader to understand the results of the main analysis.

Response: We agree, and as already noted, we have moved the sensitivity test were we have allowed for hemispheric differences in the climate sensitivity to a new section in the discussion: “*Inter hemispheric differences*“. In this section we discuss hemispheric differences in warming related to aerosol forcing, differences in climate feedbacks, differences in response time and differences in internal variability.

Also the text regarding the HadCRUT4 sensitivity test is moved to a separate section in the discussion as noted above. Only results from the main analysis are now presented in the Results section.

Specific Comments

The authors use a uniform prior for ECS – is there a possibility to test the results also for a non-uniform prior to test their sensitivity to the choice of ECS prior?

Response: We have tested this using the following two non-uniform priors for ECS. This has been done by simply re-weighting the MCMC samples from the estimation of our main model. This can be seen as a variant of importance sampling. The two alternative priors are

- Hegerl’s prior for ECS. Hegerl et al. (2006) calculated a PDF for the climate sensitivity that was a combination of PDFs from several authors, all on the basis of reconstructed temperature data before 1850. Since this PDF is based on other data than we use in our work, it can be reasonable to use this PDF as an informative prior. The median of this prior is around 3.5 with a 90 % C.I from 1.2 K to 8.6 K.

- Uniform prior for 1/ECS, which is equivalent to a prior for ECS that is proportional to $1/ECS^2$. This prior was discussed in Frame et al. (2005). As we stated in our previous paper (Aldrin et al., 2012), “this prior is strongly informative towards low climate sensitivities with 76 % probability for ECS being lower than the pure black-body radiation of 1.1 K” , and it is perhaps not very realistic.

In the table below we show estimates for ECS using the two non-uniform priors, in addition to the uniform prior that was used in the main analysis in the paper.

	Mean	Median	90% C.I.	95% C.I.
Uniform prior for ECS	1.84	1.67	(0.92, 3.18)	(0.78, 3.82)
Hegerl's prior for ECS	1.91	1.80	(1.10, 3.11)	(0.99, 3.56)
Uniform prior for 1/ECS	1.31	1.27	(0.46, 2.27)	(0.33, 2.55)

We observe that the estimates when using the Hegerl's prior are slightly larger than those obtained with a uniform prior, while the credible intervals are slightly narrower. When using a uniform prior for 1/ECS, the PDF is shifted considerable towards lower values.

We have included these results in the Appendix E6 The role of ECS prior.

An important and interesting approach is the use of multiple observational timeseries simultaneously (putting them into a vector). The observations are strongly correlated - is this correlation across the observations taken into account in the parameter estimation process (in a co-variance matrix or similar)? If not, it would be like having three independent observational constraints (both for temperature and ocean heat uptake), which would put a too strong constraint on the model parameters. A short clarification would be appreciated.

Response: We account for correlations between observations, and this is now better explained in Section 2.1.

The authors employ anthropogenic RF series of Skeie et al. (2011). How are they related to the data for the historical Representative Pathway Concentration (RCP) data that other models use? Maybe a figure (in the Supplementary) comparing the two could help the reader to see the size and evolution of the two forcing datasets.

Response: The emission inventory used in the study by Skeie et al. (2011) is from Lamarque et al. (2010). That is the same emission data as used in the CMIP5, however different models calculate different forcing, especially for the cloud indirect effects.

As suggested we have included the figure below in Appendix D: "Prior and posterior distributions for RF time series". This figure shows the prior and posterior for the anthropogenic radiative forcing time series. The dashed line is the historical radiative forcing time series from the RCP database (http://www.iiasa.ac.at/web-apps/tnt/RcpDb). The RCP4.5 value is used for year 2010 as in Skeie et al. (2011). The RCP RF time series do not include the effect of land albedo changes, as is included in our prior. In our prior we also include indirect and semi-direct aerosol effects that are not RF according to the definition in AR4, and are probably not included in the RCP historical RF time series.

The three error bars in the figure are from IPCC AR5 summary for policymakers. These error bars include the total aerosol effects and are comparable to our forcing time series. Our prior has a weaker forcing (stronger aerosol forcing) than IPCC AR5, but our posterior is more in agreement with the AR5. Our posterior have larger uncertainty in the 1950s, lower mean value in 1980 and do not span the upper range in 2010 compared to AR5 (AR5 values are 2011).

