We thank the reviewer for their careful review and for providing thoughtful feedback. Below, we first respond (in blue) to the general comments, after which we provide detailed replies to the specific comments.

General comments

This study uses a conceptual box model of the sub polar gyre to perform a bifurcation analysis and study noise-induced tipping. They find this model to be bi-stable, and test tipping from different initial conditions in phase space. As expected, the initial parametrisation and noise strength contribute to the likelihood of collapse in the time window observed. While an interesting model of the SPG, the model details and the implementation of noise needs to be much better explained and justified. I find that the results presented can be interpreted as simply confirming already known results of noise processes, and may not have such high impact. The grammar, equation, and some figure presentations need to be revised.

Reply: We agree with the reviewer's comments about the description of the model and the implementation of the noise – in the current manuscript, this is indeed somewhat lacking and we will make the required changes to explain and motivate the model and noise better. We provide answers to the reviewer's specific comments about the noise and model parameters in the detailed comments below.

We also agree that many of our results can be expected from already known results of more general noise processes. However, we do believe that our results can be of use for the SPG modelling community. The collapse of convection in the SPG region is often presented as permanent (e.g. Armstrong McKay et al. (2022), Sgubin et al. (2017), Swingedouw et al. (2022)). This conclusion is based on modelling studies with Earth System Models (ESMs). We hope that indicating the possibility of non-permanent collapse in a conceptual model can provide useful context to these model studies. We will make this more clear in the revised manuscript.

Specific comments

Section 2.1 I have difficulty understanding how the noise has been added to the model equations. First, it is not clear to me why there are not dynamical equations for boxes 2 and 4- this should be explained.

Reply: In the Born and Stocker model (2014, hereafter: BS14), convection can only occur in the inner two boxes (1 and 3) of the model, that is, it assumes that convection occurs in the center of the gyre. The temperature and salinity in the gyre center are affected by A) the hydrographic properties of the surface current, B) the hydrographic properties of the deep current, and C) the atmosphere. A, B, C are treated as prescribed conditions for the variables in the center of the sea and are set constant at values of T, S, T_0 and F that are estimated from observations (BS14). Box 2 and 4 thus provide boundary conditions for boxes 1 and 3 and are therefore not represented by dynamical equations, much like the atmosphere is not represented by dynamical equations, but only by the terms T_0 and F.

Changes in text: We will clarify that the values of T, S, U_s and U_d in boxes 2 and 4 are boundary conditions and not dynamic variables in the model.

Then, since there is no dynamic equation for box 2, the noise on S_2 is added to Equation 1d. I understand that since the noise is applied as an OU process, these noise processes have different time scales and will represent different physical processes. However, the way the noise is presented in the equations, this still appears to me as two noise processes added to box 1, not one on each box (1 and 2), therefore all noise processes would affect just box 1.

Reply: The reviewer is correct in noting that both noise processes affect box 1. In our adjusted model, we apply noise to different variables. Ultimately all noise processes affect box 1, but due to the different nature of the parameters F and S_2 (see also our answer to the question below), we apply the noise to these two variables separately.

Changes in text: We will clarify that the noise is added to the parameters S_2 and F and not to the boxes as such.

The amplitude of the noise term with zeta_S will also be heavily influenced by the pre-factors, and it is not clear to me that this is fully taken into account.

Reply: This is correct. The prefactors are a result of the relative importance of the physical mechanism between the box-2 salinity and freshwater terms in the salinity budget of box 1. Varying F directly increases or decreases the amount of freshwater that's added to box 1. By comparison, varying S_2 only indirectly changes the salinity of box 1 by first changing the strength of the baroclinic current U_s, and then the magnitude of horizontal eddy transport mu_H*U_s*(S_2 – S_1). This is a highly nonlinear process which reflects the physical mechanism by which salinity anomalies are transported from the boundary current to the convective core of the gyre. To some extent, we take this difference into account by choosing different values for the noise amplitude sigmaS and sigmaF (Table 2).

As the noise is added to the parameters S_2 and F, all pre-factors that apply to S_2 therefore also apply to the noise term zeta_S. We discuss this to some extent in in Sect. 4.2 (l. 336 – 346).

