Response to anonymous Referee #2

We thanks Reviewer 2 for the interesting and extensive comments on the manuscripts. Below we will provide a detailed response to all individual comments.

1 General Comments

1.1 Introduction

The introduction does not sufficiently reflect the state of the literature on AMOC and in particularly not on conceptual models such as the Stommel-Model to study it.

Our understanding of AMOC dynamics has advanced considerably over the last years thanks to ongoing observations e.g. in the Rapid array (see Srokosz & Byrden, Science 2015). In addition a recent studies has suggest that the AMOC might already be in decline (Rahmstorf et al. 2015). While of course not directly relevant for the emulator itself, such observational findings need to be discussed in an approach that emulate AMOC behaviour over the next centuries. This should also include a discussion of atmospheric imprints on the AMOC e.g. such as atmospheric blocking events. It should also allow to assess the performance of GCMs in relation to the observational record.

We cannot agree with the points raised above. We are presenting a new modeling framework in a journal for model development. As such, we agree that some insights as to why we think a new modeling framework is needed is called for, but a discussion of observed AMOC changes, AMOC fingerprints or the performance of GCMs in relation to the observational record does not seem appropriate in this journal and is beyond the scope of this paper..

Much more important though is the discussion in relation to the emulator approach taken. Stommel type models have been used since quite some time and might be able to capture key dynamics of the AMOC (e.g. bistability). However, they at the same time have faced a lot of criticism and alternative models describing AMOC behaviour exist. This is in particular related to the relevance of Southern Ocean upwelling reflected in a conceptual model by Gnanadesikan (1999) related to changes in the pycnocline depth. A dynamic that is completely missing in the Stommel approach.

This has been explored further in conceptual models and attempts exist to unify pycnocline and freshwater-feedback dynamics. In this context, the authors should consider the work of Sijp et al. (2012) that they may find helpful.

Another question directly relating to the physical plausibility of the Stommel model relates to the relationship of circulation strength and meridional density gradient in a geostrophic ocean. The authors should consider work by Gregoy & Tailleux (2010) that present a kinetic energy approach essential providing a physical explanation for the (empirically supported) meridional density gradient outlining the relevance of the Western Boundary Current in modelling AMOC dynamics.

These comments should not be seen as undermining the Stommel model approach taken here, but they need to be addressed. In short, the authors should show motivate their approach in the light of the most recent literature.

Thanks for pointing this out and we agree that a more thorough discussion of the pro's and con's of the used Stommel model is called for. An important caveat of using a Stommel model is that Southern Ocean upwelling, the role of Southern Hemisphere mid-latitude winds and other processes are neglected, a point that we have added to the discussion of this manuscript. Our choice to use the Stommel model was driven by two considerations. Firstly, to our knowledge no unified simple AMOC model exists and as such it is not clear if other models are better or worse than the Stommel model in relating surface temperature and freshwater flux changes to the AMOC strength. Secondly, the Stommel model allows for rather straightforward inclusion of temperature and freshwater forcings based on GCM simulations, while for other models like the ones mentioned above it is not clear to us how this could be done. Finally, it is important to note

that we did not set out to construct a new simple model that describes the main dynamics of the AMOC, but rather to use an existing model and build a framework around it that can easily be applied to GCM climate change and AMOC projections.

Following the above, we have updated the introduction to read (lines 14-23 page 2) "At the center of our approach is the assumption that changes in AMOC strength are linearly related to changes in the Atlantic meridional density contrast. Since Stommel (1961) a large number of studies have provided evidence for an important role of the Atlantic meridional density contrast in driving AMOC changes (e.g. Rahmstorf, 1996; Gregory and Tailleux, 2011; Butler et al., 2016). Nonetheless, it neglects several important processes, like the role of Southern Ocean upwelling, winds and deep water formation (e.g. Gnanadesikan, 1999; de Boer et al., 2010) and a unified theory describing the fundamental mechanisms driving and sustaining the AMOC lacks to this date (Lozier, 2010). Using a Stommel model to emulate AMOC changes driven by surface temperature and freshwater forcings seems appropriate in the light of present-day knowledge and the apparent leading role of surface buoyancy changes in simulated future AMOC weakening (IPCC Climate Change, 2013). Moreover, the model is easy to use, interpreted and can be forced directly with GCM-based forcing fields. Nonetheless, the processes that have been omitted and the simplicity of the model should be considered when interpreting the results."
To the discussion section we have added (lines 20-24 page 12) "...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."

