We thank both referees, Dr. Mottram and Dr. Kylander-Clark, for their insightful and useful comments, which will improve the manuscript. We have implemented the changes that were needed and below, we reply to their comments.

This is an interesting study that shows that some (but maybe not many?) evaporites have the potential of dating by U-Pb LA-ICP-MS dating. It is not a ground breaking study, and it is a shame that the samples were so young as to yield relatively poor analytical uncertainties, but nevertheless it is significant for those who might want to understand evaporite formation, and it is an appropriate contribution for this journal. It needs some organizational improvements and additional discussion before final publication; the main issue (matrix effects discussion) is discussed below, but also the methods section is somewhat incomplete, and it isn't until later in the data reporting and discussion that the reader understands what experiments were run, how they were configured and why they were changed. The figures and tables are mostly complete and legible but the Tera-Wasserburg diagrams could be simplified and thus easier to understand.

The main issue apparent in this paper: Why did the authors decided to use calcite as the reference material as opposed to some other material? This should be stated in the introduction. As it stands the introduction only states that new minerals are rapidly being introduced for U-Pb, non matrix-matched standards aren't reliable, and the authors decided to use calcite. But there is no mention of why calcite was chosen. Did the authors reach into a bag of RMs and pull out calcite or did they suspect that calcite should behave more like gypsum when ablated and ionized in a plasma? Calcite reference materials are not nearly as well characterized as zircon, and require a two-step process that also requires the accurate measurement of NIST glass.

Beranoaguirre et al.: Many thanks for the suggestions, we will modify the text accordingly and try to clarify all problematic issues. Regarding standardization, we are aware of matrix effects during LA-ICPMS analysis and decided to use calcite as it behaves very similar during ablation (e.g., drill speed, U/Pb downhole fractionation etc.) and ionization in the plasma (Ca2+ as the main cation) compared to sulfate. In contrast, silicate matrixes such as zircon show variable U/Pb downhole fractionation depending on laser fluence and spot diameters probably due to melting effects at the crater rim. In addition, the much higher U content (30-100 times) of zircon reference will saturate the $10^{13} \Omega$ detector ($^{238}\text{U}$) and ion counters (Pb isotopes) of the Neptune MC-ICPMS during
analysis when using similar ablation conditions as for sulfates (130 µm spot size as U content is commonly < 1 µg/g).

If this manuscript is about using non matrix-matched reference materials, there needs to be more discussion within on the differences between glass and calcite, and zircon and glass etc. In fact, there is data in the paper that can be explored: for some reason (which should be discussed) the instrument conditions produced considerably different U/Pb fractionation factors between NIST and WC-1. Sometimes there was little to no offset, and other times there was 8% offset. This is interesting (albeit also a bit troubling), that two different matrices produce varying U/Pb fractionation depending on instrument parameters. It is important because the same could be true for gypsum vs. calcite - one day the offset could be negligible, the next day, the offset could be 8%. Not surprising, as mentioned in the introduction, zircon can be used to normalize garnet under some operating conditions, and other times it can't (just like the different sessions of NIST and calcite in this study).

**Beranoaguirre et al.:** The manuscript is not devoted to the study of different non-matrix-matched reference materials in general, but to the possibility of using calcite RMs for sulfate dating. As stated in the previous comment, the chemical composition and observed ablation behaviour are very similar. Other reference materials like zircon, are not even considered, based on the bias values already published (Parrish et al., 2018). Besides, the U and Pb concentrations of zircon and carbonate or sulfate are extremely different.

Regarding the different offset factors among the sessions, we are not the first ones observing this effect. That is the reason why the accurate U-Pb dating of carbonates, or sulfates in that case, requires a two-step correction (Roberts et al., 2017): (1) $^{207}\text{Pb}/^{206}\text{Pb}$ mass bias correction based on a homogeneous reference material (usually a NIST glass) and (2) a U/Pb inter-element fractionation correction using a matrix-matched RM. The ablation behaviour of the NIST glass and carbonate are different (i.e., ablation plume, pit depth, etc.), and not comparable, even if the spot diameter is the same. This offset variability depends on various factors like slightly different plasma temperatures (RF power) or gas flows (we suspect on N$_2$) in each session. The influence of all these parameters is far beyond the objective of this study and will require a lot of work to understand the contribution of each factor.

