Dear Reviewers,

thank you very much for your time and suggestions for the improvement of our manuscript.

Please see your original comments below (**in bold**) and our responses (*in italics*); comments of Reviewer 1 are listed first, under numbers 1–25, and comments of Reviewer 2 are listed under numbers 25–35:

1) Despite understandable references to reanalyzes names, the less usage of abbreviations may improve readability of the study. Generally, in comparison to a few other reanalysis-related papers, this study is partially challenging to read due to abbreviations and many values without a context.

We have used fewer abbreviations in the revised manuscript (also according to suggestions in the comment number 14).

2) The discussion of (simplified) representation of snow and ice in the reanalysis may be improved. For the average Arctic ice and snow thickness of 2.0 m and 0.2 m, their equivalent total thickness could be 2–3 times higher than assumed 1.5 m. These parameters are quite important for surface heat balance (for high SIC), and there are a few studies comparing surface temperatures from in-situ observations with reanalysis (showing a substantial difference in IST). It would be useful to mention, how is the difference between different models in comparison to ERA5 bias in comparison to measurements, as the study also covers analysis of areas with relatively high SIC values. This is especially vital as there are known issues of strong surface temperature biases of reanalyzes in comparison to observations (Zampieri, 10.1175/MWR-D-22-0130.1).

We have added a new subsection to the revised manuscript: 3.4 Thin ice on leads and snow pack on top of sea ice (Line 331). In this subsection, we have carried out calculations on the above-mentioned effects on LHF and SHF using winter data from the SHEBA campaign to study them, and comment on the results.

We have also added a subsection to the Discussion of the revised manuscript on simplified representation of sea ice and its impact on turbulent surface fluxes (subsection 4.2).

3) The effect of cloudiness was not discussed (often reanalysis works better with clear sky conditions, for example, following Herrmannsdörfer, 10.1525/elementa.2022.00085).

We have added a paragraph to the Discussion section of the revised manuscript on the warm surface-temperature bias in clear-sky conditions in cold seasons in the Arctic (Line 442).

(There is possibly a typo in the reviewers comment – reanalyses (numerical weather prediction models) usually work worse (simulate too warm surface temperatures) in winter clear-sky conditions in the Arctic.)

4) Since the largest effect on flux variability comes from the SIC, it would be useful to discuss a bit more on flux measurements above leads, lead definition, lead fraction measurements, etc.

We have added a text on lead definition and flux observations over leads to the Introduction of the revised manuscript (Line 37). We have also added more discussion on observations of the lead fraction (SIC) and freezing of leads to the Discussion (subsection 4.2).

5) Line 8: It would be great to add the time interval for given sensitivity values.

Added 'using daily means of data' to the Abstract (Line 8).

6) Line 11: Is it possible to distinguish decrease from the warming and from the reanalysis improvement? Can you quantify that difference between 1980-2000 and 2001-2021?

We don't have evidence that more observations assimilated in reanalyses would improve the datasets in a way that the sensitivity of turbulent surface fluxes to SIC would decrease between 1980–2000 and 2001–2021. Even if the amount of observations has somewhat increased, the assimilation scheme and model are the same for the entire period covered by each reanalysis.

7) Line 14: Can you specify what is the effect of SIC on wind speed? Is it a physical effect?

We have added 'via surface roughness and atmospheric-boundary-layer stratification' to the Abstract (Line 15).

8) Line 25: Is it always the case, even for relatively thin ice without snow, assumed in some of the reanalysis?

According to observations, it is the typical case in winter, which we have clarified in the revised manuscript (Line 25).

9) Line 46:since there is no direct mention of 20 % difference in the reference, it would be helpful to explain how it was calculated. In addition, it would be great to mention the scale of those observations: when/at which conditions SIC can be as different as 20%? For the current study, SIC concentration is a key parameter, and it would be helpful to have a more detailed overview of SIC data, algorithms, scales and uncertainties. For example, one would expect that 20% difference in SIC would give around 20% difference in turbulent flux differences between various reanalysis while the actual difference is way larger.

The differences are shown in Figure 7 (right panel) of Valkonen et al. (2008), which we specify in the revised manuscript (Line 51).

They were calculated as differences in sea-ice concentration based on passive microwave data processed applying Bootstrap and NASA Team processing algorithms (specified in the Figure description).

The assumption of 20 % difference in SIC resulting in 20 % difference in turbulent surface flux between reanalyses is not entirely correct – the effect of SIC on turbulent surface fluxes is larger between e.g. SIC 1 changing to 0.8 than SIC 0.2 changing to 0, the former case having larger impact on a change in upward turbulent flux than the latter. The results of our study (in Figures 4 and 6) show the modelled change in turbulent surface flux if the SIC change all the way from 0 to 1 within 1 day.

