This document is the author's response to review os-2016-105-RC1.

First off, we apologize that this reviewer expected this paper to be about the "reliability" of the long-term trend in ocean heat content. We explicitly state that we quantify the reliability of variability ***about*** the trend and seasonal cycle--i.e., not the trend, but the variations after detrending. There is extensive literature on the reliability of the observed warming of ocean heat content (e.g., *Domingues et al. 2008, Levitus et al. 2009*). Despite our efforts to clarify, there has been repeated confusion about the distinction between our study on the ocean heat content variability versus other studies on the ocean heat content trend.

As such, near the lines indicated by the reviewer as suggesting otherwise, we will add the additional clarifying statement:

"The focus of our study is *not* the reliability of the Ocean Heat Content trend over the whole record, which has been studied exhaustively elsewhere, but the reliability of variability once the trend has been removed--i.e., variability on timescales shorter than the record length."

Since the reviewer's conception of the paper was that it addressed the reliability of the observed trend, rather than of the variability of the observed trend, we found it difficult to apply all of the comments given by the reviewer. Nonetheless, the enumerated points by this reviewer are addressed below, with the reviewer's comments bolded.

1.

The model resolution is 340km at the equator and 40km near the north polar. This is a low-resolution model, which means the model mainly simulates large-scale ocean variabilities, the smaller-scale variabilities such as meso-scale eddies are not resolved. This is a major difference between modeland observation, since observation contains variations at all scales. Therefore, using models will significantly under-estimate the uncertainties.

We discuss this issue in Section 4.4, page 9, lines 14-19. We would love to have a high-resolution model output spanning thousands of years. Unfortunately, such models are too expensive computationally to be available this decade. In light of this, we used trusted model output that was already shown in previous studies to do a good job representing larger scale variability. This model was immediately available for our use. In other words, our choice of model was one of carefully-considered convenience.

This main deficiency makes the uncertainty values provided in this study useless.

We disagree, as the cited papers make it clear that the model possesses subannual variability consistent with observations (see Section 4.4, page 9, lines 12-14).

2.

Page 2, line1-4. The authors said the focus of this study is to quantify the reliability of the warming trend.

This is incorrect, as already mentioned. As is written on page 2, line 2-4; “Our focus here is to quantify the ability to faithfully measure the variability <u>about</u> the warming trend using an (EOSSE)”.

There are two key questions related to the observed OHC trend: (1). Is the observed trend biased or not? This is a really intriguing question. (2). What is the uncertainty of the calculated trend?

As stated previously, the focus of this paper is not on the OHC trend, but on OHC anomaly variability, where we define ‘anomaly’ in this context to be about the all-time trend and seasonal cycle. Bias implies that the trend is either too fast or too slow, rather than variable in both directions on shorter timescales than the whole record, which is our focus.

3.

The section-2 is to review the literatures, but it is incomplete, lack of many recent literatures about OHC estimation based on observations.

We kindly ask the reviewer to provide references to some of the “many recent literatures” they mention here. We encourage others as well to suggest papers that may be beneficial to include in this section. We have read widely, and found the cited references to be of most direct relevance. This is, after all, not a review paper.

4.

And, the review of the literatures in the section-2 is chaos, ...

We attempted to reference previous estimates of global OHC uncertainty, not just from OSSEs but from other sources as well, such that we could compare our uncertainties to the uncertainties due to other factors in ocean observation processing. We apologize that the reviewer found this presentation to be disorganized.

5.

Page 4, line 16-17. What the authors mean by “statistical error propagation methods.”?? I think the authors are not clear at all about what the referred literatures did by so-called “objective analysis”. Almost each objective analysis method dealt with uncertainty or error in their analyses.

In the context of our referenced papers, the error estimates were either provided by the objective analysis method or based on similar mathematics. These methods are detailed exhaustively in the cited references (e.g., *Kaplan et al. 2000*), as well as exemplified in standard data analysis textbooks (e.g., *Emery & Thomson, 2001*). Our “errors” are constructed as differences from the given observations to the a-priori assumed climatological means and variances weighted by the number of observations within each specified region. Our method of uncertainty estimation does not depend on such a-priori assumptions.

