***Reviewer 2 (anonymous):***

The manuscript by Harning et al. presents the results obtained from 13 surface sediments collected in the largest Arctic polynya (North Water Polynya, NWO). The authors analysed HBIs, sterols, and GDGTs which are used to calculate sea ice- and temperature-related indices. Based on their data, the authors discussed the utility of the paleoproxies and introduced two local calibrations for TEX86-L and RI-OH. Although their attempt sounds reasonable, the dataset is very small with the very narrow temperature range of 2°C and the correlations between indices and temperatures are moderate (R2 <0.5). Nonetheless, the data are valuable since there were no GDGT data from the study area in the global dataset published before. Some issues are listed below, which should be better addressed before the manuscript is accepted.

*We greatly appreciate the reviewer's time and consideration of our manuscript and thank them for a constructive critique that will lead to a stronger paper. Below we address each comment individually.*

**Major comments:**

The authors suggest that all HBIs are derived from sea ice diatoms in Baffin Bay and thus cannot be used to distinguish sea ice and open water conditions. Although they all might be produced by sea ice diatoms, their concentrations and distribution patterns are different. Potentially, they might be derived from different diatom species. Discussion about potential biological sources of individual HBIs can be added based on the literature, although there are no direct information on specific species from Baffin Bay.

*Thank you for the suggestion. We will add some further discussion on the potential sources of HBIs to the discussion section. However, as the reviewer notes, there is no direct information on HBI sources for Baffin Bay, so our discussion will mostly rely on the modern HBI distribution in this study and that of Kolling et al. (2020).*

The TEX86-L calibration was based on 0-90 m water temperatures. But the R2 values are similar to those in 40-90 m water depth as shown in Fig. 7. Although the p values are >0.05 below 90 m water depth, this might be due to insufficient instrumental data. So it will be interesting to show how the calibration based on 0-200 m water temperatures does look like as well.

*Thank you for the suggestion. As can be seen in Figure 2a, temperatures remain relatively isothermal below ~80 m, so it follows that $R^2$ values and calibrations for the deeper depth integrations are similar. However, for the sake of simplicity we choose to focus our discussion on the calibration that features the highest correlation coefficient (i.e., 0-90 m). We will add some text in the discussion to expand upon and clarify this.*

I see that there are no GDGT data previously published in the study area. However, there are HBIs data previously published. I feel that the discussion about HBIs is in general based on 13 samples, not well integrating the previous data from 70 sites. These data are not even incorporated in Fig. S1 to S3. It is not clear what might be the reason.

*We have now added the previously published dataset from Kolling et al. (2020) to our supplementary figures for comparison and will expand our discussion to better integrate their dataset.*

**Other comments:**

Line 177: an Agilent DB-1MS GC column (60 m x 250 μm x 250 μm)? Is the column information correct?

*Apologies for the typo regarding the film thickness – it is in fact 0.25 um. This has now been corrected for both the DB1 and DB5 columns.*

Line 182-185: Concerning to the response factors for HBIs quantifications, it is not clear how the approach used in this paper is comparable to that used in the paper by Belt et al., 2012.

*Apologies for any confusion. During our analyses we did not have access to the internal standards used by Belt et al. (2012, e.g., 7-HND and 9-OHD, Belt et al., 2012), and therefore 3-methylheneicosane as our internal standard for the aliphatic hydrocarbon fraction. To account for the varying response factor of our internal standard and those of Belt et al. (2012), and make our datasets comparable with other HBI studies, including Kolling et al. (2020), we obtained the 7-HND and 9-OHD standards and ran a 5-point external dilution series along with 3-methylheneicosane. We then calculated sample HBI concentrations using the response factor of our internal standard (3-methylheneicosane) after correction for the difference in response factors of 7-HND and 9-OHD. We have now explained this in more detail in the text and have also added figures for our external HBI and sterol dilution series to the supplement.*

Line 194-196: Similarly, concerning to the response factors for sterols quantifications, it is not clear why cholesterol is used instead of target sterols directly. The standard samples for ß-sitosterol, brassicasterol, and campesterol are available in the markets, except for dinosterol.