There is considerable time spent regarding pit depths of the gypsum and calcite. No pit depths were mentioned for the NIST glass, and neither was there any discussion about the other factors that yield discrepancy in U/Pb ratios in different matrices. Why does pit depth have to be the most important? Certainly, if the matrix is similar, then we might care most about pit depth, but when one is introducing ablated hydrated calcium sulfate vs. calcium carbonate into an Ar plasma, how are the U and Pb ionized differently in such environments? This may be much more important than the pit depth.

**Beranoaguirre et al.:** Following the reviewer’s comment, the NIST pit depths are now addressed in the manuscript. However, we think that the pit depth profile is an important factor when one wants to compare different samples. Guillong et al. (2020) postulate age deviations of up to 20% depending on the degree of crater geometry mismatch, although they did not measure this directly. We propose that for using calcite RMs for sulfate dating, the spot geometry should be comparable.

Finally, unfortunately, in this study only young sample with relatively low U/Pb ratios were measured (the best 238U/206Pb ratios are only 1/4 of concordant values). This limits the ability for the authors to test their hypotheses; if the analyses do not yield better than 10% uncertainty, how do we know that calcite is a better reference material than anything else? In fact, for all but one of the sessions, the NIST glass would have worked just as well. This is worth a comment, though I do not expect the authors to find older, high-U/Pb
The main problem we faced in this study was the unavailability of old sulfate with known age. The gypsum-anhydrite can easily (de-)hydrate and transform. The issue is: if we obtain a 150 Ma age from a sample that is stratigraphically 250 Ma, what is it representing? Is this difference due to a standardization mismatch or is it representing a subsequent geological event? However, as this review is public, we can show here the age obtained for a gypsum sample from the Zechstein Unit (North Germany). This sample was found in the University of the Basque Country storage, in the old samples collection. Zechstein is assumed to have been formed at 250-260 Ma. We obtained an age of 244 ± 10 Ma, which is roughly the expected age, with a precision of 4%, even though the U/Pb ratio is also low. Unfortunately, we do not have more samples from the area.

![Zechstein](image)

Several notes on specific line items:
86: here it should be stated clearly that the data was collected over 4 XX-long sessions, from XX date to XY date, and which sessions used the SC and which used the MC. It should also be mentioned that the MC was used in the latter sessions because it was deemed necessary because of poor results in the first session. As pointed out in a few cases below, this section could stand some better organization and clarification to better set up the results and discussion.

Beranoaguirre et al.: following the suggestions of both reviewers, the method section has been reformulated and extended.

98: what was the spot size and depth drill rate to get these sensitivities? Is the XR more sensitive than the Neptune Plus? This would imply so. But it looks like in tables 1 and 2 that different spot sizes were used. Either this should be normalized (and with the same units so the reader can compare them) to a specific volume ablation rate or the spot size and rep rate should be given herein for clarity.

Beranoaguirre et al.: This is now corrected, and better explained in the extended Methods section.
104. I believe the authors mean mV, not V.

**Beranoaguirre et al.**: The sensitivity is now expressed as counts per second.

107: These tables should be referenced earlier.

**Beranoaguirre et al.**: Following the reviewer’s suggestion, the tables are now mentioned in the first paragraph of the Methods chapter.

110: This is confusing because both instruments are sector-field instruments.

**Beranoaguirre et al.**: now, single collector (SC-ICPMS) and multicollector (MC-ICPMS) are used to distinguish both instruments.

111: Wow - big difference between sequence 1 and 3. Why? And if there is no difference in sequence 3 between NIST and calcite, why bother using calcite at all?

**Beranoaguirre et al.**: It is precisely because we observe different offsets between the sequences that we need the reference material. One explanation is that the ablation conditions between NIST glass and calcite were not completely the same between the different sessions. Nevertheless, the cause of this offset is not well understood and requires a detailed study that is beyond the scope of this paper.

112: So you did a matrix correction and a down-hole correction? Does this mean the difference between the surface and the bottom of the hole was 3%? Did you make a data point by data point correction? This section is confusing and it sounds like there was a double correction made on the sulfate data.

**Beranoaguirre et al.**: Yes, we do both matrix and down-hole corrections. The downhole correction is calculated for the common-Pb corrected WC-1 and then, the fixed calculated value is applied to all the unknowns. Usually, the WC-1 gives a downhole fractionation of ca. 3%. And the matrix is also corrected to the 254 Ma value (Roberts et al., 2017) and validated with the secondary RMs. In any case, now it is reformulated in the extended Methods section.

