Reply to Anonymous Referee #1

Thank you for these constructive comments and suggestions that have led to a clear improvement of the manuscript. We have addressed all your comments and suggestions (repeated in boldface), as detailed below. The full references of the publications referred to but not found below, are found in the manuscript.

Page 3, line 30: I do not understand the meaning of "a vehicle for advanced training of Earth System researchers". Trying to keep the same ideas in the sentence, I think that it would be more appropriate to write "an advanced tool for Earth system researchers".

The suggestion of the referee improves the text and we will change the text accordingly.

Page 7, line 14: change "determmined" for "determined".

This has already been corrected in the typesetting/proof-reading process of the discussion paper.

Page 9, line 20: I suggest to change "for layers to exist that are unstable" for "for layers that are unstable to exist".

We partly agree to the suggested reformulation but in order to ensure that it is clear that stability is measured with respect to potential density, we suggest changing "… allowing for layers to exist that are unstable with respect to potential density." with "… allowing for layers that are unstable with respect to potential density to exist.".

Page 9-10, section 2.4: the new TKE model that you've chosen for NorESM seems to simulate a mixed layer depth in high latitudes that is more in agreement with the observations than two previous TKE options available in MICOM. This is well explained in the text but it would be highly interesting to illustrate this improvement on a figure.

We believe the modified TKE model is an important improvement of the ocean component of NorESM but, in our experience, so are modifications done to e.g. the evaluation of the pressure gradient, eddy diffusivity, and diapycnal mixing. Thus, we feel it would be somewhat arbitrary to focus on the TKE model development in this manuscript. Further, all results in this paper are devoted to CMIP5 experiments, thus results from sensitivity experiments related to specifics in one component of the NorESM might be confusing to the reader and make the length and scope of the paper excessive. Since there is no published description of the recent developments of the ocean component, we did feel it was necessary to provide a fairly detailed description of this model component. However, we think an even more detailed description along with an evaluation of all the recent development should be published and we plan to do so in a separate article focusing on the ocean component only. This will then include a more thorough assessment of the impact of changing the TKE model (plus modifications to the pressure gradient, eddy diffusivity, diapycnal mixing, etc.).

For your information Fig. 1 shows the impact on the mixed layer depth (MLD) of modifying the TKE model from Gaspar (1988) to Oberhuber (1993) in fully coupled NorESM experiments with indentical initial conditions as for the pre-industrial spinup described in the paper, but with constant present day aerosol emissions and greenhouse gas concentrations. The NorESM results are monthly means for years 40-49 after initial condition. Figure 1 shows that the bias of MLD at high latitudes in late winter is reduced with the new TKE

model. Note that the NorESM version shown here is as earlier version compared to the CMIP5 version presented in the paper, but the results should be valid for this newer version as well. The Oberhuber (1993) model is extended with a parameterization of mixed layer restratification by eddies (Fox-Kemper et al., 2008) but the main reduction in the MLD bias is due to the different TKE model.