*While we are aware of these standards' availability, we used cholesterol as an external standard as the study we directly compare with (Kolling et al., 2020) also used cholesterol.*

HBI III and HBI IV: It would be good to show HBIs chemical structures as a supplementary figure.

*Thank you for the suggestions, we will add HBI structures to the supplement.*

Line 236-237 & Fig. 4a: Looking at Fig. 4A, the standard deviations of mean concentrations of all HBI compounds appear to be overlapping. To better demonstrate the difference, some additional statistics should be done.

*Agreed, we will now add t-test and p-value results to better support the statistical differences or not. Description of these methods is now also added to the Methods and Materials section to more clearly lay out how we statistically assess our datasets.*

Line 240-246 & Fig. 4c: The sample number can be added. In addition, some statistical analyses should be done to better demonstrate whether the datasets between NOW and non-NOW are different. Although it is written in the text like "Although the standard deviations of

mean dinosterol and campesterol concentrations overlaps between NOW and non-NOW sites, the standard deviations of ß-sitosterol and brassicasterol for the two regions is statistically different (Fig. 4c and S2).", Figure 4c rather shows that ß-sitosteroal and dinosterol are different between NOW and non-NOW while the standard deviations of meanbrassicasterol and campesterol overlap.

*We will add the sample numbers and perform t-tests to better demonstrate the statistical differences between the NOW and non-NOW datasets. The latter issue was a typo. The reviewer correctly notes the proper biomarkers that we deemed different or not, and this will be corrected.*

Line 249: The balance factors were obtained based on a combination of previous and current datasets. However, later on, the major conclusions related to the PIP25 indices were based on the current study (i.e. n=13). How are the balance factors if they are calculated only based on the current study? Are the resulting PIP25 values similar?

*Considering that the balance factor is derived from mean IP25 to mean "phytoplankton biomarker" concentrations, and that the concentrations of these biomarkers are relatively similar between the two studies, the c factors are similar whether one relies solely on Kolling et al. (2020) or this study. However, to be more inclusive and take advantage of a larger dataset, we merged the two data sets to use all the local available data.*

Line 259: "….the standard deviation of mean values between these regions is not statistically different (Fig. 4e and S3)." – What kind of statistics were done? The statistical results were not shown to compare both NOW and non-NOW datasets.

*Sorry for any confusion. We used the range of standard deviations as a test for significant differences between NOW and non-NOW datasets. We have now conducted t-tests to ascertain the statistical differences more robustly and will amend any changes to the text accordingly.*

Line 266-280: The authors show the R2 values but it is not clear on how many samples these are based. Please provide the number of samples.

*Apologies for any confusion – the GDGT data is from all 13 samples presented in this study. This has been clarified in the Methods and Materials and sample numbers added to all figure captions.*

Line 281-286: There is no Fig. 9a and 9b.

*Thank you for catching this typo. It should read Fig. 8 and has now been corrected.*

Line 320-322: In the Kolling's dataset, HBI III is correlated with dinosterol and brassicasterol which is not observed in the current study. What would be the reason?

*Per Figure 5b, there are indeed correlations, albeit weak, between HBI III and dinosterol ($R^2$=0.17) and brassicasterol ($R^2$=0.21) in the Kolling dataset. Similarly weak (but insignificant) correlations are also observed between HBI III and dinosterol ($R^2$=0.30) and brassicasterol ($R^2$=0.25) in our dataset suggesting consistency between the two studies. However, ours are not plotted in Figure 5a as the p value was >0.05, which may be partially attributed to the small size of the dataset (n=13) compared to Kolling (n=70). In any case, we do not interpret these*

*weak regression coefficients as indicative of important correlations. We will make this clearer in the main text.*

Line 325-326 & Fig. 4: There are also some differences in sterols between two datasets. Is there any possibility that this is due to the different quantification methods applied?

*The only difference in the analysis of the two studies sterol datasets is our external dilution series that corrects for the different response factors of internal standards. Our lab's quantification protocol is very thorough and should account for these different internal standards. Therefore, we are not sure why there are some differences in sterol concentrations between the two datasets beyond what we originally posited as a possible geographic control in the main text.*

Line 345-347: It is somewhat confusing to see the difference for PBIP25 and PDIP25 between two datasets. How does it look like if the data from the sites in front of fjords in the Kolling's dataset are removed? If so, is the difference less then?

