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 is 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.

Answer:

Dear reviewer, thank you very much for your helpful comments.

We have put a certain emphasis on the evaluations in the Baltic Sea, as this is the area where the quality of the current NEMO setup is most critical.
However, the main research area is not the Baltic Sea. As stated in the abstract, "our target area [...] are the German national waters", including the German North Sea and Baltic Coasts. Since the model encompasses a larger area, we are demonstrating that the mean state and extremes of the whole NBS area are adequately represented so that the results may be used for the mentioned main area of interest.
We make it clearer now that there must be more improvements in the Baltic Sea by modifying the last sentence of the abstract and adding a new one (omitting some other details before to keep it sufficiently concise):
“Overall, the coupled simulation demonstrates **adequate** performance for both the atmosphere and the ocean, and the setup will be used to produce coupled regional climate projections for Europe. **However, bias correction for the deeper Baltic layers remains necessary for further applications, and future work will focus on refining the setup for this region.**”
With that, we omit the phrase “ready for production” in the revised version and replaced “good performance” with “adequate performance”.
Also in the conclusions, we added the sentence “For further applications, it is noteworthy that biases prevail particularly in the Baltic Sea.”

However, we cannot re-calibrate the system at the moment, otherwise we would be too late for CORDEX-CMIP6. And this article is meant to describe the quality of the evaluation simulation, which is a requirement for CORDEX, in a transparent way.
Additionally, the production simulations may teach us further lessons - we do not know yet about the performance of all the historical simulations, they might have different biases again. Of course, we will address the biases revealed by this quite broad analysis of our evaluation simulation in the future and apply further calibration. Different options are mentioned in the conclusions and at other locations in the text (see answers to the specific comments).

Ultimately, we can show that the Baltic inflows are qualitatively simulated, even if they are quantitatively underestimated. The salinity time series that you have suggested and which we have added in the appendix of the new manuscript's version (see the Figure below in the answers to the specific comments) confirm that.

I will try to explain my concerns below along with the text.
See answers in bold grey.
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.

The setup uses a pseudo σ–z* grid. This is the setting in the model's main running namelist. However, during the creation process of the grid file, specifications for the stretching of the layers had to be done, which implement the pseudo z-coordinate at the bottom besides the σ–coordinate above.
The original sentence was modified:
"This vertical coordinate allows for the representation of the deeper and shallower regions simultaneously and consists of a predominantly σ-coordinate with a hyperbolic transient transition between the top and bottom layers following Madec and Imbard (1996). While creating the grid file specifically used by NEMO with a smoothed bathymetry, a slope is determined at which the terrain-following σ-coordinate intersects the sea bed and becomes a pseudo z-coordinate. This transition leads to a smaller bottom-level index in the areas with steep slopes like the North West Shelf."

L191: 2x respective → **done**

L197: Please be consistent with SI^3 or SI3 → **done, now SI3**

L199: How is the albedo over the water set?

A short explanation was added: “Over the ocean, ICON uses a formulation by Taylor et al. (1996) for the direct albedo and the value of 0.06 for diffuse albedo as in ECMWF’s IFS model. Additionally, an albedo increase by whitecap cover after Séférian et al. (2018) is considered.”

Table 1: Why is rain bold?

The whole formula for EMP is given to denote how precipitation is used in the coupling interface of NEMO. An explanation was added in the Table caption: “Variables that are combined into other quantities directly in the coupling interface are written in bold font”

