Responses to Referee 1

General Comments:

This paper is a useful contribution to paleoclimate literature. The authors present the usage of a novel breakpoint analysis technique on a record of Cenozoic climate from Westerhold et al. (2020), and adequately support the usage of this technique on said data via sensitivity tests.

My major piece of feedback is that this paper currently lacks substantive earth science/ paleoclimate motivation, and is far too terse. While it's an impressive piece of work on the application of this changepoint technique to CENOGRID, additional context is necessary for broader application in the paleosciences. That is, as a methods paper that is designed to introduce a new approach to changepoint analysis within the paleosciences, further work is necessary to show where and when this method can be used to answer other paleoclimate questions. For example, some discussion of age uncertainty is essential, as this is an issue of fundamental importance in paleoclimate. Additionally, CENOGRID is an interesting dataset, but its length and completeness aren't typical of paleoclimate records, which complicates its being the sole non-synthetic example used in demonstrating the application of a novel technique. Additional explanation of the choice of CENOGRID as well as potential edge cases not covered by CENOGRID needs to be done. As a data scientist I'm left feeling confident that this technique is useful, and I'm intrigued by the idea of applying this method to my own work. However, as an earth scientist I'm left not quite understanding the breadth of problems that it is well suited for, nor what the significance of the additional breakpoints that were identified in the Cenozoic is.

We appreciate the reviewer's recognition of our work as a useful contribution to the paleoclimate literature. The reviewer raises an important point about strengthening the paleoclimate motivation and further demonstrating the broad applicability of our method. We acknowledge that the manuscript primarily focuses on the technical aspects of our approach, and that its relevance within the paleosciences can be better contextualized. To address this, we will expand our discussion to clarify where and when this technique can be applied to other paleoclimate problems. Specifically, we will:

- Strengthen the paleoclimate motivation and terminology.
- Discuss specific limitations of CENOGRID and provide a more detailed justification for its selection as the primary dataset.

Furthermore, we can include an additional application to further demonstrate the breadth of problems where the methodology can be applied:

- A shorter time series of δ^{18}O from Greenland ice cores (Seierstad et al., 2014), where our approach allows for the detection of breakpoints while including orbital factors (eccentricity, obliquity, and precession) as explanatory variables. This application will illustrate the method's utility in detecting transitions associated with Dansgaard-Oeschger events also considered by Livina et al. (2010). However, we plan to include this in an appendix to keep the focus on the CENOGRID in the main text. See preliminary findings in Section 1.

Specific Comments:

The need for this method in particular within paleoclimatology should be discussed in more detail. This paragraph "*Our approach contributes to the existing breakpoint detection methods in paleoclimate research by applying well-established econometric tools in the time-domain, developed in Bai and Perron (1998, 2003), to identify climate states in the paleo record. It enables the estimation of multiple breakpoints along with confidence intervals and provides procedures to estimate the number of breakpoints*" should be built upon. I see how confidence intervals might be useful, but what are the other strengths and weaknesses of this approach when compared to other methods? Some discussion of other approaches is offered in the preceding paragraph, but it is somewhat superficial. As a reader, I need better context for breakpoint analysis in paleoclimate studies: its historical usage, current applications, and future potential.

We will strengthen our discussion of our method's advantages and positioning within existing breakpoint detection techniques in paleoclimate research. Specifically, we will do the following:

- Provide a concise historical overview of breakpoint analysis in paleoclimatology, summarizing previous methods and applications.
- Clarify why the Bai and Perron (1998, 2003) approach is well-suited for paleoclimate applications.

- Highlight the key advantages of our approach and compare it with the existing methods in the paleoclimate literature. Also, we will emphasize the possibility of including explanatory variables in our approach.
- Acknowledge limitations of our approach, including assumptions of piecewise linearity, computational requirements, and the handling of irregularly spaced time series.
- We will also include a discussion on future applications and possible extensions.

This sentence "*The paleoclimate variable δ18O measures the ratio of 18O to 16O in the shells of benthic foraminifera obtained from ocean sediment cores, relative to a standard sample.*" is incorrect (or at least misleading), as δ18O is not exclusive to benthic forams, which the sentence seems to be suggesting.
We thank the reviewer for spotting this. We will correct the statement to avoid confusion and ensure accuracy.