6.

Figure 7 is a horrible figure. It does not make any sense that the OHC time series should be like that!! The large jump around 1995 and 2003 must be spurious, so it is meaningless to show such a figure.

We apologize that the large jumps in global OHC around 1995 and 2003 were not discussed in this paper. We have edited the caption of the figure, to reflect our conversations with Dr. Tanguy Szekely, the scientist in charge of the CORA dataset, copied verbatim below:

"The 1996 and 2003 jumps are associated to changes in the ocean sampling (see the PCTVAR field). The Indian ocean measurements almost disappear in 1996 leading to a decrease of the solution accuracy. In 2003, the begin of the worldwide ARGO deployment change the solution in numerous badly sampled zones (pacific ocean, South Atlantic Ocean, Antarctic Ocean, etc....) This problem will be solved in the next version of CORA (to be released on April 2016° A paper with details on the dataset description and validation should been published then."

Perhaps using the updated data set would 'fix' these jumps, but we are not addressing here how 'correct' the OHC time series is. The goal of this paper is to quantify the uncertainty due to the ISAS13 observing strategy, or construct the one-sigma confidence interval shown in the figure. The reason for including this figure was to compare the relative size of the confidence interval we find to features of the data, such as these jumps.

The related discussion makes no sense as well.

Without more specificity, we are unable to address this point.

Moreover, the error bar is too small to be believable, it makes no physical sense that the error bar is so small.

It must be emphasized here that the error bar is purely derived from the expected difference between the 'observed' and 'true' oceans from usage of the ISAS13 observing strategy. It does not take into account instrumental errors and biases, errors intrinsic to the objective analysis, mesoscale eddy activity, and other sources of error. It is clear from this comment that the reviewer has lost track of our argument by this point, for which we apologize.

And also, why not provide OHC anomaly rather than OHC?

Generally, OHC is considered always to be an anomaly as it has no meaningful baseline in a context such as this one; a zero-value of OHC has no dynamical or thermodynamical meaning. In the context of our paper, we define the OHC anomaly to be the OHC signal about the overall trend and seasonal cycle. To help clarify this point, Fig. 7 will be changed such that the all-time mean is removed.

7.
I suggest the authors give a clear definition about 'sampling strategy', 'observing strategy' and 'mapping method'.
An observing strategy in the context of our work contains two steps; the sampling strategy and the mapping method. The sampling strategy defines the locations and times of the observing strategy's set of observations. The mapping method describes the methodology used to process the resulting observations from the sampling strategy into "maps" with standardized locations and times (e.g., 1x1 degree grid of monthly means). We describe this in section 4.2, but will revise the discussion to clarify further.

8.
Figure 1 is another horrible figure. I doubt the authors make it right. The profiling floats in Fig.1 are more than 90%, but it is not!!!!! See the figure below from NCEI. And the moored buoy should be much more than this figure shown. So it is not a trustable plot for me.
In Figure 1, it is stated that our source of observations is the ISAS13 data collection. The ISAS13 dataset and the NCEI dataset do not contain the same observations, although there is some overlap. We invite others to reconstruct this figure and double-check our Figure 1.

9.
Figure 2 is useless. I don't see the point of figure 2. It seems to me Fig.2a has very good global data coverage.
This statement is subjective, as is the definition of "global coverage", but we understand that our small plot size in combination with the relatively large data points could incite this opinion. However, redrawing these figures to be larger with smaller data points would merely dwell further on subjectives and aesthetics. Many other references attempt a rigorous, objective definition of "global coverage" by ocean observations; it is not our intent to estimate such a quantity or precisely compute "global coverage" based on this figure, only to illustrate that the observations in the Argo era sample much more of the global ocean much more evenly than the pre-Argo observations.

10.
Another major confusion is about section 5.2. This section and related experiment is to investigate the dependence of timescale of errors due to sampling. However, the major variabilities of the model runs are on inter-annual scales (shown in Fig.3): i.e. there are no meso-scale signals, and no long-term trends.
This reviewer is again misled by his preconception that the focus of the paper is on the reliability of the trend. The model is run to near-equilibrium, so there is no long-term trend. We agree that mesoscale signals would add variability, which is impossible to assess based on our chosen model, but the models have significant annual and shorter timescale variability as it is a fully coupled model with synoptic winds, seasons, etc.

