Response to anonymous Referee #3

We thanks Reviewer 3 for the interesting and extensive comments on the manuscripts. Below we will provide a detailed response to all individual comments.

2 The model formulation
2.1 On using a conceptual model
I don't understand why this kind of ad-hoc box model is preferred. One would resort to Stommel's box model if one knows almost nothing about AMOC. Stommel's model was just conceptual, whose sole purpose is to get very rough ideas on how AMOC might work, and is not designed for the kind of quantitative modeling that the present authors pursue.

Stommel's box model was presented in 1961, and we know a lot more today. I would think that a simple dynamical model like Gnanadesikan's (1999, Science vol. 283, pp. 2077–) is much better because even uncertain parameters are based on (that is, constrained by) clearly-identified dynamics. In contrast, some of the present authors' "parameterizations" aren't adequately defendable; they are based on hand-waving arguments. For example, how can one defend the parameterization that the AMOC strength is proportional to the interspheric surface-density difference? We know a lot better than that.

Perhaps even better, the models of Schloesser et al. (2014, Prog. Oceanogr. Vol. 120, pp. 154–) and McCreary et al. (2016, Prog. Oceanogr. vol. 123, pp. 46–) provide constraints among integral quantities, such as AMOC strength, thermocline depth, and meridional density difference, and hence can be utilized as a "box model". Those constrains are derived as solutions to dynamical equations rather than assumed on the basis of hand-waving arguments.

In short, I don't see any advantage today in utilizing an old conceptual model for quantitative prediction.

2.2 AMOC proportional to interhemispheric density difference?
The authors says that "the assumption that the meridional Atlantic density contrast between the North Atlantic and the South Atlantic is the first order driver of the AMOC" is debatable, but I think that's off the mark. The current wisdom is that the Southern-Ocean winds (and perhaps vertical diffusivity) are the first-order driver. They cite Butler et al. (2016) as the other side of the "debate" but Butler et al. do not argue that the surface meridional density gradient "drives" the AMOC. The just use density integrated twice in the vertical as a "diagnostic" of the AMOC.
It is clear from ocean GCM studies that the meridional density gradient is not the first order driver of AMOC. When the sea-surface density is restored toward a prescribed profile in an ocean-only GCM and windstress is changed in the Southern Ocean, the AMOC strength changes roughly linearly to the windstress. See Toggweiler et al. (1995, Dee-Sea Research vol. 42, pp.477–) and the series of studies that follow. This is evidence enough that the interhemispheric density difference does not drive AMOC.
Of course, this evidence is based on ocean-only models, and it is possible that the interhemispheric density difference is correlated with the AMOC strength through atmospheric feedbacks, but to use a one-to-one correspondence like (1) needs justification based on atmosphere-ocean coupled dynamics.
By the way, I found that Butler et al. (2016) still use the traditional hand-waving parameterization px=Lx / py=Ly. See Schloesser et al. (2012, Prog. Oceanogr. Vol. 101, pp. 33–) for a better parameterization based a lot more on dynamics.
Thanks for describing in detail your view on conceptual models and the drivers of the AMOC. Indeed there is a wide variety of models of different complexity that describe (some aspect of) the AMOC, some more based on dynamic considerations than others, some including parameters

that are more easily constrained by observations than others. In this study we have chosen to use one of the simplest and most established of such models around which to build our emulator framework, this is done for various reasons: i) it can easily be forced by temperature and Greenland melt outputs from a GCM; ii) it is easy to implement; iii) very fast to run; iv) easy to understand/diagnose the results; v) has several free parameters that can be tuned towards the behavior of a GCM in terms of the sensitivity of the AMOC to changes in heat and freshwater fluxes. We acknowledge that a different model could have been chosen, perhaps one that would turn out to perform better, however, we are not able try every single one of them so that remains unknown.
We do agree that when using the Stommel model to make the connection between changes in temperature and freshwater and changes in the AMOC strength implies that certain processes are not taken into account, like the role of changes in Southern Ocean winds, upwelling and deep water formation when projecting future changes in the AMOC. This is now more clearly described in the manuscript with the following in the introduction (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, this model 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." and the following in the discussion section (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."

However, we do not agree with the reviewers view that it is now well established how the AMOC works, what drives future AMOC changes and that the Stommel model has been shown to be wrong. Although we are not experts in the field of conceptual AMOC models and we do not plan to provide a discussion of all literature on the subject, we are not aware of any such consensus in the field. As becomes clear from the references cited above by the reviewer and by the references in the manuscript, many things are still debated and the Stommel model, that is the relationship between AMOC strength and meridional density differences, is still widely used to discuss the mechanisms and stability of the AMOC in complex GCMs.