1.2 The Emulator model
Here, the work dominantly builds on a previous model by Zickfeld et al. (2004) plus a representation of the Bjerknes feedback. It does however not become sufficiently clear, why this addition will represent a substantial advancement. The authors show the differences in Fig. 9 and describe that this will represent a negative feedback on the AMOC dynamics. But it's not clear, if Figure 9 shows two sets calibrated individually (with and without atmospheric feedback) or just from the optimal parameter set with this feedback switched on and off. Therefore, I cannot judge if the conclusion drawn by the authors on the importance of the effect are due to their specific parameter set or not.
It would add merit, if the authors could show that the model including the Bjerknes effect will in the end outperform the no-atmospheric feedback model in the fitting procedure. This would also justify, why there model is actually better than the one presented in Zickfeld (2004).
Thanks for providing this comment. Firstly, the main improvement with respect to the Zickfeld et al. (2004) model is that we provide a framework that allows one to use limited GCM AMOC and climate projections to tune an AMOC-emulator in order to perform an uncertainty analysis. The Zickfeld et al. (2004) approach used a full ~20.000yr long hysteresis simulation to tune their emulator, not feasible for most IPCC-type GCMs. Moreover, they did not force their emulator with GCM-based temperature changes or consider inter-GCM differences in regional temperature changes. Those are the features we see as most important changes with respect to earlier work, a view that is now better reflected in the introduction of the manuscript by (line 23 page 1 to line 3 page 2) "To this end we developed an AMOC-emulator framework. It entails a simple box model that uses physical relationships to represent the most important mechanisms and feedbacks that govern the AMOC's response to changes in regional surface temperatures, freshwater fluxes and enhanced melting of the GIS. The AMOC-emulator can be forced by temperature and melt water fluxes from any GCM, and using AMOC time series the free

parameters of the box model are tuned to mimic the GCM's AMOC sensitivity to future climate change." and in later on in the introduction it reads (lines 11-13 page 2) "the approach described here is designed specifically to allow future studies in which a limited number of climate projections from multiple GCMs, limited in the simulated forcing scenarios and simulation length, to be combined into a Bayesian framework of century time-scale probabilistic AMOC projections."

With respect to the added stabilizing Bjerkness feedback, it indeed appears from Figure 9 that it's impact is limited. Figure 9 shows results for the same parameter sets with this feedback switched on and off, allowing for a direct investigation of its impact. Nonetheless, we deem the model including this feedback more realistic. Moreover, the effect is non-negligible (lines 1-6 page 11) "The impact of including atmospheric meridional heat transport is a small, but non-negligible ~1Sv strengthening of the control state of the AMOC (not shown) and, more importantly, a slightly lower sensitivity to changes in radiative forcing and GIS melt (Fig. 9). This confirms our understanding of atmospheric meridional heat transport acting as a negative feedback to AMOC changes. The simulations with the atmospheric feedback included have on average a stronger AMOC by 8.1±1.9% (μ ± σ; calculated over all 10 best fits and over all five forcing scenarios)."

Furthermore, the model includes 5 atmospheric boxes. Why are 5 boxes needed and not 3 to resolve the meridional heat transport? I think that can be easily motivated and maybe I missed it. Maybe it's worth considering to restructure the approach by moving subsection 2.3 further up to discuss the setup of the atmospheric forcings.

Including high latitude atmospheric boxes allows us to have a closed energy budget and more realistic meridional atmospheric heat transport.

Thanks for the suggestion to rearrange this section. We have accordingly switched sections 2.2 and 2.3.

In this context, the authors should also reflect on the limitations of the model to reproduce transient AMOC changes that relate to the assumption of well-mixed density within the boxes. This might be in particularly relevant in relation to the Greenland freshwater input. Clearly, this represents an over-simplification and may substantially limit the capabilities of this approach to emulate transient behaviour (I'll further comment on this below).

Thanks for pointing this out. We fully agree that a box model can never resolve the complexities of the interaction between Greenland meltwater and the ocean. We have experimented with an additional tuning parameter to include the GCM dependent 'efficiency' of Greenland meltwater to impact the density of the North Atlantic ocean box, but decided against it since the current 7 tuning parameters already allow for sufficient freedom to tune an AMOC-emulator towards the AMOC sensitivity of a specific GCM.

1.3 The tuning to complex model output

In the manuscript, the model is tuned to an EMIC model UVIC. I think that's generally no problem, but somehow contradicts the initial claims by the authors that this emulator could now be used to run larger ensembles. What is it exactly that the emulator provides that cannot be done with an EMIC?

In general terms, the strength of an emulator is it's capability to include projections from a range of different models. We have AMOC projections for several CMIP5 models, why is it not applied to those? In addition, there are the AMOC sensitivity studies by Gregory et al. (2005) and Stouffer et al. (2006) that would provide enough runs to calibrate the model. Why isn't it applied to those runs?

In addition, the authors mention the AMOCMIP project. Can the emulator be applied to the AMOCMIP output?

Thanks for these comments. It has become clear from the comments of the different reviewers that the aim of this manuscript is not sufficiently clear and we have changed the abstract, introduction and summary sections to improve on this. In this manuscript we want to describe a modeling framework that allows one to use limited GCM output to tune and force an AMOC box model that can in turn be used to perform uncertainty analysis. It is not the aim of this manuscript to provide future AMOC projections or provide such an uncertainty analysis. See also the responses provided above.