I don't find it has any implications for long-term trend.
Once again, the long-term trend is not our intended objective.

11.

And what is the physical meaning of the standard deviation of the cross-correlation in Fig.4 and 5.
We understand the confusion for the physical meaning of the standard deviations of the cross-correlations--this was not an obvious metric to us either when we began the project. Since the ISAS13 observing strategy only captures 23 years of the ocean, and we know the ocean has a great amount of variability on time scales longer than 23 years, the standard deviation of the cross-correlation distributions represents the sensitivity of the cross-correlations to the state of the ocean in the 23-year span of the observing strategy. This definition follows for the pre-Argo era and Argo era of observations as well.

If +/- one standard deviation means >60% confidence interval. The time variation of the mean correlation is not significant.
This statement is incorrect. In Figures 4(b,e,f), the distributions have statistically-significant changes in their means versus the time-window selection, at least to one-sigma (~68%) confidence. For Figures 5(a,b,c), the change of the 50% percentile against the 5% to 95% percentile range versus time-window is unclear without performing some significance test. However, there is a statistically significant difference between the pre-Argo era and Argo era, which is the point this Figure aims to show.

12.

Similar to my point-10/11, in section 5.3 and Fig.3, the change of the correlation is neither significant nor physically tenable.
The reviewer must clarify how this point is differs from the ones made in 11.

13.

Section 5.4 is the only section that makes some senses, but the near-zero mean error in Fig.6 is not a surprise, since the models are free-runs without any external forcing. The ensemble mean should be zero.
This is true, the OHC anomaly time series have zero-means by construction, although why the reviewer pointed this out is unclear. The important information contained in Figure 6 is the size of the standard deviation of the errors relative to the variability in the data represented by the dashed line. This gives information about the signal-to-noise ratio, an extremely important statistic in signal processing, and a primary result of this work.

The only meaningful conclusion is the uncertainty pre-Argo is larger than Argo, but it is not surprise.
It is true that this result is no surprise. However, the difference in the uncertainty in the pre-Argo versus the Argo era is rarely quantified with any precision. Some variability in OHC would be detectable even in the pre-Argo era, and some would not. Figure 6 indicates quantitatively where the boundary lies in this particular model. One of the goals of this paper was to introduce a methodology that can quantify such quantities, which we feel we've demonstrated successfully. Such a quantification is critical in determining the accuracy of reconstructions of historical variability.

14.
The authors argues that the size of the error in Argo era is 1/3 smaller than pre-Argo, I don't find it is a useful value. Because of many reasons (1). The model simulations in this study are mainly on inter-annual scales, no (weak) other variabilities, no trend etc.
See our response to point 10. Furthermore, we are repeatedly explicit that this paper presents a methodology for evaluation of these statistics which might be used with any model. Other models would have different variability which would need the same kind of quantification.

(2). The results should be specific for the ISAS mapping method and do not take account of any other errors (e.g. XBT bias, climatology issues)
This is true. The methodologies we introduced are to objectively quantify the errors associated with a specific observing strategy (mapping and sampling pattern) and do not take into account errors due to instrumental biases and errors, climatology choice, etc. One of the goals of this work was to quantify this specific source of uncertainty. Other sources of uncertainty have been studied in other papers, but we believe attention to how well variability is observed is a novel question.

Bibliography
[1] Domingues et al. 2008, Improved estimates of upper-ocean warming and multi-decadal sea-level rise, *Nature* **453**, 1090-1093.
[2] Levitus et al. 2009, Global ocean heat content 1955-2009 in light of recently revealed instrumentation problems, *Geophysical Research Letters* **36**(7).
[3] Kaplan, A., Kushnir, Y. & Cane, M. A. Reduced space optimal interpolation of historical marine sea level pressure. J. Clim. 13, 2987–3002 (2000).
[4] William J. Emery and Richard E. Thomson, Data Analysis Methods in Physical Oceanography, Elsevier Science, Amsterdam, 2001, ISBN 9780444507563.