Journal of Climate

# A Theory for Self-sustained Multicentennial Oscillation of the Atlantic Meridional Overturning Circulation. Part II: Role of Temperature

--Manuscript Draft--

| | |
|---|---|
| Manuscript Number: | JCLI-D-22-0755 |
| Full Title: | A Theory for Self-sustained Multicentennial Oscillation of the Atlantic Meridional Overturning Circulation. Part II: Role of Temperature |
| Article Type: | Article |
| Corresponding Author: | Haijun Yang<br>Fudan University<br>Shanghai, CHINA |
| Corresponding Author's Institution: | Fudan University |
| First Author: | Kunpeng Yang |
| Order of Authors: | Kunpeng Yang |
| | Haijun Yang |
| | Yang Li |
| Abstract: | In the first part of our research on self-sustained multicentennial oscillation of the Atlantic meridional overturning circulation (AMOC), we used a hemispheric box model considering only the salinity equations. In this follow-up paper, we consider both thermal and saline processes in the box model, so as to investigate the role of temperature in multicentennial AMOC oscillation. The thermal processes exert mainly three effects: shortening the oscillation period, stabilizing subpolar stratification and thus the oscillation system. These three effects are caused by the fast surface temperature restoring process, the stabilizing subpolar temperature stratification, and the negative temperature advection feedback, respectively. Nonlinear restraining effect from enhanced subpolar mixing, or a nonlinear relation between AMOC anomaly and meridional difference of density anomaly, is still needed to realize a self-sustained oscillation, whose mechanism can be generalized as follows: a combination of a linearly growing oscillation dominated by linear advection and a nonlinear restraining process. This study advances the theory reported in the first part of this research. Linear stability analyses reveal that the eigenmode of the system is sensitive to model geometry, flow properties, and meridional differences of sea-surface temperature (SST) and sea-surface salinity (SSS). Our theoretical results suggest that, a smaller (larger) meridional SST (SSS) difference weakens (strengthens) the negative temperature (positive salinity) advection feedback which may lead to a less stable AMOC. Such heuristic findings may be expected in the future due to more intense warming and freshwater hosing at the high latitudes of the Northern Hemisphere. |

**Replies to Reviewer #1:**

Thank you very much for all of your constructive comments. We have carefully revised our manuscript based on the advice by you and other reviewers. The following are our point-by-point replies.

*This manuscript is a follow up of a previous study published in J. Climate. Both studies used an idealized box-model to understand the stability and persistence of a centennial oscillation of the AMOC. Whereas the first study focuses solely on the active role of salinity, the present work studies the effect of including the temperature as an active variable.*

*This work tackles an interesting topic and is a nice and needed follow up of the authors' previous work. However, I feel that, unlike the Part I, the theoretical results are not aligned with the numerical ones. Hence, I feel that the inconsistency between theory and numerical simulations need to be resolved. Also, the fact that the theory developed in Part I is no longer correct when temperature is included needs to be discussed.*

*Hence, I recommend this work for major revision.*

**Responses:** Thank you very much for your invaluable suggestions, which help us improve the manuscript tremendously. Combining the comments from all the reviewers, we have revised the manuscript primarily in these following aspects:

1) We have completely rewritten the introduction. Coupled modelling studies on multicentennial AMOC oscillation are synthetically reviewed, the inconsistency among their mechanisms and the necessity for theoretical studies are disclosed. Inadequacy of previous theoretical models in accounting for sustainable multicentennial AMOC oscillation is also discussed. Finally, the potential impacts of thermal processes on AMOC oscillation are raised, justifying the inclusion of temperature effects in this study.
2) In section 2, the choices of parameters are discussed in more detail.
3) In section 3, we categorize the thermal effects more precisely. We propose that there are mainly three effects when including the temperature equations: (1) increase of the oscillation frequency, (2) stabilization of the overall system, and (3) stabilization of the subpolar stratification. These three effects are attributed to the following three processes, respectively: (1) fast surface temperature restoring, (2) negative temperature advection feedback, and (3) stabilizing subpolar temperature stratification. Now, it is easier to understand that the behaviors of the temperature-salinity system are different from the salinity-only model in LY22.

4) In section 4, we more clearly describe the self-sustained oscillation mechanism. In LY22, we proposed that the nonlinear subpolar vertical mixing is crucial for self-sustained oscillation. In the revised manuscript, we further propose that assuming a nonlinear relation between AMOC anomaly and meridional difference of density anomaly can also lead to self-sustained oscillation. We further show that the self-sustained oscillation mechanism not only agrees with that of LY22, but also advances the theory of LY22.

5) Most figures are re-plotted.

**Major Comments:**

1. *Inconsistency between theory and numerical simulations*

*Overall, the results from analytical approach presented here opposed the one of LY22. In LY22 the regime parameter is such 4S exhibits an unstable oscillation and 3S a stable one. Hence the self-sustained oscillation of [ 4S + diffusion between box 2 and 3] is interpreted as a mix of the unstable oscillation of 4S and the stable one of 3S.*

**Responses**: Thank you very much for these comments. Sections 3 and 4 are revised carefully. Our analytical results are not at odds with that of LY22. A more detailed description of the self-sustained oscillation mechanism is provided in section 4, reflecting that the self-sustained oscillation mechanism in the temperature-salinity system agrees well with and even advances that of LY22. Our analytical results and numerical results (of self-sustained oscillation) also match well with each other.

Here, we would like to emphasize that, first of all, *in both LY22 and this manuscript, the self-sustained oscillation can be never simply interpreted as a mix of an unstable oscillation of 4S (4TS) and a stable oscillation of 3S (3TS).* A self-sustained oscillation can never occur in the linear 3S, 4S, 3TS, and 4TS systems. The linear stability analyses (no matter using theoretical method or numerical approach) show clearly that in the linear system, there are three kinds of oscillations: the growing oscillation with a positive real part of the eigenvalue, the neutral oscillation with zero real part of the eigenvalue, and the decaying oscillation with a negative real part of the eigenvalue. None of them is a self-sustained oscillation. In different linear systems, the critical value ($\lambda_C$) of the linear closure parameter that sets up the oscillation type is different, of course.

Second, we would like to emphasize that to realize a self-sustained oscillation in the linear system, a certain degree of nonlinearity is needed. In the salinity-only system of LY22, enhanced vertical salinity mixing (i.e., nonlinear mixing) was considered in the subpolar ocean. Therefore, under a reasonable linear closure parameter $\lambda$, the enhanced vertical salinity mixing will turn the growing oscillation of the 4S model into a self-sustained oscillation. We have stressed that this mixing in the subpolar ocean cannot be too strong; otherwise, the growing oscillation will become a damped oscillation and the 4-box model will practically become the 3-box model. There is no self-sustained oscillation in the *linear* 3-box and 4-box models. The self-sustained oscillation can only occur when a certain degree of nonlinearity is considered. It cannot be understood as "*a mix of the unstable oscillation of 4S and the stable one of 3S.*"

*I think this is the major problem of the paper. All these things need to be discussed and explain in length. I would like an alternative explanation for the existence of self-sustained oscillation despite the instability of both 3TS and 4TS (for \lambda=14 Sv Kg-1 m3).*

We are sorry for not having explained it well in previous manuscript. In the revised manuscript we state: "**The essence for a self-sustained oscillation is a linearly growing oscillation restrained by a nonlinear process, which can take the form of a nonlinear subpolar vertical mixing, or of a nonlinear relation between AMOC anomaly and meridional difference of density anomaly, or take other nonlinear forms.**"

Therefore, even in a 3-box system, if assuming a small degree of internal nonlinearity between AMOC anomaly and meridional difference of density anomaly (Fig. R1c, orange curve), a self-sustained oscillation can occur (Figs. R1a, b, orange curves). This is an important development of LY22. If the 3-box system contains no nonlinear factors, self-sustained oscillation cannot occur (Figs. R1a, b, black curves). This further suggests that the nonlinearity is the key to self-sustained oscillation.

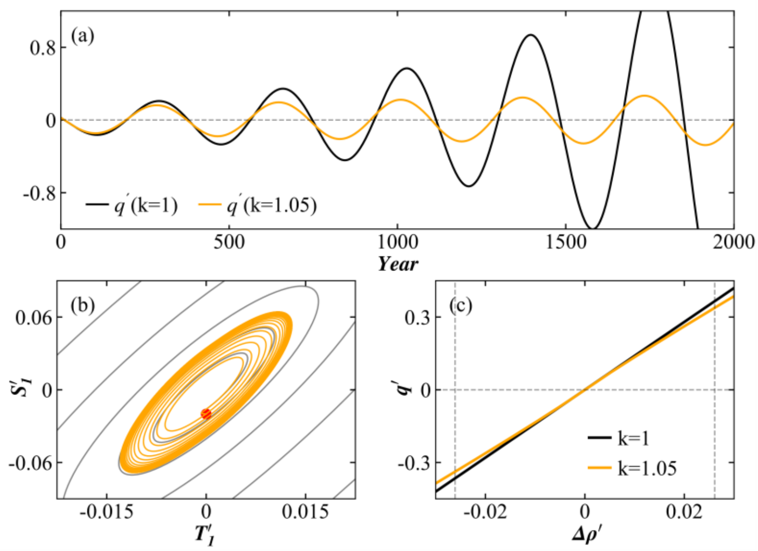


FIG. R1. Oscillations under $\lambda = 14\,Sv \cdot kg^{-1}\,m^3$ for the 3TS model. (a) Time series for $q'$ (units: $Sv$) under $k = 1$ (black curve) and $k = 1.05$ (orange curve). (b) $T_1'$-$S_1'$ phase space diagrams for years 1-10000. The red dot represents the initial location of $T_1'$ and $S_1'$. Black curve is for $k = 1$, and orange curve is for $k = 1.05$. (c) Variation of $q'$ with $\Delta\rho'$ (units: $kg/m^3$) under $k = 1$ (black curve) and $k = 1.05$ (orange curve). The intersections between the vertical dashed gray lines and the abscissa axis mark the upper and lower limits for $\Delta\rho'$ during the integration. The values of the other parameters are the same as those listed in Table 1 in the revised manuscript.