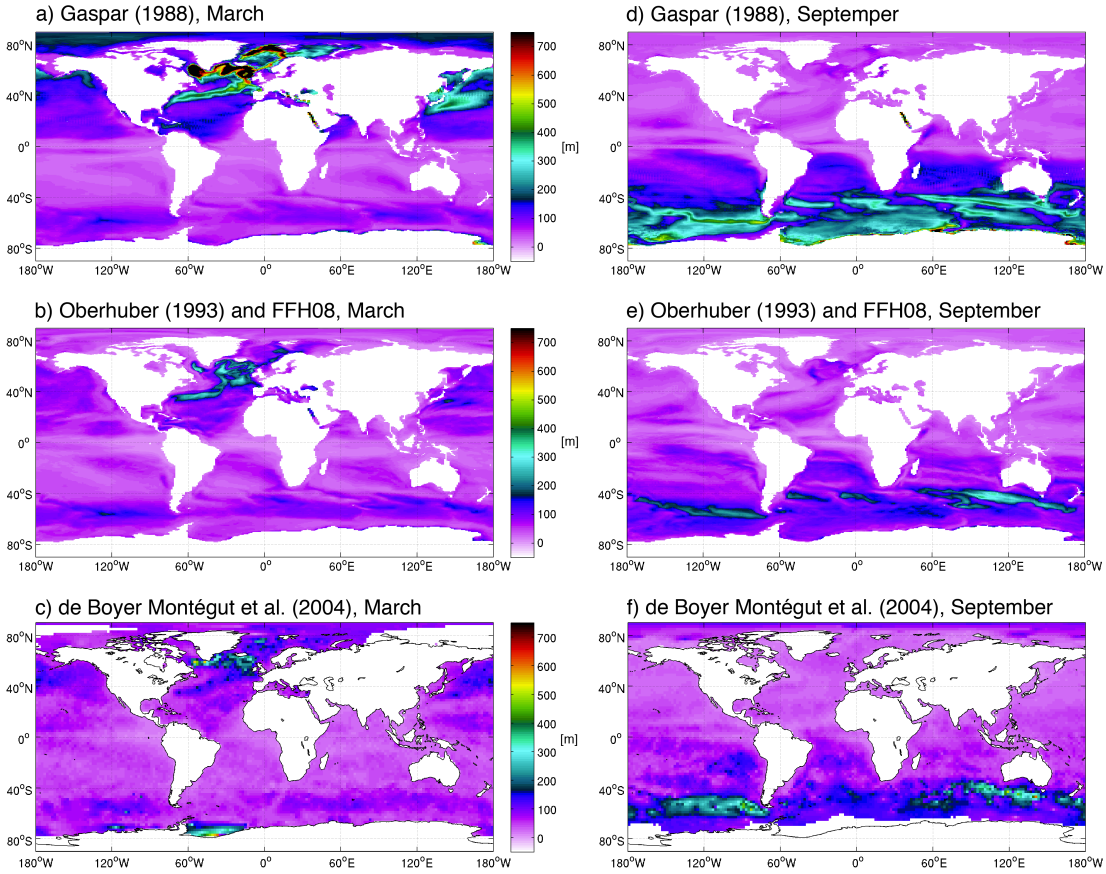

Fig 1. Monthly mean MLD for March (left panels) and September (right panels) for NorESM with Gaspar (1988) TKE model (upper panels), TKE model based on Oberhuber (1993) extended with a parameterization of mixed layer restratification by eddies (Fox-Kemper et al., 2008) (middle panels), and the climatology of de Boyer and Montégut et at. (2004) using the ΔT = 0.2 K criterion (lower panels).

References

de Boyer Montégut, C., Madec, G., Fischer, A. S., Lazar, A., and Iudicone, D.: Mixed layer depth over the global ocean: An examination of profile data and a profile-based climatology, J. Geophys. Res., 109, doi:10.1029/2004JC002378, 2004.

Page 10, last paragraph: I suggest putting this last paragraph describing the grid in second position (after the first paragraph) in section 2.4.

We will follow this suggestion and move the paragraph.

Page 12, line 1: change "CAM-Oslo" for "CAM4-Oslo"

In the paragraph where you suggest changing model name, we are discussing some cloud

micro- and macro-physical parameters that have been adjusted in CAM4-Oslo compared to values used in CAM4. It is stated that in CAM4-Oslo a specific parameter (the maximum precipitation rate at which the auto-conversion of cloud water to rain is suppressed) has the same value as used in CAM-Oslo. Thus, the use of CAM-Oslo in this context is intentional and correct.

Page 13, line 17: change "illustration for the complete" for "illustration of a complete".

We will follow this suggestion.

Page 18, line 17-18: this sentence is not clear. Revise it in such a way that we clearly understand why you average the same period in the model and in the observations (length of the observation period, type of variable). Give an example to illustrate ("For some analyses (i.e. variable1, variable2 and variable3) [. . .]").

We suggest changing the figure showing sensible and latent heat flux of NorESM compared to FLUXNET estimates to use NorESM means for the years 1976-2005 of the Historical1 experiment. The differences in the fluxes compared to the 1983-2005 means were small and do not require any change of the text. We also suggest making it clear in the legend of Fig. 8 that it is indeed means over years 1976-2005 of the Historical1 experiments that are used here. Thus, only the values of gross cycling of fresh water and the meridional overturning circulation uses a different averaging period or experiment. Thus, we propose to replace the last sentence of the first paragraph of section 5 with:

"An exception is the analysis of the gross cycling of fresh water (Table 2) using means for the years 2000-2005 of the Historical1 experiment to be more consistent with corresponding mean values from observational synthesis and atmospheric reanalysis covering the years 2002-2008. Further, the mean ocean meridional overturning circulation (MOC) is from a 30 year period of the piControl experiment."