This sentence "*The weight difference between the oxygen isotopes leads to an inverse relationship between δ18O and ocean temperatures; see for instance Epstein et al. (1951) and Shackleton (1967).*" is an inadequate description of benthic δ18O. Mention of other factors (seawater composition, ice sheet volume, etc.) needs to be included. CENOGRID is composed of many integrated signals, the makeup of which will determine what detected breakpoints are telling us about the climate.
We will revise the description of benthic δ^{18}O to include additional factors such as seawater composition and ice volume effects. Also, we will clarify that when we write δ^{18}O, we are referring to benthic δ^{18}O.

As the reader, I'm left wondering why only oxygen isotopes were considered. Carbon isotopes are also available, why not include carbon isotopes in the analysis, as was done in the original Westerhold publication?
Our method is flexible and can certainly be applied to other time series, including carbon isotopes from Westerhold et al. (2020). However, to better illustrate the broad applicability of our approach, we can include an additional analysis of oxygen isotopes from ice cores. This example highlights how the method can be used in different paleoclimate contexts and demonstrates how explanatory variables, such as orbital factors, can be seamlessly incorporated into the framework. We will consider having this additional illustration in an appendix.

The authors discuss the varying resolution of the time series at length, which is helpful. However, I would be curious as to whether or not the resolution impacted the detection of break points. Is there any correlation between the detection of new breakpoints (discussed later in the manuscript) and the resolution of the time series? For example, just by visual comparison, it seems to me that the breakpoint observed in Coolhouse 1 in later sections might be related to a large change in resolution that occurs nearby. While this breakpoint isn't heavily interpreted here, this is an important point to understand if other researchers are to apply this method to their own data.

The resolution issue essentially reflects a bias-variance trade-off, where a higher binning frequency (i.e., higher resolution) reduces variance of the breakpoints but may introduce some bias, whereas a lower binning frequency is more likely to increase variance. This means that the primary effect of changing the resolution is on the constructed confidence intervals rather than on the location of the breakpoints themselves. This is evident in Figure 3 of the manuscript, where a lower binning frequency generally leads to wider confidence intervals. Additionally, we observe strong stability in the estimated breakpoints across different binning frequencies, suggesting that resolution does not systematically impact breakpoint detection.

Regarding the estimated breakpoint in Coolhouse 1, it is indeed correct that this breakpoint is fairly close to a change in the resolution of the data. We agree that binned data reflect the quality of the original dataset, which can make it more challenging to accurately estimate breakpoints near or within data gaps. However, from a visual perspective, the breakpoint in Coolhouse 1 could also be influenced by the hump-shaped pattern that follows shortly after, between 17 and 14 million years ago. We plan to add a more insightful discussion on this issue.

This sentence "*Furthermore, we recommend using binning frequencies 10 and 25 kyr as they result in the most consistent outcomes.*" strikes me as rigid and somewhat unhelpful, as many records will not share the general time axis properties of CENOGRID. Is there a different way to describe your binning recommendation that's more flexible and/or applicable to other datasets?

We acknowledge the referee's concern and will clarify that the choice of binning frequency should be tailored to the characteristics of the time series at hand. We will recommend selecting a frequency that maintains sufficient observations per bin while considering the dataset's length and resolution. If the time series is already equidistant, we suggest keeping the original resolution. We will revise the sentence to specify that our recommendation applies to this particular dataset and provide guidelines for other applications separately.