*However, in the present study, according to Figure 2c, 3TS and 4TS both exhibit stable oscillation under the standard value (\lambda=12 Sv Kg-1 m3) and both exhibit unstable oscillation under the other tested value (\lambda=14 Sv Kg-1 m3). In either case the explanation of LY22 failed. The addition of mixing in the subpolar region leading to a self-sustained oscillation cannot be interpreted as a mix of a stable oscillation in 3TS and an unstable one in 4TS. This makes the entire purpose of the theoretical model useless.*

Since the temperature equations affect the behaviors of the system, it is natural that the critical values of $\lambda$ (i.e., $\lambda_C$) in the 4S (3S) and 4TS (3TS) models differ. This does not necessarily suggest the failure of the explanation of LY22, since a self-sustained oscillation is not a result of "*a mix of a stable oscillation in 3TS and an unstable one in 4TS*." A self-sustained oscillation emerges from a growing oscillation that is restrained by a nonlinear process.

The theoretical models used in LY22 and this manuscript are extremely useful. The linear stability analyses on the linear model give us the conditions of stability and the oscillation of the system. Particularly, the 3S model of LY22 can be solved completely theoretically, so that we can see exactly which model parameters and how these parameters determine the stability and the oscillation of the system (see section 4c of LY22). This allows a fundamental understanding of the model behaviors.

*Finally, the regime with \lambda ~= 13 Sv Kg-1 m3, appear quite interesting since it provides a stable oscillation for 4TS and an unstable oscillation for 3TS (Fig.2c). This is the opposite to LY22 for \lambda ~= 12 Sv Kg-1 m3. How would you explain the increase stability of the 4-box model over the 3-box model?*

We rephrase your above concern as "*why the 4TS model is more stable than the 3TS model, while in LY22 the 3S model is more stable than the 4S model, regardless of the value of $\lambda$ (Fig. 2c)*?" We have a detailed explanation in lines 388-420 of the revised manuscript, though here we would like to provide a more straightforward explanation in the following context.

Suppose there is an initial perturbation of the AMOC ($q'$), when $q' > 0$ ($q' < 0$), the subpolar salinity and temperature perturbations $S_2' > 0, T_2' > 0$ ($S_2' < 0, T_2' < 0$). In the salinity-only 4S system of LY22, $S_2' > 0$ ($S_2' < 0$) leads to a stronger (weaker) downward motion in the subpolar ocean, so that it in turn reinforces the initial $q'$. This process can be simply sketched as $q' > 0 \rightarrow S_2' > 0 \rightarrow q' > 0$, or $q' < 0 \rightarrow S_2' < 0 \rightarrow q' < 0$. In other words, the initial $q'$ can be reinforced by the subpolar vertical salinity perturbation; or the subpolar salinity stratification has potentially a destabilizing effect on the oscillation. This destabilizing effect is absent in the 3S model of LY22. Therefore, the 3S model is more stable than the 4S model.

In a *temperature-only* box model, we have the following process: $q' > 0 \rightarrow T_2' > 0 \rightarrow q' < 0$, or $q' < 0 \rightarrow T_2' < 0 \rightarrow q' > 0$. That is, the initial $q'$ can be damped by the subpolar vertical temperature perturbation. In other words, the subpolar temperature stratification has potentially a stabilizing effect on the oscillation. This stabilizing effect is absent in the 3T model. Therefore, the 4T model is more stable than the 3T model.

In the temperature-salinity box model (4TS), the resultant effect of the subpolar temperature and salinity stratifications on the oscillation behavior is determined the relative effects of the two. We show in the manuscript that under a reasonable surface temperature restoring timescale, the stabilizing effect of temperature stratification will overcome the destabilizing effect of salinity stratification; therefore, the 4TS model is more stable than the 3TS model.

This does not necessarily suggest inconsistency between the TS models and the LY22 models. We can also show that the 3TS model is more stable than the 4TS model, given some unrealistic surface temperature restoring timescale (Fig. R2).

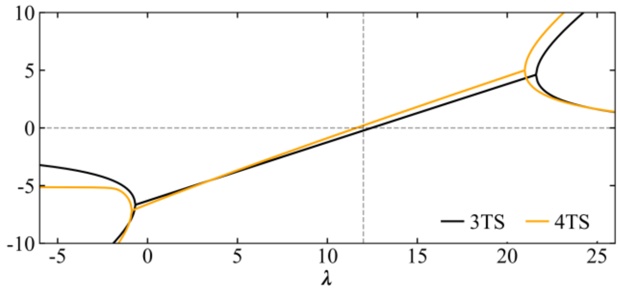


Fig. R2. Dependences of real parts of eigenvalue $\omega$ on $\lambda$ in the 4TS (orange curve) and 3TS models (black curve) under the temperature restoring coefficient $\gamma = (1\ month)^{-1}$. The units of the ordinate are $10^{-10}\ s^{-1}$. The values of the other parameters are the same as those listed in Table 1. The vertical dashed line denotes the situation under the standard value $\lambda = 12\ Sv \cdot kg^{-1} m^3$.

2. *Inconsistency with hypothesis of LY22 when Temperature anomaly are allowed*

*I think this is a major problem of the paper. What you demonstrated here is that the self-oscillation does not work at all like LY22.*

*The addition of T creates a regime where both 4TS and 3TS are both stable (under \lambda = 12 Sv kg-1 m3) or unstable (under \lambda = 14 Sv kg-1 m3). This is a huge discrepancy with 3S and 4S model - and the explanation of the self-sustained oscillation in LY22.*

**Responses**: Thank you very much for these comments. The self-sustained oscillation mechanism in LY22 and that in this work are consistent, and can be concluded as follows: **"a linearly growing oscillation restrained by a nonlinear process,"** instead of "a mix of a stable oscillation in 3TS and an unstable one in 4TS." Here, we would like to emphasize again that the results from the 4TS (3TS) model is mostly consistent with the results from the 4S (3S) model. The only difference is that the 3S model is more stable than the 4S model, while the relative stability of the 3TS and 4TS models is not straightforward, depending on the strength of surface temperature restoring. The addition of temperature will naturally affect the stability of the models, thus the relative stability between 3-box and 4-box models. If there is no difference in system stability after including temperature, our current work "role of temperature" would become useless. Moreover, the more stable 4TS model than the 3TS model suggests exactly that the temperature effects make the 4-box models more stable than the 3-box models.

The difference in relative *linear* stability between 3-box and 4-box models is not important, because as far as the self-sustained oscillation is concerned, the fundamental mechanism is the same for all the models; that is, **the self-sustained oscillation emerges from a linearly growing oscillation that is restrained by a nonlinear process."** The nonlinear process can be enhanced vertical mixing in the subpolar ocean, or a nonlinear relationship between AMOC anomaly and the meridional difference of density anomaly, or be other nonlinear forms. Therefore, the self-sustained oscillation mechanism in LY22 and that in the TS models are consistent, and the change in relative stability between 3-box and 4-box models after including temperature is reasonable.

*One perfect example is the use of the dashed orange curves in Figs. 8, 9, and 10. There definition was clear in LY22 ("end" of the self-oscillatory regime, through the instability of 3S). Here it is not explained.... (Note that the orange computation of the solid line is also quite unclear - Is it the instability of the 4TS, 3TS, or both?). It looks like a lucky guess. This is not acceptable. This is highly problematic. How did you compute these two lines and how do they relate to 4TS and 3TS stability? If it does not relate to it, this suggests that the computation of the stability of 4TS and 3TS is not useful. And that the system does not behave equivalently to LY22 when T is introduced.*

*This raises questions regarding the robustness of LY22. My personal conclusion after reading the manuscript is that LY22 results cannot be generalized to the presence of temperature. A detailed* *discussion on that is needed. Qualitative inconsistencies with LY22 should be clarified (i.e., difference between S and TS model).*

Sorry for not having clearly described the meaning of the solid and dashed orange curves in Figs. 8-10. The solid and dashed orange curves in both LY22 and this manuscript stand for the lower and upper stability limits within which the self-sustained oscillation can occur, when including an enhanced subpolar vertical mixing. They are all calculated by numerical integration; no self-sustained oscillation occurs when the system is less stable than the dashed orange curve in the 4TS (4S) models, no matter how strong the subpolar vertical mixing is. The stability threshold of the 3S model is also higher than that of the 4S model, and can be well represented by the dashed orange curve.

The stability threshold of the 3TS model under the parameters of Figs. 8-10 is lower than that of the 4TS model, as revealed in Fig. 2; thus, it will never become the upper stability limit for the self-sustained oscillation. Therefore, the dashed orange curve in Figs. 8-10 of the original manuscript stands for the upper stability limit for the self-sustained oscillation when subpolar vertical mixing is included, instead of the stability threshold of the 3TS model. Even so, the calculation of 3TS stability is meaningful, since the change of relative stability between the 3-box and 4-box models after including thermal effects can reflect the effects of thermal processes. The more stable 4TS (3TS) model than the corresponding 4S (3S) model also suggests that the thermal effects have stabilizing effects on the system. In this way, the topic “role of temperature” is justified. The system behavior will naturally differs from that of LY22 since the thermal effects must play a role in the system stability, but the underlying self-sustained oscillation mechanisms in LY22 and that in our current work are still consistent. We made it clear in this revised manuscript that the solid orange curve stands for both the stability threshold and the lower stability limit for the self-sustained oscillation in the 4TS model. For simplicity, we removed the dashed orange curve and related discussion.

