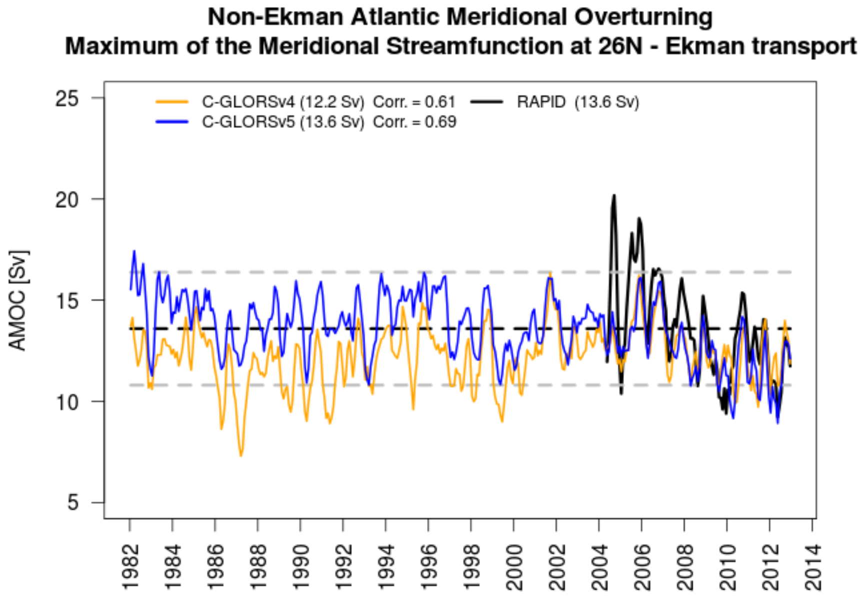
Response to Reviewer 2

We thank both reviewers for the very careful reading and the many suggestions they proposed on how to improve the quality of the manuscript, and in particular Reviewer 2 for pointing out ways to better discus the AMOC validation, further to many technical corrections. Below, we answer in details to all the points arisen from Reviewer 2. Sentences in italic are the reviewer's comments, while our reply is in bold, and includes the reply and the proposed modifications to the revised version of the manuscript.

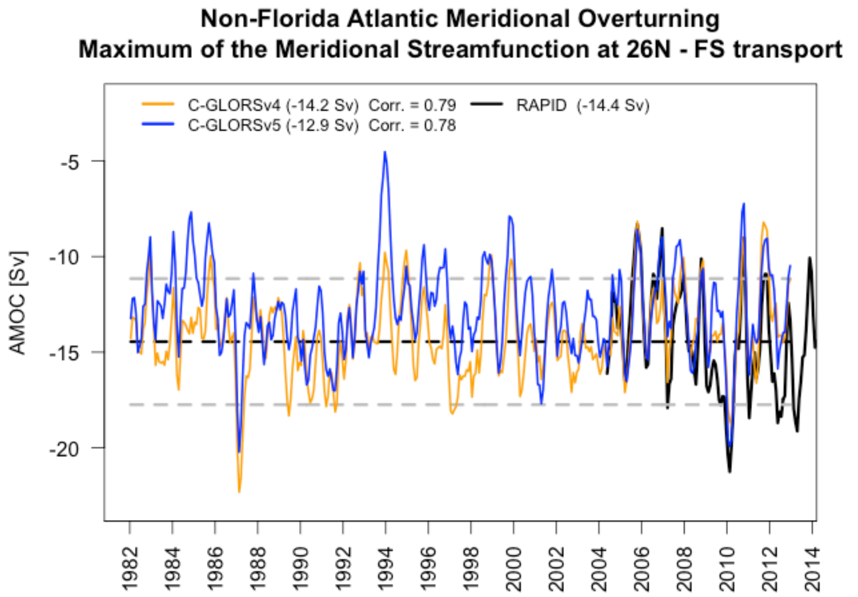
My largest concerns were to do with the presentation and discussion with regards to the AMOC in Section 3.3. Firstly, it is somewhat difficult to make out the C-CGLORYS overturning underneath the RAPID observations in my paper version of Figure 9 – although I suppose that is one of the virtues of electronic media, as it is much easier to see in a zoomed in electronic version. It would also be illuminating, but not necessarily convenient, for the authors to show the non-Ekman component of the overturning streamfunction. All models, pretty much by definition, would replicate the Ekman component of the overturning derived by the RAPID observations, since that is solely determined by the wind stress forcing which is typically identical, or near identical to that used in the RAPID calculation. Replicating the density driven circulation, on the other hand, is more difficult — and ideally it would be the correlation between the non-Ekman portions of the overturning in the RAPID observations and C-GLORYS re-analysis that would be most interesting [Roberts et al., 2013]. In the absence of that calculation, however, it is noteworthy that neither C-GLORYS analysis appears to pick up the early period peak in the RAPID observations — although they do appear to pick up this 2005-2010 decrease in the RAPID overturning, and subsequent increase after 2010 from about 2007 onward — which coincidentally would be when the ARGO float array is fully deployed, and their analysis potentially morphs from being largely driven by the SST nudging to one where the sub-surface profiles are playing a substantial role. Perhaps the authors may wish to substantiate on that further.