Certain technical choices need to be better explained given the audience of this journal. For example: "*To address these issues, we use the autocorrelation and heteroscedasticity consistent (HAC) covariance matrix estimator with prewhitening in our implementations.*". Perhaps this is standard fare in breakpoint analysis literature, but most paleoclimatologists won't be familiar with this procedure. Some explanation as to why this approach is suitable for this data is warranted here, as is mention of alternatives that were considered. In the same vein, it would be helpful to spend a little bit more time explaining information criteria. That is, expand upon "We use information criteria to estimate the number of breakpoints". What does this mean, why are they used, have they been used in paleoclimate contexts before, etc.
We appreciate the referee's feedback and agree that additional explanation is needed to make these methodological choices more accessible to a paleoclimate audience. In the revised manuscript, we will expand on the rationale for using the HAC covariance matrix estimator with prewhitening, explaining its role in addressing autocorrelation and heteroscedasticity, which are common in paleoclimate time series. We will also briefly discuss why our chosen approach is particularly suited for this application. The presence of autocorrelation and heteroscedasticity in paleoclimate time series has been considered in previous work focusing on ice core records, including Davidson et al. (2015), who considers heteroscedasticity, and Keyes et al. (2023), who investigate autocorrelation. In contrast, our manuscript primarily extends focus on paleoclimate records from ocean sediment cores, where similar challenges arise.
We will also clarify the use of information criteria for breakpoint estimation, explaining their role in model selection and previous applications in paleoclimate studies. Furthermore, we will include the R^2 as a complementary goodness-of-fit measure, to enhance accessibility for a broader readership.

Age uncertainty needs to be addressed somewhere in this paper. It doesn't need a full treatment, in that the method doesn't need to be modified to account for it, nor does it need to be included in the analysis, but discussion of how to include it in future studies is essential. Specifically I would be curious as to how this technique might be expanded to include the usage of age ensembles and how choice of age modeling method might impact breakpoint detection. However, discussion of including age uncertainty directly would also be acceptable here.
We will include a brief discussion on age uncertainty and age ensembles. While this is an important issue, we are not aware of any methodology in the paleoclimate literature that

explicitly treats a time stamp as a random variable, with its variance representing the uncertainty of the time stamp. This challenge has also not been addressed within our framework, and incorporating the time-stamp uncertainty would require significant methodological developments. Such an extension involves advanced statistical techniques that are beyond the scope of this study.

That said, addressing age uncertainty is a crucial direction for future research. We will highlight this in the manuscript and reference relevant studies that discuss the issue (Telford et al., 2004; Franke and Donner, 2019), as well as papers that explore aspects of transition detection in the presence of age uncertainty, such as Goswami et al. (2018).

The simulation study is a particular strength of this work. The authors thoroughly test their method across different data-generating processes, demonstrating its robustness to various forms of non-stationarity and serial correlation. This kind of rigorous testing is essential for establishing the reliability of statistical methods in paleoclimate contexts. I just wanted to make a note of that.

We appreciate this positive feedback.

When analyzing the possible presence of multiple breakpoints, I'm left desiring some kind of prescription as to how I should set the number of breakpoints. Certainly the claim that there are more than 5 statistically significant breakpoints in CENOGRID seems robust. However, the current analysis feels somewhat hand-wavy, with seven breakpoints being settled upon in a rather arbitrary way. In particular this statement needs to be expounded upon: "*The estimation results based on information criteria justify dividing the climate states Warmhouse II and Coolhouse II into two substates each at approximately 39.7 Ma and 10 Ma, respectively. This is supported by the presence of breakpoints estimated approximately at these time stamps in the estimations with seven or more breakpoints.*"

We recognize the need for a clearer justification of the chosen number of breakpoints. In the revised manuscript, we will provide a more detailed explanation of how information criteria guide breakpoint selection and why seven breakpoints were ultimately chosen. Specifically, we will elaborate on the evidence supporting the subdivision of Warmhouse II and Coolhouse II, highlighting how breakpoints at approximately 39.7 Ma and 10 Ma consistently emerge in models with seven or more breakpoints.

The ending of this manuscript is far too abrupt. Potentially new breakpoints are discovered when varying numbers of breakpoints are allowed, but what do they mean? A few climate events are referenced, but events themselves may or may not justify entirely new regimes. Much context is needed here, interpreting and explaining the presence of these novel breakpoints. While the authors are free to choose how to address this comment, I might suggest including a "Discussion" section, in which the primary results are emphasized, and an explanation/interpretation of these results is offered. Some of my other comments could probably be folded into this section as well.