Please also refer to our reply to your 1st question. This paper is a further development of LY22. The self-sustained oscillation mechanism raised in LY22 can be generalized to cases with both temperature and salinity. Figure 2 in the manuscript shows clearly the curves for 3S, 4S, 3TS, and 4TS are almost identical, except that the values of $\lambda_C, \lambda_1$ and $\lambda_2$ differ in different models. Note that the curves for the 3S model are obtained purely theoretically (LY22), while the curves for the other models are obtained numerically. The consistency between these curves is not of coincidence, but of certainty, because of the consistency of the physical fundamentals in these models. Different models and different physical processes lead to slightly different sensitivity of the AMOC anomaly to the meridional difference of density anomaly.

*3. Lack of information*

*The manuscript suffers from a lack of relevant and clear information. This affects both the understanding of the study and its reproducibility. This needs to be fix.*

*It should be made clear that you have two models. In my understanding, since you impose the mean state, you have an only-salinity *active* variable model and a both TS *active* variable model. Where active mean that anomaly can exist (i.e., \dot{T/S}'). Clarifying this point in the text would help the reader.*

*If you want to set "any" background state. You should discuss the equilibrium (2a-b) with care, since they are not used in full. To my understanding you need to set the same background (and not the independent equilibrium of 3TS 4TS, 3S, and 4S) to have a fair comparison between the two versions of the model TS and S. Otherwise it is impossible to use the exact same background state for each version. Do I understand correctly? If so, could you add this rationale in the text?*

*In equation (7b) and (7e) T2, S2, and V2 does not have the same meaning. For instance V2 is a larger volume in 7 than in 5. This should be reflected by the use of different symbols.*

*Without the explanation or inclusion of the equations for the S model (derived in LY22). It is hard to follow. Could you confirm that the S model is actually the (5 e-h) and (7d-f)?*

**Responses**: Thank you very much for these suggestions. We clarified in the revised manuscript that we use models with only salinity equations (4S and 3S models) and models with both temperature and salinity equations (4TS and 3TS models).

We demonstrated more clearly in this revision that Eqs. (5e-h) and Eqs. (7d-f) are the models of LY22, while in the 4TS and 3TS models both temperature and salinity are included.

The equilibrium states in Eqs. (2a-b) have been fully used in Eqs. (5a-h), but might be confusing due to the linearization from Eqs. (1a-h) to Eqs. (5a-h). Take Eqs. (1e) and (1g) as examples, linearizing them gives $q'\left(\overline{S_4}-\overline{S_1}\right)+\overline{q}(S_4'-S_1')+F_w$ and $q'\left(\overline{S_2}-\overline{S_3}\right)+\overline{q}(S_2'-S_3')$ on the right-hand side, which can be finally reduced to Eqs. (5e) and (5g). Therefore, some equilibrium values like $\overline{T_3}$ and $\overline{S_3}$ are cancelled, and all the equilibrium values have been used exactly. The equilibrium values for temperature are calculated after imposing the same pair of restoring temperatures $T_1^*$ and $T_2^*$ for the 4TS and 3TS models. Based on Eq. (2a), the equilibrium values for temperature must be different in the 4TS and 3TS models since the 3TS model has a larger $V_2$. Other standard background parameters are the same in the 3TS, 4TS, 3S, and 4S models, while the corresponding rationales have been added in the revised manuscript. In conclusion, we have set the same background states for the 3TS, 4TS, 3S, and 4S models, while the calculated equilibrium temperatures must differ.

To be consistent with LY22, we still use $V_2$ in the 3-box models; we stress that the $V_2$ in the 3-box models equals to the sum of $V_2$ and $V_3$ in the 4-box models.

4. *Physical description*

*I have several issues with the physics described in the text.*

**Responses**: Thank you very much for these comments.

*I disagree that the feedbacks could be spotted/illustrated by the lag of the timeseries. The lag is mainly related to the oscillation which is driven (as well as its timescale) by the mean advection.*

We agree that the lag is mainly due to the mean advection. In the manuscript, we state that the positive (negative) correlation coefficient at ***lag 0*** is a further illustration of the positive (negative) feedback.

We would like to emphasize that the feedbacks we discussed in the correlation figures are based on the deterministic equations, not just deduced from the correlations. If there is no equation between different processes, the correlation between different processes suggests exactly some kind of relationship and one can never say there is causality between them.

In the correlation figures of the manuscript, all processes are connected by equations. Therefore, we can say that the positive (negative) correlation coefficient at lag 0 is a further illustration of the positive (negative) feedback, because the bases are in the equations, thus the underlying physics, instead of correlation coefficients alone.

*A description of what the nonlinear AMOC-density relation physically represents would be useful. Also, a description of how it is parameterized in your equation (10) would be useful.*

This nonlinear relation stands that if the meridional difference of density anomaly is larger than certain threshold value (like $\rho_{cri}$ in section 4b), the growth of AMOC anomaly could show certain degree of nonlinearity. This parameterization follows the one used in Rivin and Tziperman (1997; RT97). In our work, only very weak nonlinearity (k=1.05) is considered, much weaker than that used in RT97 (k=3). This also suggests that this kind of nonlinearity is quite efficient at turning a linearly growing oscillation into a self-sustained one.

More detailed description is added in the revised manuscript.

*I had a hard time understanding what you called the "positive restoring-advection feedback".*

*For instance, in l.243-245, the fact that T'2 is decreased at the end suggests that you are illustrating a negative feedback. I wonder if what you are referring to is not the action of temperature restoring on a density anomaly dominated by salinity and partially compensated by temperature. Here the action of thermal restoring will reduce the temperature anomaly hence intensifying the density anomaly. This is some kind of positive feedback for the density. However it cannot be described using only temperature. Also this is more an oddity than a positive feedback. Indeed this mechanism still only leads to a transient increase of the density, but asymptotically it is still removing perturbation (because it is driven from the negative feedback induced by the temperature surface restoring). This should be clarified*

Thank you very much for pointing out that the “positive restoring advection feedback” had not be described well in the original manuscript.

In line 237 of the original manuscript, we stated: “There are mainly two feedbacks between the thermal processes and the AMOC.” The positive restoring advection feedback describes a relation between temperature restoring and AMOC (advection), not between temperature restoring and temperature itself.

As you point out above, “*the action of thermal restoring will reduce the temperature anomaly hence intensifying the density anomaly. This is some kind of positive feedback for the density*.” This is correct since the density anomaly determines AMOC anomaly, and they are positively correlated. It is also true that the density anomaly is not only determined by temperature. Here, mainly temperature-related density anomaly and thus temperature-related AMOC anomaly is discussed.

In the revised manuscript, this feedback is discussed in more detail. Take box 2 as an example, by limiting the increase of subpolar temperature anomaly and thus the negative temperature-advection feedback, the restoring effect manifests as the positive feedback between the restoring term and AMOC anomaly. In other words, such restoring advection feedback is to increase AMOC anomaly, i.e., to amplify the initial AMOC perturbation.

Note that the temperature feedbacks in the 4TS model should be viewed as a combination of a negative temperature advection feedback and a positive restoring advection feedback. The latter is driven by the former, which in turn hampers the former. Their total effect is a negative feedback. Nevertheless, the restoring advection feedback should still be termed as a positive one.

*Mean advection described as a feedback can be misleading. This is not exactly the same kind of feedback as the positive salinity feedback, for instance. Mean advection will conserve the anomaly (even in a linear framework), it just moves things around. It leads to an oscillation of period*

*~1/\bar{q}. This can be seen by the action of the mean advection in the equations for the anomaly: it acts as a skew-symmetric component of the Jacobian operator.*

We totally agree with your comments. Figure 3b shows that the mean advection term has the smallest contribution to the temperature anomaly, although this term has good positive correlation with temperature-related AMOC anomaly (Fig. 3e). Therefore, practically the positive mean advection feedback can be neglected.

In the revised manuscript, we tone down the mean advection feedback to avoid misleading our readers.

5. *Choice of parameters*

*The choice of the parameter is not well discussed.*

**Responses**: Sorry about this.

*For instance, the choice between \lambda = 12 Sv kg-1 m3 (as in LY22) or \lambda 14 Sv kg-1 m3 here is not discussed. Also the value of 13 Sv kg-1 m3 seems more interesting since it provide a regime where 3TS and 4TS are unstable and stable, respectively. All these choices need to be explained.*

In the revised manuscript, more explanation on the choice of $\lambda$ is provided.

The choice of $\lambda = 12\ Sv \cdot kg^{-1}m^3$ as the standard parameter is to make it the same as in LY22. Under $\lambda = 12\ Sv \cdot kg^{-1}m^3$, the 4S model is unstable while the 4TS model is stable, **reflecting that the overall thermal effect is to stabilize the system**.

The choice of $\lambda = 14\ Sv \cdot kg^{-1}m^3$ is to make the 4TS model unstable, since the self-sustained oscillation has to be based on a linearly unstable regime.

Because the subpolar temperature stratification effect is a stabilizing effect and can overcome the destabilizing effect of subpolar salinity stratification under realistic parameter ranges, the 4TS model can be more stable than the 3TS model. Therefore, the choice of $\lambda = 13\ Sv \cdot kg^{-1}m^3$ is to make the 4TS model stable and the 3TS model unstable. However, we cannot realize self-sustained oscillation in the 4TS model under $\lambda = 13\ Sv \cdot kg^{-1}m^3$ since the system stays in a linearly stable regime.

*Also the value of \bar{q} = 10 Sv is quite low. (GT95 is more consistent to observations.) I would use a value way closer to 20 Sv. Maybe 18 Sv? Do you have any argument to do otherwise?*

The choice of $\overline{q}$ is critical to the oscillation timescale. In the observation, the maximum mean AMOC is about 20 Sv. This mass transport includes water in the upper 1000 m and the water from the Southern Hemisphere. In a one-hemispheric box model with the upper ocean depth set to 500 m, the mean mass transport should be remarkably smaller than the realistic value. Otherwise, the turnover timescale for a one-hemisphere box model would be unrealistically short. In Griffies and Tziperman 1995 (GT95), $\overline{q}$ is much larger than that in ours, and their subpolar boxes are much smaller than ours, so that the dominant timescale in the GT95 box model is the decadal timescale, instead of the centennial timescale.