3 Tuning and validation
3.1 GCMs for tuning
Why aren't multiple coupled GCMs used to tune the parameters? Do the authors recommend that the AMOC-emulator be tuned differently for each model?
The present manuscript aims to describe a modelling framework that can be used in combination with any GCM. Since the AMOC sensitivity to changes in heat and freshwater fluxes is strongly

GCM-dependent, we indeed recommend that the AMOC-emulator is tuned separately for every GCM to reflect these differences.

The emulator is based on equations that represent physical processes in the real world. Then, if at all possible, the parameters should be tuned on the basis of reality. Granted that there is not enough data for the deep ocean. Then the second best thing is the publicly-available collections of coupled GCM runs. I think that studies have indicated that a multi-model ensemble is usually better than a single model to mimic reality. So, the tuned parameters would be more likely better if they are based on multiple models.
We agree that ideally one would tune the free parameters within bounds provided by observational data. However, since it is such a highly simplified and conceptual model, the parameters are not easily obtained from observations even if that data would be abundant. For future applications, we indeed recommend that a large number of AMOC-emulators is used, tuned towards different GCMs, in order to provide a range of parameter values that is hopefully as close to reality as possible.

3.2 Variables for tuning
The variables (salinity, ocean temperature, etc.) of the emulator should be compared with those from the GCM. It is possible that the state of the emulator is very different from that of the GCM even when the AMOC strength m agrees.
On a more basic note, have the authors made sure that all the variables of the emulator take reasonable values? I don't think it would be okay if, say, salinity takes a value of -100 psu even if the value of m is reasonable!

If the atmosphere and ocean states aren't realistic, how can we trust the emulator?
One approach to cope with this problem would be to include other variables than AMOC strength in the cost function. Another approach would be to compare various variables between the runs of the tuned model and those of the GCMs (part of validation). I think both are necessary.
Thanks for providing this interesting idea. First of all, we fully agree that one has to make sure that the values for salinity and temperature in the different parts of the box model are realistic. This has been checked. Indeed one could include the comparison between GCM and AMOC-emulator temperatures and salinity in the tuning procedure. However, there does not seem to be a reason why an AMOC emulator that has temperatures and salinities that are closer to the GCMs, would perform better in terms of the AMOC behavior, in fact the tuning of the parameters would very much be steered by getting the right temperatures and salinities and less so by the AMOC, thus likely deteriorating the capabilities of the AMOC-emulator in terms of mimicking the AMOC in the GCM. Since our focus is purely on providing a computational efficient method to provide uncertainty estimates of GCM AMOC projections, we deem the current approach most suited.

3.3 Models for validation
Moreover, I think the tuned emulator should be validated against another set of different models. Otherwise the validation isn't robust.
As discussed above, the sensitivity of the AMOC differs from one GCM to the next, therefore the AMOC-emulator should be tuned separately for every GCM and thus not be validated with results from a different GCM.

I also wonder if an ensemble of runs are necessary for the GCM (UVic) for tuning and validation. For *example, there is only one run for each case in Figure 8, but doesn't the AMOC strength differ from realization to realization? I don't know how chaotic the GCM is (because it uses a low-degree-of-freedom atmospheric model), but isn't the reality more or less chaotic?*

Thanks for this question. This is indeed an interesting point. There are two points to this question. Firstly, like mentioned by the reviewer, Uvic is a low resolution GCM that is know to have less variability than higher resolution and complexity models. However, for high and low resolution models, the forced response of the AMOC to strong changes in temperature and freshwater is much stronger than internal variability, which is on the order of 1Sv on decadal and longer timescales. There are indications from observations that in the real world AMOC variability is substantially larger, however, those time series are currently too short to make any robust statements about this and, furthermore, this does not impact our GCM-based methodology.

4 Forcing

I may be missing something, but it's not clear to me what forces the emulator. The solar flux S seems constant in time (Table 1), but then how is the increase in green-house gas represented?

Thanks for this comment. The section on AMOC-emulator forcings describes the way we use GCM output to force the AMOC-emulator. Indeed the solar flux is constant in time. The GCM regional temperature changes, including the impact of increased green-house gasses and all feedbacks, are included in the AMOC-emulator through the so-called 'total atmosphere effect' parameter. A parameter of the atmosphere, "a temporal and spatial varying parameter that effectively combines atmospheric emissivity, the greenhouse effect and all other processes included in a GCM that cause regional temperatures to differ from global temperature changes". This way it is ensured that the ocean is regionally forced by almost the same temperature changes as in the GCM. Furthermore, the AMOC-emulator is forced by the GCM-based F_{GIS} forcing.