This figure and the comparison with the AR5 values are also discussed in section 4.1 “Uncertainties in radiative forcing”:

“Limited information regarding the uncertainty in rate of change of RF is available. In Fig. D2 the prior and posterior anthropogenic RF time series is plotted together with RCP historical RF time series and AR5 forcing estimates with uncertainties (IPCC, 2013) for the years 1950, 1980 and 2011 (See Appendix). Our posterior is in good agreement with the AR5 values, however we have larger uncertainty in 1950s, lower mean value in the 1980s and we do not include the upper range of IPCC AR5 for the year 2011 estimate. This indicates that the low estimate for the ECS in our main results is not due to unreasonable high RF estimates, however large changes to the historical RF path (e.g. due to indirect aerosol effects) may change the ECS estimate. “

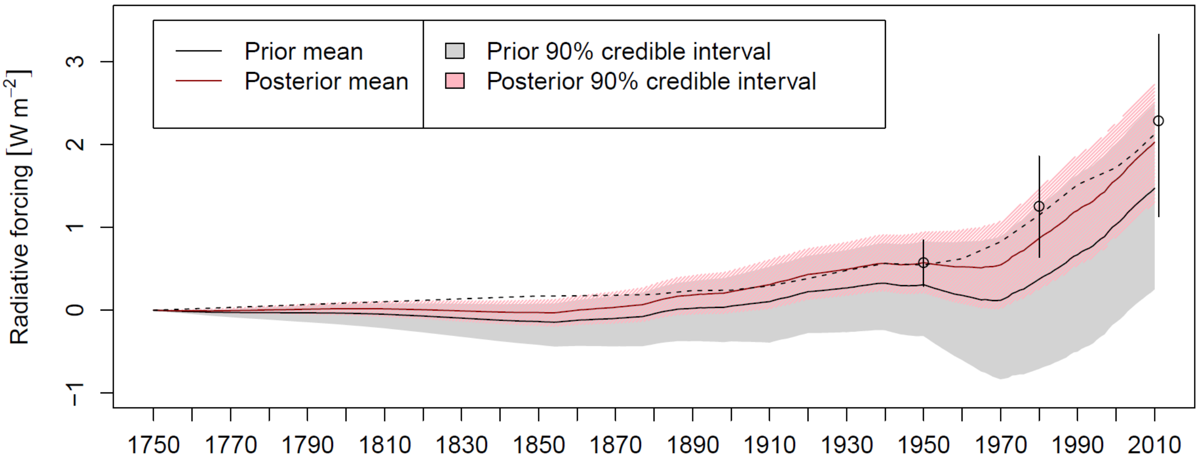

Figure R1: Prior and posterior distribution of the RF time series and PDF of RF in 2010 for anthropogenic RF from the main analysis. Red color for the posterior distributions and black lines and gray shadings for the prior distribution. The dashed line is the historical RCP total RF (http://tntcat.iiasa.ac.at:8787/RcpDb/) and the three error bars is the total anthropogenic RF in 1950, 1980 and 2011 from IPCC AR5. The error bars is the very likely range (>90%).

Regarding the standard deviations of all stochastic terms: the authors note that in constrast to other studies, they are estimated here from the data. Is there a chance of compensation between different uncertainties?

Response: Yes, the different uncertainty terms may be mixed, and we have included a modifying sentence about this at the end of Section 2.1:

“However, even we think it is conceptually useful to divide the errors into separate terms, each with a distinct interpretation, it is of course a possibility that the estimated error terms may be mixed, so one should perhaps be careful with a too strict interpretation of each term.”

However, the model formulations are made such that the decomposition of uncertainties into four terms should be possible: 1) The correlation structure (in time and between components) of the term representing long-term internal variability (and possibly other slowly varying model errors) has been estimated from external sources (runs of GCMs). 2) The short-term internal variation is directly coupled to the ENSO index. 3) Since there is three observations for each physical component, the system learn in fact a lot of the structure of the observational errors. Assume we have two observational series y_{1t} and y_{2t} for the same truth g_t. When taking the difference y_{1t}-y_{2t}, the unknown truth disappears, and the differenced series may give some information on both their standard deviations and the auto correlation of the observational errors. 4) We use also the reported temporal profiles of the standard deviations reported by the data providers. 5) The term for short-term model error ($\mathbf{n}_t^m$) is common for all measurements of the same physical components, which helps in separating it from the observational errors.

Regarding the comparison between the observed temperature and simulated temperatures with the posterior model parameters: the authors mention in section 3.1. that observations are outside the 90% C.I. of the fitted temperature increase in the Northern Hemisphere over the last two decades. The authors also note that this uncertainty range includes only deterministic terms. This could be a very important point in the physical interpretation of the results. How can the climate model used in the study reproduce the last decade of only very small temperature increase while there is a positive forcing?

Response: The fitted values shown in Fig. 3 are based on the deterministic terms of the model

$$\boldsymbol{m}_t(\boldsymbol{x}_{1750:t}, ECS, \boldsymbol{\theta}) + \boldsymbol{\beta}_1 e_t$$

i.e. it is not only driven by the forcings, but also by the variation in El Nino/La Nina, through the Southern Oscillation Index (e_t). The model's ability to give only a very small warming over the last decade is a combination of natural forcings contributing to a reduction in the rate of increase in the forcing (figure 1, top panel) and a mostly positive SOI index since 2000. The parameter β_1 is negative, so higher values of SOI slow down the warming. Figure R2 below shows the SOI index (http://www.bom.gov.au/climate/current/soihtm1.shtml). After about 2005 the SOI has been mainly in a positive phase contributing to a decreased rate of warming (cf. Fig. 4).