Our single-hemispheric model incorporates only the AMOC recirculating in the Northern Hemisphere, so that a smaller mean AMOC is reasonable. This is also consistent with the choice of mean AMOC in Nakamura et al. (1994).

**Reference:**


Nakamura, M., P. H. Stone, and J. Marotzke, 1994: Destabilization of the thermohaline circulation by atmospheric eddy transports. J. Climate, 7, 1870-1882.


*1 year for 1/\tau is quite long. It is set to be 60 days for 50 m is NEMO handbook (i.e., −40 W m-2 K-1), for example.*

We admit that it has been prevailing to set the temperature restoring timescale for a surface layer with a few tens of meters to be 1-2 months in models with higher complexity (Marotzke and Willebrand 1991; Weaver and Sarachik 1991; Mysak et al. 1993; Pierce 1996). However, there is no such thin surface layer in our theoretical model, due to its simplicity. As a substitute, we permit the temperature restoring to happen over the entire depth range of the upper boxes (0-500 m). Such thick surface layer clearly necessitates a much longer restoring timescale. The rather deep restoring depth and long restoring timescale are common in theoretical studies (GT95; Roebber 1995; RT97; Scott et al. 1999; Lucarini and Stone 2005).

**References**:


Griffies, S. M., and E. Tziperman, 1995: A linear thermohaline oscillator driven by stochastic atmospheric forcing. J. Climate, 8, 2440-2453.

Lucarini, V., & Stone, P. H. (2005). Thermohaline circulation stability: A box model study. Part I: Uncoupled model. J. Climate, 18(4), 501-513.

Marotzke, J., and J. Willebrand, 1991: Multiple equilibria of the global thermohaline circulation. J. Phys. Oceanogr., 21, 1372-1385.

Mysak, L. A., T. F. Stocker, and F. Huang, 1993: Century-scale variability in a randomly forced, two-dimensional thermohaline ocean circulation model. Climate Dyn., 8, 103-116.

Pierce, D. W., 1996: Reducing phase and amplitude errors in restoring boundary conditions. J. Phys. Oceanogr., 26, 1552-1560.

Rivin, I., and E. Tziperman, 1997: Linear versus self-sustained interdecadal thermohaline variability in a coupled box model. J. Phys. Oceanogr., 27, 1216-1232.

Roebber, P. J. (1995). Climate variability in a low-order coupled atmosphere-ocean model. Tellus Ser. A-Dyn. Meteorol. Oceanol., 47(4), 473-494.

Scott, J. R., Marotzke, J., & Stone, P. H. (1999). Interhemispheric thermohaline circulation in a coupled box model. J. Phys. Oceanogr., 29(3), 351-365.

Weaver, A. J., and E. S. Sarachik, 1991: Evidence for decadal variability in an ocean general-circulation model - An advective mechanism. *Atmos. Ocean*, **29,** 197-231.

6. *Introduction*

*The context within the literature is unclear and extremely short. The multi-centennial variability of the AMOC in model, theory, and observations is a wide research topic. I expect a more in-depth discussion of our current knowledge and of gaps in our current understanding. Introduction should be re-written with clear scientific questions and clear discussion of the large literature on the topic.*

**Responses**: Thank you very much for this suggestion. The introduction has been rewritten. We reviewed the mechanisms for multicentennial AMOC oscillation in state-of-the-art coupled models, and realized that their discrepancy calls for theoretical studies. On this account, we further reviewed the related theoretical studies, and found that they each reflects a certain degree of inadequacy in accounting for the sustainable multicentennial AMOC oscillation. Finally, we stated the prospective effects that thermal processes have on AMOC oscillation, and came up with our improved model of LY22, namely, temperature variation is permitted now.

**Specific Comments:**

1. *l.31: Change "the thermal process exerts mainly" to "the thermal processes exert mainly".*

Revised.

2. *l.33: Change "stratification, which are contributed by" to "subpolar stratification. These thermal processes are composed of"*

**Responses:** Thank you very much for this comment. The abstract is revised; and this sentence is rewritten. We have made it clear: the three thermal effects (shortening the oscillation period, stabilizing the overall system, and stabilizing the subpolar stratification) are separately caused by three thermal processes (the fast surface temperature restoring process, the negative temperature advection feedback, and the stabilizing subpolar temperature stratification).

3. *l.34: Change "feedback and subpolar" to "feedback, and the subpolar"*

**Responses:** Thank you very much for this suggestion. This sentence is rewritten.

4. *l.35: Remove ", respectively"*

**Responses:** Thank you very much for this suggestion. This sentence is rewritten.

5. *l.54-55: It would be more accurate to state that only the anomalous salinity evolution was kept. Indeed in these studies the temperature was acknowledged within the background state (making them more "realistic"). Unlike the salt-oscillator of Huang and Dewar (Journal of Physical Oceanography, 1996), for instance.*

**Responses:** Thank you very much for this suggestion. The introduction has been rewritten; and this sentence is deleted. In the revised manuscript, we clarified that in LY22 model only salinity variation (or evolution) is possible while in our current model both salinity and temperature variations (evolutions) are permitted.

6. *l.65-67: Describe like that it is a negative feedback. In general restoring is a negative feedback (acting along the diagonal of the Jacobian Matrix). Please clarify.*

**Responses:** Thank you very much for this comment. We have rewritten the introduction; and here we give a clarification on this problem. The feedback between restoring and temperature is indeed negative, which could be termed as the negative restoring temperature feedback, while the restoring advection feedback (which we actually focus on in the paper) depicts the relation between restoring and AMOC perturbation.

7. *l.90: add a coma after "properties".* Revised.

8. *l.98: Change "Model formulae" to "Model formulation".* Revised.

9. *l.107-108: A bit odd to use \tau (which is often use for time "t") as an inverse of a time scale. I suggest to use \gamma, for instance, or to write as 1/\tau.*

**Responses:** Thank you very much for this suggestion. In the revised manuscript, we replaced $\tau$ with $\gamma$ everywhere.

10. *l.122: remove "diagrams".* Revised.

11. *l.124: Change "lower tropical oceans" to "deeper tropical ocean boxes".* Revised.

12. *l.124-125: Change "lower subpolar oceans" to "deeper subpolar ocean boxes".* Revised.

13. *l.125: Change "lower ocean depths" to "deeper ocean box depths".* Revised.

14. *l.129: remove "easily".* Revised.

15. *l.157: add a coma after "\alpha".* Revised.

16. *l.158: add a coma after "expansion".* Revised.

17. *l.217: Would "increase frequency" be better than "acceleration of the oscillation"?*

**Responses:** Thank you very much for this suggestion. We rephrased the sentence accordingly.

18. *l.219: Change "for" to "of".* Revised.

19. *l.275: Change "very front part" to "initial part"*

**Responses:** Thank you very much for this suggestion. This sentence is revised; and the paragraph containing this sentence is moved to the end of section 2a.

20. *l.279: It would be more interesting to (also) discuss the ratio of \tau over \bar{q}, rather than \tau alone. I.e., It would be more physical to compare the relative action of two processes.*

**Responses:** Thank you very much for this comment. The primary object of this study is the role of temperature in multicentennial AMOC oscillation. Thus, we first discussed the effects of temperature feedbacks (of course including the restoring advection feedback whose strength is determined by $\gamma$, the substitute for $\tau$ in this version of manuscript) in section 3. As for $\overline{q}$ (related to flow property), we put it in the part of sensitivity studies (section 5). Additionally, we treated the model parameters as being independent of each other in our model; thus, we tested eigenmode's sensitivity to each parameter, instead of to multiple parameters like $\gamma/\overline{q}$.

*21. l.296-297: This is because the restoring for tau->0 acts essentially as a flux. So you are back in a system equivalent to the salinity-only one. (I.e., flux boundary condition for both T and S and not mixed boundary condition.) In this context everything could be written with a single variable (i.e., density).*

**Responses:** Thank you very much for this comment. Under $\gamma = 0$, $\overline{T_1} = \overline{T_2} = \overline{T_3} = \overline{T_4}$; thus, the temperature advection becomes null. Therefore, the density variation will be exclusively controlled by salinity variation at this point.

*22. l.299: It is even stronger. You have: T1=T*1 and T2=T*2.*

**Responses:** Thank you very much for this comment. We clarified it in the revised manuscript.

*23. l.337-341: I am not sure to follow... Are you suggesting that the action of restoring (what you call "they are removed") is acting more efficiently in 4TS than 3TS? I wonder if this does not come from the fact you use the same \tau for both 4TS and 3TS. This is quite unphysical that a restoring will act as quickly on a 500 m layer and on a 4000 m layer. Maybe \tau should be a function of the thickness of the layer - to be closer to a constant flux (as often hypothesized in more advanced numerical model: −40 W m-2 K-1, as suggested in the NEMO handbook, for instance). Overall there is a problem in the lack of equations/explanations of the 3TS model. One have to guess your treatment of box2 in 3TS...*

**Responses:** Thank you very much for this comment. Our explanation of the 3TS model was not clear, which led to such misunderstanding. The "removed" actually denotes the advection of anomalies out of the subpolar region. Since there are no temperature and salinity differences among boxes 2, 3, and 4, it is the mean advection of anomalies that transports the anomalies out of the subpolar region in the 3TS model. Since boxes 2 and 3 are well mixed, the transportation time of anomalies between boxes 2 and 3 is saved.