If I understand it correctly, forces the model (equation 19) toward one particular GCM solution, but wouldn't it damp the emulator's variability? especially when the emulator is to simulate a state that is very different from the GCM state used for ? Doesn't this amount to building the solution into the simulation?

The GCM-based regional temperatures that are used as forcing, are used in the tuning phase. If one would like to use the AMOC-emulator to simulate a temperature forcing that is different (as is done in the section "Predictive power of the UVic-based AMOC-emulator"), the changes in the 'total atmosphere effect' parameter can be changed accordingly.

5 Minor points: math notations

5.1 Arrays The authors define boldface math symbols to mean "arrays", but I recommend avoiding this unconventional convention. For example, a multiplication of two "arrays" can mean several different things in conventional mathematics. Equation (13) includes the multiplication of the arrays S and p, which is meant to represent (S1p1; S2p2; : : :), which is hardly conventional. K and Ta have the same problem in (14). Also, equation (11) includes 1=z, by which the authors mean (1=z1; 1=z2; : : :), but which is not widely used in math.

5.2 Subscripts

I recommend using an upright font for multi-character math symbols such as "start" and "gcm"; or avoiding them. In particular, the subscript "it" looks as if it represented two subscripts i and t. I recommend using a single-character subscript, such as " i1" or if you insist on multi-character subscript, you may want " it□1" using an upright font.

Thanks for pointing out the issues with math notations and subscripts. Indeed the notations that are used are confusing and in some places wrong. We have updated the manuscript following the

recommendations of the reviewer.

6 Point by point comments

Some of the following comments support my arguments above, some raise other concerns, and others point out minor, mostly editorial, problems. I wrote many of them as I read the manuscript for the first time, and as a result, they include some redundancy. I leave them as they are, because they often reflect difficulties or problems the reader may encounter as she reads the text.

6.0.1 p. 1, l. 19:

"due to climate sensitivity, polar amplification, GIS melt and model dependent sensitivity of the AMOC . . . "—I'm confused. Doesn't "climate sensitivity" include all the remaining items in the list? Why is it listed in parallel with the rest?

Thanks for pointing this out. Indeed taking climate sensitivity as the the global temperature change for a doubling of CO2, this term is mostly taken to include all other processes that are listed. However, in the model world this is not always the case. For instance, ice-sheet-climate interactions are mostly not considered and thus GIS melt not taken into account. Moreover, one can have the same climate sensitivity, but different polar amplification and the latter can result in a different AMOC response because of the sensitivity of the AMOC to latitudinal temperature differences. Why we prefer to list all of them separately in this context, is because GCMs differ in all those terms, and all those uncertainties can be tested individually with the AMOC-emulator.

6.0.2 p. 2, l. 5:

"(Rahmstorf and Willebrand, 1981)"—As the reference list indicates, this should probably be Rahmstorf and Willebrand (1995).

Thanks for pointing this out, it has been corrected.

6.0.3 p. 2, ll. 5 & 31:

"the so-called Bjerknes feedback"—Probably this is because I'm not much versed in climate research, but isn't the "so-called Bjerknes feedback" restricted along the equator? A direct overturning circulation occurs connecting cooling in the eastern Pacific, say, and warming in the western Pacific only along the equator, where the Coriolis force vanishes, and the surface windstress associated with this zonal overturning circulation enhances the upwelling of sea water, which further lowers the sea-surface temperature in the eastern Pacific—a positive feedback, which is "the so-called Bjerknes feedback".

The authors cite Rahmstorf and Willebrand (1995) for "the so-called Bjerknes feedback", but Rahmstorf and Willebrand proposed a negative feedback due to heat transport within the atmosphere, I think.

Indeed this topic is somewhat confusing as indeed the Bjerknes feedback often refers to the feedback described by the reviewer, but this term (or Bjerknes compensation) is also used to describe the compensation between meridional heat transport by the ocean and atmosphere as first proposed by Bjerknes in 1964. The latter indeed provides a stabilizing or negative feedback to AMOC changes.

6.0.4 p. 2, l. 7:

"tuning a numbrer of free parameters"—They aren't "free". They represent specific physical processes and hence must be ultimately determined by physics, even though it's in practice difficult to derive their values purely from physical principles.

