Response to anonymous Referee #1

We thanks Reviewer 1 for the interesting and extensive comments on the manuscripts. Below we will provide a detailed response to all individual comments.

I don't understand the reasoning for why you generate 100 reasonable fits and then select only the 10 best fits (P7 L15). Firstly, this emulator/box-model should be cheap to run, so why choose such small numbers? Surely ensembles of order 10,000 or 100,000 are more reasonable. Secondly, the choice of 10 best fits seems to narrow the ranges of several parameters (e.g. V4, F1, h1). By doing this you rule out large regions of parameter space that give perfectly reasonable fits, and could behave differently under different forcing scenarios. If the primary aim is to assess uncertainty in AMOC projections I would expect to see a rigorous analysis of the uncertainty. By discarding large areas of parameter space uncertainty will certainly be underestimated.

Thanks for raising this valid point. Our aim here is not to provide an uncertainty assessment of AMOC projections, it is to provide a method with which one could do this, as for example done in the manuscript by Bakker et al. under review in GRL, now also pointed out in the last line of the main manuscript (lines 1-2 page 13) "The AMOC-emulator is a valuable tool to study the uncertainty in GCM-based AMOC projections, such as the one recently being performed on the results from the AMOCMIP project (Bakker et al., 2016)." The assessment referred to here is based on multiple GCMs, decreasing the need for a large number of AMOC-emulators for a single GCM.

With regard to the point that the AMOC-emulator is cheap and could thus be run for tens or even hundreds of thousands of times. This is very true, however, it takes many time steps and iterations to find a single reasonable fit. We have now included the following description to the manuscript (line 30 page 8 to line 2 page 9) "To provide an idea of the computational expenses of the model we provide a back of the envelope calculation. This shows that a single run over all scenarios takes 10^5 time steps which are done in about 5 seconds. You need on the order of 400 iterations (in which parameter values are perturbed) to find a single reasonable fit, resulting in approximately half an hour to calculate a single reasonable fit on a normal desktop computer." This shows that by using more powerful computers and or running in parallel the number of reasonable fits could be enhanced, but it shows that 10,000 to 100,000 reasonable fits is ambitious nonetheless.

In the four scenarios not seen by the emulator (Fig. 8) the behavior of UVic is clearly not captured in two cases (lower left and upper right). There are no confidence intervals plotted (or computed as far as I can tell), but I believe the GCM would lie well outside 2 standard errors in those two cases. Therefore, the GCM would still need to be run for any untested scenario. I would not trust the emulator in its current form.

We don't agree with the general notion given by the reviewer. Firstly, it is important to realize that the values given in figure 8 are anomalies with respect to the time series given in figure 7. Thus even the largest mismatch between GCM and AMOC-emulator (~1-2Sv in lower left panel) is 'only' a mismatch of 10-20%. We have added an objective assessment of the predictive power of the AMOC-emulator by comparing the results with a null-model that assumes that the emulator has no predictive power; it doesn't know if an additional forcing on top of the ones used in the tuning procedure would further increase or decrease the AMOC and would thus result in zero anomalies. This assessment shows that in three out of four cases the AMOC-emulator has substantial predictive power. We discuss this assessment in the manuscript (lines 20-30 page 11) "This is quantified by comparing the AMOC-emulator results with a null-model

that assumes an AMOC-emulator with zero skill, meaning that it simply reproduces the original calibration data. The results from these experiments are shown as anomalies relative to the original scenario, the original being RCP8.5-GIS for RCP8.5x0.5-GIS, RCP8.5x1.5-GIS and RCP8.5-GISx1.5, and RCP4.5-GIS for RCP4.5-GISRCP8.5x1.5. We find that for large changes in the GHG forcing the Uvic-based AMOC-emulators are well capable of predicting the AMOC evolution of UVic in terms of sign and amplitude and perform better than the null-model (upper panels Fig. 8). For large changes in the applied GIS melt forcing the picture is more complex (lower panels Fig. 8). A strong increase in GIS melt under a low GHG scenario shows an excellent performance of the AMOC-emulator and a RSME that is much lower than for the null-model (RCP4.5-GISRCP-8.5x1.5 in Fig. 8), but for the high GHG scenario, a 50% increase in GIS melt leads to a deterioration of the fit between UVic and AMOC-emulator with consequently a larger RSME than that provided by the null-model (RCP8.5-GISx1.5 in Fig. 8). The latter shows that the UVic-based AMOC-emulators tend to overestimate the impact of GIS melt on the AMOC strength under high-end GHG scenarios. Summarizing, in all four cases the emulator predicts the correct sign of the AMOC response to changes in the forcings, and in three out of four cases the predictive power of the AMOC-emulator is better than of the null-model.". Nonetheless, it is important to acknowledge that using an emulator will introduce a new type of error in any assessment, pointed out by the following text in the manuscript (lines 5-7 page 12) "It is clear that using an AMOC-emulator introduces a new type of uncertainty into AMOC projections, however, for which level of added uncertainty an AMOC-emulator is still useful is a question that is difficult to address."