We thank the reviewer for suggesting several ways to improve the discussions on C-GLORS with respect to the AMOC performances.

First, we certainly agree on the idea of evaluating the non-Ekman component between the reanalyses and RAPID in order to assess the density driven overturning circulation. To this end, we have evaluated the non-Ekman component, and we added a third (middle) panel in Figure 9 showing it (reported below). While there is a slight decrease in the correlation with RAPID, due as expected to the withholding of the Ekman transport, qualitative results on the larger correlation and increased mean value of the AMOC in v5 still hold.



Second, the reviewer questions why the reanalyses fail in capturing the first peaks occurring in the RAPID timeseries, which are due to the under-estimation of the western boundary current contribution (Florida Strait, FS, defined as the total transport at (80.1°W–77.4°W; 26.5°N)). The figure below reports the AMOC-FS (AMOC minus the transport across the Florida Strait), showing in particular that the peaks at the beginning of the RAPID time-series are not captured because of the FS under-estimation, due to poor observing network in 2005 along with the fact that the ¼ degree model resolution does not allow to perfectly resolving the Florida Strait geometry. Note that only the mismatch in the first two peaks of AMOC can be explained by the Florida transport under-estimation. We have explicitly added this explanation in the revised version of the manuscript.



1. *p. 7, l. 2: 9% of the observations affected by the bug seems large. Are you inferring that 9% of the profiles were actually only surface measurements? Furthermore, is this 9% of the profiles, or 9% of all the profile observations at all levels?*

   We thank the reviewer for pointing this out: this was a typo as the percentage of mistakenly rejected surface in-situ observations is equal to 0.9 % on the average. We have corrected the value, also specifying that the number of mistakenly rejected surface in-situ observations is equal to about 3% at the beginning of the reanalysis period (ie at the beginning of the 1980s, when there are less profiles sampling deep waters).

2. *Figure 3: There is a large spike in the monthly inflation value in 1993, very close to the coming on line of the altimeter data. Coincidence? Spurious?*

   Thanks for pointing this out: although it is not straight-forward to prove, we also agree that the spike might depend on altimetry data ingestion, and suggested it as a possible cause in the revised version.

3. *p. 8, l 18: "namely the floats used represent a fairly independent dataset." Firstly, tacking this onto the end of the previous sentence does not make grammatical sense, but more importantly, you are being unduly brief with what in my mind is a fairly complex statement. What I believe you are saying is that because you are comparing observation minus background for floats (as opposed to say moored buoys), measurements at any given point can be consider independent, since (in principle) no one float makes repeat measurements at the same location. Perhaps it would be better to expand your statement somewhat – making it a complete sentence while you are at it.*

   The reviewer is correct in understanding what was implied by the sentence: we have reformulated the phrase to explicitly state the independence of floats observations given their spatial misplacement along time, and corrected the sentence.

4. Figure 4 and discussion p. 8, ll. 20-23. Only the global average observations minus background stats are shown. It would be worth at least showing the tropical (possibly Tropical Pacific) statistics that you note as significantly improved. Can this be attributed to the decrease in background-error standard deviation in the tropical Pacific. Conversely, observations minus background for the North Atlantic where the background-error has been increased and the skill decreased could be illuminating. Is the Gulf Stream more misplaced in the non-assimilated version of v5 compared to v4? Finally, you attribute the decreased skill at high latitudes to differing sea ice cover – but the sea ice cover should be largely constrained by the sea ice concentration observations. Is not simply that you have increased the background error in this region as well? Note that the SST rmse is reduced at high latitudes as well.

We have added a panel with the skill scores in the Tropical Pacific and North Atlantic oceans, showing the significant attenuation of Tropical biases in v5. A discussion on the reasons for different scores has been added to the revised version of the manuscript, as also suggested by Reviewer 1 (see also the Response to Reviewer 1). SST Skill scores in the Gulf Stream get worse in v5 mostly due to the increased background-error variances that also leads to a northward shift of the Gulf Stream separation. However, skill scores against in-situ profiles in the North Atlantic Ocean do not look worse than in v4. We have also better detailed the changes in scores at high latitudes, which may be due not only to the re-tuned background-error covariances but also to the use of uncorrected atmospheric forcing. In the first version, we meant that under sea-ice, NOAA SST data are extrapolated by sea-ice concentration data: SST analyses are affected by large uncertainties and the RMSE are not really "meaningful". We have however rephrased the sentence to make it clearer.