Page 19, second paragraph, with Figures 5 and 6 : for convenience, I suggest to make only one figure with Figures 5 and 6, since the vertical scale is the same and both are surface fluxes. For example, on the left column would be the sensible heat fluxes, and on the right column the latent heat fluxes. As well, the maps of the differences between NorESM and FLUXNET-MTE would be of great help. You say on line 17 that the latent heat fluxes are better represented than the sensible heat fluxes from the distribution point of view. However, you do not show any distribution. Your point could be easily supported with a spatial RMSE score, centered (you've already calculated the mean biases) and normalized by the spatial standard deviation of the observations (score written in the upper corner of the differences maps), or a spatial correlation coefficient. This will give an objective measure of the model-observations agreement.

We will combine Figs. 5 and 6 as suggested and provide maps of the differences and propose to replace Figs. 5 and 6 with the Fig. 2 of this reply. As suggested, normalized RMSE score and spatial correlation coefficients will be provided. We propose the following modified second paragraph of section 5.1:

"In Figs. 5 the annual mean sensible and latent heat fluxes from NorESM Historical1 are compared to the FLUXNET Model Tree Ensembles (MTE) estimates (Jung et al., 2011). FLUXNET-MTE estimates are restricted to vegetated land surface and this is the reason why no fluxes are estimated for the desert zones. The NorESM simulated annual mean sensible

heat flux (Fig. 5a) is in the same range as the FLUXNET-MTE estimations (Fig. 5b). As seen in Fig. 5c, NorESM underestimates sensible heat flux in most of the African continent south of Sahara, in the west coast of India, in Australia, and in the western part of the United States. The model overestimates sensible heat flux in the extreme eastern part of South America. Comparing NorESM and FLUXNET-MTE estimates the root mean square error (RMSE) normalized by the standard deviation of the FLUXNET-MTE estimate is 1.01 and 0.65 for sensible and latent heat flux, respectively, and the spatial correlations are 0.52 and 0.82 for sensible and latent heat flux, respectively. Thus, from the distribution point of view, the simulation of annual mean latent heat flux (Fig. 5d) compares better with the FLUXNET-MTE estimate (Fig. 5e). Figure 5f show that NorESM generally overestimates latent heat fluxes compared to FLUXNET-MTE, but with clear underestimations in the extreme eastern part of South America. As listed in Table 1, the global mean surface sensible heat flux for the years 1976–2005 of Historical1 is 17.8 Wm^{-2} and within the observational range of 13.2–19.4 Wm^{-2}, while the global mean surface latent heat flux of 81.7 Wm^{-2} is slightly below the observational range of 82.4–89.1 Wm^{-2} which is in contrast to the general overestimation compared to FLUXNET-MTM."