*Per the reviewer's suggestion, we tested whether removing sites closer to the fjords in the Kolling et al. (2020) dataset would bring the PIP indices into closer alignment with the samples from our study. Unfortunately, this did not substantially change the results. However, upon statistical analysis of the NOW and non-NOW sites using t-tests following this reviewer's earlier suggestion, none of the PIP mean values in the NOW are statistically different from mean values outside the NOW. Therefore, this section will be amended, and the differences as noted between the two datasets by the reviewer in terms of PBIP25 and PDIP25 may simply be the result of the different number and distribution of samples between the two studies. However, the mean value differences should not be viewed as statistically different.*

Line 401: Besides the regression analysis, other statistical analyses, such as PCA and RDA would be helpful to better illustrate the impact of the main environmental factors on the GDGT distributions.

*While we appreciate the suggestion by the reviewer, we originally opted not to conduct this analysis to the large number of environmental variables across different seasons and depth integrations. In our opinion, plotting all these variables, along with the 13 samples would produce plots that are too cluttered and difficult to read as well as add more figures that may overwhelm the reader. In addition, we believed that the illustration of our sample and environmental variable relationships would not reveal anything that is not already apparent in our figures. To be open, we show below a PCA analysis for the surface (25 m depth) using annual temperature, salinity, DO and nitrate. As can be seen in Figure 7, annual SST plots closely with the RI-OH index at 25 m depth, and not the other GDGT indices. The RI-OH index also plots closely with DO, as can be seen in Figure S5 at this depth. Therefore, we respectfully intend to leave the regression analyses as is for our GDGT and OH-GDGT datasets.*

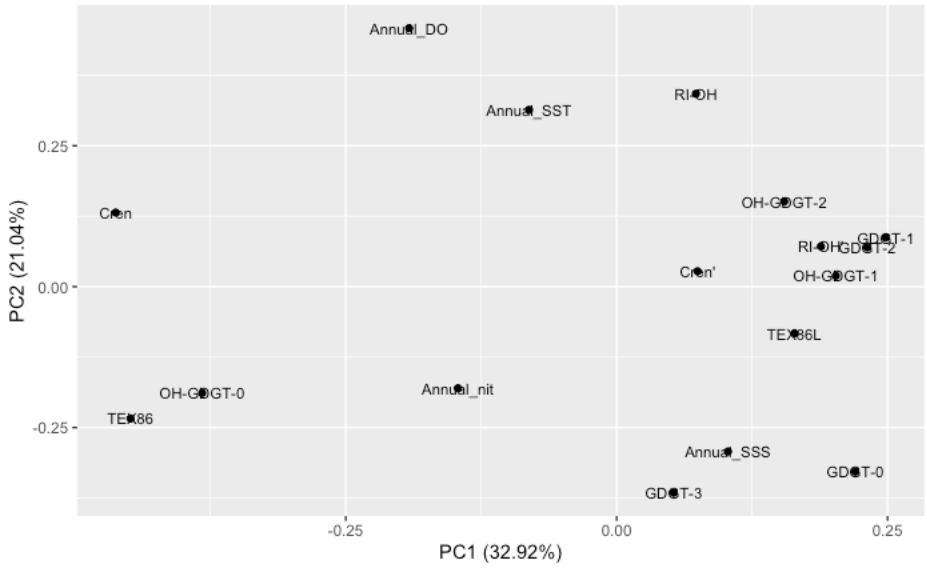


Fig. 5: It is a little bit confusing to see the correlations between the same compounds. It is obvious that they have the value of 1. It would be better to remove them.

*Agreed, and will remove. Thank you.*

Fig. S1, S2, and S3: The color bar scale is not visible.

*We apologize for any difficulties and have adjusted the scale bars to be more visible.*

Table: It would be beneficial for other researchers to present data of individual HBIs, sterols and GDGTs as an Exel file or tables in Supplementary Information, although they can be deposited in a website later on.

*We absolutely agree and plan on submitting our datasets to the PANGAEA online repository upon acceptance of our manuscript. However, we can also include the data as supplemental material for easier access to the reader.*