On multidecadal timescales the emulator is plagued by sensitivity to surface temperature oscillations. These seem to have arisen from the addition of the atmospheric boxes to the ocean box model published by Zickfeld et al., 2004. Can the authors confirm that this is the case, and if so can they control this sensitivity, e.g. by introducing a damping/mixing term?

The multidecadal AMOC oscillations result from the UVic-based regional temperature forcings of the AMOC-emulator and thus in turn to internal variability of UVic. Zickfeld et al. (2004) applied highly idealized linear temperature increases of global temperature, thus not including any multi-decadal variability. On the contrary, in our approach we directly use regional GCM-based temperature time series to force the AMOC-emulator. In this way the forcing not only takes into account the GCMs global climate sensitivity, but also mechanisms like polar amplification etc. that cause regional temperature change differences. This method also introduces any multi-decadal internal variability that might exist in a GCM into the AMOC-emulator when expressed in regional temperature time series. We acknowledge this feature, but do not see it as an issue.

If the authors have good reason to retain this behavior they need to test the sensitivity to the phase of the variability. For all of the scenarios, the chosen start date (2006) appears to be shortly after a peak in the strong multidecadal variability, so the AMOC is preconditioned to decline at this time. Under all scenarios the AMOC in the 'best' emulators appear to decline faster than the UVic model. Consequently, the SA tuning and the cost function used may be adversely affected by this multidecadal variability.

Indeed, following from the usage of the Stommel model to emulate the AMOC, multi-decadal temperature variability and its phasing impact the projected AMOC changes, in the AMOC-emulator, in UVic and most likely also in reality. Perhaps the AMOC response in the AMOC-emulator to regional temperature changes is too direct (as mentioned in the manuscript) and thus the importance of multi-decadal variability overestimated, but we don't see this as a major issue. It seems to us that the years before 2006 represent in fact a time of relatively weak AMOC, not

strong, thus preconditioning the AMOC-emulator to a somewhat weaker response to global change. We don't agree with the notion of the reviewer that the decline in the emulators is faster than in UVic, they seem very similar to us. Finally, multi-decadal AMOC variability only impacts the absolute value of the cost function, not the resulting optimal fits.

On centennial timescales the emulator (as currently presented) does not capture crucial features of the AMOC response to the forcing (Fig. 7). In particular I would draw attention to the RCP4.5 scenarios, in which the GCM exhibits a strong reduction followed by a steady recovery. The emulator fails to identify either the timing or amplitude of the AMOC minimum and it fails to identify the recovery phase. In addition it appears to show signs of a recovery phase under RCP8.5 when UVic shows none. The authors state (P9 L28) that the fit can be improved, but that this would entail a higher overall cost function for the SA tuning method. Is this indicative of a poor choice of cost function? Does it mean that the box model should be tuned separately for each scenario?

The failure of the AMOC emulator to capture the slight recovery of the AMOC under RCP4.5 is indeed an issue and shows the limitations of the simple box model to capture all complex feedbacks in the GCM. Indeed, as mentioned in the manuscript, the AMOC-emulator does allow for an AMOC recovery under RCP4.5, but that would mean a large deterioration of the fit of the AMOC emulator to the AMOC in RCP8.5 and thus it would increase the value of the cost function, for which reason this solution is not found through this approach. It is an interesting point if the AMOC-emulator should be tuned separately for each scenario. We added the following to the manuscript to cover this issue (lines 23-33 page 10) "It is also worth noting that the fit for an individual simulation could be improved, for instance the AMOC-emulator does allow for a partial AMOC recovery as UVic shows for RCP4.5, but such an AMOC-emulator is not found through the SA tuning methodology in this example, because it would degrade the fit for the other scenarios and thus lead to an overall higher cost function." More discussion on this topic follows in Sect. 4 of the manuscript (lines 7-13 page 12) "Another important consideration when using the AMOC-emulator is the spread in GCM climate forcing scenarios that is included in the tuning process. When using only a single climate change scenario, a better match can be obtained between the AMOC evolution given by the GCM and AMOC-emulator, however, the reliability of the AMOC-emulator will quickly decrease for different climate forcings. On the other hand, one could use a large number of climate change projections in the tuning process to obtain a lesser fit for individual scenarios, but an AMOC-emulator that is applicable to a much larger range of climate change scenarios. The best strategy to be follow strongly depends on the research question in mind."