Changes in text: In l. 135 add "We note that S_2 and consequently the noise term zeta_S are influenced by nonlinear prefactors in a way that F and zeta_F are not. This is discussed in Sect. 4.2."

Similarly, why would precipitation only affect the surface gyre current and not the surface core box? What does 'precipitation upstream' mean (l.175)?

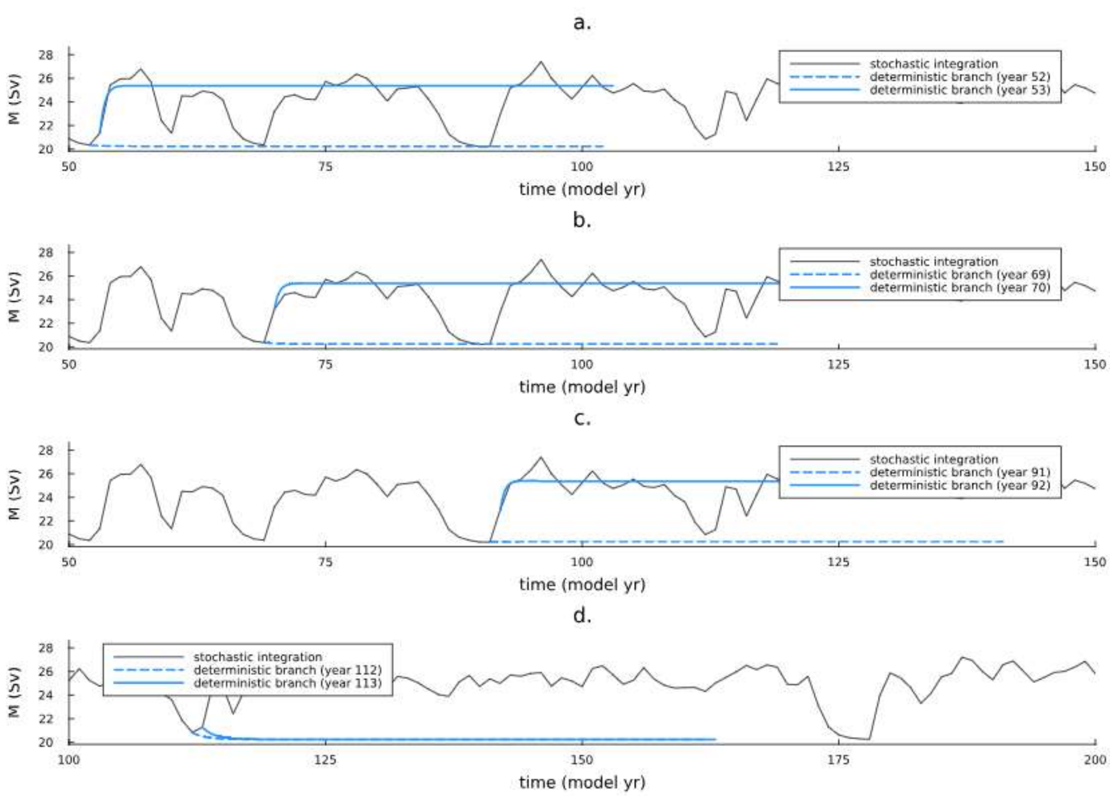
Reply: Precipitation affects both the surface gyre current and the surface core box. In fact, the freshwater flux F mostly represents precipitation (l. 89). The salinity of the surface gyre current S_2 is influenced by the salinity of the currents in the Nordic and Irminger Seas and the Baffin Bay, sea ice melt and export from these regions, and Greenland Ice Sheet meltwater. 'Precipitation upstream' refers to precipitation that affects the salinity of the aforementioned currents, but we agree that this term is unclear.

Changes in text: We will revise the text throughout to make clear what physical processes contribute to (variability in) the values of F and S_2. We will replace the term 'precipitation upstream' with 'precipitation'.