We appreciate this suggestion and agree that a more structured discussion will improve the manuscript. As time series analysts, we do not claim to have the expertise to define entirely new climate regimes in the Cenozoic Era. However, from a statistical perspective, our results strongly indicate that these time points mark shifts in the underlying dynamics of the time series, distinguishing them from other periods.

To address this, we will add a "Discussion" section where we highlight and interpret our primary findings and their implications for paleoclimate time series analysis and for the understanding of the Cenozoic climate. This section will also include a brief discussion of age uncertainty and the limitations of the CENOGRID dataset. Additionally, we will provide insights for other researchers working with similar data, ensuring that our methodological contributions are framed within a broader paleoclimate context.

Technical Comments:

The paper currently is a bit undercited. I suggest the authors go back through with a fine toothed comb and make sure they're citing existing literature wherever possible. In particular, all sections discussing δ18O interpretation should be thoroughly cited, particularly regarding ice volume effects, temperature relationships, etc. A couple of other key spots (non-exhaustive) that need citations include:

- "*The climatic transitions contain important information about variations in Earth's climate system*" (here Tierney et al. 2020 is referenced, but some explanation of what is contained in that review along with additional citations is called for)
- "*Our approach contributes to the existing breakpoint detection methods in paleoclimate research*" (cite breakpoint analysis in paleoclimate literature)
- "*This breakpoint aligns with the Middle Eocene Climatic Optimum, a known climatic event*" (cite original papers describing this event)

- "Some of these breakpoints coincide with other climatic events, for instance, the Latest Danian Event at 62.2 Ma and the onset of the Miocene Climatic Optimum at 16.9 Ma" (cite original papers describing these events, not just Westerhold 2020)

Thank you for this detailed feedback. We will carefully review the manuscript to ensure that the relevant literature is cited appropriately. In particular, we will strengthen the citations in sections discussing δ^{18}O interpretation, providing references on ice volume effects, temperature relationships, and other key aspects. Here, we will acknowledge that benthic δ^{18}O reflects of many signals including deep-ocean temperatures, ice volume, and sea-water salinity. As a citation, we will consider Waelbroeck et al. (2002) who provided a comprehensive analysis of sea-level and deep-ocean temperature changes derived from benthic δ^{18}O records, and Oerlemans (2004), who proposed corrections to the Cenozoic δ^{18}O deep-sea temperature record to account for Antarctic ice volume. For a detailed discussion on disentangling these signals, we will refer to Berends et al. (2021). However, we will also emphasize that our goal is not to disentangle the individual signals embedded in the record, but rather to estimate climate states that reflect the aggregated climate signals.
Furthermore, we will ensure that:

- The discussion of climatic transitions includes a more detailed explanation of Tierney et al. (2020) along with more citations speaking to the importance of Cenozoic climate states. As argued by Burke et al. (2018), we will note that the Cenozoic climate states are valuable for finding analogs for modern warming scenarios. Also, we will mention that several studies find that the climate sensitivity is state-dependent (Caballero and Huber, 2013) and for this the Cenozoic climate states play an important role. Furthermore, it will be noted that planktonic foraminifera records show that Cenozoic climate shifts have greatly influenced marine ecosystems (Swain et al., 2024).

- Our contribution to breakpoint detection in paleoclimate research is contextualized by referencing the review studies of Mudelsee et al. (2014) on benthic δ^{18}O time series analysis in the Cenozoic Era and Marwan et al. (2021) on nonlinear time series analysis of paleoclimate data, along with relevant applications discussed in these reviews.

- Statements about specific climatic events, such as the Middle Eocene Climatic Optimum, the Latest Danian Event, and the Miocene Climatic Optimum, are supported by original studies rather than secondary sources. Specifically, we will cite Bohaty and Zachos (2003) for the Middle Eocene Climatic Optimum, Bornemann et al. (2009) for the Latest Danian

Event, and Flower and Kennett (1994) and Zachos et al. (2001) for the Miocene Climatic Optimum.