A far more substantial summary is required. For example, the emulator's limitations need to be clearly stated (and whether/how the authors think these can be addressed). For what purposes are the emulator suitable in its current form, and for what purposes might it be useful subject to further work? With the current analysis, I disagree with the statement that "the UVic-based AMOC-emulator captures well the overall characteristics of the multi-centennial response of the AMOC".

Thanks for this comment. We agree that are more substantial and clear discussion is needed to make clear what the model can and cannot do. We have added the following to the discussion section (lines 1-22 page 12) "Overall, the predictive power of the AMOC-emulator is reasonable when one considers the simplicity of the AMOC box model, but for forcing scenarios that are increasingly far away from the forcings that are used in tuning the AMOC-emulator, the predictive power decreases. A large advantage of using a physics-based AMOC-emulator that is tuned with large climate forcings, over the use of for instance a statistical AMOC-emulator, is that it projects the point after which the AMOC collapses and switches to an off state, as this is an integral part of the physics of the Stommel model. It is clear that using an AMOC-emulator

introduces new uncertainty into AMOC projections, however, for which level of added uncertainty an AMOC-emulator is still useful is a question that is difficult to address. Another important consideration when using the AMOC-emulator is the spread in GCM climate forcing scenarios that is included in the tuning process. When using only a single climate change scenario, a better match can be obtained between the AMOC evolution given by the GCM and AMOC-emulator, however, in this case the reliability of the AMOC-emulator will quickly decrease for different climate forcings. On the other hand, one could use a large number of climate change projections in the tuning process to obtain a lesser fit for individual scenarios, but an AMOC-emulator that is applicable to a much larger range of climate change scenarios. The best strategy to be followed strongly depends on the research question in mind. The assumptions behind the AMOC-emulator presented here, limit it to projecting AMOC changes on multi-decadal and larger timescales. Therefore, the applied GCM-based climate forcings and AMOC strength time series should best be filtered to exclude high frequency variability. Moreover, an AMOC-emulator that is tuned to specific GIS melt experiments is likely not applicable to experiments in which melt water is applied to a different geographical region or with a different seasonal cycle. This is not to say that the presented AMOC-emulator framework cannot equally be applied to other sources of melt water input. Finally, many processes that are known to impact the AMOC are not considered in the AMOC-emulator, for instance the impact of winds, gyre circulation, Southern Ocean upwelling or deep water formation outside of the North Atlantic (see Sect. 1). If such processes would prove to dominate the AMOC response to future climate change, a different AMOC box model should be considered that places emphasis on that particular process."

Minor comments:-
Page 3 Line 12: Prescribed FW fluxes: F1 and F2 are tuned parameters. I would have expected these to vary as a function of the forcing/climate. What is the justification for fixing them?
This part was not sufficiently clear in the manuscript and has now been updated. The total freshwater fluxes F1 and F2 are not part of the tuning procedure, but F01 and F02 (the combined wind-driven oceanic and atmospheric meridional freshwater fluxes for the reference state are). The text should have read (line 18 page 4) "Freshwater fluxes F_{01} , F_{02} and coefficients h_1 and h_2 are included in the tuning procedure (Tab. 2)"

Page 5 Line 10: What you also fail to consider are nonlinearities between these parameters. Co-varying the parameters in Tables 1 and 2 could yield very different behaviours.
The parameter fitting method we employ, simulated annealing, randomly varies the individual parameters, thus considering (although not explicitly) both linear and nonlinear relationships between parameters. Moreover, by including Figure 6 we perform a first order test to see whether relationships exist between parameters, which indeed is the case for several of them.

Page 5 Line 30: algorith > algorithm
Thank you, it has been corrected.

Page 6 Line 4: I find the arbitrary choice of +/- 200% rather strange. What is the justification for this?
We agree that this choice is arbitrary. Our approach has been to take this arbitrary value, perform the analysis and then to analyze whether or not all parameter values that resulted from the fitting procedure were well within the +-200% range (see also figure 6). From this it was decided to keep the +-200% value. This point is clarified by adding (lines 14-15 page 7) "The appropriateness of this arbitrary range of initial parameter values is later verified by ensuring that all final parameter values are well within the initial range."

Page 6 Line 8: analogues > analogous
Thank you, it has been corrected.

Check typesetting in Tables (e.g. Table 1 column 2)
Thank you, typesetting is checked.

Table 1: (typo) dependend > dependent
Thank you, it has been corrected.

Check typesetting on Figure 8: it appears corrupted.
Thank you, typesetting is checked.

Figure 4 caption: (typo) relatvie > relative
Thank you, it has been corrected.

Figure 8 caption: (typo) calculate > calculated
Thank you, it has been corrected.

Figure 8 caption: (typo) rigth > right
Thank you, it has been corrected.