Section 3.3 It is not clear to me that the collapses discussed are full tipping events? Are these full transitions to the alternative state? In the discussion lines ~250-270. It is noted 'even in the least stable case… the gyre never becomes fully nonconvective'. It is therefore not clear to me that a full tipping event has taken place? What does the alternative stable state correspond to, and therefore is this an expected result? Given the bifurcation diagram is found, could one not check if the system has actually transitioned to the alternative state and is in the alternative basin of attraction?

Reply: We thank the reviewer for questioning this. To confirm that the model is indeed transitioning to the alternative state, we ran additional simulations. We ran time integrations with noise in the gyre salinity S_2 (identical to those described in the current manuscript) and identified events when the gyre transport M was below a threshold of 21 Sv (see discussion below) for multiple consecutive years.

We then branched off from this simulation twice for each event. First, we ran a deterministic simulation with no noise using the values of T1, S1, T3, S3 from the last year of the event as initial conditions. We also ran a deterministic simulation using the values of T1, S1, T3, S3 from one year after the first branched-off simulation as initial conditions. These simulations are shown below for four events (for parameter set B):



Clearly, when the gyre transport is below the threshold value, it is in the alternate basin of attraction and therefore moves to the constant non-convective (low-transport) state. We conclude from this analysis that the collapses we discuss are full transitions between states.

We note that defining the threshold is not easy. In our manuscript, we used a threshold value of 22 Sv. However, when running the branched deterministic simulations, we found that this value is too high; when $21 < M < 22$, sometimes the gyre is still in the basin of attraction of the convective state and therefore returns to high values of M. This can be seen in the bifurcation diagram (Fig. 2a). Finding the exact value of the threshold is, unfortunately, computationally expensive. We propose erring on the side of caution and taking a threshold value of 21 Sv. This means our statistics will underestimate the amount of years in which the gyre is in a non-convective state (see panel d in the figure above), but we expect this difference to be minor, since only one or two years without convection will not be counted in only a few cases.

A last note is that for the deterministic simulations for parameter set A (see Fig. 3) the gyre always returns to the convective high-transport state. This is expected since for these parameters the gyre is in the mono-stable regime, and the non-convective regime is unstable. Noise on S2 can temporarily allow for a transition to the bi-stable regime, but as soon as it is removed the system returns to the convective state.

We thank the reviewer for this feedback, since it has improved our analysis.

Changes in text: We will run the analysis in Sect. 3.3 again with a threshold of 21 Sv instead of 22 Sv and change the figures, text and tables accordingly. We will clarify that branched-off simulations show that we observe full state transitions for parameter sets B-F, but not for A and interpret the variability for parameter set A as noise events followed by a recovery.

The results discussed on page 11, I think are expected mathematically. As you move the initial conditions in phase space, you essentially start the simulations with different effective potential barrier heights for transitions. Therefore, with the same noise amplitude and a fixed time, a different percentage of transitions will take place according to large deviation theory (Freidlin & Wentzell, 1984, Bouchet & Reygner, 2016).

Reply: We agree with the reviewer that these results are expected mathematically. However, we do believe that presenting these results in the context of the SPG can be insightful for the SPG modelling community.

Changes in text: We will frame these results in the context of large deviation theory as described by Freidlin & Wentzell (1984) and Bouchet & Reygner (2016). We will also add a reference to Kuhlbrodt & Monahan (2003) who discuss this in the context of convection.

Additionally, in some of these 'short excursions' if there is not a full transition to the alternative state, this would then be an example of a noise event followed by a noise-induced recovery (Chapman, Ashwin et al 2024). This could be checked, and the tipping (or not) mechanisms could be identified since the bifurcation diagram is known.

Reply: See discussion above. We will discuss the noise event followed by noise-induced recovery we observe for parameter set A in the context of Chapman, Ashwin et al. (2024).

Grammar and phrasing needs revising throughout.

Reply: We will critically revise the grammar and phrasing throughout the manuscript.

Model equations need to be defined more rigorously, not all variables are defined, should be defined immediately after the Equations.

Reply: We will define all variables immediately after the Equations.

Where do the values of the model parameters come from? Literature, GCMs, physical estimates from observations?

Reply: All model parameters are taken from BS14, who estimated the values based on observations. We will clarify this in the text.