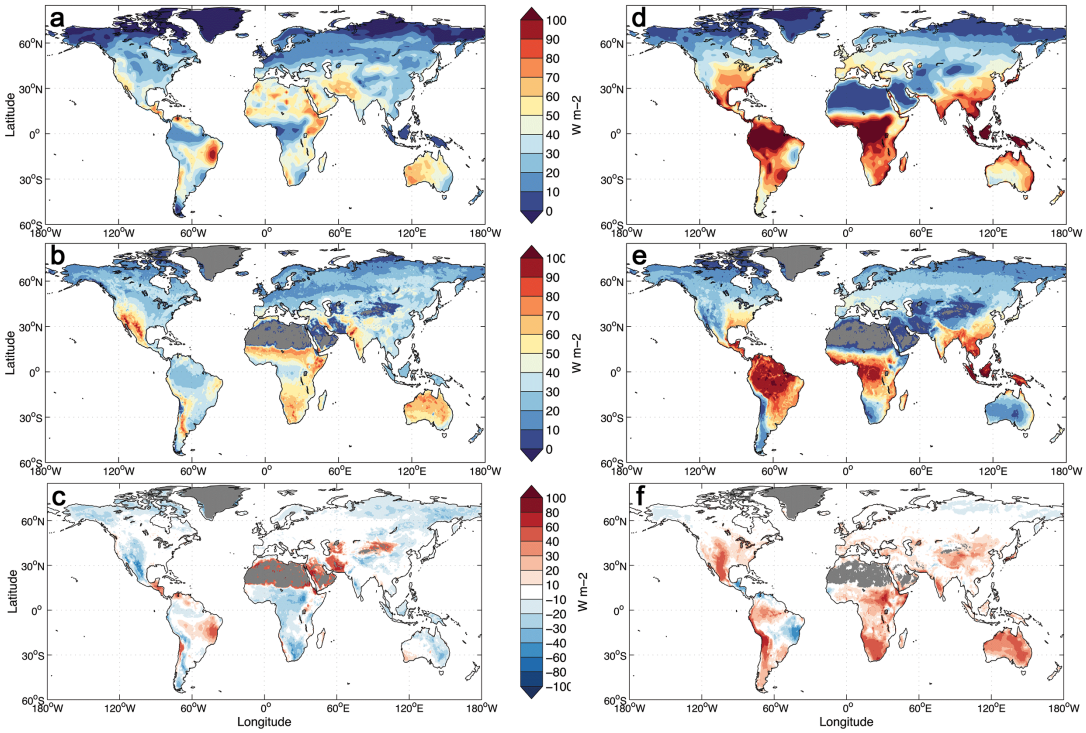

Fig. 2. The left panels show sensible heat flux from **(a)** NorESM, **(b)** FLUXNET-MTE estimates, and **(c)** the difference **(a)**–**(b)**. The right panels show latent heat flux from **(d)** NorESM, **(e)** FLUXNET-MTE estimates, and **(f)** the difference **(d)**–**(e)**. The NorESM fluxes are means for the years 1976–2005 of the Historical1 experiment and the FLUXNET-MTE fluxes are means for the years 1982–2005. Areas with missing observations are shaded with dark grey color.

Page 19, line 25: there is no dataset labeled "IPCC" at the Climatic Research Unit. Thus remove IPCC in this line, and in the legend of Figure 7. Furthermore, you cite two different papers for the CRU dataset you use, and you use only one, so which one is it? Cite only one paper, and precise the name of the dataset (apparently CRU TS 2.1).

Thank you for notifying us that the use of the acronym IPCC is misplaced here. We will

correct this and specify in the text that we are using the CRU TS3.1 dataset. Following the recommendation at http://badc.nerc.ac.uk/data/cru we only cite Mitchell and Jones (2005).

Page 20, line 9-10: although the meaning of this sentence is understandable, say what are the dynamical factors and geographically determined feedbacks. Be more precise.

In the companion paper by Iversen et al. these details are further discussed. We agree, however, that the text also needs to be strengthened, mainly by introducing some references. We propose to replace the last two sentences of section 5.1 with:

"Note that the global pattern of this underestimate (see Figure 7) reflects dynamical factors such as changed occurrence of modes of variability or flow regimes (Palmer, 1999; Branstator and Selten, 2009) and geographically determined feedbacks in the climate system associated with strong interactions between the atmosphere and the ground surface (e.g. sea-ice and snow cover) as discussed by Boer and Yu (2003). Hence, given that there is a slightly too cold climate, it is natural that the amplitude is larger over continents than the ocean (e.g. the cold-ocean warm-land pattern, Wallace et al., 1996) and at high latitudes."

The added references are:

Boer, G. J. and Yu, B.: Climate sensitivity and response, Clim. Dyn., 20, 415–429, doi:10.1007/s00382- 002-0283-3, 2003.

Branstator, G. and Selten, F.: "Modes of Variability" and Climate Change, J. Climate, 22, 2639–2658, doi:10.1175/2008JCLI2517.1, 2009.

Palmer, T. N.: A Nonlinear Dynamical Perspective on Climate Prediction, J. Climate, 12, 575–591, 1999.

Page 21, line 12-13: please cite a paper or give objective arguments to support the fact that CCSM4 is according to observations. Moreover, "in better agreement with the observations" would be more appropriate.

We agree that the statement "according to observations" needs to be softened. The reference to observational data was already in place, but we appreciate that this can be made in a more clear way. The amended text of page 21, lines 12-16 thus becomes:

"In comparison, CCSM4's estimate of the flux of water vapor from ocean to land is very close to the observationally based estimate by Trenberth et al. (2011). However, the oceanic evaporation in CCSM4 is, also according to the estimates of Trenberth et al. (2011), exaggerated by 8%, while the re-cycling is overestimated by almost 1% in that model. It is believed that these differences between NorESM and CCSM4 are linked to aerosols and the tuning of cloud properties."

Page 24, second and third paragraphs: it would be much more convenient for the reader to have the thickness values cited directly in the text rather than having to find them in the cited papers. For each area that you comment, give the thickness value in NorESM and in the observations.