1 δ^{18}O from ice cores extracted in Greenland

This section shows some preliminary results for estimation of breakpoints in an ice core record. We consider the NGRIP dataset (Seierstad et al., 2014) and aim to estimate Dansgaard-Oeschger events. This is inspired by Livina et al. (2010) who use potential analysis to determine the number of breakpoints in a paleoclimate time series of δ^{18}O stemming from ice cores from Greenland.

Context:

- The dataset consists of Greenland ice-core δ^{18}O data from Seierstad et al. (2014), an updated version of the time series analyzed by Livina et al. (2010). The dataset spans from 60,000 years before present to today. The Last Ice Age lasted from approximately 115,000 to 11,700 years ago.

- Dansgaard-Oeschger (D-O) events are rapid warming periods during the Last Ice Age (Dansgaard et al., 1993). The vertical lines in the plots on the next page indicate the four D-O events mentioned by Livina et al. (2010). There are of course more D-O events, which we will consider in future versions of this analysis.

We aim to detect D-O events using breakpoint detection with the same AR model specification as before.

- First, we estimate breakpoints and confidence intervals without accounting for orbital factors. The results are shown in Figure 1 and indicate that the estimated breakpoints align well with known D-O events.

- We are then incorporating orbital factors, namely eccentricity (E_t), obliquity (O_t), and precession index (Pr_t) when detecting the breakpoints. The results are shown in Figure 2, where we observe similar breakpoints as in the previous case, but with slightly better alignment to the D-O events.

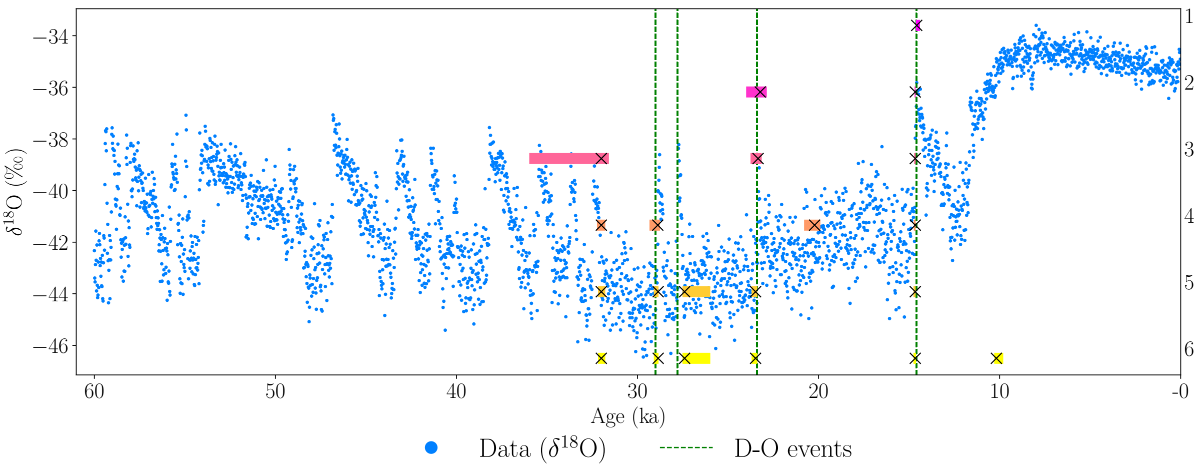

Figure 1: Estimated breakpoints using the AR model for one to six breakpoints on the NGRIP data. Black $\times$'s indicate estimated breakpoints, while colored shaded rectangles represent 95% confidence intervals. The results overlay the δ^{18}O dataset from Seierstad et al. (2014), with vertical dashed lines marking the D-O events considered in Livina et al. (2010).