10) Line 72: it would be helpful to add explanation which data in this reanalysis is based on which measurements or models in addition to Table 1. For example, more details about snow and ice thickness. Or stating that surface temperature is calculated from the surface energy budget.

We have added the information about the observations (satellite input for sea-ice concentration) in each reanalysis to Table 2. Snow and ice thickness representation of the sea ice is presented in Table 2 in the original version of the manuscript.

Surface temperature over the water and both snow covered and bare sea ice is calculated from the surface energy budget in each reanalysis, which we have specified in the revised manuscript (Line 109).

11) Line 83: It is vital for future analysis to give a better overview of algorithms behind different models. What is calculated, what is measured among parameters important for turbulent fluxes.

The turbulent fluxes are prognostic variables in each reanalysis (calculated according to Eq.1 and Eq.2).

We have added that 'turbulent exchange coefficients depend on the roughness lengths for momentum, heat and moisture, and on the stratification of the atmospheric surface layer' (Line 110).

12) Line 131: It would be helpful to comment on the background of these SIC algorithms, not only covering their labels.

As addressed in the specific comment number 10, we have added the information about the observations (satellite input for sea-ice concentration) in each reanalysis to Table 2.

The specific algorithms for obtaining information on SIC (mostly from satellite data) are complex and changing/ evolving in time (during the 42-year study period of our work) with e.g. reanalyses using different external datasets as described in the end of section 2 Material and Methods.

We do not think that going very deep into this background information would necessarily help us interpret the differences in results between various reanalyses (but it would considerably expand the study and the length of the manuscript).

13) Table 4: The SIC in BS and GS is close to zero, yet the average LHF are typical for ice thickness of around 0.1 m, LHF would be 2-4 times larger for open water. Despite this not being your data, can you comment on that effect? It would be useful to show LHF and SHF as a function of SIC for some average conditions to quantify the potential range of their values.

Please note that the values presented in Table 4 are both time and area-averaged. The large average upward fluxes over the Barents and Greenland Seas, in particular in cold seasons, are due to the high sea-surface temperatures. In the case of 0.1 m ice layer, the fluxes would be reduced in magnitude.

SHF as a function of SIC in NCEP/CFSR data is shown for example in Figure 7 in the original manuscript (days in November–December–January 1980–2000): in Laptev Sea (point A), Beaufort Sea (point B), and Central Arctic (point C).

14) Line 177: I fully understand the reason to use some abbreviations, but some of them, lees commonly used (like MB or months) may be removed to increase readability.

We have replaced the abbreviations of seasons and Mean Biases of Daily Field Means with words (the latter as Mean Biases).

15) Line 183: Similarly, as BS is almost ice free, one would expect SHF for open water or 0.1-0.2 m thick ice of negative 100-500 W/m^{2}during winter using simple parameterizations. What could be the reason for much lower values?

In February–April, SHF over the Barents Sea is indeed -100 to -500 W m^{-2} during cold-air outbreaks originating from the Arctic. However, there are also a lot of cases of warm-air advection from the south. Due to these cases, the seasonal mean loss of sensible heat is no more than 34 W m^{-2} (ERA5) or 44 W m^{-2} (NCEP/CFSR).

Please note that in the revised manuscript, we have decided to change the 'reference' data set for subsection 3.1. While before, ERA5 was chosen randomly among reanalyses, we do think that it is more appropriate to use NCEP/CFSR as it appears to be the most realistic in terms of physical processes due to its modelled sea-ice thickness and the snow on top of sea ice (compared to the rest of reanalyses prescribing constant sea-ice thickness and no snow on top of the sea ice). However, we do not assume that it is the best reanalysis with respect to turbulent surface fluxes and use Mean Biases to present an overview and comparison of the typical values in all reanalyses.

16) Figure 4: years and values/difference titles may be also added to the figure to avoid reading the full caption. It may also be possible to present both data for 1980-00 and 2001-21, and difference with a separate color bar.

We have added the side titles and data from 2001–2021 to Figures 4 and 6.

17) Line 213: Previously it was reported that LHF for CA is negative 2-11 W/m^2, while for GS it is negative 10-36 W/m^2. Assuming these two regions are almost on different side of SIC range (0.9 and 0.1), the corresponding difference of 10-25 W/m^{2}should roughly represent a slope of monthly average LHF per unit SIC. Here the slope is one order higher, is it because of different time averaging or why is that?