24. *l.347-361: I find it hard to follow, with discussion of stratification of the subpolar (which does not exist in 3S or 3TS) and no figure of its evolution has a function of \tau. The lag correlation is not a demonstration of a stabilizing or destabilizing effect. I do not understand this argument. (Lag-)Statistical link (which are not causality) cannot demonstrate a dynamical feedback. Overall I think the problem is simpler: restoring -> stabilizing effect. Their different impact on 3TS and 4TS is not, for me, demonstrated here by this discussion.*

**Responses:** Thank you very much for this comment. We have rewritten these sentences.

The correlation in these figures can be a demonstration of a stabilizing or destabilizing effect, because all processes are connected by deterministic equations. If there is no equation connecting different processes, the correlation between different processes can only suggest some kind of relationship, and one can never say there is causality between them.

In the manuscript, we state that the positive (negative) correlation coefficient at ***lag 0*** is a further illustration of the positive (negative) feedback, because the bases are the equations and the underlying physics. Therefore, the statistical link in this work can demonstrate a dynamic feedback.

Compared to the 3TS model, the subpolar stratification in the 4TS model ($T_2' - T_3'$) increases the time that the subpolar temperature anomaly stays within the subpolar region; thus, it physically has a stabilizing effect. This process is lacking in the 3TS model. After this physical interpretation, we can use the correlation coefficient at lag 0 as a more intuitional representation for the stabilizing/destabilizing effect.

25. *l.386-388: It would be nice to add something like: "If strong enough, the mixing make the 4TS model virtually equivalent to the 3TS one".* Revised.

26. *l.400: You need to show longer simulations. It is unclear for me if the cycle has indeed saturated. The figure suggests that there is a slowdown in the rate of growth but not a stability in the amplitude of the cycle, yet. A large number of period with the same oscillation amplitude would be needed to demonstrate that.* Revised.

27. *l.411: Change "insensitive" to "almost insensitive".* This sentence is deleted.

28. *l.412: Change "dominated" to "controlled".* This sentence is deleted.

29. *l.414-415: Change "lead to" by "stabilize the unstable oscillation and lead to".* This sentence is rewritten.

30. *l.431-432: Yes, but the parameter also "are hardly deviated". So I don't think it demonstrates the "robustness". It rather shows that the stability does affect the period. Please rephrase.*

**Responses:** Thank you very much for this comment. The text on the impact of mixing and the nonlinear relation between AMOC anomaly-meridional difference of density anomaly on the period was removed.

31. *l.447-451: You should be slightly more careful and clear. Your background state is set with parameters, rather than computed as an equilibrium (as in the previously cited study). Hence, despite some advantages (as setting the parameters to what you wish), you take the risk of using un-consistent parameters. This should be highlighted.*

**Responses:** Thank you very much for this suggestion. This paragraph is deleted. We highlighted that our background state is set with parameters.

32. *l.453: I do not find it extremely useful to compare with so much details previous (rather old) studies. I feel that idealized model are useful to explain physics, rather than to validate each other.*

**Responses:** Thank you very much for this suggestion. We removed the annotations related to GT95 and RT97 in Fig. 8. However, it would be nice to see that different theoretical works are consistent.

33. *Fig8-9-10: You show us that the period and e-folding timescale can be a bit affected by the choice of 3 or 4 boxes (Fig. 4). It would need to be clear which one you choose for the plots. A comment on this issue (since GT95 and RT97 are un-consistent on this matter) would be useful.*

**Responses:** Thank you very much for this suggestion. Figures 8-10 are all based on the 4TS model. We clarified this issue in the revised manuscript.

34. *l.499-500: Technically if a term is used to build the period (\bar{q}) it could not affect the growth. This could be easily shown by computing the eigenvalues by hand. Terms that end up in the imaginary part cannot contribute anymore to the real part of the eigenvalues.*

**Responses:** Thank you very much for this comment.

We plotted the dependence of real and imaginary parts of the eigenmode on $\overline{q}$ in Fig. R3 under the standard parameters in Table 1. We find that $\overline{q}$ influences both the oscillation period and e-folding time of the system. $\overline{q}$ affects the system stability through the mean advection feedback, and affects the oscillation period through its influence of the overturning rate. Overall, although we are unable to solve the analytical solution of the 4TS model theoretically, $\overline{q}$ still influences both the e-folding time and period of the eigenmode, as revealed by linear stability analysis (Fig. R3).

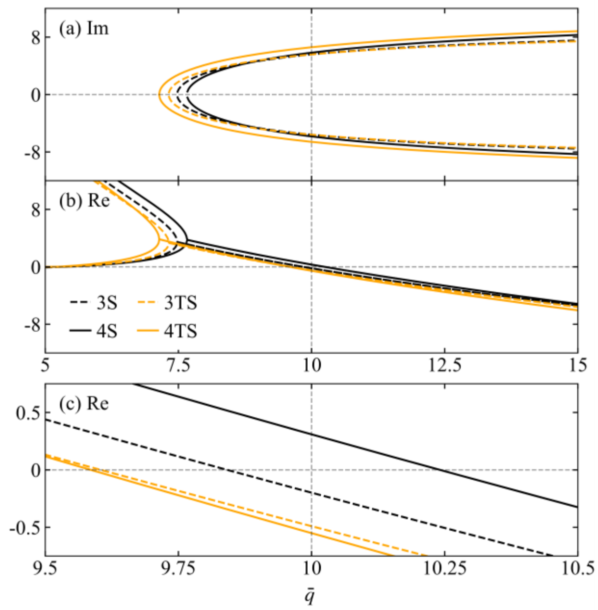


FIG. R3. Dependences of (a) imaginary parts and (b) real parts of eigenvalue $\omega$ on $\overline{q}$ in the 4TS (solid orange curves), 3TS (dashed orange curves), 4S (solid black curve), and 3S (dashed black curves) models under $\lambda = 12\ Sv \cdot kg^{-1}m^3$. (c) is the zoomed-in version of (b) near line $Re(\omega) = 0$. Results of the 4S and 3S models are from LY22. The units of the ordinate are $10^{-10}$ $s^{-1}$. The values of the other parameters are the same as those listed in Table 1. The vertical dashed line denotes the situation under the standard value $\overline{q} = 10\ Sv$.

35. *Fig.9: It is problematic that the realistic regime \bar{q}>14 corresponds to regime 1. A word on that would be useful.*

**Responses:** Thank you very much for this comment. We have removed the annotations of regime 3. We think that under realistic $\overline{q}$, the system should be unstable and there is possibility for self-sustained oscillation. The problem is that in our theoretical model, the depth for boxes 1 and 2 represents the upper ocean instead of the entire AMOC upper branch, and only the AMOC recirculating in the Northern Hemisphere is considered. Thus, the realistic value for $\overline{q}$ should be a smaller value like 10 Sv, instead of a much larger value like >14 Sv. The standard and realistic $\overline{q} = 10\ Sv$ corresponds to an unstable regime; and self-sustained oscillation has possibility to occur; thus, is reasonable.

36. *section 5c: This is a perfect example why I feel that you should stress that background mean state is treated as independent parameters: here an increase in the restoring temperature difference does not affect the mean flow (whereas in general it should).*

**Responses:** Thank you very much for this suggestion. In the revised manuscript, we clarified that the background parameters are all independent of each other.

37. *l.523: Change "gradient" to "difference"* Revised.

38. *l.530-533: Unclear, please rephrase.*

**Responses:** Thank you very much for this comment. We have rewritten the abstract and also revised section 3 of the manuscript, emphasizing that three thermal effects are introduced when considering the temperature equations: (1) an increase of the oscillation frequency, (2) a stabilization of the overall system, and (3)a stabilization of the subpolar stratification. Additionally, these three thermal effects are, respectively, caused by three thermal processes: (1) the fast surface temperature restoring, (2) the negative temperature advection feedback, and (3) the stabilizing subpolar temperature stratification.

39. *l.522: "consume" I do not understand this word in this context. Please rephrase.* This sentence is deleted.

40. *l.552: Change "system" to "oscillation".* This sentence is deleted.

41. *l.558-559: There is even more recent studies in 2 other CMIP6 models which show this kind of multi-centennial oscillation. This should be mentioned in the introduction.*

    https://doi.org/10.1007/s00382-022-06534-4

    https://doi.org/10.1029/2020MS002366

**Responses:** Thank you very much for this suggestion. We rewrote the introduction, and cited these papers.

42. *l.562: Change "easy to realize" to "included in some form"* Revised.

43. *l.569-573: I am not sure to follow the arguments... It is either quite general (and I don't see the point) or the relation to your study is not well explained.* This sentence is deleted.

*44. l.576: Change "On one hand" to "On the one hand"* We rewrote this sentence.

*45. l.578-579: Change "as revealed in this paper" to "consistently with our study"* We rewrote this sentence.

*46. l.579-580: Unclear. Please rephrase.*

**Responses:** Thank you very much for this comment. We revised this sentence as follows: "this also implies that the period of the multicentennial AMOC oscillation is likely to be lengthened in the future, which has not gained attention yet."

*47. l.581: I don't know what you mean there with "portion".* We rewrote this sentence.

*48. l.590: Change "reserved to "included"* Revised.

*49. l.594: Remove "authenticity of"* Revised.

**Replies to Reviewer #2:**

Thank you very much for all of your constructive comments. We have carefully revised our manuscript based on the comments from you and other reviewers. The following are our point-to-point replies.