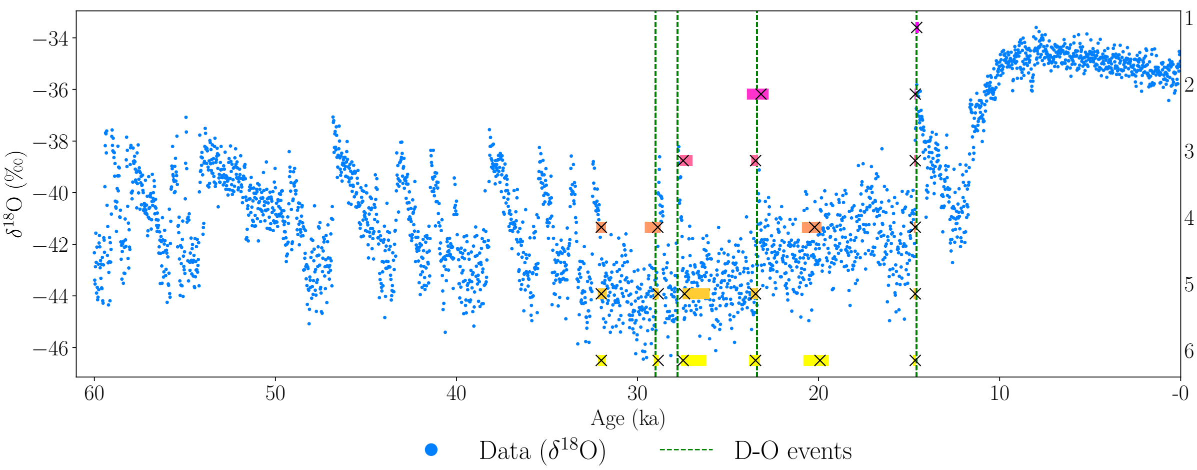

Figure 2: stimated breakpoints using the AR model for one to six breakpoints on the NGRIP data, while including **orbital factors** in the model. Black $\times$'s indicate estimated breakpoints, while colored shaded rectangles represent 95% confidence intervals. The results overlay the δ^{18}O dataset from Seierstad et al. (2014), with vertical dashed lines marking the D-O events considered in Livina et al. (2010).

References

Berends, C. J., de Boer, B., and van de Wal, R. S. W. (2021). Reconstructing the evolution of ice sheets, sea level, and atmospheric CO_2 during the past 3.6 million years. *Climate of the Past*, 17(1):361–377.

Bohaty, S. M. and Zachos, J. C. (2003). Significant Southern Ocean warming event in the late middle Eocene. *Geology*, 31(11):1017–1020.

Bornemann, A., Schulte, P., Sprong, J., Steurbaut, E., Youssef, M., and Speijer, R. P. (2009). Latest Danian carbon isotope anomaly and associated environmental change in the southern Tethys (Nile Basin, Egypt). *Journal of the Geological Society*, 166(6):1135–1142.

Burke, K. D., Williams, J. W., Chandler, M. A., Haywood, A. M., Lunt, D. J., and Otto-Bliesner, B. L. (2018). Pliocene and eocene provide best analogs for near-future climates. *Proceedings of the National Academy of Sciences*, 115(52):13288–13293.

Caballero, R. and Huber, M. (2013). State-dependent climate sensitivity in past warm climates and its implications for future climate projections. *Proceedings of the National Academy of Sciences*, 110(35):14162–14167.

Dansgaard, W., Johnsen, S., Clausen, H., Dahl-Jensen, D., Gundestrup, N., Hammer, C., Hvidberg, C., Steffensen, J., Sveinbjörnsdottir, A., Jouzel, J., and GC, B. (1993). Evidence of general instability of past climate from a 250-kyr ice-core record. *Nature*, 364:218–220.

Davidson, J., Stephenson, D., and Turasie, A. (2015). Time series modeling of paleoclimate data. *Environmetrics*, 27:n/a–n/a.

Flower, B. P. and Kennett, J. P. (1994). The middle Miocene climatic transition: East Antarctic ice sheet development, deep ocean circulation and global carbon cycling. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 108:537–555.

Franke, J. G. and Donner, R. V. (2019). Correlating paleoclimate time series: Sources of uncertainty and potential pitfalls. *Quaternary Science Reviews*, 212:69–79.

Goswami, B., Boers, N., Rheinwalt, A., Marwan, N., Heitzig, J., Breitenbach, S., and Kurths, J. (2018). Abrupt transitions in time series with uncertainties. *Nature Communications*, 9(48).