The term 'free parameter' is used here to make the distinction between parameters that are prescribed and those that are not, or in other words, those that are part of the tuning process and

those that are not. All of them represent physical processes and should (and often are) determined from observations.

6.0.5 p. 3, l. 12:

Why is F prescribed? I would expect it to change according to the state of the climate system. What do IPCC-class coupled GCMs say about the change in F under global warming, for example? . . . but, later in the text, the authors say that F is related to the global atmospheric temperature (equation 12). So, it's not prescribed after all.

Thanks for pointing this out. The manuscript is not sufficiently clear on this topic and changes have been included for clarification. F_i consists of two parts (equation 12), a part that is fixed in time (F_{0i}) and a part that is a function of global temperature changes ($h_i \Delta T_{glob}$). Both F_i and h_i c are part of the tuning process.

6.0.6 Equations (3)–(10): What does this "" mean? Is it a typo for "@"?

Thanks for pointing this out, we have updated the manuscript for clarity.

6.0.7 Equation (11):

State whether z's are fixed, and if so, give their values here or refer the reader to a table or something.

The formulation now includes a clear notation showing that *z* is a function of *i* and and a reference to Table 1 is given.

6.0.8 Equation (12):

Tglob should be defined. (How is it computed from Ta?)

This has been rewritten to read "global atmospheric surface temperature anomalies".

6.0.9 Equation (12):

Give F0's their values here or refer the reader to a table or something.

This line now reads "Freshwater fluxes F01 , F02 and coefficients h1 and h2 are included in the tuning procedure (Tab. 2)."

6.0.10 Equation (13):

I may be mistaken, but it seems that the Ta4 is the only nonlinear term. Doesn't it make sense if this term is linearized around a mean state?

Yes it could be, but we prefer to keep the current form.

6.0.11 Equation (13):

The solar flux S is a confusing notation. By the authors' own convention, S =(S0; S1; S2; S3; S4), which uses the same symbols as salinity.

Thanks for pointing this out, this is indeed confusing. The notation has been changed to read I_i.

6.0.12 Equation (13):

The solar flux S should be discussed right below equation (13). Does it depend on time? Table 1 suggests that it's constant in time but that should be stated explicitly. So, does the emulator solves only for annual averages?

The first line after equation 13 now reads "where σ is the Stefan Boltzmann constant and I_i and α_i the latitude dependent yearly mean incoming shortwave radiation and planetary albedo, respectively (see Tab. 1 for details)".

6.0.13 Equation (13):

Define this symbol precisely. (But I don't recommend this notation because a gradient of an array is a strange mathematical entity.)
Thanks for pointing this out. It should have been a Delta symbol.

6.0.14 Equation (14):
What does this "" mean? Is it a typo for "@"?
This has been corrected.

6.0.15 Equation (14):
State that Ha is defined to vanish at the northern and southern ends of the northernmost and southernmost boxes. (I guess they are so defined, right?)
The following has been added "Meridional heat fluxes are assumed zero at the northern and southern boundaries of the domain."

6.0.16 p. 5, l. 15:
I guess we need some discussion on other possible sets of tuning parameters. We have a vast range of possibilities. Then, how have we settled on these seven parameters? Have the authors tried other combinations of parameters?
We deem this discussed by the line (lines 25-27 page 6) "This selection of parameters is somewhat subjective, but it proved a good balance between, on the one hand, sufficient degrees of freedom to tune the AMOC emulator's behavior towards that of a specific GCM and, on the other hand, the efficiency to find optimal parameter fits."

6.0.17 p. 5, l. 15:
F and h are related by equation (12), and so cannot be determined independently. Moreover, if you tune F, you can forget about equation (12) and don't need to consider h.
We apologize for the confusion that has arisen because of errors in the notation. This has been corrected in the text and figures. F_{0i} and h_i are the parameters used in the tuning process. The former giving the steady state meridional freshwater transport by both the atmosphere and the wind driven ocean part, while the latter controls the changes in atmospheric transport as a function of global temperature changes.

6.0.18 Equation (15):
Why try to optimize m alone? It's conceivable that widely different states have similar m values. Because we have other variables like salinity, we could choose better sets of parameter values, if we include other variables in the cost function, couldn't we?
As discussed above, we don't think that including more variables in the tuning process would lead to a better behavior of the AMOC-emulator in terms of its capacity to mimic the GCMs AMOC sensitivity to changes in temperature and freshwater.