I checked the project homepage and understood that the AMOCMIP will explictly resolve different Greenland basins separately. Is that correct? If so, and following recent findings that it actually matters a lot for North Atlantic dynamics where the freshwater is actually applied, will this emulator be the best tool to reproduce these dynamics? Or should it maybe consist of a subpolar (Labrador Sea) and North Atlantic box? And/or should conceptual models of convection in marginal seas e.g. by Spall (2004) and Straneo (2009) be integrated?
Thanks for this question. Indeed the aim of the simulations in AMOCMIP is to provide 'realistic' Greenland melt scenarios and to apply those to IPCC-type climate change projections. This includes explicitly resolving spatial and seasonal differences in the meltwater flux. Such details cannot be captured by the AMOC-emulator. However, as described above, by tuning the AMOC-emulator to the forcings and AMOC projections of a specific GCM, we take into account the inter-GCM differences in the sensitivity of the AMOC to changes in temperature and freshwater.

1.4 Results
I've to admit I'm not impressed by the capabilities of the emulator in reproducing the model outcome. As apparent from Fig. 7, the emulator is systematically underestimating AMOC reduction for RCP4.5 and RCP8.5 no melt, while then over-estimating it for RCP8.5 plus GIS (maybe due to non-linearities kicking in here and timescale issues discussed above?). The authors discussion of this simply stating that "It is, however, to be expected that a box-model does not completely capture the behavior of the AMOC as simulated with a higher order climate model" is clearly insufficient. In particular, as there have been much simpler AMOC emulators around that actually perform much better (also and in particularly an AMOC recovery, e.g. Schleussner et al. 2014).
Thanks for pointing this out. We agree that there are limitations to the AMOC-emulator and that because of choices that have been made, it appears that previous emulator perform better. There are, however, a number of important things to take into consideration. Firstly, one could perform the tuning on a single GCM forcing scenario and the result will be a closer fit between GCM and emulator AMOC. However, when choosing that approach, one is limited to applying the emulator to forcing scenarios close to the one used for tuning. By using a larger number of scenario in the tuning process, the emulator can be used to test the AMOC for a much larger range of scenarios, albeit at the cost of having larger discrepancies between GCM and emulator. We have added text along these lines to the manuscript (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 much 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.”

Another issue to consider is the use of physics-based or statistical emulators. With a statistical AMOC emulator one could obtain better agreement between GCM and emulator, however, such a model cannot be used to extrapolate for larger forcings. With a physics-based AMOC-emulator one can have more confidence in the response to large forcings, for instance a complete AMOC shutdown, notwithstanding that also in this approach the uncertainty is likely to increase for forcings further away from those used for tuning. This is discussed in the final section of the updated manuscript (lines 1-8 page 12) “Overall, the predictive power of the AMOC-emulator is reasonable when one considers the simplicity of the AMOC box model, but 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 larger 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 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.”

The apparent oscillations in the emulator arising from a “too direct response” of the emulator towards multi-decadal surface temperature oscillations also merits more discussion.

The origin of the oscillations is already mentioned in the manuscript (lines 17-18 page 9) “The UVic-based surface temperature evolution exhibits multi-decadal to centennial oscillations that result from global climate variability originating from the Southern Ocean” and we do not deem it necessary to discuss the resulting AMOC osculations in much detail as they are a feature of the forcing based on this particular climate model and not a feature of the AMOC-emulator. In the discussion section we have added some words describing the kind of temperature forcings that are appropriate to use (lines 15-17 page 12) “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 be filtered to exclude high resolution variability.”

It is even worse for the predictions in Fig. 8. First of all, the figure is not well-labelled (no y-axis labeling, panels not clearly distinguishable, and what is given by the numbers 5,1,5?) and that there is no such thing as a top-middle panel for only two boxes.

The conversion of the figure must have gone wrong at some point because the points raised by the reviewer are difficult to understand looking at the figures we have in the manuscript. We will ensure that the figures are correct in the next version.

For none of the panels, the model actually captures key features. It fails to capture the bumps in the top-left and bottom right, and for the two other panels, it gets it wrong completely. I cannot agree to the author’s conclusions that “Overall, the predictive power of the AMOC-emulator is good for reasonable forcing scenarios when one considers the simplicity of the model.”

We don't agree with the general notion given by the reviewer. Firstly, the AMOC-emulator is not designed to emulator decadal AMOC fluctuations as simulated by the GCM. As mentioned in the manuscript, those results from internal climate variability mostly originating from the Southern Ocean and it is not to be expected that the emulator captures those. Moreover, the focus of the AMOC emulator is on multi-decadal to multi-centennial scales, something that is now specifically

mentioned in the discussion (see reply above).
Furthermore, 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' an 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."

1.5 Summary
Generally, I miss a section that reflects on the limitations and short-comings of the approach taken, given in particular the apparent limitations in reproducing the EMIC results. Furthermore, an outlook of where this can be applied and what it specific strengths are compared to other approaches should be included.
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 resolution 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."