*This manuscript is an extension of a paper by the same authors by considering the effects of temperature, and its advection and restoring feedbacks on the self-sustained multicentennial oscillation of the overturning circulation. The authors used 4-box, 3-box and diffusive 4-box models to show the condition for the existence of such oscillations and showed how temperature made the system more stable than the previous salinity-only oscillation. I found the results interesting and helpful for us to further understand the behavior of the MOC, especially in the context of ongoing climate change. However, I do have some confusions and concerns about the manuscript as it currently is. The general major comments are below, followed by minor comments on details. Comments are made in the general order of appearance in the manuscript instead of importance.*

**Responses:** Thank you very much for your invaluable suggestions, which help us improve the manuscript tremendously. Combining the comments from all the reviewers, we have revised the manuscript primarily in these following aspects:

1) We have completely rewritten the introduction. Coupled modelling studies on multicentennial AMOC oscillation are synthetically reviewed, the inconsistency among their mechanisms and the necessity for theoretical studies are disclosed. Inadequacy of previous theoretical models in accounting for sustainable multicentennial AMOC oscillation is also discussed. Finally, the potential impacts of thermal processes on AMOC oscillation are raised, justifying the inclusion of temperature effects in this study.
2) In section 2, the choices of parameters are discussed in more detail.
3) In section 3, we categorize the thermal effects more precisely. We propose that there are mainly three effects when including the temperature equations: (1) increase of the oscillation frequency, (2) stabilization of the overall system, and (3) stabilization of the subpolar stratification. These three effects are attributed to the following three processes, respectively: (1) fast surface temperature restoring, (2) negative temperature advection feedback, and (3) stabilizing subpolar temperature stratification. Now, it is easier to understand that the behaviors of the temperature-salinity system are different from the salinity-only model in LY22.

4) In section 4, we more clearly describe the self-sustained oscillation mechanism. In LY22, we proposed that the nonlinear subpolar vertical mixing is crucial for self-sustained oscillation. In the revised manuscript, we further propose that assuming a nonlinear relation between AMOC anomaly and meridional difference of density anomaly can also lead to self-sustained oscillation. We further show that the self-sustained oscillation mechanism not only agrees with that of LY22, but also advances the theory of LY22.

5) Most figures are re-plotted.

**Major comments**

1. *This manuscript shows several versions of box models in which temperature, salinity and overturning circulation interact. For me, it seems that the 4-box and 3-box models are two limits of the diffusive 4-box model introduced the latest, and the self-sustained oscillation, which is the main focus of this manuscript, is realized in the diffusive model. It will be clearer if the diffusive model is introduced first and the manuscript can discuss the focus (self-sustained oscillation) earlier. Also, it seems to me that the 3-box model is a convective version of the 4-box model. However, it makes more sense if the authors set a condition for convection in your numerical simulation (say, density of box 2 is larger or equal to density in box 3) for a transition between the 4-box and 3-box model, instead of treating them completely separately, at least in the numerical part.*

**Responses**: Thank you very much for this comment.

Both our previous paper (LY22) and this manuscript focus on self-sustained oscillation. In LY22, the self-sustained oscillation was introduced before the 3-box models. In this manuscript we investigate the role of temperature in self-sustained multicentennial AMOC oscillation. We first analyze the role of temperature via *linear* stability analysis. After we recognized the overall stabilizing role of temperature, we naturally came up with the question: will the self-sustained oscillation occur in the linear 4TS system with the stabilizing effect of temperature?

The 3-box models are indeed convective versions of the 4-box models. We use the 3TS model to analytically show that the subpolar temperature stratification has a stabilizing effect and can overcome the destabilizing effect of subpolar salinity stratification, in order to reveal the mechanisms of thermal effects and self-sustained oscillation.

Thanks for your suggestion of “including convection in numerical simulation to see the transition between the 4-box and 3-box models.” Note that the convection can only be included in the 4-box models. The 3-box models are the results after the convection.

In an earlier draft of LY22, we did set a condition for convection in the 4S model (similar to your suggestion) in the numerical simulation; and self-sustained oscillation occurred (Fig. R4). The convection processes considered included those due to both static instability and convective instability, which can be expressed as follows,

$$|S_2 - S_3| > S_b \Rightarrow \begin{cases} S_2 - S_3 > S_b^+ \,, & \mathit{Static\ instability} \\ S_2 - S_3 < S_b^- \,, & \mathit{Convective\ instability} \end{cases}$$

where $S_b^+$ and $S_b^-$ are the thresholds of vertical salinity contrast for convection, whose absolute values are set to the same for simplicity. Note that $S_b^+$ and $S_b^-$ can be different, which does not affect the conclusions of this study. In reality or in ocean-alone and coupled models, the stratification at the North Atlantic deep-water formation region is such that fresh and cold water lies on top of saline and warm water (The Lab Sea Group 1998), especially in winter. Under this background stratification, deep convection can occur.

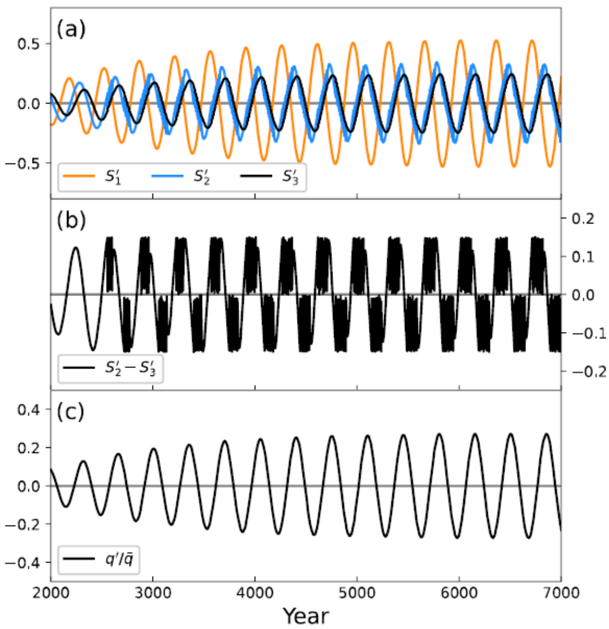


FIG. R4 Self-sustained stable oscillations in the 4S model with both types of deep convection considered. (a) Time series for $S_1'$, $S_2'$, and $S_3'$ (units: $psu$); (b) time series for $S_2' - S_3'$ (units: $psu$); (c) time series for $q'/\bar{q}$. Before the convection occurs, the evolution of the system in the first 2500 years shows growing oscillation. In (b), the packed black curves show the happening of deep convection.

In the final version of LY22 and in this study, we decided enhanced vertical mixing is physically more reasonable than the "one-step" convection in the numerical integration. At least, the changes in salinity and temperature in the subpolar ocean will be much smoother. The convection is just one situation of the enhanced vertical mixing.

2. *Also, the authors discuss eigenmode sensitivity in section 5, while some sensitivities have already been discussed before (e.g., the effect of $\lambda$). The structure of the manuscript will be clearer if such discussions can be combined.*

**Responses**: Thank you very much for this suggestion. The discussion of eigenmode on $\lambda$ in section 2 is to reveal the thermal effects (accelerating the oscillation, stabilizing the overall system, and stabilizing the subpolar stratification). This led to the investigation of thermal processes in section 3. To reveal the sensitivity of eigenmode to flow properties, we analyzed the sensitivities of 4TS eigenmode to $\lambda$ and $\overline{q}$ in section 5b. These two parameters are inseparable since $q'$ is jointly determined by $\lambda$ and $\overline{q}$. That is why the discussion of $\lambda$ occurred twice.

3. *In the discussion of the 4-box and 3-box models, my impression with the results is that the effect of temperature can be described as "marginal" at best (change of period and its contribution in q'), even though it indeed makes the system more stable. This might be due to the fact that the four- and three-box models are introduced before the diffusive version in which self-sustained oscillation is finally emphasized. The manuscript as it is now leaves me the impression that temperature is not a crucial factor, even though including temperature is more realistic.*

**Responses**: Thank you very much for this comment. Yes, the temperature effects are marginal. However, since in reality there is always temperature, we have to investigate thoroughly the role of temperature. After seeing "temperature is marginal" clearly, we can safely neglect the temperature effects and be more focus on the salinity effects in future studies on this issue.

**Minor comments**

1. *line 33: "caused by" instead of "contributed by".* Revised.
2. *line 38: remove "can".* We rewrote this sentence.
3. *line 39: word "thus", the causality is not clear in this sentence.* We rewrote this sentence.
4. *line 40: "which is less stable" instead of "with less stability".* We rewrote this sentence.

5. *line 49: "its thermohaline circulation portion" does not need to include "circulation".* The introduction has been rewritten.

6. *line 58: remove "following the Newtonian law".* The introduction has been rewritten; and this sentence is deleted.

7. *line 65–67: the description of how restoring-advection feedback is not clear. Do you mean lead to more artificial heat loss due to restoring after a positive MOC perturbation? Will this actually make the SST lower such that it will hinder the AMOC from recovering or just partially hinder warming effect? If there is still warming, AMOC should decrease anyway, maybe to a different extent. Some references may be helpful.*

**Responses**: Thank you very much for this comment. As you pointed out, more heat will be lost due to restoring after positive MOC perturbation, which was illustrated in Zhang et al. (1993) and Lucarini and Stone (2005). Therefore, the net effect of restoring itself is to hinder the warming effect, although as a whole there is still warming and the AMOC will decrease in strength.

The positive restoring advection feedback describes a relation between temperature restoring and AMOC (advection), not between temperature restoring and temperature itself. In the revised manuscript, this feedback is stated clearly with more detail. Take box 2 as an example, by limiting the increase of subpolar temperature anomaly and thus the negative temperature-advection feedback, the restoring effect manifests as a positive feedback between the restoring term and AMOC anomaly. In other words, such restoring advection feedback is to increase AMOC anomaly, i.e., to amplify the initial AMOC perturbation.

We would like to emphasize that the temperature feedbacks in the 4TS model should be viewed as a combination of negative temperature advection feedback and positive restoring advection feedback. The latter is driven by the former, which in turn hampers the former. Their combined effect is negative feedback. Nevertheless, the restoring advection feedback should still be termed as positive feedback.

**References**:

Lucarini, V., & Stone, P. H. (2005). Thermohaline circulation stability: A box model study. Part I: Uncoupled model. J. Climate, 18(4), 501-513.

Zhang, S., R. J. Greatbatch, and C. A. Lin, 1993: A reexamination of the polar halocline catastrophe and implications for coupled ocean atmosphere modeling. J. Phys. Oceanogr., 23, 287-299.