The time averaging is the same throughout our study – daily means of data.

However, there is area averaging in the presented ERA5 (NCEP/CFSR) values in Table 3 and Table 4; these are calculated as mean of values from all days from respective season (~ 2000 days) in all the grid cells of the respective arctic basin (tens to hundreds of thousands of grid cells in case of GS and CA in ERA5 or NCEP/CFSR). Therefore, the slope using just these two heavily averaged values is not directly comparable to the slope coming out of the linear bilateral ODR model using all data from all days of each season in each grid cell (outputs of which are shown in Figure 4 and Figure 6).

18) Line 241: Despite this being not your assumption, it may be mentioned later that leads in winter refreeze extremely fast (Petrich, 10.1029/2006JC003466) and such surface temperature assumption could lead to flux overestimation.

We have addressed the impact of refreezing of leads in the new subsection '3.4 Thin ice on leads and snowpack on top of sea ice' (Line 331).

19) Figure 7: Maybe the cells could be chosen to be larger and cover wider range of SIC and SHF.

The ODR model works with SIC/SHF data from original-sized grid cells of each reanalysis (apart NCEP/CFSR, details explained in the original manuscript) in each season and time period (~ 2000 days) and returns slope of the regression line (and other information).

Including more grid cells to make the area larger would require averaging of the variable values and therefore we would be showing different SIC/SHF data than those that the model works with to compute the slope of the regression line.

As mentioned in the text of the original manuscript, the main purpose of Figure 7 is to show the failure cases of the regression algorithm.

20) Line 299: Great you mention the effect from snow and ice thickness. Yet, it would be very useful to present what would be an expected bias in LHF and SHF

purely from snow/ice thickness inclusion. Zampieri et al. (2023) might be a useful reference 10.1175/MWR-D-22-0130.1 , especially in the context of MOSAiC expedition, mentioned in line 415. Additionally, there are products including ice and snow thickness (for example, KARRA), which may be used to reduce some of uncertainties or quantify their importance.

We have carried out calculations on the effect of snow pack on LHF and SHF. The results are presented in a new subsection '3.4 Thin ice on leads and snow pack on top of sea ice' (Line 331). In this subsection, we used the SHEBA campaign data on average winter conditions to study the above-mentioned effects on LHF and SHF and comment on the results.

We have not added CARRA results in the present study, as CARRA only covers a small part of the Arctic. The treatment of snow pack is at least equally good in NCEP/CFSR (included in the study) as in CARRA.

21) Line 329: Please add that NWP stands for Numerical Weather Prediction (or something else).

Explanation of the acronym added.

22) Line 377: It would be useful to specify for each condition this is the case as SIC directly changes surface temperature from seawater freezing point to close to air ambient. What would be the bias if the lead is refrozen (as in winter they typically do in just several tens of minutes), which gives surface temperature lower than seawater.

We have carried out calculations on the effect of thin ice on leads on LHF and SHF. The results are presented in a new subsection '3.4 Thin ice on leads and snow pack on top of sea ice' (Line 331).

23) Line 408: The sentence should end with a dot.

Period added.

24) Line 416: I would suggest having a bit longer and clearer conclusion, underlining what you achieved, what reanalysis capture accurately and what can be still improved. And why this type of intercomparison work is important.

We have expanded the Conclusion in the revised manuscript (Section 5).

25) Line 484: The correct link to the study is https://doi.org/10.1002/qj.3803

Link was changed.

26) Lines 51-52: "uniform spatial and temporal resolution" is unclear. Suggest rewording this.

We have revised this as '...spatial and temporal resolutions that are uniform around the globe…' (Line 55).

27) Lines 113-117: I'm not sure why here switched back to OLSR here. Since ODR seems more appropriate, I believe it should be continued to make the subsequent text clearer and the conclusions more coherent. If the only reason for using OLSR is that it requires fewer computing resources, I think that reason is not compelling enough.