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

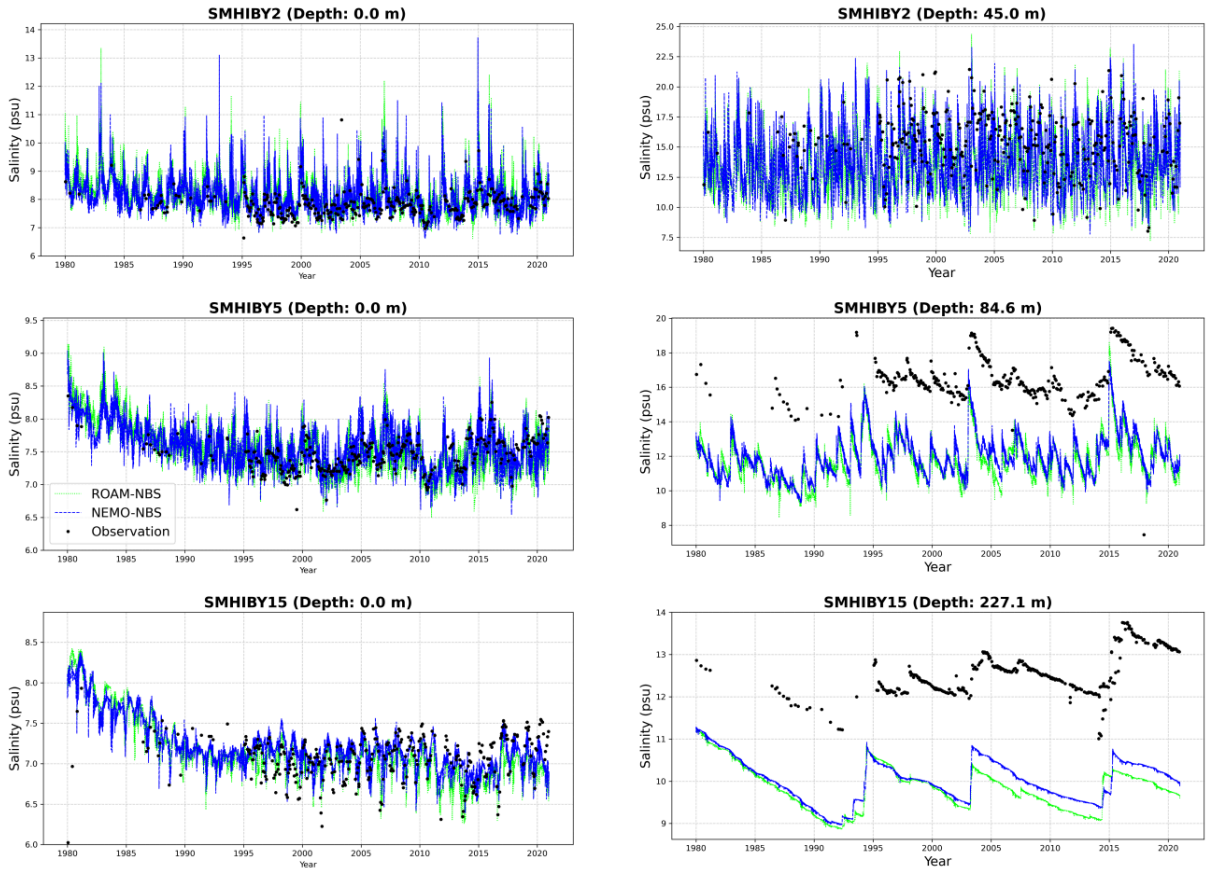
We apologize for the confusion. This point was also raised by the other reviewer. We added an overview table (new Tab. 2) and the sentences “ [... we are overall evaluating and comparing three simulations for the years 1979 - 2020] , which is the minimum period required for CORDEX. However, especially for the ocean part, many reference data are available for shorter time periods only: SST and sea ice data are available from September 1981, salinity from 1993 and station data are very sparse before 1993. For the atmospheric part, satellite data used for the evaluation of surface radiation were available from 2001 only. Therefore, the evaluated time periods had to be adapted in these cases. An overview is given in Tab. 2. For evaluations where statistics from hourly data were calculated, shorter time periods were used, partly due to limited data availability as well, partly to reduce the computational costs.”

Table 2. Overview of datasets used for evaluation; references are given in the respective sections; the full years were used if not denoted otherwise.

Dataset	Evaluated variables	Used in Sect.	Evaluated time period
Copernicus ESA SST CCI and C3S reprocessed SST analyses	SST, sea ice	3.1, 3.3.1, A2	1981–2020
ERA5	SST	3.2.1	1979–2020
E-OBS	tas, tasmin, tasmax, precipitation	3.2.1, A1	1979–2020 (2001–2020 in Fig. A1)
meteorological stations	hourly wind speed (10 m)	3.2.3	2011–2020
FINO1 wind measurement	hourly wind speed (100 m)	3.2.3	2004–2010
Copernicus Baltic Sea- In Situ Near Real Time Observations	ocean temperature and salinity (profiles)	3.3.2, 3.3.3, A2	1979–2020
Copernicus Multi Observation Global Ocean Sea Surface Salinity and Sea Surface Density	surface salinity	3.3.3	December 1993–November 2020
GESLAv3.0 observational data	sea surface heights (SSH)	3.3.4, 4.2, A2	2015–2019 (or selected events)
Baltic thalweg level 4 dataset	temperature and salinity (thalweg)	4.1	November 2014, February and March 2015
Copernicus Baltic Sea Physics Reanalysis	SST	4.3	1989–2020
CERES	surface radiation	A1	2001–2020
IMERG	precipitation	A1	2001–2020
Copernicus Atlantic- European North West Shelf-Ocean Physics Reanalysis	SST	A3	1989–2020

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).

Thank you for suggesting the salinity time series figures at stations BY2, BY5, and BY15. They are presented in the Appendix in the new version of the manuscript for both surface and bottom layers. The simulated surface salinity agrees well with observational data, while the interannual variability of the bottom salinity is generally well captured, albeit with a consistent background bias. Major inflow events in 1993, 2003, and 2015 are also reasonably well reproduced. We agree that the ocean model employed a relatively short spin-up period, as acknowledged by the authors in the conclusions (line 593 of the submitted version). This limitation will be addressed in future production runs to improve model performance.
To avoid duplication of information and save figure space, the salinity hovmoeller diagrams are taken out of the manuscript and are replaced with the described time series.



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?

We have modified the text to make it clearer:
"To ensure that the spatial mean of the modeled SSH fits the observed SSH, the boundary conditions for the SSH from ORAS5 were bias corrected with respect to the SSH at Helgoland from the GESLAv3.0 observational data (Haigh et al., 2023). This bias was calculated on the basis of a 5-year test simulation with uncorrected boundary conditions."

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.

The colorbar in Fig. 2 was adjusted to 2 K. Additionally, the corresponding sentence was adapted to "ROAM-NBS shows a cold bias of locally up to 2 K".
The colorbar with discrete steps for the SST bias was tested as well. It was concluded that more details are visible with the "normal" color bar.