*8. line 68–69: references needed to verify that the restoring feedback never fully overruns the other feedback.*

**Responses**: Thank you very much for this comment. We stated that the restoring advection feedback never overruns the temperature advection feedback, instead of the other feedbacks. The temperature feedbacks in the 4TS model should be viewed as a combination of negative temperature advection feedback and positive restoring advection feedback. The latter is driven by the former, which in turn hampers the former. However, their combined effect is negative feedback.

Physically, the most extreme form of restoring feedback is that it completely fixes the variation of temperature thus making the total effect of temperature equations null. Therefore, it cannot overrun the temperature advection feedback since the total effect of temperature equations would not be destabilizing. In Nakamura et al. (1994) and Marotzke (1996), the authors described the feedback between meridional atmospheric heat transport and AMOC, which has a similar effect of restoring advection feedback in ocean models employing mixed boundary conditions. It works as follows: an initial positive perturbation of the AMOC lowers meridional temperature difference, weakens meridional atmospheric heat transport, thus lowering high-latitude temperature; the end result is a net strengthening of the AMOC via this feedback alone. Marotzke (1996) further stated that: “since the change in atmospheric transports is a response to anomalous meridional temperature contrasts, which at most can eliminate its cause completely but can never overshoot, this feedback can’t be stronger than the temperature advection feedback.” This supports that the restoring advection feedback cannot overcome the temperature advection feedback.

**References**:


Marotzke, J., 1996: Analysis of thermohaline feedbacks. Decadal Climate Variability: Dynamics and Predictability, D. L. T. Anderson and J. Willebrand, Eds., Springer, 333-378.

Nakamura, M., P. H. Stone, and J. Marotzke, 1994: Destabilization of the thermohaline circulation by atmospheric eddy transports. J. Climate, 7, 1870-1882.


*9. line 75: the comma before the word “whose” is not needed.* The introduction has been rewritten.

*10. line 76–79: How do you relate period to the degree of stability of the system? How stable the system is determined by the real part of the eigenvalue while period is the imaginary part. And some people may intuitively think an oscillation of longer timescale is more stable. Stability and* *period are often discussed together throughout this manuscript, but their relation (if any) seems confusing to me.*

**Responses**: Thank you very much for this comment. We rewrote the introduction, including these lines. In those lines of the original manuscript, we regarded the restoring advection feedback as a fast process and linked it to period therein, although we also discussed its destabilizing effect elsewhere in the manuscript. Actually, throughout the paper, we separate the discussion of temperature's effects on period from that on stability.

*11. line 86: change the word "unravel".* The introduction has been rewritten.

*12. line 106: define abbreviation before use.* This abbreviation is defined in the revised manuscript.

*13. formulation of the box models: it is assumed that deep water completely upwells in the interior of the ocean. However, people have also found that upwelling in the Southern Ocean is also, if not more, important than the interior diffusive upwelling. How important the authors think the Southern Ocean is for their results here?*

**Responses**: Thank you very much for this comment. In the rewritten introduction, the Southern Ocean and even the Arctic Ocean may play a role in the multicentennial AMOC oscillation. Therefore, these two oceans definitely play a part. The box models can only grab the most fundamental physics, the importance of the Southern Ocean cannot be specified in the one-hemisphere model.

Actually, we are working on a two-hemisphere model, in which the Southern Ocean is included. We believe the Southern Ocean plays a role in the multicentennial oscillation of the AMOC, and also in the millennial oscillation of the climate system. We are also using a coupled climate model to study the role of the Southern Ocean.

Our guess is that including the Southern Ocean in the box model should prolong the oscillation period of the AMOC. The system stability might also be affected, since $q'$ now is likely to be determined by the density difference between the North Atlantic and the Southern Ocean. However, the temperature and salinity advection feedbacks related to the original box 2, together with the northern subpolar vertical mixing process, will still be at work; thus, the basic self-sustained oscillation mechanism should remain the same.

14. *Usually $\tau$ is a symbol for timescale. Here it is a reversal of a timescale. Using another letter may be less confusing to some people.* Revised. $\tau$ is replaced with $\gamma$ throughout.

*15. Fw: directly say that it is an artificial salt flux representing the effect of freshwater flux, in context and in the table.* Revised.

*16. the choice of 10 Sv for q is not very characteristic of time-averaged AMOC. Why is this value chosen. In the discussion part of the paper, we see some references in which the timescale of MOC oscillation is dependent on q. How different the results of this manuscript will be if q is more realistic? Also, $\bar{q}$ is directly given, is it consistent with the time-mean large-scale meridional density gradient as specified by the formula (related by $\lambda$)?*

**Responses**: Thank you very much for this comment.

We admit that the choice of $\overline{q}$ is critical to the oscillation timescale. In the observation, the maximum mean AMOC is about 20 Sv. This mass transport includes water in the upper 1000 m and water from the Southern Hemisphere. In a one-hemispheric box model with the upper-ocean depth set to 500 m, the mean mass transport should be remarkably smaller than the realistic value. Otherwise, the turnover timescale for a one-hemisphere box model would be unrealistically short. $\overline{q}$ in GT95 is much larger than ours; and their subpolar boxes are much smaller than ours, so that the dominant timescale in GT95 box model is the decadal timescale, instead of the centennial timescale. Our single-hemispheric model incorporates only the AMOC recirculating in the Northern Hemisphere, so that a smaller mean AMOC is reasonable.

Table R1 lists the eigenvalues for the 4TS and 4S models under the parameters in Table 1, but with $\overline{q} = 18\ Sv$. The e-folding time is 36 years for decaying oscillations in the 4TS model, while it is 40 years for decaying oscillations in the 4S model, both are more stable than their counterparts under $\overline{q} = 10\ Sv$. The period is 208 years in the 4TS model, while it is 220 years in the 4S model, both shorter than their counterparts under $\overline{q} = 10\ Sv$. The results suggest that a larger $\overline{q}$ leads to a more stable system with a faster oscillation, consistent with the results in section 5b of this revised manuscript. However, the overall results of this paper are not substantially altered with a larger $\overline{q}$ and there is still the potential for multicentennial oscillation.

TABLE R1. Eigenvalues ($10^{-10}\ s^{-1}$) for the 4TS and 4S models under the parameters of Table 1, but with $\overline{q} = 18\ Sv$.

| 4TS | 4S | Physical Significance |
|---|---|---|
| $-8.79 \pm 9.57i$ | $-7.91 \pm 9.07i$ | Oscillatory mode |

| | | |
|---|---|---|
| 0 | 0 | Zero mode |
| −393 | — | Damped mode |
| −330 | — | Damped mode |
| −65.5 | −65.5 | Damped mode |
| −9.84 | — | Damped mode |
| −1.47 | — | Damped mode |

In our parameterization, it is $q'$ rather than $\overline{q}$ that is proportional to the meridional difference of density anomaly $\Delta\rho'$. Therefore, $\overline{q}$ is directly given instead of being related to the large-scale meridional difference of the mean density.

*17. The upper layer where restoring happens is as thick as 500 meters, which is too deep for restoring to be justifiable. Also, the total thickness of ocean in these box models is 4 km, but these models are only for the upper cell which may not penetrate so deep in the ocean.*

**Responses**: Thank you very much for this comment. We admit that in reality the restoring can only happen within the surface layer with a depth of a few tens of meters, which is also seen in studies using models of higher complexity (Marotzke and Willebrand 1991; Weaver and Sarachik 1991; Mysak et al. 1993). However, there is not such surface layer in our theoretical model due to its simplicity.

For model conciseness, we permit the restoring to happen within the whole depth range of boxes 1 and 2, which is commonly adopted in box model studies. As compensation, we chose 1-year restoring timescale, which is significantly long due to the rather thick restoring surface layer ($D_1$, 500 m). In GT95, the authors permitted the surface restoring to take place over the 300-m thick surface layer, and chose 180-day restoring timescale. Also, in RT97 the surface restoring can happen over the 1000-m thick surface layer; they also chose 1-year restoring timescale to compensate such thickness.

Boxes 1 and 2 are for the upper ocean, while boxes 3 and 4 are for the deeper ocean including the southward NADW, not the upper cell. This is also evidenced by the southward mean AMOC between boxes 3 and 4 in our theoretical model. Therefore, the 4-km depth for our box model is reasonable.

**References**:

Griffies, S. M., and E. Tziperman, 1995: A linear thermohaline oscillator driven by stochastic atmospheric forcing. J. Climate, 8, 2440-2453.

Marotzke, J., and J. Willebrand, 1991: Multiple equilibria of the global thermohaline circulation. J. Phys. Oceanogr., 21, 1372-1385.

Mysak, L. A., T. F. Stocker, and F. Huang, 1993: Century-scale variability in a randomly forced, two-dimensional thermohaline ocean circulation model. Climate Dyn., 8, 103-116.

Rivin, I., and E. Tziperman, 1997: Linear versus self-sustained interdecadal thermohaline variability in a coupled box model. J. Phys. Oceanogr., 27, 1216-1232.

Weaver, A. J., and E. S. Sarachik, 1991: Evidence for decadal variability in an ocean general-circulation model - An advective mechanism. Atmos. Ocean, 29, 197-231.

