Reviewer 1:

David Harning and colleagues present new biomarker data from 13 surface sediment samples in northern and western Baffin Bay, focusing on the region of the North Water Polynya. Biomarker data comprise HBIs, sterols, alkenones, and GDGTs, representing a comprehensive study of the use of biomarker-based sea-ice and temperature proxies in Baffin Bay and for the characterization of polynya dynamics. They conclude that the pelagic and sympagic productivity is an order of magnitude higher at the polynya sites, compared to sites outside of the NOW region. This translates to low (higher) PIP_{25} indices within (outside) of the NOW, in line with satellite-derived sea-ice concentration. Harding et al. recommend using sterols rather than HBI III to calculate PIP_{25} indices in the Baffin Bay region, due to uncertainty with regard to HBI III producing species. Further, they propose regional temperature calibrations for both alkenone and GDGT indices, but also show that other environmental variables might contribute to the variability in alkenone and GDGT assemblage recorded in sediments.

The manuscript is very well written and follows a clear storyline with a logical succession of scientific arguments. The presented data fits within the remit of Biogeosciences and I only have a few minor comments regarding comparison with previously published data, and correlations of the environmental and biomarker data. Thus, I recommend publication in Biogeosciences following minor revisions.

> We kindly thank Reviewer 1 for their thoughtful review of our manuscript and offering valuable suggestions to improve its overall quality. Below we provide responses to each of their comments and suggestions and look forward to submitting a revised and stronger manuscript.

General comments:

1. Please comment on why the concentrations of HBIs in the Baffin Bay surface sediments (in part from the same samples) vary by orders of magnitude between the data presented here and the data presented in Kolling et al. 2020.

   Please see our detailed reply to a similar comment by Reviewer 2. We plan to double check our quantification approach using synthetic standards that are structurally more similar to HBIs, which have been requested from colleagues.

2. The Baffin Bay data presented in Kolling et al. 2020 do not show lower P_BIP_{25} and P_DIP_{25} indices in the NOW region. Please comment on this in the manuscript. Is it possible to integrate your data with previously published data to strengthen/confirm the arguments made?

   Given that we may be underestimating the concentration of sterols (please see detailed reply to Reviewer 2), the values of P_BIP_{25} and P_DIP_{25} are subject to change as we test different standards. Therefore, we refrain from making any conclusions about the similarity or lack thereof between the PIP indices of Kolling et al. (2020) and those in our manuscript at this time. However, once our data is finalized, we will make more detailed and explicit comparisons between the two datasets in the revised manuscript.

3. For the data presented in Figure 6 and 9:
    1. Why are alkenone indices not compared to autumn conditions?

       This was an oversight in the data comparison, and we appreciate the notice. We will add a comparison of alkenones with autumn conditions in the revised manuscript.

    2. What is the significance level for the presented R^2 values? Can you include confidence intervals?

       Calculating *p* values is a common method for determining the significance of a statistical tests, which we will assess for all R^2 values in the manuscript (i.e., correlations between alkenones and GDGTs and the corresponding WOA18 environmental variables).

    3. I also wonder how the uncertainty of the WOA18 data for each given depth interval and station influences the significance of R^2. However, I am unsure of the best way to test this, maybe you could consider determination of R^2 confidence intervals using bootstrapping where the environmental data (e.g. temperature) is based on random sampling of normal distributions characterised by the mean and standard deviation of the data for a given depth interval and station.

       This is an excellent point by the reviewer and nicely follows the previous comment. In addition to *p* values, performing bootstrap resampling would allow us to calculate confidence intervals for our R^2 values. We plan to take this advice and perform these analyses with our datasets.

    4. Considering the fragmentary availability of winter and spring WOA18 data, do the data summed up in the annual datasets (Fig. 6 and Fig. 9) cover the same interval for each year or is it an average of all data available for a given year? If the latter, does it make a difference if the data is restricted to the same seasons for every year?

       This is another excellent point by the reviewer, and we are glad to have the opportunity to expand upon the dataset. The annual WOA18 datasets are an average of the data available for the 12 months, so yes, the annual value would be influenced by how complete the given year is. In this regard, we do note in L236-238 that the annual data is more reflective of the ice-free months as winter data is unavailable due to seasonal ice cover. However, we can expand this sentence in the main text to also state that the availability of monthly data during the winter and spring months will ultimately impact the annual mean and introduce an added uncertainty when comparing annual correlations.

4. Line 459-460: This is a somewhat circular argument, as the temperature calibration is based on correlation with the WOA18 data.

   We agree with the reviewer and appreciate this highlight. We will remove this sentence in the revised version of the manuscript.

Visual presentation:

- In Figure 1, could you highlight the samples from where bulk geochemical proxies are available (e.g. differing fill/line colour)?
- Figures S1-S3: Please add the same style of overview maps for $C_{37:2}$, $C_{37:3}$, and $C_{37:4}$ to the supplementary information, so the reader can see which samples had alkenone concentrations over the limit of detection.

All figures will be edited accordingly, thank you for the suggestions.

Typographic comments:

- Line 25: Remove bvc from Kuhlbrodt et al., 2017
- Check spelling of sea ice vs. sea-ice for consistency. The dominantly used spelling is sea ice, but sea-ice is used in 4 instances (line 45, 63, 98, 436).
- Lines 73-74: 'In contrast, the NOW has anomalously low concentrations of thin ice, even during winter months.' Maybe rephrase this sentence for clarity (anomalously low concentrations of thin ice can also be read as meaning thick sea ice).
- Line 204: There is a full stop missing after '…for 20 min.)'
- Lines 501-502: check nitrate vs. nitrite
- Revise formatting of the reference list, to comply with the style of Biogeosciences.

All typographic comments will be corrected, thank you.