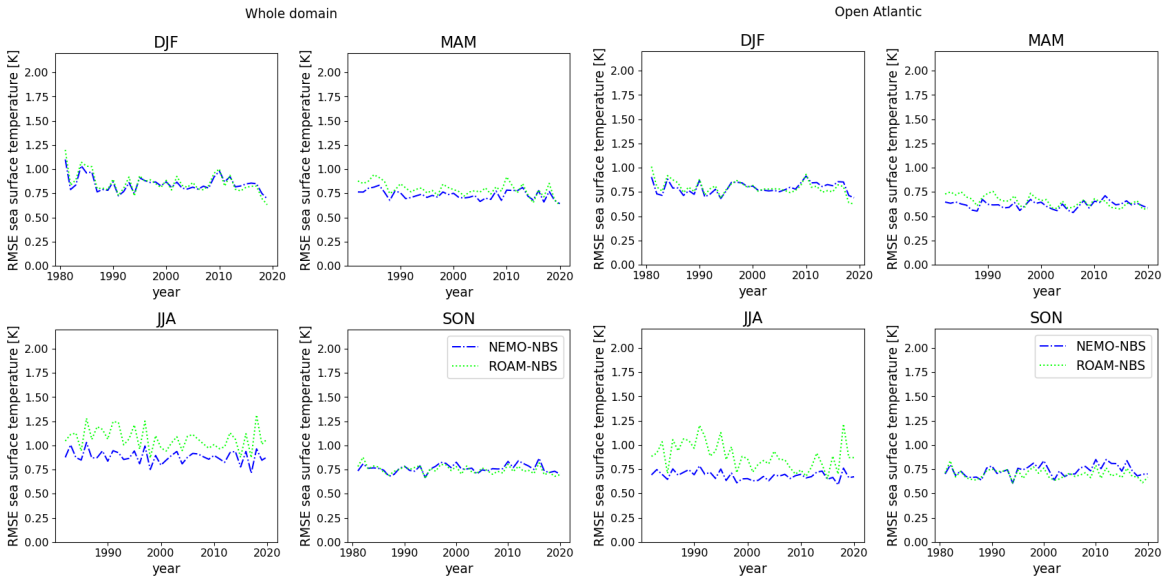
L251: I am not exactly sure why the bias of the SST in the northern part cannot be discussed. If you mean that you cannot compare it because ROAM-NBS is

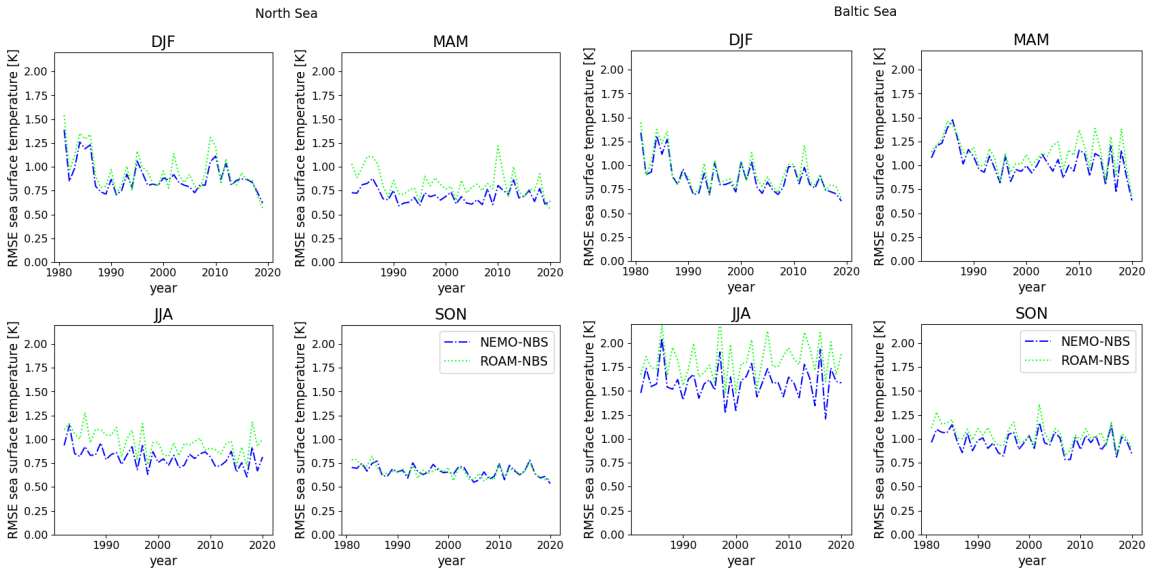
overestimating the sea ice, fixing the SST whereas the OBS show different temperatures, please just mask this area.

Thank you for this remark, it led us to re-consider the treatment of sea ice points for the SST evaluation. The contours of the sea ice extent were removed in the final Fig. 2, since we decided to apply the time-evolving sea ice mask of the observations for calculating the SST bias between the model experiment and the observation. This is described now in the manuscript:
"During the winter (DJF) and spring (MAM) seasons, the area in the Northern Baltic is covered by sea ice with varying extent over time. In the processed Copernicus observations, the SSTs are artificially set to -1.8°C over the regions covered by sea ice. The points where these artificial SST values are found are masked out for the calculation of the mean differences."

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.

The cancelling out of positive and negative biases by a spatial average is particularly true for the Baltic Sea. For the main part of the article, we decided to keep the mean biases to give an idea of their sign. Additionally, we included the RMSE time series in the Appendix (Fig. A4; the DJF and JJA bias time series were duplicates of Fig. 3 anyway) and refer to them in Sect. 3.1:
"As the spatial averaging of biases may cancel out positive and negative values, the time series of the RMSE are additionally shown in Fig. A4. Especially in the Baltic Sea, the RMSE is, with values of about 1.5 K in summer, higher than the absolute values of the mean bias. In the Open Atlantic and the North Sea, it is comparable to the mean bias, with about 0.75 K in all seasons."





