General Comments

The study provides the assessment of the latest Copernicus Marine Service Baltic Sea Physics reanalysis product (BALTICSEA_MULTIYEAR_PHY_003_011) for 1993/94 to 2020/21. They adopt the satellite data and SAR & ice charts as the validation and continue finding that a significant decline in sea ice fraction and thickness, particularly during the melting phase, was observed, with the Bothnian Bay and Gulf of Finland. The study also emphasizes the recent period (2007/08–2020/21) exhibits a shorter ice season and reduced maximum sea ice extent compared to the preceding period (1993/94–2006/07).

I appreciate the criteria for assessment and the clear objectives listed in the Introduction. However, I have several concerns regarding the usage of validation data, the methodology protocol, and some unclear explanations. Therefore, I recommend that the paper undergo major revisions before it can be considered for publication.

Here are my major comments:

1. I do have major concern in the period split: why you choose 2007 as the threshold for the date division, please provide some explanation.
2. Another concern is the data usage in SST_BAL_SST_L4_REP_OBSERVATIONS_010_016 (satellite product) and SEAICE_BAL_SEAICE_L4_NRT_OBSERVATIONS_011_004 (SAR & ice charts-based product). How do you consider the uncertainty in the satellite and SAR & ice charts-based product considering you use the satellite product to determine/correct the sea ice fraction threshold in the reanalysis data, it is important to know the accuracy or uncertainty of the satellite product. And when I look at the Table 1, I am also wondering how is the RMSE and Bias look like during 0.15 and 0.25? What about other thresholds, such as 0.18 or 0.23? And since you've showed two criteria for threshold selection, how do you coordinate them together, such as in RMSE, 0.20 reanalysis threshold has the lowest value while in Bias, 0.25 seems to have the lowest value. And I am quite lost in Line 151, when you mentioned, "TH_SIF of 0.15 for the model dataset, provides more accurate estimates of maximum SIE", can you provide more clearly and statistically evidences in why 0.15 the accurate estimate of maximum SIE is. In Section 4.2, when you are trying to correct the reanalysis sea ice thickness based on three years SAR images and ice chart product, I am not sure if it is statistically robust. Given that samplings for grid is large, but when you consider the annual changes, 3 years is quite short, and not long enough to support your ice thickness correction statements. When I look at the Figure 5, it is quite obviously that Model SIT has the saturation stage in high value compared with the SAR images and ice chart. Then (1) how to explain this condition; (2) instead of the linear relationship, how about using the exponential lines to picture the fitting? And I don't understand how to apply the correction coefficient in Line 168, did you overall divide the values by the 1.81?
3. My next concern is the motivation behind the assessment of the Baltic Sea ice product, which is missing in the Discussion section. For example, what are the limitations of using the current data? What insights can be provided to modelers for improving models? Which updates have improved the product performance compared to previous versions? The current discussion lacks depth and does not provide the audience and the community with sufficient information beyond the assessment results.

Detailed comments:

1. Line 65, Dataset part: please provide detailed information on the temporal resolution and time span of the three products. Additionally, explain how you coordinate these products with different resolutions and specify the interpolation methods used.
2. Line 80, please fill in the reference.
3. Figure 2, I suggest moving either panels (a) and (b) or (c) and (d) to the appendix, as they seem to replicate information.
4. Figure 10: specify the units in panels (a) and (d). I'm quite interested in the fitting process in panels (c) and (f). When focusing on the density plot, consider showing how the linear fitting looks when focusing on high-intensity values or averaging bin values, and then performing the linear fitting.
5. Line 231: could you provide an interpretation of why the Gulf of Finland sub-basin exhibits the most significant reduction during the melting season compared to the freezing season?
6. Figure 11 and 12, could you overlay the trend with the 95% significance level? For example, use stipples to indicate the 95% confidence level or plot only the trends that are above the 95% confidence level.
7. Figure 12(b): verify the values around 55°N, 21°E, and explain why this area shows the largest reduction in ice thickness.
8. Line 261, wrong reference format.