Changes in response to AE recommendations

AE comments are in black font and specific author changes to the manuscript are in red font. Please also read the detailed explanation of changes in the final replies to the referees that accompany this document. Additional revisions were made in response to internal reviewer, Jeremy Powell.

One of the referee comments from Rich Ketcham suggests that you discuss the additional cost of elemental analysis in multi-kinetic AFT workflow. Your response contains a detailed, useful quantification of costs and EPMA analysis times, factors that influence their variability, and an assessment of whether these costs are justified by the data produced. I believe those who read only your article will share the same interest, even if it's only a snapshot of your costs at the present time. Please add a sentence or three on this topic to your revised manuscript.

See our final response to Ketcham's comments that describe the changes we made to the revised manuscript (next section below). We added three sentences that summarize EPMA analysis times, cost reduction strategies, and the relative additional cost of adding EPMA to the workflow.

Likewise, Ketcham's questions about the logistical considerations of first performing laser ablation vs. compositional analysis elicits four or five pages of explanation. Condensing and summarizing these remarks into a paragraph of discussion would benefit those considering whether and how to adopt or adapt your methods. I had the same inclination as Ketcham, around line 340, to add an inline reference to potential downsides of the procedure in section 2.3.

See our final response to Ketcham's comments that outline the changes we made to the manuscript. We have made significant additions to the text in section 2.1 to deal with these concerns.

First, in the second paragraph (line 103, p. 4) under "2. Multikinetic AFT methodology", we give an overall rationale for our chosen methods.

"*The order of steps is based on (1) efficiency and speed of analysis, (2) maximizing the number of track lengths, (3) minimizing selection bias for age grains, and (4) obtaining replicate elemental data. The method can be modified to optimize for other factors or to deal with particular sample conditions but this may increase the cost or the time for analysis.*"

The following text was also added in section 2.1 starting at line 190.

"*Results indicate that compositional zoning is not a common problem for the samples we have studied which is important because the grains are not probed at the exact point of age measurement (laser ablation precedes EMPA). This could contribute to the occasional compositional outlier in kinetic parameter space if elemental zoning is present but kinetic populations are still better resolved compared to when conventional parameters D_{par} or Cl are used (see below). Changing the order of steps so that EPMA is done before laser ablation may help alleviate this problem and reduce the number of AFT grains without elemental data but it could delay sample analysis time by up to several weeks because samples must be transferred back and forth between labs and schedules need to be coordinated. Our current method is efficient and works well for the majority of our samples but it can be modified as needed to deal with more problematic samples. For example, if compositional zoning is a significant issue, it might be better to do laser ablation after EPMA and obtain track lengths from age grains only.*"

We added an inline reference to section 2.1 on line 337 in section 2.3 and eliminated the following two sentences:

"Although elemental measurements can be made with high precision, they are the last step after the grains have been subject to multiple treatments (e.g., etching, laser ablation, Cf-irradiation and etching) and it may not be possible to find a clean spot for EPMA that is close to where the age was measured. Therefore, if compositional zoning is present, the grain may appear as an r_{mr0} outlier."

For Figure 2, reducing the marker size to reduce point saturation should suffice, and I also prefer not to see the correlation coefficient. Like Ketcham, I did not notice the "N =" in the plot titles and would suggest adding it to the statistical summary inset text, but I leave the final decision to the authors.

See our final response to Ketcham's comments. Points were reduced in size and "N=" was added to each panel in Figure 2.

The explanation of the implementation of the CRS method on pages 12-13 of the authors' response to Ketcham looks fine. I see broad parallels ramping alpha and simulated annealing algorithms, and similarities between its smoothing and other inverse problem approaches. As you mention on your related answer on page 14, though, please do make clear where you believe your AFT data do not constrain the thermal history.

See our final response to Ketcham's comments that outline the changes we made to the manuscript. We added some text and dashed lines to Figures 6 and 7 that indicate the approximate upper temperature limit that can be constrained by the AFT data.

Finally, a revised manuscript would benefit from a brief summary of the more detailed discussion around the AFTINV and QTQt differences on pages 16-18 of the responses to Ketcham's review. My feeling is that (4) plays the largest role here. The mechanics of the AFTINV and QTQt fits are discussed elsewhere in the manuscript, but a separate paragraph in the discussion would draw these threads together.

See our final response to Ketcham's comments that outline the changes we made to the manuscript. We added more text and a third paragraph in section 4.2 on QTQt to elaborate on factors that contribute to the differences in model results.

To address Karl Lang's initial comment, please seek maximum clarity about "detrital" vs. "sedimentary" nomenclature at the beginning of the paper. The use of "detrital sample" throughout the paper is ok with me.

See our final response to Lang's comment that outline the changes we made. We use "sedimentary" in place of "detrital" in several places near the beginning and retained "detrital" where we believe the meaning is clear.