To provide thickness values of specific areas mentioned in the text, we suggest the following reformulation of the last two paragraphs of section 5.4:

“The thickest ice found in the simulations is north of Greenland and Ellesmere Island, in agreement with observational climatologies (Rothrock et al., 2008). Also, in the Central Arctic, the thickness is comparable with estimates based on submarines from the late 1970s (Rothrock et al., 2008; Kwok and Rothrock, 2009). At the North Pole modelled March values are close to 4.5 m, while estimates from Rothrock et al. (2008) for this month are very similar for the years 1975–1979 (4.4–4.7 m). However, the model also shows a maximum north of the East Siberian coast that is not realistic (above 5 m). Satellite estimates from 2006–2007 (Kwok and Cunningham, 2008) give values near the North Pole of 2–2.5 m, while thickness near the East Siberian coast is 1–2 m. Clearly, the modelled ice is too thick compared with these estimates. However, as discussed by Kwok and Rothrock (2009), there was a considerable loss of Arctic ice volume after year 2000. The modelled Central Arctic sea-ice thickness therefore seems to be more similar to that observed during the 1970s, than after 2005.

The Antarctic sea-ice thickness shown in Fig. 12 is comparable with observations (Worby et al., 2008) in large regions, with thin first-year ice with thickness less than 1 m over large regions, and with the thicker ice in the Western Weddell Sea, close to the coast. Spring values reported by Worby et al. (2008) indicate mean thickness of 0.89 m and 1.33 m in Eastern (sector 45°W–20°E) and Western (sector 60°W–45°W) Weddell Sea, respectively, while modelled mean values are 0.5–1.5 m in the Eastern, and from 1.5 m to more than 4 m in the Western Weddell Sea, respectively. Thus, modelled ice is too thick in the thickest regions and does not melt during summer, consistent with Fig. 4.”

Page 30, second paragraph: I’m not comfortable with the way you describe your EOF analysis (which is probably relevant at the end). The EOFs (spatial patterns) are obtained by the decomposition in eigenvectors of the covariance matrix; then, the principal components (the time series) are calculated as the projection of the total anomaly field (spatio-temporal) onto the eigenvectors. Here, you describe the contrary. To compare the amplitude of the EOF of two different datasets, I suggest following this protocol: first calculate the EOF/PCs of both datasets then normalize the first principal components PC1 so they have unit variance (divide by their respective standard deviations, refered to as σ_PC1) and multiply the EOF1 by their associated σ_PC1 this way, the EOFs of both datasets show the patterns associated with a unit deviation from the mean in their associated PCs, and can thus be reliably compared. This may be the protocol that you have followed, but I did not understand it in the text. Eventually, although this protocol should be cleaner, I don’t think it will dramatically change your findings. I thus suggest revising your methodology or just explaining it more clearly.

Your suggestion is essentially the protocol we have followed. First, we used singular value decomposition (SVD) on the (spatio-temporal) matrix of area weighted anomalies to find both the EOFs and PCs. These PCs were then normalized to unit standard deviation as you suggest. The EOFs displayed in Figure 22 were then calculated by projecting the original anomalies (not area weighted) onto these standardized PCs. Figure 22 thus shows the spatial patterns associated with a one standard deviation change in the index time series. This is the method for instance used by Thompson and Wallace (2000) to display the annular modes. We propose to add the following text to section 6.2 to explain the method better:

"The NAM is defined as the first empirical orthogonal function (EOF) of the Northern Hemisphere (20–90° N) winter SLP anomalies. Prior to the EOF analysis the data was weighted by the square root of the cosine of latitude so that equal areas are afforded equal weight. The principal components (PCs) were scaled to unit standard deviation and projected

on the original (not area-weighted) SLP anomalies to obtain corresponding EOFs. Figure 22 shows the leading EOF, associated with a one standard deviation change in the corresponding PC (index time series), for Historical1 and NCEP-2 data (Kanamitsu et al., 2002)."

Page 30, same paragraph, line 12: it is not obvious for me that the amplification of the centers of action in NorESM than in NCEP-2 can be the cause of more variance explained by the first EOF in NorESM than in NCEP-2. This is very likely the case, but saying that implies that the SLP variance fields are the same in NorESM and NCEP-2. And because you do not show this, I thus suggest replacing "As a consequence" with "This likely explains why" (or a similar expression removing the call of a "cause-consequence" phenomenon).

