We thank the referee#2 (Dr. M. Balmaseda) for the comments on our manuscript submitted to *ocean science*. We appreciate the thoughtful and constructive feedback on the paper. We have addressed all concerns in the revised manuscript, as documented below in our point-by-point responses (in blue) to the comments (in black)

The manuscript presents a new estimates of global ocean heat content temporal evolution for the period 1970-2005, and compare them with previous observational estimates. The authors then use an ensemble of observational estimates to evaluate the OHC trends in an ensemble of CMIP5 model integrations. They find that the median of the CMIP5 ensemble agrees well with the observational estimates of OHC global trends, both of them showing an acceleration of ocean warming during the period 1992-2005. They and propose to use OHC as a metric to evaluate climate models. The paper is clear and well written: the problem in question is well introduced, the results clearly presented, and there is a levelled correspondence between the numerical findings and the interpretation given.

I have some questions and comments that the authors may want to take into account

1. Abstract: "We suggest that OHC be a fundamental metric for climate model validation and evaluation". The current study only deals with trends of global OHC for a given period of time. Maybe this should be the specific metric proposed. Otherwise the current statement in the last sentence of the abstract is far too generic, and open to miss-interpretation.

Reply: Yes it is a good point. We modified the last sentence in the abstract to "*We suggest that OHC be a fundamental metric for climate model validation and evaluation especially for forced changes (decadal time-scales)* "

2. Why the validation period does not extend beyond 2005? The period post-2005, when the so-call hiatus started, is of large interest. Can the authors comment on their choice of period? Would the choice of period change their conclusions?

Reply: There are several reasons not extending the period beyond 2005 in the current study: (1). Observational-based OHCs from Durack et al 2005 end at 2005. (2). CMIP5 historical runs end at 2005. Post-2005 runs are projections, which are not correctly forced by the climate forcings: such as greenhouse gas emissions and

volcanic eruptions. Also there is often a discontinuity in observational estimates of ocean heat content due to transition from mostly XBT measurements to mostly Argo measurements

3. In the observational estimates, the corrections by Durack etal (2014) seem to be included in some of the ensembles. Those use CMIP5 model information to fill the gaps. Then, these corrected estimates are used to validate CMIP5 models. It seems to me like a circular argument. How would the results be influenced by removing the Durack etal (2014) corrections?

Reply: It is a good point. We have now tested this. We have explicitly discussed this point in the main context (page8 line19-25)

"*Because the Durack et al. (2014) global OHC adjustments are partly based on heat uptake in the CMIP5 models, they should not be used to then evaluate the models. When removing Durack et al. (2014) estimates, the median change within 1970-2005 is 0.56* $\times 10^{22}$ *J/yr for OHC0-700m and 0.75* $\times 10^{22}$ *J/yr for OHC0-700m, both of which are nearly identical with the results in Table 2, suggesting that including Durack et al. (2014) does not influence the main conclusion of our study.*".

4. If the median is chosen against the mean in recognition of the non-gaussianity of the distribution, the use of Gaussian estimations for the confidence levels (twice the standard deviations) to evaluate the significance of the median seems inconsistent. Are there any other ways of estimating confidence levels for the median using nonparametric distribution?

Reply: To calculate the model ensemble results, we used median to be consistent with the observation-based results. This is to remove the impact of outliers, since the sample size is small for both observation and CMIP5 based OHC estimates. For the confidence intervals, there is no a priori reason for the statistics to be non-Gaussian other than there is a small sample and the likelihood that there are some outliers.

For model results, we test the difference of the results by using the following three strategies: (1) Median (used in this study); (2) Mean; and (3) Mean after removing the minimum and the maximum of the model results. The third method is to estimate the percentiles by ranking the model trends and reading off the percentiles (E.g. if there are 24 models, the 10^{th} percentile is the 2.4^{th} model, so we remove the

minimum and the maximum of the models to get 5^{th}-95^{th} percentiles).

OHC0-700m

1970-2005: Median: 0.42×10^{22} J/yr

1970-2005: Mean: 0.42×10^{22} J/yr

1970-2005: Mean (remove the minimum and maximum): 0.42×10^{22} J/yr

OHC full-depth

1970-2005: Median: 0.68×10^{22} J/yr

1970-2005: Mean: 0.66×10^{22} J/yr

1970-2005: Mean (remove the minimum and maximum): 0.66×10^{22} J/yr

This test indicates that it makes no much change on the results when using the three different strategies.

Based on these discussions, we still decided to use the standard deviation to characterize the spread and the median to characterize the ensemble average – essentially because we do not have enough models to have a good statistics.

We found it could be helpful to include a discussion with respect to this point in section 3, see page13line17:

"*Furthermore, the OHC for models show a non-Gaussian distribution (Figure 3), potentially challenging our method of the use of Gaussian estimations for the confidence levels. However, there is no a priori reason for the statistics to be non-Gaussian other than there is a small sample and the likelihood that there are some outliers. The non-Gaussian nature of the distribution (Fig.3) may be partly due to the small sample size. The use of the median reduces the impact of outliers and then enables us to use the standard deviation to characterize the spread.*"

5. It is said in the text that the estimate of OHC by Smith and Murphy is discounted because the values are smaller than the others. This is quite an adhoc reason. Can the authors provide a more solid motivation for excluding the estimation?. The estimate is not removed from figure 3, which is misleading

Reply: (1). Fully evaluating an OHC estimate (such as Smith&Murphy2007) requires a more comprehensive study in the future (to understand the mapping methods etc.).

Our decision to remove Smith&Murphy is based on fig.2, because apparently it is an outlier. Why this is so is beyond the scope of this study.

(2). We still keep Smith&Murphy2007 in Fig.3. And we note in the main context that including Smith&Murphy2007 value does not impact our results, since we use median rather than mean (Page6line11).