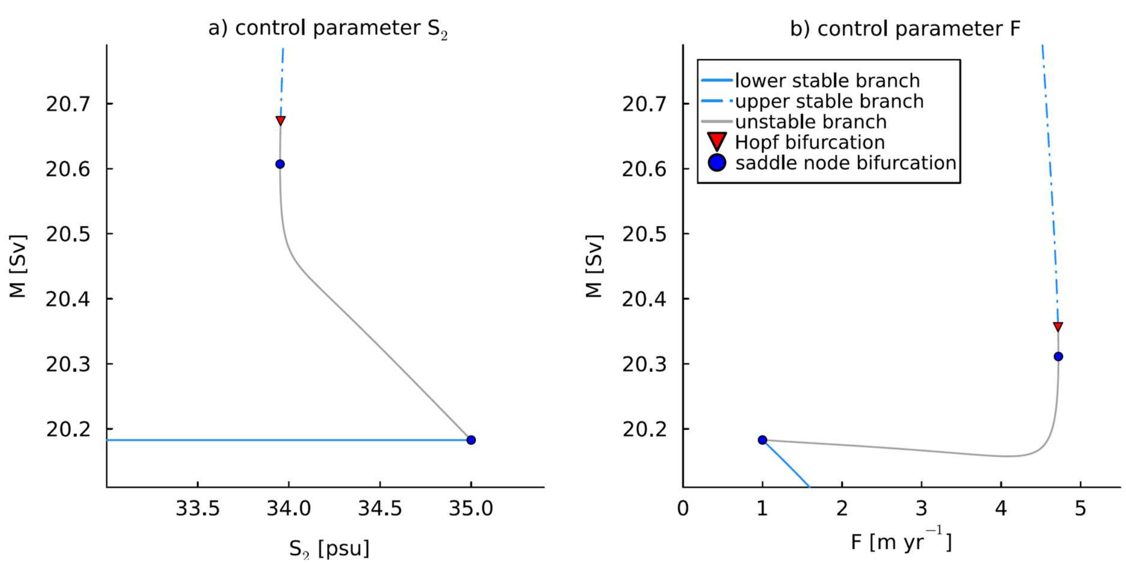
Changes in text: In the description of Table 1 change "The values were calculated from the default model parameters as outlined in Table 1 of Born and Stocker (2014) and Appendix A. No values are given for the parameters mu_h, mu_C, and mu_A, as these values do not have an intuitive interpretation."

to

Suggestion: "We use the parameter values in Born and Stocker (2014) (their Table 1, here Table A1), to compute the parameter values for our non-dimensionalised model. Their parameter estimates are based on observations and expert assessment. The non-dimensionalisation introduces some additional dimensionless parameters (eta, mu_h, mu_C and mu_A), for which no dimensional values are given for lack of interpretability."

Figure 2 Labels need to be much larger, and a lot of white space can be removed from both subfigures. Could a 'zoom in' panel be provided near the hopf point to allow the detail there to be seen.

Reply: We thank the reviewer for this suggestion. We will increase the font size in figures throughout the manuscript and remove white space where possible. We will add a "zoom in" figure (see below) in the appendix for interested readers, as we think that adding such a figure in the main text will distract from the main message.



Line 195 was other continuation software tested as well? Is the hopf super- or subcritical? It is concerning to me that this cannot be identified/ is not a robust result.

Reply: We found the Hopf bifurcation in the experiments with S2 to be control parameter as supercritical and the Hopf bifurcation in the experiment with F as control parameter to be subcritical. Although we searched for periodic orbits for both points, we did not find them, possibly because their amplitudes are very small. It is not possible to dismiss this bifurcation point as a numerical error. It is found in almost all continuations with S2 and F, and in codim-2 continuations (manuscript Fig. 3) the presence of a Hopf point close to a saddle node is extremely constant. Since we did not find any periodic orbits in our time integrations, these Hopf bifurcations do not affect the results. For this reason, no other continuation software was tested.