Regarding using this point to derive a better physical interpretation of the results, we have been thinking along these lines ourselves. Two of the sensitivity tests were designed to see if specific aspects of the simple EBM were the cause of the underestimation of the NH warming heating during the last decades, and if this had a significant impact on the ECS estimate. We first allowed for different feedback strength in each hemisphere, and then in a second experiment different depths of the mixed layer in the ocean were allowed for each hemisphere. Either of these sensitivity test changed the interhemispheric temperature difference or the ECS estimate significantly. Cf. Revised section (4.5) on Interhemispheric differences in the revised manuscript, and reply to referee #2 (comment to page 794, line 3).

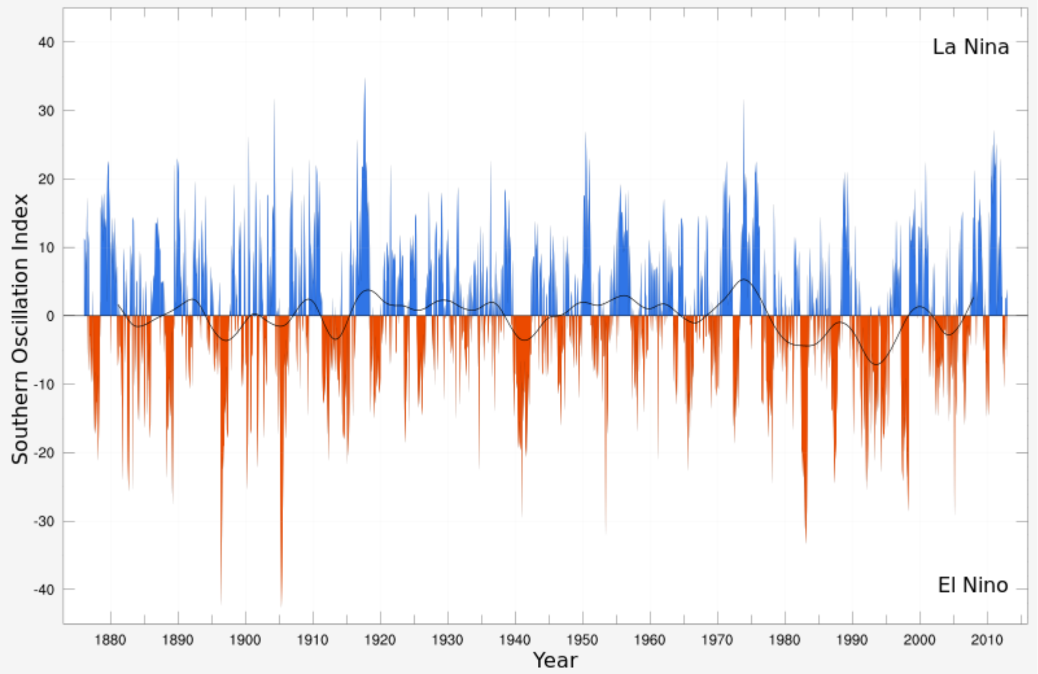

Figure R2: The SOI index.

Is there a chance that the low climate sensitivity value is a statistical effect - in the sense that if there is a positive forcing, the climate model responds by tuning down its climate sensitivity. A discussion of this would be very helpful and appreciated.

Response: Well, yes. This is a statistical method and indeed that climate sensitivity is estimated to give the best fit between the observed changes in atmospheric temperatures and OHC with the estimated radiative forcing. However, there are physical constraints to the relation between the air temperatures, the OHC and the radiative forcing, i.e. that there must be an energy balance globally. That is that the forcing must be balanced by sum of heat stored in the oceans and the increased outgoing radiation. The latter is defined as the product $\lambda \cdot \Delta T_a$, where λ is the climate sensitivity parameter ($Wm^{-2}K^{-1}$). The reviewer is correct in that if e.g. the negative radiative forcing from aerosols was better constrained at a lower (more negative) value (i.e. that the scientific knowledge about this forcing would allowed a more narrow a priori estimate of this forcing), our method would almost certainly have given a lower net forcing and a higher estimate of the climate sensitivity. However, this is the main purpose of using this method (i.e. taking into account the uncertainties through the a priori estimates) and should not be considered as a statistical artefact of the method. Indeed the comparison with our net posterior RF with the RCP RF histories applied in the IPCC AR5 report (cf. Fig. R1), does not indicate that the net forcing is too positive.

The introduction of the long-term variability term is very interesting. A comparison between its posterior estimates and the model error term in Fig. 4. suggests that the model error is largely dominated / represented by this long-term variability. I would just have a little comment regarding its interpretation: there could be a similarity between structural uncertainty of the climate model used here in reproducing the observed record and the notion of unforced internal climate variability (and Fig. 4 also shows that they are correlated). In terms of attribution, the authors note that "during the *period 1910-1940 and 1970-2000 a warming of about 0.2 K can be attributed to internal variability". Or could it be that this warming could be due to climate processes not represented in the climate model ? A short comment regarding the interpretation of the internal variability term and structural model limitations in an attribution framework would be much appreciated.*

Response: We have modified the interpretation of the term for long-term internal variability, and state now in the text that it also may represent potential other slowly varying model errors.

