Nelson et al. quantified skill and uncertainty in observing the ocean heat content (OHC) in the upper 700-m layer using observing system simulation experiments (OSSEs). This is a highly recommended method to evaluate both historical observation system and the current mapping strategies in OHC studies, so the topic is very interesting and useful. However, when I read this paper through and observe the conclusion of this study, major deficiencies are found. The conclusion is rather shallow: the authors find that the uncertainty of OHC estimate pre-Argo era is larger than the Argo era. This is a broadly known issue, so it is not novel contribution. And the uncertainties values quoted in this study are not reliable, because this is a model-based study and the uncertainty depends largely on the model performance and model resolution (see my comments below). Besides, there are some other major flaws in this study, so I recommend rejection.

Major

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 model and observation, since observation contains variations at all scales. Therefore, using models will significantly under-estimate the uncertainties. This main deficiency makes the uncertainty values provided in this study useless. This is my first major concern.

2. Page 2, line1-4. The authors said the focus of this study is to quantify the reliability of the warming trend. But I don't think the paper did that. 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? This study provides some clues for the second question based on OSSE method, but understanding the real uncertainty in OHC estimate is difficult and is not (and can not) sufficiently done by this model-based study. So the current study contributes little to the community.

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

4. And, the review of the literatures in the section-2 is chaos, mixing observation-based studies and OSSE-based analyses together. For instance, Lyman et al. 2006 and 2008 are not OSSE. And the paragraph (lines around 20, page 3) seems strange in the context and von Schuckmann and Le Traon (2011) is not OSSE as well. Page4, the 2rd and 3rd paragraph, after two paragraphs discussing OSSE, those two go back to observation-based analyses again. Though I agree von Schuckmann et al and Lyman et al dealt with the climatology, did this paper have any contributions to clarify how the choice of climatology impacts the OHC calculation??? I think it will be helpful if

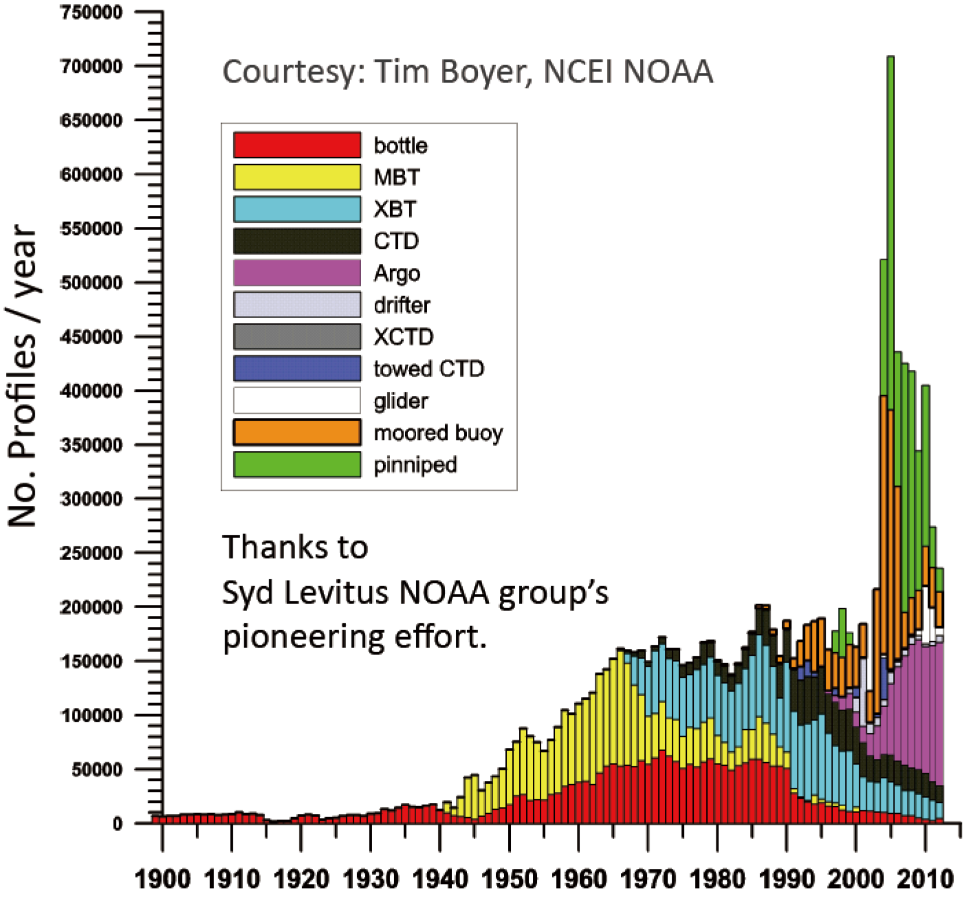
the authors could deal with this issue in the future using OSSE analysis (similar to Good et al. 2015), but it is not in the current manuscript.

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.

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. The related discussion makes no sense as well. Moreover, the error bar is too small to be believable, it makes no physical sense that the error bar is so small. And also, why not provide OHC anomaly rather than OHC?

7. I suggest the authors give a clear definition about ‘sampling strategy’, ‘observing strategy’ and ‘mapping method’.

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.



Courtesy: Tim Boyer, NCEI NOAA

Thanks to
Syd Levitus NOAA group's
pioneering effort.

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..

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. So it doesn't make any significant difference when using different window sizes ranging from 1 to 24 months. The small changes may reveal the uncertainty due to the small fluctuations added to the large inter-annual variation. I don't find it has any implications for long-term trend.

11. And what is the physical meaning of the standard deviation of the cross-correlation in Fig.4 and 5. If +/- one standard deviation means >60% confidence interval. The time variation of the mean correlation is not significant.

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.

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. The only meaningful conclusion is the uncertainty pre-Argo is larger than Argo, but it is not surprise.

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. (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)