As a larger point, I do share some of Lang's concern that this manuscript includes a "vigorous preference of LA ICPMS over [the] EDM approach." I understand that this is a current community-wide discussion and has been for some time now. This manuscript makes a meaningful contribution to this discussion in that it demonstrates the success of a workflow that includes LA-ICPMS in the AFT workflow. As such, I believe the paragraphs from lines 114 to 146 represent an out-of-place, discursive discussion in the Methods portion of the paper, and they should be omitted. These thoughts would be better included, I believe, in a separate manuscript that has space and scope for a more complete discussion of the "different methods with different advantages and disadvantages" described in the response to Lang. The advantages of the LA-ICPMS method in your workflow are well documented elsewhere throughout the paper. There is also plenty of additional discussion, outlined above, to add to a revised manuscript that is more germane to your novel method and its application to examples from the northern Yukon.

We have eliminated much of the text from the first two paragraphs in section 2.1 and replaced it with a single paragraph (see final response to Lang's comment). We agree that the subject is best left to a more focused paper in the future that deals with both LA-ICP-MS and EDM data. We have quality EDM data with clearly defined multikinetic populations and we believe that both methods can yield good results. As suggested by Karl Lang, if there are difference in results, it is more likely to be related to how users implement the methods rather than to intrinsic differences in the methods.

Finally, this paper received a series of contributed comments and replies from Ian Duddy and Paul Green, with a short clarifying response also from Rich Ketcham. The broad-ranging discussion includes questions about the LA-ICPMS method in general, use of the r_{mr0} metric, and the quality of the EPMA analyses in this manuscript in particular. I appreciate the discussion and replies, which will remain available as part of the format of the GChron journal. I agree with Ketcham's overall assessment (RC3) of the EPMA data in that "I don't think perfect is necessary for this to be a useful contribution." I acknowledge the analytical difficulties in measuring compositions in heterogeneous apatite populations, and I find the discussion in AC6 of the low elemental totals, the effect of including vs. excluding OH in elemental totals to bring them up, and the effect of Fe on track retentivity interesting. A revised manuscript should include a sentence or two that summarize your reasoning that your conclusions are not adversely affected by low totals.

See our final response to Ketcham's comments that outline the changes we made to the manuscript. We added the following text to the beginning of the paragraph on lines 187-195 (p. 8) at the end of section 2.1:

"*We try to obtain EPMA measurements on a smooth "clean" surface but this is not always possible for small grains with many etched tracks and other imperfections. Missing elements and track void space in the electron beam excitation volume can result in elemental totals that are less than the 97-100 wt% expected for good analyses (see Issler et al. (2021) data tables) and this can happen for samples with AFT data acquired by LA-ICP-MS or EDM. Fortunately, we have observed that suboptimal analyses still yield an elemental signature that allows for discrimination of different kinetic populations and replicate elemental analyses with good and lower elemental totals yield similar results for the study samples.*"

Review 1 (Richard Ketcham) Comments

Reviewer comments are in black font and specific author changes to the manuscript are in red font.

At the same time, to be effective in doing so (or at least transparent in trying), it would be good to better document the costs.

We discussed this in detail in our reply (AC1) but we don't think that absolute costs should be mentioned in the paper because they depend on different factors such as the laboratory doing the work and the number of samples being processed at a given time among other things. However, we included information on relative costs for additional EPMA work (see point immediately below).

For example, how long does the EMPA protocol take per spot?

See detailed response (AC1). We added the following three sentences (in italics) (line 185) to provide some information on analysis time and relative cost with and without EPMA. "*The raw analysis time for our samples (Peter Hooper Geoanalytical Lab, WSU) is approximately 3.7 minutes per spot (16 grains per hour), excluding the time for setup (1-2 hours), standardization (8 hours) and point picking (150*

grains per hour). Hourly billing is capped at 12 hours per day so it is advantageous to run samples in large batches for 48-72 hours straight which can reduce the hourly rate by 30-40 %. EPMA increases the average cost of AFT analysis by approximately 20% for our samples."

What considerations went into decision(s) of whether to do compositional analysis first versus laser ablation? Does doing laser ablation first save time, by figuring out which grains work and providing evidence of whether there is kinetic dispersion, and does this outweigh the disadvantage of not getting the analysis precisely where the tracks were measured?