We do believe that our choice of using OLSR for the multilateral regression analysis is justified. To be clearer for the reader about our choice, we have revised the respective paragraph in section 2 Material and Methods as follows (Line 112):

'For the bilateral-relationship analysis, we utilised the orthogonal-distance regression (ODR; Boggs, 1988). Because all variables in reanalyses include uncertainties, we theoretically considered the ordinary-least-square regression (OLSR), which assumes no errors in the independent variable, not optimal for this case. Additionally, we carried out tests on bilateral ODR and OLSR performance using data from several grid cells from each reanalysis and while we found 'nearly identical' (at least five decimal numbers identical) coefficients of determination (correlation coefficient squared, R^2) for both regression methods, importantly, the slopes of the regression lines varied considerably. This is attributable to the above-mentioned OLSR's assumption of no errors in the independent variable (x, in our case SIC) and therefore minimising the distance only for x data to the regression line, whereas ODR minimises the orthogonal distance between both x and y data (in our case y is LHF or SHF) and the regression line. Utilising the same above-described tests comparing ODR and OLSR performance for multilateral regression analysis, however, we found 'nearly identical' values for all slopes of the regression lines between LHF (SHF) and SIC, Q_{diff} (T_{diff}), and WS_{10m} for both ODR and OLSR. Values of R^2 for all and individual components of the multilateral regression were 'nearly identical' using both ODR and OLSR as well. Based on the findings that both methods yielded 'nearly identical' results for the multilateral regression analysis (using our reanalyses data), we decided to use OLSR for the multilateral regression analysis in our work, as it requires much fewer computing resources to perform.'

28) Lines 118-119: Further explanation on this point is needed.

We have revised the sentence as follows (Line 125):

‘We used linear model for both ODR and OLSR as we evaluated it as the most applicable for our purposes, being aware of some non-linearity in the SIC effect on Q_{2m} (T_{2m}) and LHF (SHF), as shown for near-surface air temperature in e.g. Lupkes at al. (2008), their Figure 4.’

29) Lines 149-150: You need to specify which figure or table is being described here.

We have revised the sentence as follows (Line 163):

‘The mean SIC in NCEP/CFSR ranged from 0.01 in Baffin Bay in August–September–October in 2001–2021 to 0.96 in the Central Arctic in February–March–April in both 1980–2000 and 2001–2021 (Table 3).’

Please note that in the revised manuscript, we have decided to change the ‘reference’ data set for subsection 3.1 to NCEP/CFSR (more details in response to comment number 31).

30) Figure 2: “Mean biases of daily field means of sea-ice concentration between ERA5 and JRA-55 (grey), ERA5 and MERRA-2 (black), and ERA5 and NCEP/CFSR (light grey).” Which one is subtracted from which? Is it ERA5 - JRA-55 or JRA-55 - ERA5? This needs to be clarified.

We have revised the part of description of Figure 2 (and analogically the description of Figure 3 and Figure S2) as suggested, to avoid reader’s confusion:

‘Mean Biases of Daily Field Means of sea-ice concentration: ERA5 minus NCEP/CFSR (light grey), JRA-55 minus NCEP/CFSR (grey), and MERRA-2 minus NCEP/CFSR (black). Horizontal axis refers to Arctic basins as seen in Figure 1...’

Please note that in the revised manuscript, we have decided to change the ‘reference’ data set for subsection 3.1 to NCEP/CFSR (more details in our response to comment number 31).

31) Lines 154 -200: I noticed that both in the Tables/Figures and the texts, you are comparing other reanalysis data to ERA5, even though you didn't assume ERA5 to be the best in your previous description. I don't think this is appropriate. These descriptions and graphics overly emphasize ERA5 and neglect the inter-comparison between other reanalyses, for example, JRA55 vs CFSR. I believe this is neither fair nor accurate. Please modify the text description and figures to express "inter-comparison" in a more equitable and intuitive manner.

We do believe that the comparison using Mean Bias of Daily Field Means allows us (and the reader) to compare each reanalysis to the ‘reference’ and, at the same time, the other reanalyses between each other.

We did, however, reconsider the selection of the ‘reference’ dataset. While before, we chose ERA5 randomly (as indicated in the original manuscript under ‘We do not assume that ERA5 is the best reanalysis with respect to turbulent surface fluxes…’), NCEP/CFSR *appears to be the most realistic in terms of physical processes due to its modelled sea-ice thickness and the snow on top of sea ice.*

Still, comparisons between other reanalyses (e.g. JRA-55 and MERRA-2) are clearly visible in our Figure 2, 3, and S2 – e.g. in cold seasons and most basins in 1980–2000 Mean Bias in sea-ice concentration (Figure 2, top row) JRA-55 minus NCEP/CFSR is positive, while Mean Bias NCEP/CFSR minus ERA5 or NCEP/CFSR minus MERRA-2 is negative, therefore, we know that the sea-ice concentration prescribed in JRA-55 is the highest of all reanalyses considered.