119. Might be worth noting that the Pagel paper only reports LA data for B-6.

**Beranoaguirre et al.**: Following the reviewer’s suggestion, this is now mentioned in the text.

119. What was the in-house RM? Calcite? How old? Has it been analysed by TIMS?

**Beranoaguirre et al.**: This is a calcite that has been measured several times in our lab. The data is highly reproducible (ca. 36 Ma) and it is under consideration for becoming a potential calcite RM. TIMS analyses are still to be performed.

121: This is worse than the 1.5% added in quadrature. So maybe 2% should be the minimum expanded uncertainty.

**Beranoaguirre et al.**: This sentence has been removed. The expanded uncertainty added is based on the long-time reproducibility of the secondary reference materials.

125. It looks like there was only one secondary RM in the SC experiments. This should be stated in the methods text.

**Beranoaguirre et al.**: Following the suggestions of the reviewer, this point has been
Table 2 Spot shape and size:
Why use a different spot size for 614? In every session? Maybe this explains some of the variability in corrections between NIST and WC-1 in the different sessions.

Beranoaguirre et al.: As it has been stated before, the NIST is used for correcting the $^{207}\text{Pb}/^{206}\text{Pb}$ ratio and the drift of the $^{206}\text{Pb}/^{238}\text{U}$ during each session (instrument drift). We do not aim to match the ablated volume of NIST and WC-1, which is needed to estimate the Pb/U offset between both matrices. Instead, we used the WC-1 for standardization and thus used the same ablation parameter. Even if the spot diameter was the same, the crater on NIST glass is always shallower and thus, precludes its comparability.

Is this why the "sensitivity" was lower for the Neptune?

Beranoaguirre et al.: There was an error in the sensitivities expressed in the text. Now they are corrected, and of course, the sensitivity of the MC-ICPMS is higher than the one in the SC-ICPMS.

T2 QC. Generally, people use the term session, not sequence. There are only 3 mentioned in the text, but there are 4 here.

Beranoaguirre et al.: This is now corrected. We have performed 4 sessions with the MC instrument.

Table 3. Maybe good to highlight the samples that actually worked. This table probably belongs in a repository.

Beranoaguirre et al.: The table is modified and the successful/unsuccessful samples and distinguished. Likewise, the table will be added as additional material and will not be within the main text.

Figure 3: 3 significant figures is overkill (harder to read). Please reduce to 2 significant figures.
Figs 3 and 4: Why not have 1 figure for each sample, and plot the data from the different sessions in different colors (and a legend somewhere to indicate sessions)? It would be much easier to compare them. That way, one could even calculate an age and MSWD for all sessions, as the data should be equivalent. The text is difficult to read at this resolution.

Beranoaguirre et al.: Following both reviewer’s recommendations, Figs. 3 and 4 are now gathered in a single figure (new figure 3).

213: what does "low salinity" here mean? Relative to sea water? Relative to other evaporites? Salinity increases during evaporation, no?

Beranoaguirre et al.: Yes, the low or high salinities are always compared to the seawater (33-37 grams per litre) and it increases during evaporation. However, this "low salinity" refers to the initial values estimated for the Messinian Salinity Crisis (Clauer et al., 2000; Grothe et al., 2020).

217: The header of this section is "high common-Pb content." This paragraph doesn't reflect that header.
This section is now reformulated, and the header has been changed to "low success rate". Likewise, the section has been divided into two different sub-sections, one dealing with the high-common Pb content (or low spread on X-axis) and the other refers to the improvement in the results by using the MC-ICPMS.

219: maybe poor, but not meaningless.

224: This is misleading: n = 34 and 17 in the SC session, whereas n = 66, 75 and 35, 43 in the MC session (twice as much data in latter sessions). Still better with the MC, but not as much better as this sentence would suggest.

234: sulfate not sulfates

234: specifically calcite - not just carbonate.

236: what does hardness have to do with light absorption and ablation? Did you measure the zircon pits? How deep are they? Does ablation depth relate to hardness in other materials?

238: But fluorite isn't very hard compared to zircon. I suggest removing the hardness argument unless there is some scientific evidence that indicates it is important.

239: What is the point here? Fluorite is similar to calcite or different? What does this have to do with gypsum ablation? Clarify or remove.

264: are these single crystals or multi-grain conglomerations? Is their texture described anywhere?

280: It would be nice to mention this in the methods (along with mentioning the specifics of each session).
Beranoaguirre et al.: following the suggestion, we have mentioned it in the method section.

282: Four of them were indistinguishable, not lied.

Beranoaguirre et al.: This is now corrected.