RMSE time series per season and domain.

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

Yes, it is a kind of expert tuning. The *allow_overcast* parameter, which was introduced by colleagues of Israel Met Service for a more direct control on cloud cover, is decreased (which increases the steepness of the distribution function of total water used to determine cloud cover when relative humidity is above a certain threshold) to increase cloud cover.
In the text, we added ".. a tuning of the cloud cover **scheme (monthly varying *allow_overcast* parameter)** is applied **..."**

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

L274: Many repetitions of time series

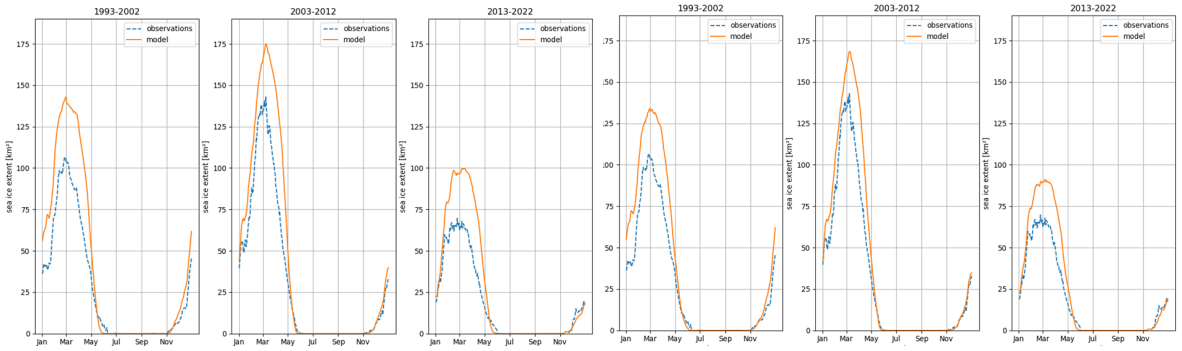
The whole section about the SST bias time series was revised:
"The time series of the spatial mean seasonal SST biases against the Copernicus observations are shown for different regions in Fig. 3. The outlines of these regions (Baltic Sea, North Sea, Open Atlantic, and the whole domain) are shown in Fig. 1. As for the bias maps, points with ice cover in the observations were masked out for the calculation of the biases. For the whole domain, the area-averaged bias is about -0.5 K for both simulations in winter (DJF, Fig. 3a, upper panel). In summer (JJA, Fig. 3a, lower panel), the bias is slightly higher for NEMO-NBS (about -0.75 K), but smaller for ROAM-NBS. However, this smaller bias for ROAM-NBS in summer is due to the higher warm bias in the Atlantic (Fig. 3b, lower panel), combined with a negative bias in the Baltic Sea (Fig. 3d). The magnitudes of the biases for the Open Atlantic region and the North Sea are similar to those for the whole domain, or even smaller. In the

Baltic Sea region, the SST biases in both simulations fluctuate around zero in winter while reaching -0.75 K to -1 K during the summer season."

Figure 6: Typo at the beginning: It should be Seasonal → **done**

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

The misleading expression was corrected and a more complete explanation added: "While the start of the ice growth season and the growth rate in the end of December are well captured, the sea ice extent is overestimated during the peak in the late winter (February-March) by both simulations (mean annual time series of sea ice extent not shown). Furthermore, the sea ice season is prolonged towards May compared to the observations."

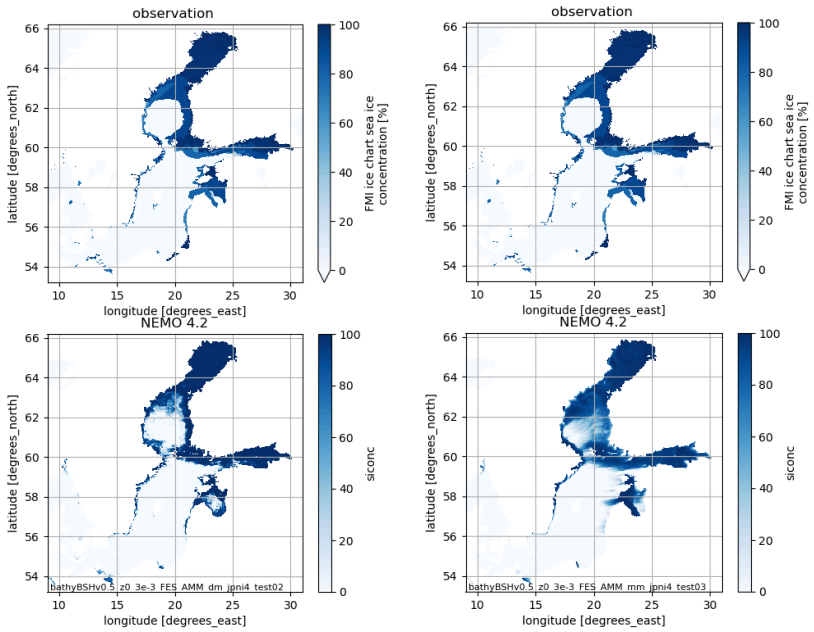


These mean annual cycles of sea ice extent in km² (averaged over three 10-year windows, with ROAM-NBS on the left-hand side and NEMO-NBS on the right) were included in the first draft of the article. Amongst other plots, we omitted them for submission to keep a reasonable length of the manuscript.