In the method section we now write: "The long-term internal variability is represented by the term $\boldsymbol{n}_t^{liv}$, which also accounts for potential other slowly varying model errors."

In the discussion section, we have discussed the possibility that AMO is forced, a question asked by referee #2. In the new section "Interhemispheric differences" we have discussed structural uncertainties in the climate model.

To highlight the uncertainties we now write in the beginning of the discussion section: "In addition to inclusion of uncertainties in parameters of the deterministic model, and in the observations used that together is propagated to give the pdf of the climate sensitivity, there might be other limitations in the method and sources of uncertainties that is not quantified in the estimated pdfs. This includes e.g. uncertainties due to the simplified structure of the EBM or the a priori estimates."

Technical Comments

Both the size and font of the Figures could be enhanced to improve readability.

Response: We have improved the figures.

Anonymous Referee #2

This ms estimates the equilibrium climate sensitivity, using a method similar to one published in a precursor paper, Aldrin et al., but applying updated estimates of radiative forcing, and an improved estimate of internal climate variability as well as a combination of data for ocean heat uptake. The ms finds a quite narrow estimate of ECS and TCR, and explores the origin of this tighter constraint. The work is interesting and important, although it does have some weaknesses, some of these need to be at least discussed more clearly in the ms:

1) *the statistical method description is very terse - for readers familiar with Aldrin et al it is clear, but it would be better to explain a bit more clearly how the pdfs are derived.*

Response: The description of the statistical method in Section 2 is extended, also to be able to clarify the worries in point 2) below. We have also in that regard shortened the description of the statistical method in the Appendix.

Also, can you remind a reader: is the Aldrin method tested using GCM historical runs with known sensitivity? that would be a powerful evaluation of the method.

Response: Such tests were performed previously with satisfactory results which were presented in Aldrin et al. (2012), and we now refer to these tests at the end of Section 2.1:

"Our model is of course an extreme simplification of the real climate system. Therefore, to investigate if the model is useful for estimating the ECS from observations, we have previously validated its performance on artificial data generated from GCMs in the CMIP3 experiment. The estimates of ECS were, in light of their corresponding uncertainties, comparable with the "true" values of ECS for two different GCMs (Aldrin et al., 2012)."

2) *my major worry is the use of 3 ocean heat uptake datasets within the observational vector. I am not quite sure what this means statistically (doesnt this weight the ocean data more?) and also, I worry that this is interpreted as the three datasets providing three independent samples of the true values - in reality, the three datasets share similar data, and sample the uncertainty in processing only - does the result really reflect this? However, the difference of the 3 OHC data approach to the 1 OHC data approach seems reasonably small, but a clearer explanation as to why the combination of OHC data would increase confidence should be given.*

Response: In Section 2, we now give a better motivation for using several observational data series for each physical component (OHC and temperatures on the two hemispheres), not only for OHC. The main reason is that this reduces the influence of observational errors (i.e. the combined sampling and analysis errors). We argue that this doesn't necessarily give observations on one physical component more influence than observations on another component, since we take into account the measurement errors of these correlations and also correlations between them. However, since the discrepancies between the OHC series are larger than between the temperature series, indicating larger observational errors for OHC, we believe that it is more important to use several data series for OHC than for the temperatures. This is pointed out in subsection 4.3 in the Discussion. In fact, if the number of observations was the main factor for the influence of the OHC data vs. the temperature data, the temperature data would be much more important than the OHC data, since there are six of them per year (compared to three for OHC) and they are observed for about 150 years (compared to about 60 years for OHC).

The problem is illustrated some by figure 3, where the posterior interval does not include the larger trends in one of the datasets (unless I am missing something) - this strong conclusion, which rejects one of the datasets as far outside the posterior uncertainty, is hard to justify just based on statistics only - or am I missing something?

Response: Yes, the CSIRO data series fall outside the OHC posterior the first 7-8 years, but note that the standard errors for the observations reported by CSIRO for these years (Figs. B3 and B4) are larger than the difference between our estimate and their data. We think in fact this is reasonable, see our answer below.

3) *the data model comparison figures in the paper suggest that the model with best fit underestimates the recent warming (in other words, attributes some of it to internal variability), for all temperature datasets for the recent time, and for the change in OHC over*

time in at least one dataset (figure 3 shows undersimulations). this should be clarified and the uncertainty this highlights should be discussed.

Response: Yes, we underestimate the recent temperature warming on the northern hemisphere and therefore also globally, but less, and this is in our model mainly attributed to the term that accounts for internal variability (and other slowly varying model errors). The estimated internal variability term has decreased over the last 10-12 years (Fig. 4) before 2010. The temperatures for 2011 and 2012 (that we have not included in our analysis) are around the average for the period 1998-2010, which strengthens (but of course not proves) the hypothesis that we are in a period with a downward trend in the internal variability.

