The preprint represents an evaluation of ROAM-NBS (ICON + NEMO v4.2 + SI3, coupled via OASIS3-MCT, alongside the uncoupled components for the period 1979 - 2020. The mean climate in the atmosphere in generally well reproduced with remaining issues concentrating on an observed SST bias, sea-ice overestimation in the Baltic Sea and too low salinities in the deeper basins in the Baltic Sea impacting stratification and inflow representation. The study itself is timely and highly relevant. The paper is well written. Find suggested revisions below:

Major concerns:

In the beginning of the paper you stress that the availability of regional ocean projections is sparse and emphasize that the focus of the evaluation will be on the Baltic Sea region.

However, I find that the representation of the Baltic Sea dynamics is not yet satisfactory. In my view, there are two possible ways forward:

1. Recalibrate the model - which I know is a long and stressful approach.
2. If the model cannot yet capture key dynamics (beyond surface variables) in the Baltic Sea, I suggest explicitly stating that the system is not yet production-ready in that regard and that further improvement is needed.

Alternatively, if I am mistaken, I welcome further clarification. My intent is not to criticize the effort but to ensure the results are accurately contextualized. This is in no way an offense to the great work that is being presented here.

I will try to explain my concerns below along with the text.

Line references:

L148: Please clarify the implementation of the σ–z* grid. At which depth or criterion switches the model between σ and z coordinates.

L191: 2x respective

L197: Please be consistent with SI^3 or SI3

L199: How is the albedo over the water set?

Table 1: Why is rain bold?

L218: Maybe some overview figure/table addressing the different evaluation periods would be good. I lose track throughout the article.

L222: Is 4 years of spin up really enough for the Baltic Sea. I would be really interested to see the timeseries of stations BY2, BY5, BY15 for surface salinity and bottom salinity (similar to Hordoir et al., 2019 their Figure 8).

L234: The sentence is not so easy to understand. It sounds strange to calibrate fixed reanalysis data. I guess you mean to calibrate the initial/boundary conditions extracted from the reanalysis?

Figure 2 (a) if the bias of the SST is locally up to 1.5K, please adjust the colorbar. Please also consider using discrete steps. In addition, the coloring of the contour lines are hard to follow.

L251: I am not exactly sure why the bias of the SST in the northern part cannot be discussed. If you mean that you cannot compare it because ROAM-NBS is overestimating the sea ice, fixing the SST whereas the OBS show different temperatures, please just mask this area.

Figure 3: I am not sure if an average is suitable here, because you mix positive and negative anomalies. I feel that something like a RMS would be better here, but I am not convinced myself.

L263: What kind of tuning is applied. You mean expert tuning to adjust certain cloud parametrization schemes? Please name it shortly.

L271: What is the reason for the decrease? If not discussed, you can also delete it.

L274: Many repetitions of time series

Figure 6: Typo at the beginning: It should be Seasonal

L362: But isn’t March also the peak of the ice season?

L368-370: If the solution is obvious I am curious why it was chosen to not use the dynamical model in NEMO as it can be turned on with a simple namelist switch.

L380: At least in Figure 10 there is no seasonal signal depicted. I think Figure 10 is too cluttered. The mean profile already gives a good impression about the mean state and the stratification. Instead of plotting SD i would appreciate a second Figure where seasonal temperature profiles are shown (JJA, DJF). Maybe this could be a new Figure A5 because as of right now it is impossible to see differences. The quality of the figure should be improved. I also think the markers just make it hard to see the curve.

L384: I would say the immediate layer is not there at all. If I remember correctly this is a hint that the small inflows are not captured correctly. This would also add up to the negative salinity bias.

L393: The evaluation is something the authors do. I would propose: The SST bias fluctuates around zero …

L395: What is the reason for the shallow depth of the Landsort Deep? I get that you need to probably define a max depth, but why is it not at least the same depth at station BY15 at the Gotland Depth?

L402: It is not shown how momentum fluxes compare in both simulations setups.

L413: What do you mean by too strong prescribed inflow of fresh water? Aren't these observations? Again time series plots of bottom salinity at different stations (BY2, BY5, BY15) could help to see that the problem is related to missing inflow intensities.

L421: The underestimation seems like that the model is roughly 50 percent off at the bottom? Or am I mistaken, maybe I am misinterpreting the colorbars.

L424: I fear that this won't be enough. I think a careful investigation of the conditions at the NS BS interface driving the inflows will be necessary and potentially with subsequent recalibration of vertical and horizontal mixing. May be a slightly longer paragraph here is needed.

Figure 12: I think the figure hints that the inflow dynamics are probably not really well resolved. Again time series plots would help to clarify this. Also the stratifications seem to be underestimated significantly. For future studies processes such as oxygen transportation into the Baltic Sea won't be correctly estimated.

L430: Why only a short period?

Figure 14: I think the colorbar makes it hard to see the differences. At a first glance it looks like that the inflow is not reaching the deeper basins. Would anomalies help here or a different coloring? Maybe you have a better idea.

L582: Freshwater is from observations? How can it be too strong?

L585: I am not sure if the stratification is reproduced as well as the inflows.

Overall, this is a valuable and well-structured contribution, and I appreciate the effort that went into both the coupled setup and the detailed evaluation. Addressing the points above will further strengthen the scientific robustness and clarity of the study.