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.

We did several test experiments to find a well working configuration. When switching on sea ice dynamics, the result was an extended sea ice area but with lower ice concentrations in the vicinity of the Bothnian Bay and Gulf of Finland, i.e. enhanced advection and spreading of the produced sea ice.
In the manuscript in the same section we state that there are lower SST and underestimated salinity during the ice season compared to observations, which favor more production of sea ice rather than melting, when advected to the open water.
We modified the text to make it clearer:
"Since only thermodynamical processes are modeled within ROAM-NBS and NEMO-NBS, further enhancements in results could be obtained by also considering

ice dynamics. First tests with ice dynamics did not improve the overestimation of sea ice in spring. They will need to be carried out in the future in combination with further parameter tuning and recalibration to eliminate the temperature and salinity bias in the Northern Baltic."

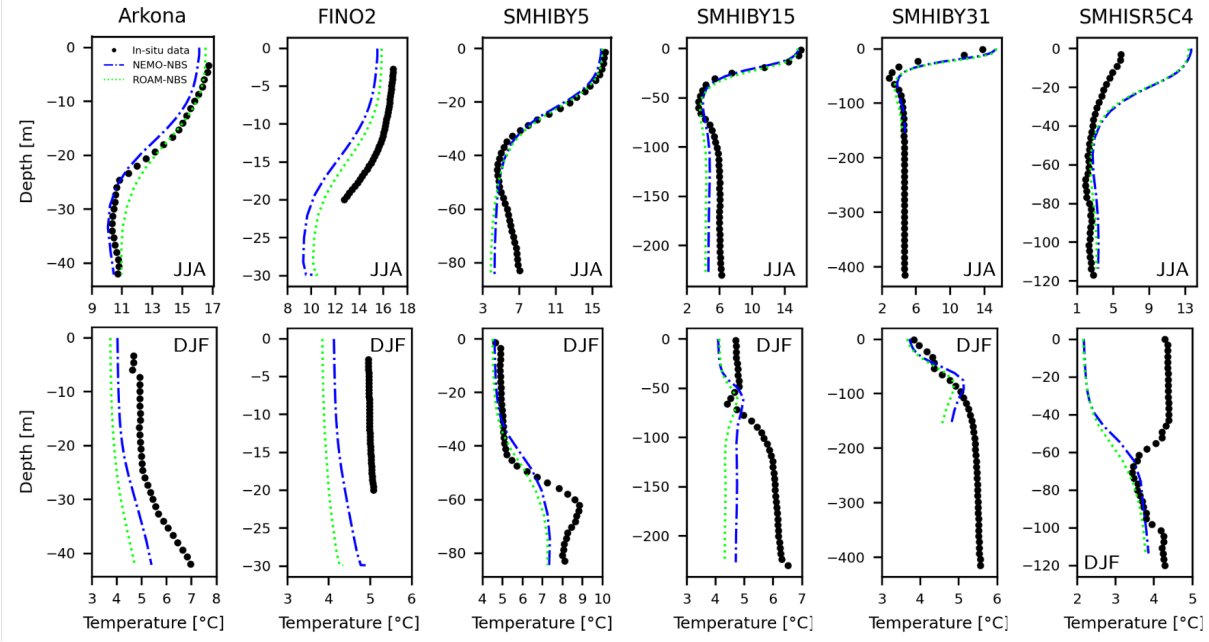


Additional figure: Sea extent, without ice dynamics on the left-hand side (top: observation, bottom: simulation), test with ice dynamics on the right.

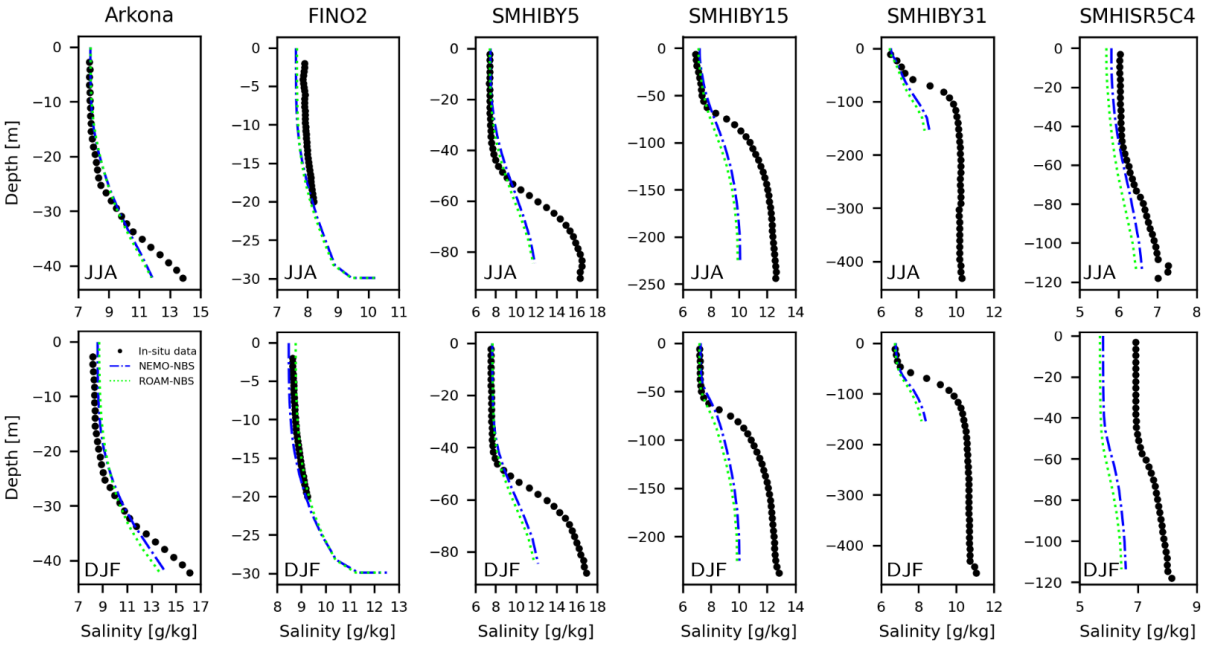
L380: At least in Figure 10 there is no seasonal signal depicted. I think Figure 10 is too cluttered. The mean profile already gives a good impression about the mean state and the stratification. Instead of plotting SD i would appreciate a second Figure where seasonal temperature profiles are shown (JJA, 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.