Regarding OHC: Since the three OHC series are so different, it is of course not possible to fit all three well by one common curve. It is correct that our model fit does not show the same increase as the CSIRO series from 1950 to 1958 (Fig. 3). This is reasonable, since both the reported (Fig. B3) and the posterior estimate of the measurement standard deviations (Fig. B4) are very high in this period, and the Ishii and Kimoto series is decreasing in this period (Figure 3) with much lower standard deviations, whereas the Levitus series is only observed from 1955. If we compare the model fit with the three series of observations from 1958 to 2010, we see that the increase in the model fit curve is almost exactly the same as in the Levitus series, around 3 x 10^{22} J lower than in the CSIRO series and around 3 x 10^{22} J higher than in the Ishii and Kimoto series.

We have rewritten the text in the result section regarding the data model comparison to clarify these issues:

"The fitted posterior mean and the observed hemispheric temperatures and OHC are compared in Fig. 3. For the SH the model is reproducing the long-term trend of the observations. In the NH, the fitted temperature increase over the last two decades is not as rapid as in the observations, leaving the observations just outside to 90% C.I. of the fitted temperatures. The fitted temperature and OHC include only the results from the deterministic model ($\boldsymbol{m}_t(\boldsymbol{x}_{1750:t}, ECS, \boldsymbol{\theta})$) and the effect of ENSO ($\boldsymbol{\beta}_1 e_t$) at the right side of Eq. (1). In the NH, much of the discrepancies are accounted for by the long-term internal variability represented by the term $\boldsymbol{n}_t^{liv}$ (Fig. 4 left panel). This is further discussed in section 4.4. Figure 4 also shows the posterior estimates for the ENSO term $\boldsymbol{\beta}_1 e_t$ and the model error $\boldsymbol{n}_t^m$.

For the OHC the model is reproducing the long-term trend of the observations (Fig. 3). In the 1950s one of the observational time series is outside the 90% C.I. of the fitted OHC. The CSIRO group report a large standard error of up to 10 x 10^22 J in this period (Fig. B3), and the posterior estimates of the standard errors are of the same magnitude (Fig. B4), so the observational error term ($\boldsymbol{n}_t^o$) will explain this discrepancies. This is further discussed in section 4.3."

Other models, eg CMIP5 models on average reproduce the warming better over much of the period, although less recently

This claim is only partially correct. Figure 10.7 in the final draft of IPCC AR5 (published online 30.September 2013) shows the yearly averages of CMIP3 and CMIP5 model simulations from 1860 to 2012, together with a global HadCRUT4 series for near-surface temperatures. This is an updated version of figure 4 in an article by Jones et al. (2013). The global HadCRUT4 series used here has been calculated by the authors based on regional HadCRUT4 values. (This is similar to, but not exactly the same, as the global HadCRUT4 series (= mean over the two hemispheres) that is published on the Met Office Hadobs website which is the one we have used in our work.) We have received the data used in AR5 by Gareth Jones, the first author of the article.

We have investigated how good our model and the CMIP3/CMIP5 average fit the global HadCRUT4 series shown in Figure 10.7 in AR5 by calculating the root mean squared errors (RMSE) over the common period 1860-2010. The RMSE of the CMIP3/CMIP5 average is 0.115 ˚C, whereas it is 0.122 ˚C for our main model estimate which is based on HadCRUT3 (together with two other temperature series and three OHC series) and 0.120 ˚C for our alternative model estimate which is based on HadCRUT4. So, yes, the CMIP3/CMIP5 average gives a better fit to the HadCRUT4 temperature observations, but the improvement is only between 4 % and 6 %.

However, is it true that the CMIP3/CMIP5 average reproduce the warming better? Our model fit shows less warming than the observed temperatures the last fifty years, but on the other hand, the CMIP3/CMIP5 average shows more warming than the observations (Figure 10.7 in AR5). To investigate this more formally, we have calculated the temperature increase over the fifty last years in our data period (1961-2010) for the models and the observations by a linear regression with yearly temperatures as response and time (year) as predictor. Our model fits based on HadCRUT3 and HadCRUT4 have decadal temperature increases of 0.108 ˚C and 0.111 ˚C in this period; whereas the decadal increase for CMIP3/CMIP5 average is 0.176 ˚C. The global HadCRUT4 series used in Figure 10.7 has a decadal increase of 0.143 ˚C in the same period, whereas that HadCRUT4 series we used (downloaded January 2013) has an increase 0.138 ˚C. Based on this analysis, we disagree in the statement that the CMIP3/CMIP5 average reproduces the warming better than our model.

- it is intuitively not clear why this paper suggests a better fit with small sensitivity even if this doesnt fully resolve some of the observed trends. I wonder if this finding may be sensitive to model error, such as in the ratio of SH to NH aerosol response, or response time or shape of forcing. To my mind, this suggests a further role of structural model uncertainty in the result - if you agree, it would be good to mention this further uncertainty which would probably widen your pdf some more.