It's worth nothing that bifurcation analysis of conceptual models describing the AMOC often shows the existence of Hopf bifurcations that are close to saddle node bifurcations. Titz et al. (2002b) found that when freshwater flux is increased in the four-box interhemispheric model described by Rahmstorf (1996), the upper stable branch loses its stability in a Hopf bifurcation, which is followed by a saddle node linking two unstable branches. It can be shown that this Hopf bifurcation is always subcritical and that as such, all periodic orbits emerging from this point are unstable (Titz et al., 2002a). These results were replicated in the five-box model described in Wood et al. (2019) by Alkhayuon et al. (2019), who also found that the distance between the Hopf and saddle node bifurcations increases with increasing atmospheric CO2 concentrations. Furthermore, the presence of a subcritical Hopf point near a saddle node is also found in hosing experiments with ESMs (van Westen et al., 2024), indicating that this result is not unique to models of low and intermediate complexity.

Since the Hopf bifurcations do not affect our results, we will keep the discussion of these points to a minimum in the manuscript, but elaborate on the above mentioned points in the SI for the interested readers.

Changes in text: We will add the zoomed figure and a discussion on the Hopf bifurcations to the appendix and in ll. 193-196 we change

"In our analysis, we found a Hopf bifurcation very close to a saddle node bifurcation for a wide range of parameters. However, attempts at finding the corresponding periodic orbits were unsuccessful and it is not clear if these Hopf bifurcations are real characteristics of the model or rather artefacts of the used continuation software. Consequently, these points are shown in the results, but will not be discussed further."

to

Suggestion: "In our analysis, we found a Hopf bifurcation very close to a saddle node bifurcation for a wide range of parameters. This feature together with zoomed-in figures of the region is discussed in the SI. Since these Hopf bifurcations do not affect the results, they will not be discussed further.”

Figure 3 Is there not a 4th region between the hopf and saddle? However small, I think this should be acknowledged.

Reply: The reviewer is correct in noting that this region exists. However, we would argue that the effect of this small region is negligible, also since no periodic orbits have been found. Nonetheless, we agree it is good to acknowledge its existence and will add a note on its existence, without “naming” it like the others.

Changes in text: In l. 225 we will add “Formally, a fourth region can be distinguished between regions II and III. This region is demarcated by the Hopf bifurcation and the lower saddle node bifurcation. However, this region is very small and since we do not find period orbits associated with the Hopf bifurcation, we do not consider it in our analysis.”

Line 240 Why were no points with higher-than-reference salinity tested? Is there a physical justification?

Reply: No points with higher-than-reference salinity were tested, because nearly all external changes to the gyre current (increase in meltwater from the Greenland Ice Sheet, increases in precipitation, increase in Fram Strait sea ice export) serve to decrease its salinity. In addition, observations also indicate that the North Atlantic is freshening (de Steur et al., 2018).

Moreover, points with higher-than-reference salinity mostly fall into the monostable regime (Fig. 3), so we don’t expect an analysis of such points to add much to the results.

Changes in text: We will add a short discussion on mechanisms that can decrease the salinity of the gyre region in the introduction. In l. 242 add

“No points with higher-than-reference salinity were tested, since nearly all external changes to the gyre current serve to decrease its salinity.”

Line 400 The two sentences at the start of this section seem to contradict each other? Of course, with enough time and noise, any system would collapse/ recover. If the system is tipping because of noise fairly often (even for short times), I would say it is fairly unstable, and possibly near to a tipping threshold.

Reply: We thank the reviewer for pointing this out, the phrasing we used here is indeed unclear. The point we want to make is that that the system can change between a state with and without convection quite easily, and that a collapse of convection does not have to be permanent. This is in contrast with the common result that a collapse of convection in the SPG is a more or less irreversible change with grave consequences. This conclusion is of course based on several assumptions. Most notably, in deriving these results we have assumed that the atmospheric temperature T0 remains constant. It is possible that the stability of convection can change in this model as other climate parameters (e.g. T0, Ubtp) vary. This is an interesting avenue for future research.

Changes in text: In l. 400 change

"Based on the results presented here, it can be concluded that convection in the North Atlantic subpolar gyre is quite stable under current oceanographic conditions in a simple model"

to

"Based on the results from the simple model presented here, we conclude that a permanent collapse of convection in the North Atlantic subpolar gyre is unlikely under current oceanographic and atmospheric conditions."