Thank you for this remark. We have updated Figs. 10 and 12 to display now the mean profiles for summer and winter witout the standard deviation (see below). We have also revised the layout to improve clarity and better meet the reviewers' expectations. Correspondingly, the text in Sections 3.3.2, 3.3.3, Conclusions and A2 has been revised to reflect these new figures.
For transparency, all mean profiles and seasonal mean profiles across the four seasons are displayed below for temperature and salinity.

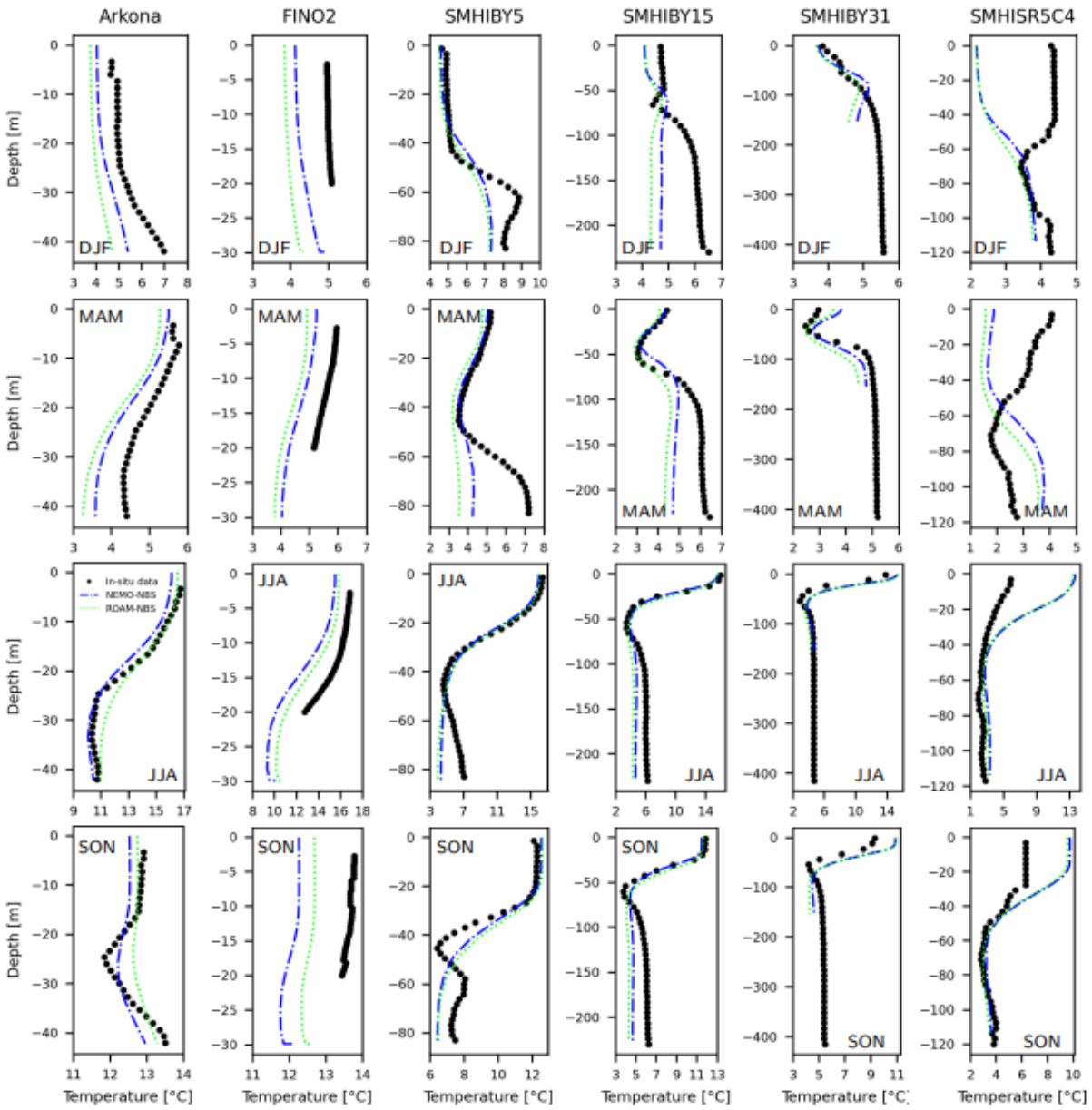
(b)