We agree. The expression "This likely explains why" will be used instead.

Figure 20: please add the label of the X axis.

Label will be added.

Figure 21: the figure is too small, the arrows can barely be seen. Please provide a figure at least twice as large. Add the X and Y axis labels.

The arrows will be enlarged and labels added.

Legend of Figure 22: add "in" between "(20-90°N)" and "Historical1".

We will follow this suggestion.

Page 31, discussion on the results on the AMO: a recent paper (Booth et al., 2012) claims that the AMO is actually a forced variability due to the indirect effect of anthropogenic aerosols during the industrial period. This reference can put some additional value to discuss your results on the AMO.

Thank you for pointing us to the findings of Booth et al. (2012) in the AMO context (we already refer to this paper in the section 7 "Modelled climate evolution of the 20th century"). We suggest replacing the sentences of line 2-6, page 31 with:

"The larger AMO index standard deviations of historical experiments (0.07–0.09 K) compared to the control might be due to the prescribed temporal variability in the external forcing, an interpretation that is supported by Booth et al. (2012), who found a strong impact of aerosol emissions and volcanic activity on the multidecadal variance in North Atlantic SSTs for the years 1860–2005. The AMO index standard deviation of Historical1 is still smaller than the observations indicate and the multidecadal fluctuations of the observational based index seem not to be reproduced by the historical experiments."

Also we believe the potential added variance of North Atlantic climate indices, due to variations of the external forcing after 1850, might mask the 20 yr variability of the subpolar North Atlantic we found in the pre-industrial control experiment. Therefore we suggest to change the sentence in lines 28-30, page 31 to:

"Possible explanations are that variability or trend in the external forcing of the historical experiments either disrupts some feedback mechanism causing the signal in the control simulation or masks the 20 yr variability by increasing the variance of North Atlantic climate

indices."

Page 31, line 17-31: as an element of discussion, Escudier et al. (2012) have thoroughly described a 20-yr cycle at play in the IPSLCM5A coupled model in a recently accepted paper. This could help supporting your discussion.

Indeed study of Escudier et al. (2012) is highly relevant in the discussion of the 20 yr cycle we find in the pre-industrial control experiment. Related to this we feel it is appropriate to cite a few more studies documenting variability in the North Atlantic with similar time scales. Thus, we propose to add the following text after "…possibly involving the subpolar gyre." on page 31, line 26:

"Variability in the North Atlantic with similar time scales has been documented in several studies of climate proxies, observations, and climate model simulations (e.g. Frankcombe and Dijkstra, 2009; Frankcombe et al., 2010; Chylek et al., 2011). Further, Escudier et al. (2012) attributed a 20 yr cycle found in the IPSL-CM5A-LR model to a coupled oscillatory mode involving propagation of temperature and salinity anomalies in the subpolar gyre, sea ice changes in the Nordic Seas, and changes to the strength of the East Greenland Current across the Denmark Strait due to modified regional atmospheric circulation."

The added references are:

Chylek, P., Folland, C. K., Dijkstra, H. A., Lesins, G., and Dubey, M. K.: Ice-core data evidence for a prominent near 20 year time-scale of the Atlantic Multidecadal Oscillation, Geophys. Res. Lett., 38, doi:10.1029/2011GL047501, 2011.

Escudier, R., Mignot, J., and Swingedouw, D.: A 20-year coupled ocean-sea ice-atmosphere variability mode in the North Atlantic in an AOGCM, Clim. Dyn., doi:10.1007/s00382-012-1402-4, published online, 2012.

Frankcombe, L. M. and Dijkstra, H. A.: Coherent multidecadal variability in North Atlantic sea level, Geophys. Res. Lett., 36, doi:10.1029/2009GL039455, 2009.

Frankcombe, L. M., von der Heydt, A., and Dijkstra, H. A.: North Atlantic Multidecadal Climate Variability: An Investigation of Dominant Time Scales and Processes, J. Climate, 23, 3626–3638, doi: 10.1175/2010JCLI3471.1, 2010.

Page 33, line 23: replace "A" with "a".

This has already been corrected in the typesetting/proof-reading process of the discussion paper.

Legend of Figure 26: I do not understand the difference between the two panels. They are labeled a and b, but nothing identify them in the legend.

In the typesetting/proof-reading process of the discussion paper, the labels have been addressed in the legend.