Given the presence of a limit cycle, has the possibility of phase tipping been considered?

Reply: Indeed, mathematically the limit cycle should exist, but as we noted above, we did not find any periodic orbits, possibly because the amplitude is very small. Hopf bifurcations close to saddle node bifurcations are frequently found in such box models of the ocean, and clearly have something to do with the destabilization process. However, it is difficult to attach physical mechanisms to this behavior in the current context. Therefore we prefer to not overinterpret the mathematical findings in this model.

References

Alkhayuon, H., Ashwin, P., Jackson, L. C., Quinn, C., & Wood, R. A. (2019). Basin bifurcations, oscillatory instability and rate-induced thresholds for Atlantic meridional overturning circulation in a global oceanic box model. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *475*(2225), 20190051. https://doi.org/10.1098/rspa.2019.0051

Armstrong McKay, D. I., Staal, A., Abrams, J. F., Winkelmann, R., Sakschewski, B., Loriani, S., Fetzer, I., Cornell, S. E., Rockström, J., & Lenton, T. M. (2022). Exceeding 1.5°C global warming could trigger multiple climate tipping points. *Science*, *377*(6611), eabn7950. https://doi.org/10.1126/science.abn7950

Born, A., & Stocker, T. F. (2014). Two Stable Equilibria of the Atlantic Subpolar Gyre. *Journal of Physical Oceanography*, *44*(1), 246–264. https://doi.org/10.1175/JPO-D-13-073.1

de Steur, L., Peralta-Ferriz, C., & Pavlova, O. (2018). Freshwater Export in the East Greenland Current Freshens the North Atlantic. *Geophysical Research Letters*, *45*(24), 13,359-13,366. https://doi.org/10.1029/2018GL080207

uhlbrodt, T., & Monahan, A. H. (2003). Stochastic Stability of Open-Ocean Deep Convection. *Journal of Physical Oceanography*, *33*(12), 2764–2780. https://doi.org/10.1175/1520-0485(2003)033<2764:SSOODC>2.0.CO;2

Li, H., & Fedorov, A. V. (2021). Persistent freshening of the Arctic Ocean and changes in the North Atlantic salinity caused by Arctic sea ice decline. *Climate Dynamics*, *57*(11), 2995–3013. https://doi.org/10.1007/s00382-021-05850-5

Rahmstorf, S. (1996). On the freshwater forcing and transport of the Atlantic thermohaline circulation: *Climate Dynamics*, *12*(12), 799–811. https://doi.org/10.1007/s003820050144

Sgubin, G., Swingedouw, D., Drijfhout, S., Mary, Y., & Bennabi, A. (2017). Abrupt cooling over the North Atlantic in modern climate models. *Nature Communications*, *8*(1), 14375. https://doi.org/10.1038/ncomms14375

Swingedouw, D., Bily, A., Esquerdo, C., Borchert, L. F., Sgubin, G., Mignot, J., & Menary, M. (2021). On the risk of abrupt changes in the North Atlantic subpolar gyre in CMIP6 models. *Annals of the New York Academy of Sciences*, *1504*(1), 187–201. https://doi.org/10.1111/nyas.14659

Titz, S., Kuhlbrodt, T., & Feudel, U. (2002a). Homoclinic bifurcation in an ocean circulation box model. *International Journal of Bifurcation and Chaos*, *12*(04), 869–875. https://doi.org/10.1142/S0218127402004759

Titz, S., Kuhlbrodt, T., Rahmstorf, S., & Feudel, U. (2002b). On freshwater-dependent bifurcations in box models of the interhemispheric thermohaline circulation. *Tellus A*, *54*(1), 89–98. https://doi.org/10.1034/j.1600-0870.2002.00303.x

Van Westen, R. M., Jacques-Dumas, V., Boot, A. A., & Dijkstra, H. A. (2024). The Role of Sea Ice Insulation Effects on the Probability of AMOC Transitions. *Journal of Climate*, *37*(23), 6269–6284. https://doi.org/10.1175/JCLI-D-24-0060.1