Response: We agree that it is intuitively not easy to understand and decompose the role of the different factors that contribute to the overall results in our study. Some of the factors can be analyzed through sensitivity studies (cf. specific replies below), while some of the factors and their potential contributions to additional uncertainty only can be discussed qualitatively in the discussion section at this point. For the future, it should be addressed through controlled multi-model studies where structural issues (mainly the choice of the simple climate model applied) were analyzed in a controlled way.

SH/NH aerosol response: This is discussed in the new section in the discussion “Interhemispheric differences”. See also specific reply to comment to p.794, l 3: SH/NH aerosol response below.

Response time: The response time of the system is mainly determined by the depth of the mixed layer of the ocean in the EBM. This is one of the parameters in the EBM (in the vector, containing the physical parameters of the EBM) that is estimated using the MCMC technique. It is interesting to note that the estimate for the mixed layer depth decreases gradually (from 80 m to 58 m) as more observational data is used to constrain the model (Table A3). We have done one sensitivity test allowing for different mixed layer depths in the two hemispheres. This had very little impact on the ECS estimate (cf. Fig. E2).

Shape of forcing (geographical and historical): The shape and magnitude of the forcing have been compared to the estimate in AR5 (cf. specific reply to one reviewer comment). Uncertainty related to the simplified treatment of indirect aerosol forcing which affect both the geographical (NH/SH) and the historical shape of this forcing, have been addressed. However, as for the structural uncertainty discussed below, it is difficult to assess how uncertainty in the shape of the forcing would affect our results. See also specific reply to comment by referee #1 regarding forcing in the RCP timeline.

Structural model uncertainty: We agree that this is an important point, and we have to some extent tested this through the two sensitivity experiments (estimating separate ECSs and mixed layer depth in each hemisphere). However, there are many more aspects relating to the structure of the model that would have been interesting to test through a multi-model comparison experiment. However, this is beyond the scope of the current study, but we have noted this in the discussion.

In the discussion we conclude that: "Overall it is difficult to determine which factors are responsible for the discrepancy between the observed and fitted ITA, and thus it is also difficult to assess the impact of this short-coming on the estimated ECS."

5. Lastly, now that the AR5 is out, it would be good to crosslink to it rather than AR4.

Response: We have added some references to the relevant chapter of AR5 where appropriate.

Minor comments: 1. it took me a while to understand that the last sentence in the abstract refers to the main findings range, not a widening of it - maybe clarify.

Response: This is rewritten in the abstract:

"The analyses show a significant contribution of internal variability on a multi-decadal scale to the global mean temperature change. If we do not explicitly account for long-term internal variability, the 90% C.I. is 40% narrower than in the main analysis and the mean ECS becomes slightly lower, which demonstrates that the uncertainty in ECS may be severely underestimated if the method is too simple. "

l 14, p 787: the references to LGM estimates are a bit aged, maybe refer to some newer ones (Schmittner, Hargreaves).

Response: We will include these two references: Schmittner et al. (2011) and Hargreaves et al. (2012).

the introduction is well written.

p. 794, l 3: but the temperature response and interhemispheric difference to the same aerosol forcing in models can be quite different, as seen in some of the detection attribution work - this also relates to the hemispheric gradient - so isnt there a source of uncertainty that is not accounted for in this one *simple model? if the authors agree, this should be mentioned - model uncertainty might widen the estimated narrow pdf*

Response: It is true that there are differences in the calculated interhemispheric difference in the aerosol experiments in CMIP5. However, the range is actually smaller than for the GHGs only experiments (cf. figure 3 of Friedman et al. (2013), panels d and e). Also the intermodel differences in the forcing (in particular in the effective RF which is the most relevant here) is very likely larger than for the GHGs since it also depends on the aerosol lifecycle schemes and the cloud microphysics of the models.

The CMIP5 models also indicate that the aerosol forcing has not contributed to the interhemispheric temperature Asymmetry (ITA) after about 1975, and that it is the contribution from GHGs that has led to the growth of ITA since 1980 (Friedman et al., 2013). Also in the CMIP5 historical simulations (ensemble mean for combined GHG + aerosols + natural forcing) there are indications of less variability in the ITA than in the observations.

Thus, if the low ITA in our fitted temperatures towards the end of the period is due to simplifications in our simple climate model we believe that it is more likely due to hemispheric differences in the feedbacks and possibly the thermal inertia. The latter should be taken into account by different land-fractions in the two hemispheres in our simple climate model, although there could be contributions from differences in the mixed layer depth (stronger winds and more waves in the SH mid-latitudes could give a deeper ML).

We carried out an experiment to test the impact of allowing for differences in the feedbacks for the two hemispheres (cf. panel e in figure 2). A separate feedback parameter was estimated for each hemisphere. This estimated feedbacks from the NH was indeed larger than for the SH, however, the estimated difference was not large (10%), and did not have a huge impact on the ITA (figure below). The ITA increased by 0.015K over the last 20 years of the simulation.