We have also revised the respective part of the manuscript to be more clear about this issue as follows (Line 153):

'NCEP/CFSR appears to be the most realistic in terms of physical processes due to its modelled sea-ice thickness and the snow on top of sea ice (see more in subsection 3.4), however, we do not assume that it is the best reanalysis with respect to turbulent surface fluxes and use Mean Biases to present an overview and comparison of the typical values in all reanalyses. Mean values (temporal together with spatial) of NCEP/CFSR variables in Arctic basins, seasons, and periods are shown in Tables 3, 4, and S1. The mean values of NCEP/CFSR variables in these Tables are not directly comparable with the values of Mean Biases of Daily Field Means between NCEP/CFSR and other reanalyses presented in Figures 2, 3, and S2 as the method of their calculation differs. However, looking at the Tables 3, 4, and S1 together with the Figures 2, 3, and S2 can provide an estimate of absolute values of SIC, LHF, and SHF in ERA5, JRA-55, and MERRA-2.'

32) Lines 214-215: What caused the higher sensitivity of LHF to SIC in this region? It is not explained here.

In the revised manuscript, we have clarified that this matter is 'further addressed and explained in subsection 3.4' (Line 221).

This is a new subsection in our revised manuscript, where we made calculations for typical winter conditions observed in the SHEBA campaign to study the effects of thin ice on leads and snow pack on top of sea ice on LHF and SHF and comment on the results.

33) Line 246: The explanation here is not clear - Why would an increase in SIC variability lead to an increased statistical relationship between SIC and LHF? And where is the literature supporting the increase in SIC variability? Please add a reference.

In our data, we clearly see the increase in SIC variability in some regions of the Arctic (by variability in this case, we mean 'more days in the season with SIC other than 1', which could have been understandably confusing without an explanation).

While it is natural, that 'more days in the season with SIC other than 1' increase the statistical relationship (significance) between LHF/SHF and SIC (variability in SIC is *needed for a statistical relationship between SIC and LHF), which evidently occurred in some regions of the Arctic, this mechanism is probably more often related to ODR model not converging in 1980–2000 but solving a value of the regression slope between SIC/LHF (or SIC/SHF) in 2001–2021.*

Upon further inspection of the differences of SIC/LHF or SIC/SHF relationships between 1980-2000 and 2001-2021 in single grid cells, we found that in cases where the ODR model converged in both study periods and returned steeper slope of the regression line between LHF/SHF and SIC in the latter period, the reason for stronger statistical relationship wasn't as much caused by 'more days in the season with SIC other than 1' but rather just values of SIC/LHF or SIC/SHF forming a steeper slope (shown in Figure S3 – NCEP/CFSR data).

We have revised the possible explanation of the larger sensitivity of LHF/SHF to SIC as follows (Line 249):

'Mostly in the Central Arctic, however, we found some areas of increased SIC effect on LHF between 1980–2000 and 2001–2021… This increased SIC effect on LHF may be explained as follows. As mentioned before, the effect of SIC on near-surface air temperature (and specific humidity) is not linear, but it is usually the strongest with leads opening in SIC very close to 1. As indicated in Table 3 and shown in our representative grid cells (Figure S3), SIC in some areas of the Central Arctic increased between 1980–2000 and 2001–2021 (possible reasons discussed in subsection 4.5). Therefore, there have been mostly very high SIC in 2001–2021, where even very small decrease in SIC has a strong effect on near-surface air temperature and specific humidity. We cannot be sure, however, whether SIC increased in reality in these parts of the Central Arctic in 2001–2021 compared to 1980–2000, and only comment on possible physical and statistical explanations of the phenomena as represented in reanalyses data.'

34) Line 291: There's an extra space here.

The following line should have been new paragraph, we have fixed this.

35) Discussion and Conclusions: This section is too verbose for me and lacks clarity in its organization. I believe the authors can add subheadings to make the structure clearer, such as 4.1, 4.2, etc. Some of the content in this section is repetitive with the previous section; I suggest simplifying it. At the same time, separating the discussion and conclusion into two parts would make the structure clearer and more specific.

In the revised manuscript, we have divided Discussion and Conclusion into two sections (4 and 5), and used subdivision of the Discussion section as following:

4.1 Differences between reanalyses, their importance, and consequences,

4.2 Simplification of the sea ice in reanalyses and its impact on surface turbulent fluxes,

4.3 Other uncertainties in parameterization of surface turbulent fluxes, *4.4 Role of sea-ice concentration and meteorological variables on surface turbulent fluxes,*

4.5 Decadal changes

Some of the subsections or paragraphs were added based on the comments of the other Reviewer, however, we have also tried to simplify the Discussion section where possible.