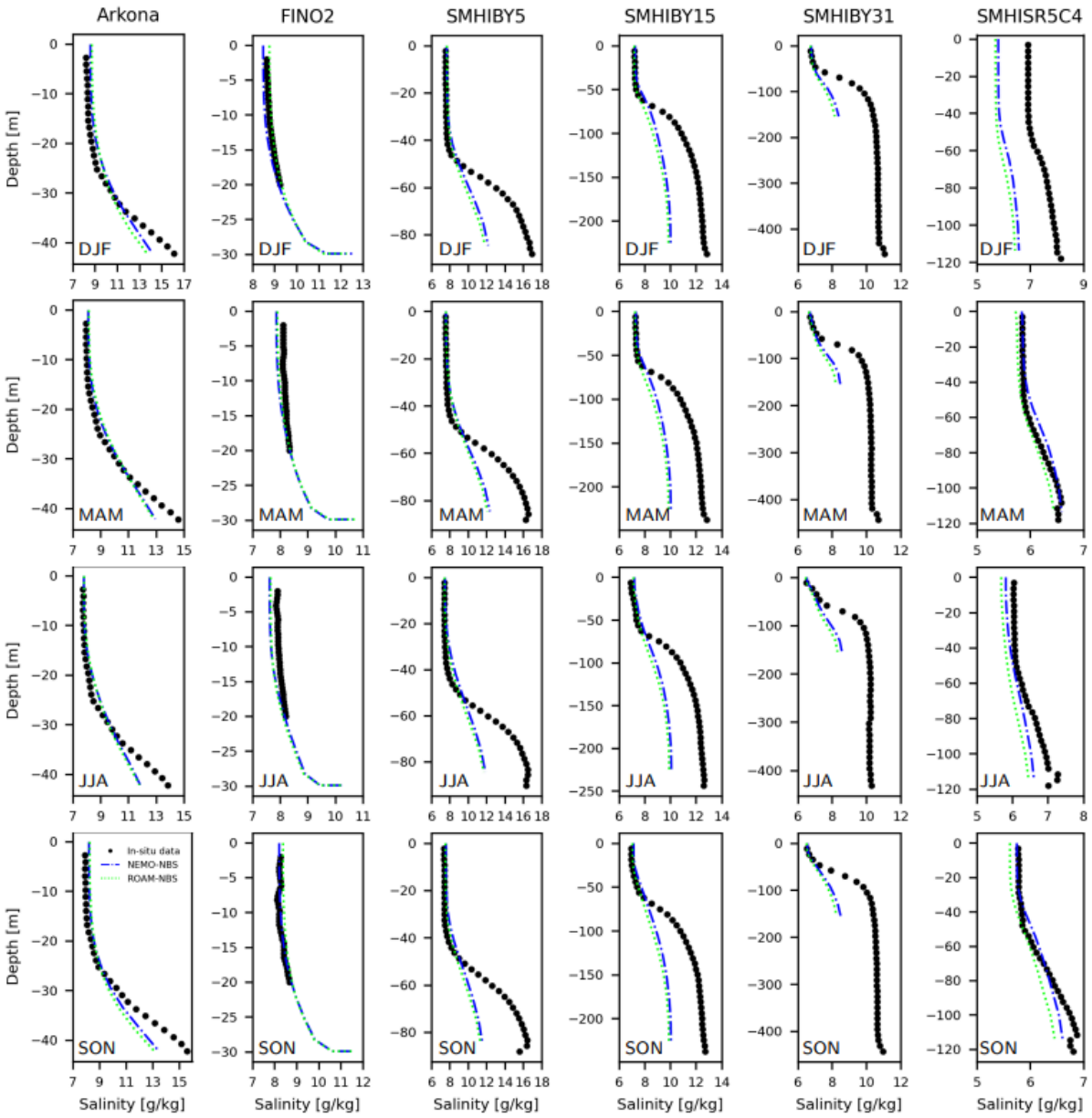
New Fig. 10b.



New Fig. 12.



Additional Figure: mean temperature profiles for all seasons.



Additional Figure: mean salinity profiles for all seasons.

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.

The comment regarding the standard deviation in line 384 has been removed from the manuscript, as the corresponding information is no longer shown in the figures. The mean temperature profiles indicate that an intermediate layer is present, although its magnitude is underestimated. It is possible that this feature was not clearly depicted in the previous figures, which may have contributed to misunderstandings. Furthermore, the seasonal analysis demonstrates that at stations without ice cover, the intermediate layer can be identified, but again appears underestimated. Corresponding remarks have been added to Section 3.3.2, including the sentence:

“In summer, the temperature profiles also display an intermediate layer, although its magnitude is underestimated in both model runs.”

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

Modified to “In Sect. 3.1., it was shown that the SST bias …”

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?

The shallow depth of the Landsort Deep in the simulations results from two factors: the Laplacian smoothing applied to the EMODNET bathymetry and the use of the nearest grid cell for the station-based analysis. In the smoothed bathymetry, the deepest grid cell in the Landsort Deep reaches 370 m; however, this cell does not coincide with the grid point nearest to station SMHIBY31. An explanatory sentence is added in the manuscript.