This discussion is included in the new section: "Interhemispheric differences "

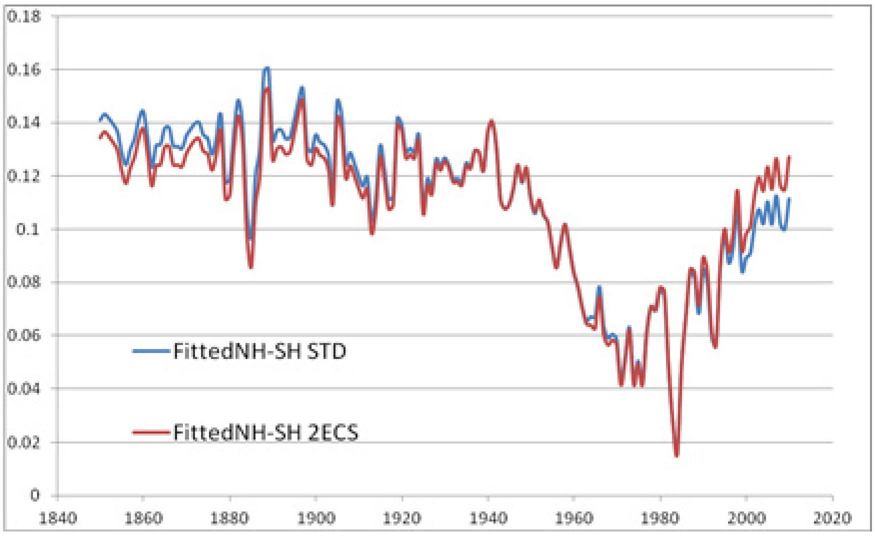

Figure R3: Estimated ITA for the standard and 2 ECS cases.

p 795, l 19-21 has some typoes

Response: We have rewritten the sentence: “To investigate the influence of the last ten years’ data on our model’s estimate, we have re-estimated the model using only data up to year 2000.”

p 796 top: doesnt the ipcc report conclude that the recent forcing is smaller (at least over the 15 yrs) than the long term trend because of negative natural forcing? I find these lines hard to reconcile with that can you explain please? Similar, p 798 top paragraph.

Response: In summary for policymaker AR5 it was stated that: ”The observed recent warming hiatus, defined as …, is attributable in roughly equal measure to a cooling contribution from internal variability and a reduced trend in external forcing”. And further that: “The forcing trend reduction is primarily due to a negative forcing trend from both volcanic eruptions and the downward phase of the solar cycle.” So it is a reduction in the forcing trend, and not a reduction in the total forcing.

In the discussion we have now summarized the increase in total and anthropogenic RF prior and posterior to clarify: “In comparison the net anthropogenic RF increased by 0.44 Wm^{-2} over the last decade for our prior assumptions (0.33 Wm^{-2} from LLGHGs). For both the prior and posterior the increase in total RF over the last decade is less than the increase in anthropogenic RF (Fig. 1), but show a clear increase of 0.29 Wm^{-2} for the prior mean and 0.31 Wm^{-2} for the posterior mean.”

Overall the external natural forcings taken into account in our study is of similar magnitude as in AR5.

p. 799 last paragraph: the carbon feedback is not really included in ECS, as ECS describes the temperature response to CO2 concentrations - please revise.

Response: We agree that in a formal definition of ECS where it is defined as the response to a doubling of CO_2 the carbon-cycle feedback is not included. However, a more reasonable definition is the response to a forcing with the same magnitude as for a doubling of CO_2, and thus all feedbacks, including carbon-cycle feedbacks should be included. For application to policymaking this is the relevant quantity. We have rewritten the paragraph with this discussion.

p. 804 discussion of AMO: other authors (e.g. Booth et al.,. although it is not uncontroversial) discuss that some of the AMO response may be forced. Can you mention this too?

Response: We have now added a new paragraph addressing this issue at the end of the Multidecadal Oscillations section in the discussion:

“There are studies indicating that there might be a forced component in the long-term variability, e.g. the AMO (Booth et al., 2012). In the climate system (and in GCMs) this would mean that the forcing would affect the variability of mixing of OHC from the surface layers to the deeper ocean. With the simple structure of our EBM we cannot represent this possibility since the parameters of the EBM are fixed over time (although the values are estimated from the data), and long-term variability not explained by variations in forcing will be attributed to internal variability. There are large uncertainties in historical forcings, and in the temporal development of the RF. However, the forcing histories from Skeie et al. (2011) applied here are based on recent estimates of historical emissions and detailed modelling of atmospheric chemistry and is significantly more detailed than RF histories applied in many previous studies. Shortcomings in temporal development of the historical RF could lead to either too much or too little of the response attributed to the internal variability term. If too little of the response attributed to the internal variability we would expect that the uncertainty in the ECS estimate is underestimated and vice versa.”

References:

Abraham, J. P., Baringer, M., Bindoff, N. L., Boyer, T., Cheng, L. J., Church, J. A., Conroy, J. L., Domingues, C. M., Fasullo, J. T., Gilson, J., Goni, G., Good, S. A., Gorman, J. M., Gouretski, V., Ishii, M., Johnson, G. C., Kizu, S., Lyman, J. M., Macdonald, A. M., Minkowycz, W. J., Moffitt, S. E., Palmer, M. D., Piola, A. R., Reseghetti, F., Schuckmann, K., Trenberth, K. E., Velicogna, I., and Willis, J. K.: A review of global ocean temperature observations: Implications for ocean heat content estimates and climate change, Rev. Geophys., 51, 450-483, doi:10.1002/rog.20022, 2013.