Keyes, N. D. B., Giorgini, L. T., and Wettlaufer, J. S. (2023). Stochastic paleoclimatology: Modeling the EPICA ice core climate records. *Chaos*, 33(9):093132. Special Collection: Theory-informed and Data-driven Approaches to Advance Climate Sciences.

Livina, V. N., Kwasniok, F., and Lenton, T. M. (2010). Potential analysis reveals changing number of climate states during the last 60 kyr. *Climate of the Past*, 6(1):77–82.

Marwan, N., Donges, J. F., Donner, R. V., and Eroglu, D. (2021). Nonlinear time series analysis of palaeoclimate proxy records. *Quaternary Science Reviews*, 274:107245.

Mudelsee, M., Bickert, T., Lear, C. H., and Lohmann, G. (2014). Cenozoic climate changes: A review based on time series analysis of marine benthic δ18O records. *Reviews of Geophysics*, 52(3):333–374.

Oerlemans, J. (2004). Correcting the Cenozoic δ^{18}O deep-sea temperature record for Antarctic ice volume. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 208(3):195–205.

Seierstad, I. K., Abbott, P. M., Bigler, M., Blunier, T., Bourne, A. J., Brook, E., Buchardt, S. L., Buizert, C., Clausen, H. B., Cook, E., Dahl-Jensen, D., Davies, S. M., Guillevic, M., Johnsen, S. J., Pedersen, D. S., Popp, T. J., Rasmussen, S. O., Severinghaus, J. P., Svensson, A., and Vinther, B. M. (2014). Consistently dated records from the Greenland GRIP, GISP2 and NGRIP ice cores for the past 104 ka reveal regional millennial-scale 18O gradients with possible Heinrich event imprint. *Quaternary Science Reviews*, 106:29–46. Dating, Synthesis, and Interpretation of Palaeoclimatic Records and Model-data Integration: Advances of the INTIMATE project(INTegration of Ice core, Marine and TErrestrial records, COST Action ES0907).

Swain, A., Woodhouse, A., Fagan, W. F., Fraass, A. J., and Lowery, C. M. (2024). Biogeographic response of marine plankton to cenozoic environmental changes. *Nature*.

Telford, R., Heegaard, E., and Birks, H. (2004). All age–depth models are wrong: but how badly? *Quaternary Science Reviews*, 23(1-2):1–5.

Tierney, J. E., Poulsen, C. J., Montañez, I. P., Bhattacharya, T., Feng, R., Ford, H. L., Hönisch, B., Inglis, G. N., Petersen, S. V., Sagoo, N., Tabor, C. R., Thirumalai, K., Zhu, J., Burls, N. J., Foster, G. L., Goddéris, Y., Huber, B. T., Ivany, L. C., Turner, S. K., Lunt, D. J., McElwain, J. C., Mills, B. J. W., Otto-Bliesner, B. L., Ridgwell, A., and Zhang, Y. G. (2020). Past climates inform our future. *Science*, 370(6517):eaay3701.

Waelbroeck, C., Labeyrie, L., Michel, E., Duplessy, J. C., McManus, J. F., Lambeck, K., Balbon, E., and Labracherie, M. (2002). Sea-level and deep water temperature changes derived from benthic foraminifera isotopic records. *Quaternary Science Reviews*, 21(1–3):295–305.

Westerhold, T., Marwan, N., Drury, A. J., Liebrand, D., Agnini, C., Anagnostou, E., Barnet, J. S. K., Bohaty, S. M., Vleeschouwer, D. D., Florindo, F., Frederichs, T., Hodell, D. A., Holbourn, A. E., Kroon, D., Lauretano, V., Littler, K., Lourens, L. J., Lyle, M., Pälike, H., Röhl, U., Tian, J., Wilkens, R. H., Wilson, P. A., and Zachos, J. C. (2020). An astronomically dated record of Earth's climate and its predictability over the last 66 million years. *Science*, 369(6509):1383–1387.

Zachos, J., MO, P., Sloan, L., Thomas, E., and Billups, K. (2001). Trends, Rhythms, and Aberrations in Global Climate 65 Ma to Present. *Science (New York, N.Y.)*, 292:686–93.