*18. How are the parameters T1, T2, S1, S2, ~~T1 * and T2 *~~ determined? Are they based on some dataset?*

**Responses**: Thank you very much for this comment. They are based on the CESM1 simulation analyzed in LY22.

*19. equation 3: The use of letter κ0 may confuse it with diffusivity. What is the unit of this coefficient, and how is it determined?*

**Responses**: Thank you very much for this comment. In the revised manuscript, we rewrote this paragraph.

$\kappa_0$ denotes the restoring coefficient, which reflects the strength of SST relaxation toward $T_1^*$ or $T_2^*$, with the units of $W/(m^2 \cdot ℃)$. Originally, it was determined from observations. Bretherton (1982) suggested that a large value like 100 $W/(m^2 \cdot ℃)$ be chosen for the sea surface with a few tens of kilometers across, since atmospheric advection of heat is also important for small-scale surface heat flux. He also stated that a small value like 2 $W/(m^2 \cdot ℃)$ be chosen for global SST, whose restoring is largely determined by local radiation instead of atmospheric advection. In later modelling studies, this value was usually set to a constant, leading to 1-2 month restoring timescale for a surface layer of a few tens of meters (Marotzke and Willebrand 1991; Weaver and Sarachik 1991; Mysak et al. 1993; Pierce 1996).

**References**:

Bretherton, F. P., 1982: Ocean climate modeling. Prog. Oceanogr., 11, 93-129.

Marotzke, J., and J. Willebrand, 1991: Multiple equilibria of the global thermohaline circulation. J. Phys. Oceanogr., 21, 1372-1385.

Mysak, L. A., T. F. Stocker, and F. Huang, 1993: Century-scale variability in a randomly forced, two-dimensional thermohaline ocean circulation model. Climate Dyn., 8, 103-116.

Pierce, D. W., 1996: Reducing phase and amplitude errors in restoring boundary conditions. J. Phys. Oceanogr., 26, 1552-1560.

Weaver, A. J., and E. S. Sarachik, 1991: Evidence for decadal variability in an ocean general-circulation model - An advective mechanism. Atmos. Ocean, 29, 197-231.

*20. line 157: for equation of state, how realistic is it to use a uniform value for α and β across such a wide latitudinal range? Around what temperature and salinity are these coefficients determined?*

**Responses**: Thank you very much for this comment.

The thermal expansion coefficient of $1.468 \times 10^{-4}\ ℃^{-1}$ and the haline contraction efficient of $7.61\ \times 10^{-4}\ psu^{-1}$ are derived from UNESCO (1987) around 9℃ and 35 $psu$ under 0 dbar. We admit that these two coefficients can vary not only with temperature, but also with salinity and pressure. However, we chose constant values in our study for simplicity, as the essence for a box model lies in its ability of revealing mechanisms via its conciseness instead of statistical accuracy. Constant thermal expansion and haline contraction coefficients have been adopted in a large number of theoretical models for simplicity, and some even covered a larger latitudinal extent (Scott et al. 1999; Lucarini and Stone 2005; Alkhayuon et al. 2019; Shi and Yang 2021; Wei and Zhang 2022).

**References**:

Alkhayuon, H., P. Ashwin, L. C. Jackson, C. Quinn, and R. A. Wood, 2019: Basin bifurcations, oscillatory instability and rate-induced thresholds for Atlantic Meridional Overturning Circulation in a global oceanic box model. Proc Math Phys Eng Sci, 475, 20190051.

Lucarini, V., and P. H. Stone, 2005: Thermohaline circulation stability: A box model study. Part I: Uncoupled model. J. Climate, 18, 501-513.

Scott, J. R., J. Marotzke, and P. H. Stone, 1999: Interhemispheric thermohaline circulation in a coupled box model. J. Phys. Oceanogr., 29, 351-365.

Shi, J. Q., and H. J. Yang, 2021: Bjerknes compensation in a coupled global box model. Climate Dyn., 57, 3569-3582.

UNESCO, 1987. International oceanographic tables. Tech. Paper Mar., Sci. 40, 196 pp.

Wei, X., and R. Zhang, 2022: A simple conceptual model for the self-sustained multidecadal AMOC variability. Geophys. Res. Lett., 49, 11.

*21. line 196–197: "1025 years for growing oscillations" and "576 years for decaying oscillations" instead of "positive 1025 years" and "negative 576 years".* Revised.

*22. line 199: Figure 3 is mentioned before Figure 2.* Revised.

*23. line 208: change "y=0" to "Im(ω) = 0" to be more clear.* Revised.

*24. Figure 2(a): The changes between the "S-only" and "T" models happen mainly where λ is large. but less so when it is small. Why?*

**Responses**: Thank you very much for this comment. Compared to salinity-only models, the range of $\lambda$ with complex eigenvalues is increased in the TS models, especially when $\lambda$ is large. As $\lambda$ increases, the imaginary parts of the S models will move to 0 sooner than those in the TS models; thus, the changes between the S and TS models are the most evident when $\lambda$ is larger.

*25. line 225 (Figure 2 caption): "zoomed-in" instead of "magnified".* Revised.

*26. table 3: Why would λ be a negative value? Will it actually be zero? If so, what does it physically imply?*

**Responses**: Thank you very much for this comment. Physically, $\lambda$ cannot be zero or negative since the AMOC will be stronger under a bigger meridional difference of density anomaly (positive $\lambda$). However, we decided to include $\lambda \leq 0$ in Fig. 2 so as to show both the upper and lower limits for the existence of imaginary part. This would not influence the results presented in this manuscript since the analyses are only related to $\lambda > 0$ case.

*27. line 248: But the advection feedback is related to large-scale circulation, why would it be local?*

**Responses**: Thank you very much for this comment. The local advection feedback we referred to here is the negative temperature advection feedback $q'\left(\overline{T_1}-\overline{T_2}\right)$. It influences $T_2$ without affecting $T_1$; therefore, it is a local one. Also in reality and complex coupled model, $q'$ denotes local change. When it comes to the feedback between mean advection of temperature anomalies and AMOC anomaly ($\overline{q}(T_1'-T_2')$), it influences $T_2$ through transporting temperature anomaly from box 1 to box 2; thus, it is a remote feedback (non-local).

*28. line 272–277: Move this paragraph to before the discussion of Figure 2.* Revised.

*29. line 293–294: Temperature is still advected and enters the EOS, correct? I am a little confused why the $\tau$ = 0 case should be the same as the salinity-only case.*

**Responses**: Thank you very much for this comment. Temperature will always be advected as long as the temperature equations are involved. However, when $\gamma = 0$ (or approaches 0, more specifically), from Eq. (8) we have $\overline{T_1}-\overline{T_2}=0$, thus $\overline{T_1}=\overline{T_2}=\overline{T_3}=\overline{T_4}$. Therefore, $q'\left(\overline{T_1}-\overline{T_2}\right)$, $q'\left(\overline{T_2}-\overline{T_3}\right)$, $q'\left(\overline{T_3}-\overline{T_4}\right)$, and $q'\left(\overline{T_4}-\overline{T_1}\right)$ will be 0; thus, there will be no variation in temperature at all. Consequently, the temperature advection would not warm or cool any boxes; and the temperature equations are useless.

*30. line 306: What is the reasonable range of $\tau$ since this is an artificial form of forcing? How is this reasonable range determined?*

**Responses**: Thank you very much for this comment. Typically, the restoring timescale is 1-2 months with a surface layer of a few tens of meters. However, since our theoretical model does not have such an explicit surface layer, we permit the restoring to happen over the whole depth range of boxes 1 and 2. Consequently, the surface layer now is 500 m, almost 10 times the typical value. As compensation, we should choose a restoring timescale much longer than the typical value of 1-2 months. A value of 1-year or longer is reasonable, while 9-10 months might also work. However, it should not be as short as a season or shorter.

*31. lines 359–361: Judging from Figure 4, over some range of $\tau$ 3TS is more stable. Be more specific in this sentence.*

**Responses**: Thank you very much for this comment. As we have explained earlier in our replies, the reasonable range of restoring timescale is no shorter than one year and also cannot be too long.

Therefore, although over some range of $\gamma$ the 3TS model is more stable, the 4TS model is more stable than the 3TS one under realistic $\gamma$.

*32. line 394–396: Why? It is the averaged T/S in box 2 and 3 that determines the strength of MOC, which will not be changed here since mixing only occurs between these two boxes. Is this stabilizing effect from the surface conditions?*

**Responses**: Thank you very much for this comment. A warmer (more saline) anomaly in the subpolar region has stabilizing (destabilizing) effect on AMOC oscillation. The subpolar vertical mixing moves the anomalies from box 2 to box 3 faster. Therefore, although the mixing process conserves salt and temperature, it removes the warm and salty anomalies from the subpolar region more rapidly, limiting their stabilizing and destabilizing effects, respectively.

*33. section 4b: The only purpose of this is to make self-sustained oscillation possible by using some complicated relation between density gradient and MOC, otherwise I find it distracting. Consider putting in supplementary.*

**Responses**: Thank you very much for this comment.

We have improved the logic of section 4 in the revised manuscript. Section 4a is to show that the self-sustained oscillation is able to be realized even with the presence of destabilizing vertical mixing (the destabilizing temperature mixing overcomes the stabilizing salinity mixing).

In section 4b, we use this seemingly complicated relation between meridional difference of density anomaly and AMOC anomaly (in fact this relation is physically simple; it means restraining of $q'$ is introduced when the meridional difference of density anomaly is large) to introduce a degree of nonlinearity in the absence of mixing. The realization of self-sustained oscillation in both the 4TS and 3TS models here highlights the role of nonlinearity, while the self-sustained oscillation is not sensitive to the exact form of the restraining term. Only with what in section 4b, can we generalize the LY22 self-sustained oscillation mechanism of "a combination of salinity advection and enhanced mixing" to "a linearly growing oscillation dominated by advection and a nonlinear restraining effect from restraining terms" in this manuscript. Therefore, we prefer to place section 4b in the main body of the manuscript.

*34. line 455: "much less stable than ours" instead of "stability far lower than ours".* This paragraph is deleted.

*35. line 563: What do you mean by "random components"?* We deleted discussion on this.

*36. line 568: "models of higher complexity" instead of "higher complexity models".* This sentence is deleted.

*37. line 576–579: The collapse is more related to bifurcation than stability of oscillation.*

**Responses**: Thank you very much for this suggestion. We have removed the contents about AMOC collapse.