5. *Figure 8 and Section 3.2: The more accurate (compared to the NOAA/NODS estimates) trend in 2000m heat content in the Gulf Stream region seems at face value at odds with earlier statements regarding a loss of observation minus background skill in this region. However, there is also a decrease in SST rmse here as well, presumably due to the increased background error standard deviation – although the lack of flux correction could also play a role. Is the trend in 2000m heat content largely surface driven here?*

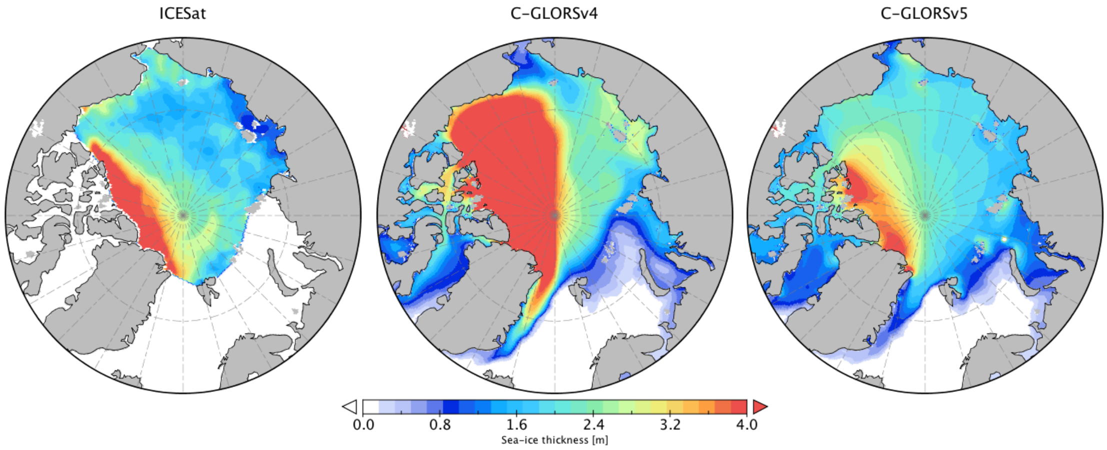
   We have now better discussed the strengths and weaknesses of the reanalysis, also according to Reviewer 1: while SST skill score results seem worse in the Gulf Stream area, here the increase of heat content trend in the central and eastern North Atlantic Ocean drives the sustained global signal in the last decade. The trend is largely driven by the surface.

6. *Figure 9 and Section 3.3: As mentioned above, a comparison of the non-Ekman component of the Atlantic overturning streamfunction would be useful, but not essential.*

   Please see the answer above to the general comment. We have added the comparison with the non-Ekman component.

7. *Sections 2.2.5 and Section 3.4: Note on using PIOMAS as data. While PIOMAS does validate well with the sea ice thickness over the period it was validated – mostly ICESat data. However, it does not validate as well over data from more recent periods, possibly overestimating March ice thickness. However, I have no citable literature to back my claim, so this amounts to hearsay. Nevertheless, the Arctic ice volumes in C-GLORYSv5 are undoubtedly more realistic then those of v4. Undoubtedly, the spatial thickness patterns are close to those that are begin imposed by PIOMAS, nevertheless, a spatial map could be useful, especially if it can be compared with a satellite observations for a particular period. Laxon et al. [2003] could be used for an early altimeter based thickness estimate.*

We agree that PIOMAS is not a validating dataset, although it has been in turn extensively validated, as also the reviewer suggested. To this end we added also the reference to Schweiger et al. (2011, JGR). While we acknowledge that an independent dataset would be useful, that from Laxon is not accessible and there are copyright issues. We compared the 2004-2008 Winter mean thickness with ICESat (reported below), which highlights that the sea-ice accumulation in v4 is now fixed in v5. However we prefer not to include the figure, but mention this comparison in the text only.



8. *Figure 10: A yearly timeseries (with collapsed vertical axis) along with a seasonal cycle climatology might be more easily decipherable than the monthly timeseries shown. It might even be possible, with dual vertical axis, to plot area and volume on the same plot so that the number of sub-figures remains the same.*

We thank the reviewer for the suggestion and have replaced the Figure 10 in a Figure with 4 panels, 2 of which showing the yearly means, and 2 showing monthly climatology.

9. *Section 3.4 and Figure 10. There is (presumably?) no ice thickness restoring performed in the Antarctic, yet the volume field in v5 is also more stable here than in v4. No mention was made of any sea ice improvements in the model, so can this be attributed to either the removal of the flux corrections, or the changes (increase?) in background error covariance for the profiles. I note the error RMS error in SST is also reduced in both the Antarctic and Arctic. Would there have been similar improvement in the Arctic volume without the PIOMAS restoring?*

This is quite difficult to prove, but we believe that other factors such as retuning of the data assimilation system and use of uncorrected forcing might also contribute to a better sea-ice reconstruction. We explicitly added a sentence on this on the revised version.

Typos and Grammatical Errors

All the 8 Typos and grammatical errors indicated by Reviewer 2 have been corrected in the revised version of the manuscript.