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

Booth, B. B. B., Dunstone, N. J., Halloran, P. R., Andrews, T., and Bellouin, N.: Aerosols implicated as a prime driver of twentieth-century North Atlantic climate variability, Nature, 484, 228-232, doi:10.1038/nature10946, 2012.

Frame, D. J., Booth, B. B. B., Kettleborough, J. A., Stainforth, D. A., Gregory, J. M., Collins, M., and Allen, M. R.: Constraining climate forecasts: The role of prior assumptions, Geophys Res Lett, 32, L09702, doi:10.1029/2004GL022241, 2005.

Friedman, A. R., Hwang, Y.-T., Chiang, J. C. H., and Frierson, D. M. W.: Interhemispheric Temperature Asymmetry over the Twentieth Century and in Future Projections, J. Clim., 26, 5419-5433, doi:10.1175/JCLI-D-12-00525.1, 2013.

Gouretski, V., and Koltermann, K. P.: How much is the ocean really warming?, Geophys Res Lett, 34, L01610, doi:10.1029/2006gl027834, 2007.

Hansen, J., Sato, M., Kharecha, P., and von Schuckmann, K.: Earth's energy imbalance and implications, Atmos. Chem. Phys., 11, 13421-13449, doi:10.5194/acp-11-13421-2011, 2011.

Hargreaves, J. C., Annan, J. D., Yoshimori, M., and Abe-Ouchi, A.: Can the Last Glacial Maximum constrain climate sensitivity?, Geophys Res Lett, 39, L24702, doi:10.1029/2012GL053872, 2012.

Hegerl, G. C., Crowley, T. J., Hyde, W. T., and Frame, D. J.: Climate sensitivity constrained by temperature reconstructions over the past seven centuries, Nature, 440, 1029-1032, doi:10.1038/nature04679, 2006.

IPCC: Summary for Policymakers. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA., 2013.

Johnson, G. C., Lyman, J. M., Willis, J. K., Levitus, S., Boyer, T., Antonov, J., Good, S. A., Domingues, C. M., Wijffels, S., and Bindoff, N.: Global oceans: Ocean heat content in: State of the Climate in 2012 edited by: Blunden, J., and Arndt, D. S., Bulletin of the American Meteorological Society, 8, American Meteorological Society, S50–S53, 2013.

Jones, G. S., Stott, P. A., and Christidis, N.: Attribution of observed historical near–surface temperature variations to anthropogenic and natural causes using CMIP5 simulations, Journal of Geophysical Research: Atmospheres, 118, 4001-4024, doi:10.1002/jgrd.50239, 2013.

Lamarque, J. F., Bond, T. C., Eyring, V., Granier, C., Heil, A., Klimont, Z., Lee, D., Liousse, C., Mieville, A., Owen, B., Schultz, M. G., Shindell, D., Smith, S. J., Stehfest, E., Van Aardenne, J., Cooper, O. R., Kainuma, M., Mahowald, N., McConnell, J. R., Naik, V., Riahi, K., and van Vuuren, D. P.: Historical (1850–2000) gridded anthropogenic and biomass burning emissions of reactive gases and aerosols: methodology and application, Atmos. Chem. Phys., 10, 7017-7039, doi:10.5194/acp-10-7017-2010, 2010.

Lyman, J. M., Good, S. A., Gouretski, V. V., Ishii, M., Johnson, G. C., Palmer, M. D., Smith, D. M., and Willis, J. K.: Robust warming of the global upper ocean, Nature, 465, 334-337, doi:10.1038/nature09043, 2010.

Schmittner, A., Urban, N. M., Shakun, J. D., Mahowald, N. M., Clark, P. U., Bartlein, P. J., Mix, A. C., and Rosell-Melé, A.: Climate Sensitivity Estimated from Temperature Reconstructions of the Last Glacial Maximum, Science, 334 1385-1388, doi:10.1126/science.1203513, 2011.

Skeie, R. B., Berntsen, T. K., Myhre, G., Tanaka, K., Kvalevåg, M. M., and Hoyle, C. R.: Anthropogenic radiative forcing time series from pre-industrial times until 2010, Atmos. Chem. Phys., 11, 11827-11857, doi:10.5194/acp-11-11827-2011, 2011.

Tomassini, L., Reichert, P., Knutti, R., Stocker, T. F., and Borsuk, M. E.: Robust Bayesian uncertainty analysis of climate system properties using Markov chain Monte Carlo methods, J. Clim., 20, 1239-1254, doi:10.1175/jcli4064.1, 2007.

Urban, N. M., and Keller, K.: Complementary observational constraints on climate sensitivity, Geophys Res Lett, 36, L04708, doi:10.1029/2008gl036457, 2009.