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

With the new Fig. 10, the whole Sect. 3.3.2 was revised and the mentioned statement was removed.

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.

The river runoff applied in the ocean model combines observational data with model outputs from the WaterGAP model, forming a comprehensive monthly dataset. I.e. it is not completely observation-based. This statement was also added in the conclusions. The observational measurements are taken further upstream than the model input locations, which may influence the freshwater distribution in the simulations. Recent model runs utilizing the E-HYPE dataset have yielded improved results, particularly in reproducing coastal salinity patterns and bottom salinity time series at Baltic monitoring stations. The bottom salinity time series at stations SMHI BY5 and SMHI BY15 show no indication of missing inflow events, as the magnitude of salinity increases aligns well with observational data from NEMO-NBS. However, a persistent background bias, already present in the initial conditions of the evaluation runs, remains evident throughout the simulations.

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.

To improve the understanding and scientific relevance, statistical values are calculated based on the bottom salinity time series at the three suggested Baltic stations. The bottom salinity bias is added in the manuscript text in section 3.3.3. As can be seen in the normalized root mean square deviation (NRMSD) the error is not 50 percent but deviates between 18 % and 29 % for bottom salinity.
Results for ROAM-NBS and NEMO-NBS are:
SMHIBY2:

Mean of observations: 15.502 psu
Bias ROAM-NBS: -1.874 psu
Bias NEMO-NBS: -1.553 psu
RMSD ROAM-NBS: 3.134 psu
RMSD NEMO-NBS: 2.838 psu
Correlation coefficient ROAM-NBS: 0.488
Correlation coefficient NEMO-NBS: 0.566
NRMSD ROAM-NBS: 0.202
NRMSD NEMO-NBS: 0.183

SMHIBY5:

Mean of observations: 16.470 psu
Bias ROAM-NBS: -4.643 psu
Bias NEMO-NBS: -4.384 psu
RMSD ROAM-NBS: 4.775 psu
RMSD NEMO-NBS: 4.484 psu
Correlation coefficient ROAM-NBS: 0.552
Correlation coefficient NEMO-NBS: 0.671
NRMSD ROAM-NBS: 0.290
NRMSD NEMO-NBS: 0.272

SMHIBY15:

Mean of observations: 12.571 psu
Bias ROAM-NBS: -2.738 psu
Bias NEMO-NBS: -2.520 psu
RMSD ROAM-NBS: 2.785 psu
RMSD NEMO-NBS: 2.551 psu
Correlation coefficient ROAM-NBS: 0.434
Correlation coefficient NEMO-NBS: 0.672
NRMSD ROAM-NBS: 0.222
NRMSD NEMO-NBS: 0.203

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.

We acknowledge that longer spin-ups may not fully resolve the issue; however, our current results indicate that they do lead to improvements in bottom salinity. As

suggested, the sentence in L424 of the preprint has been revised and expanded into a slightly longer paragraph (can be found at the end of Sect. 3.3.3 of the revised manuscript):
“Overall, an underestimation of bottom salinity in deeper Baltic basins can be observed. This underestimation could be mitigated in future simulations by employing longer spin-up periods or improved initial conditions. Additionally, efforts should focus on refining vertical and horizontal mixing within the Baltic and increasing spatial resolution in the Danish Straits. For climate projections based on the current setup, bias correction of bottom salinity values in basins deeper than 40m is necessary.”

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.

As suggested, the salinity time series plots have been included in the appendix. These plots show that, while the magnitudes of the major and minor inflows are slightly underestimated in the time series of the lower levels, the overall dynamics are well captured. We acknowledge that further improvement in stratification will be necessary for future coupling with biogeochemical models, and a corresponding note has been added to line 418.

L430: Why only a short period?

A period was selected during which all observational stations provided continuous hourly data, ensuring that the resulting statistics are fully comparable.

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.

Thank you for this remark. By applying discrete colorbars and removing the contour lines, we hope that Figures 14 and 15 are now easier to interpret.
It is indeed correct that the inflows do not reach the deepest basins, as also described in lines 477– 482 of the preprint. Nevertheless, the model setup is considered as production-ready for the EURO-CORDEX domain, provided that a post-simulation bias correction for bottom salinity and temperature within the Baltic Sea is applied. The summary sentence in Sect. 3.3.3 was changed to “Overall, an underestimation of bottom salinity can be observed and shall be taken into account for bias-correction of future climate projections.”

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

As mentioned above and in the text, the runoff data is a combination of observational and model data. Also, within the current NEMO-NBS and ROAM-NBS setups, the runoff is introduced in the upper layer and no enhanced treatment to feed it into the model throughout the water column, available in NEMO, is applied. A sentence discussing these options was added in L413-414 and L582: "Within the current NEMO-NBS and ROAM-NBS simulations, the runoff is prescribed only in the upper layer. A prescription over the complete water column and additional vertical mixing could further enhance the sea surface salinity."

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

Yes, you are correct. The stratification in the Baltic Sea is underestimated from the outset of the simulations, as is particularly evident in the salinity time series. Improving the representation of stratification could be achieved through enhanced initial conditions or by fine-tuning the vertical mixing parameterizations. The sentence in the conclusion was adjusted to: "The major inflow events are qualitatively reproduced but quantitatively underestimated."
Within the following section, the limitations of the stratification are discussed for the Baltic (lines 587-597 in pre-print).

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.