6.0.19 Equation (15):
I may well be mistaken, but it seems that the differential equations are linear in the tuning parameters and if so, the optimization problem on the new cost function ...is a quadratic function of the parameters and can be solved analytically, I think.
The optimization of the parameters cannot be solved analytically as the system is non-linear and includes 7 parameters that influence each other.

6.0.20 Equation (16):
The notation "pstart(1□z)" is confusing because it looks as if pstart(z) were a function of z. Vectors

customarily come after scalars, as in “(1 □ z)pstart”
It has been changed.

6.0.21 Step 1:
I don’t understand why we have to repeat this step. Why not choose values that are within the ranges in Table 2 in the first place? We can use a random variable whose PDF is uniform over the specified range for each parameter, can’t we? I mean, if (1 □ z)p1 is below the range, we can just use p1min for the lower bound; that is, we can use U(max((1 □ z)p1; p1min); min((1 + z)p1; p1max)) without repetition. The same argument holds for the last part of Step 3.
The solution given by the referee will not always give the same results. More specifically, the p1min and p1max values would be used much more often than other values in case random values from outside of the range are often picked. Another solution would be to split z into z_{min} and z_{max} and adjust those values for every parameter to ensure that the randomly picked values are never outside of the imposed ranges. We don't think it matters which solution is picked.

6.0.22 Equation (17):
I may be missing something, but shouldn’t (16) and (17) be written in parallel forms? If we write (1 □ z)p for (16), then we should write (1 □ it)p for (17). If we write p □ itp for (17), we should write p □ zp for (16). For a moment, I was confused with (17).
Thanks for pointing this out, using the same notation for both equations indeed improves readability.

6.0.23 p. 7, l. 16:
I think that efforts should be made to narrow the range of the parameter values. If parameters are widely different even though the cost function is similar, doesn’t that suggest that the parameters aren’t well tuned?
What about comparing variables other than m between the emulator and the GCM? Wouldn’t that tell which parameter values are bad?
It seems that the authors have forgotten that there is only one reality.
Firstly, we assume the reviewer is pointing towards 'GCM reality' in this comment, since we do not aim to work towards a single parameter set that provides the closest resemblance to the real world AMOC, no matter how much we would like to do so. However, also when we are talking about using an emulator to mimic the complex AMOC behavior in a GCM, we do not expect that there is a single parameter set that provides the perfect match between GCM and emulator; because of the highly simplified nature of the emulator, it cannot be determined which parameter set is closest to 'GCM reality'.

6.0.24 p. 7, l. 22:
What is “RCP”? (I may have missed its definition given in the text.) Because how is determined is important, it may be helpful to give a bit more information here.
This line reads “RCP4.5 and RCP8.5 (Representative Concentration Pathways; Meinshausen et al., 2011)”, which we deem sufficient information, especially since in the context of testing the AMOC-emulator, the exact imposed climate forcings are only of secondary importance.

6.0.25 Equation (19):
Does the model really use the full time-series of Tgcm? Or is that a long-term mean? State clearly how is Tgcm defined. If the emulator uses the full time-series, it may not be appropriate for other models or for other scenarios.
Thanks for pointing this out. T_{GCM} can in principle be of any temporal resolution. However, the

model aims at resolving AMOC changes on decadal and longer timescales and as such, including high frequency variability in the temperature forcing could lead to misinterpretation of the results. We have added a discussion point in the final section of the manuscript to clarify the strengths and weaknesses of the AMOC-emulator (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. ".

6.0.26 p. 8, l. 2:
"Note that the temperature forcing files need to be interpolated onto the temporal resolution used in the atmospheric component of the AMOC-emulator"—Awkward in several counts.
1. The interpolation draws the attention of the reader as if it were something noteworthy. Perhaps the results are sensitive to the method of interpolation? The reader would wonder.
2. Is the fact that the GCM data are saved in files noteworthy? (I mean, why mention the files at all?)
3. Despite this cautious tone, the interval at which the GCM data is saved is not indicated.
If the result is sensitive to the interpolation, give more details. If not, what about just saying, "The GCM variables are saved at an interval of XXX hours and interpolated on to the time steps of the AMOC-emulator", something along the lines.
The line has been removed.

6.0.27 p. 8, ll. 5–6:
A similar problem. If interpolation is so noteworthy, give more details. If it's not so big a deal, just say, "the GIS melt forcing is interpolated. . . ." instead of "Note that the GIS melt forcing needs to be interpolated. . . ."
The line has been removed.