See detailed response (AC1). We added the following two sentences (in italics) after line 103 (below Table 1) to add some context without going into too many distracting details. "*The order of steps is based on (1) efficiency and speed of analysis, (2) maximizing the number of track lengths, (3) minimizing selection bias for age grains, and (4) obtaining replicate elemental data. The method can be modified to optimize for other factors or to deal with particular sample conditions but this may increase the cost or the time for analysis*." We thought it was also important to modify another sentence in this paragraph (line 107 and 108) to read as, "The goal here is to try to use all the available data except for obviously poor analyses *that provide no useful information (i.e., inaccurate U measurements)*." This is in response to the community comment that we address in substantial detail (AC4). We strongly disagree with view of Duddy and Green that elemental data with suboptimal totals provide no useful information. Our rejection of obviously poor analyses are those where U measurement is clearly inaccurate as detailed in our comment AC4 (grains blown out during the initial stage of laser ablation and U values below ICP-MS resolution).

Are there cases where changing the order would be a good idea?

See detailed response (AC1). We are hoping the author replies to comments are an online resource for readers who want more detailed information.

We added the following text starting at line 190 to address this point.

"*Results indicate that compositional zoning is not a common problem for the samples we have studied which is important because the grains are not probed at the exact point of age measurement (laser ablation precedes EMPA). This could contribute to the occasional compositional outlier in kinetic parameter space if elemental zoning is present but kinetic populations are still better resolved compared to when conventional parameters D_{par} or Cl are used (see below). Changing the order of steps so that EPMA is done before laser ablation may help alleviate this problem and reduce the number of AFT grains without elemental data but it could delay sample analysis time by up to several weeks because samples must be transferred back and forth between labs and schedules need to be coordinated. Our current method is efficient and works well for the majority of our samples but it can be modified as needed to deal with more problematic samples. For example, if compositional zoning is a significant issue, it might be better to do laser ablation after EPMA and obtain track lengths from age grains only.*"

[line 44] Justifying the 20°C bound seems to require citing Donelick et al (1990), and optionally Tamer and Ketcham (2020).

These references have been inserted into the paper.

[line 243] Change to "thermal history modeling" (or "all model calculations").

The first recommended change has been made to the text.

[line 249] Although replicate values are indeed important for assessing the reproducibility of kinetic parameter values, they may also be taken as an indication of the presence of zoning. The authors do not specify how many spots they took per analysis, but I suspect the answer is one, and that it reflects the usual 2-µm activation zone for EMP; was this driven by the desire for a faster and/or less expensive analysis?

Efficient and cost-effective analysis is an important factor. We believe that the ability to define kinetic populations is related to the good reproducibility of eCl values from single spot analyses for samples with replicate measurements (shown in our Fig. 2) and this implies zoning is not a pervasive problem. We mention that single spot elemental analyses are used in line 256 in the paragraph immediately below. To further clarify this point, we have modified the sentence on lines 248-250 (new insertion in italic) to be "Replicate elemental *(a single EPMA spot per AFT analysis)* and Dpar analyses from separate measurements on grains with both age and length data (step 2, Fig. 1) are very important for assessing the reproducibility of kinetic parameter values (Fig. 2)."

Likewise, how many Dpar measurements are averaged for each Dpar determination? The usual procedure is to average four, which ought to make the reproducibility better than observed in Fig. 2c.

The LHA003 sample was analysed at AtoZ Inc. following protocols established by Ray Donelick and these procedures continued to be followed at GeoSep Services. The majority of D_{par} measurements were derived by averaging four D_{par} measurements per grain analysis. We modified text (line 153, p. 7; new insertion in italic) in section 2.1 to read, "…D_{par} is measured for individual apatite age grains *(average of four D_{par} measurements where possible)* …." Details such as this are also included in the separate GSC Open File 8821 that contains the sample AFT and elemental data.

Also, it's a little unfortunate that the discussion of the downsides of this procedure (lines ~326-340; might not get a compositional analysis near the counting area, or for the grain at all, I gather partly due to the LAICPMS spot) is in the next section; the authors can probably clarify and condense things by briefly mentioning these here, and then referring to them in section 2.3.

We decided to address this point in section 2.1 (all changes in italics). We also added comments concerning elemental wt% totals because the issue was raised in the community comments. We made additions and modifications to the text in several paragraphs to clarify the order of steps for acquiring AFT data. Sentences were restructured and reordered in the third paragraph of section 2.1 (lines 151-165, p. 7) to make it clear that all track count and length measurements are done before laser ablation to obtain U concentration. The first two sentences duplicated information already presented and were eliminated. The relevant changes to the paragraph are:

"Following standard mineral separation, and *grinding, polishing and etching of apatite crystals to expose spontaneous tracks, grain mounts are typically ^{252}Cf –irradiated to increase the number of confined tracks for length measurement (this may not be necessary for samples with high track densities such as Precambrian samples). Then,* spontaneous tracks are counted, D_{par} is measured for individual apatite age grains *(average of four D_{par} measurements where possible),* and grain x–y coordinates are recorded so that subsequent measurements can be linked to the age grains *(Step 1, Figure 1). The sample is re-etched to reveal horizontal confined tracks and their lengths, angles with respect to the mineral c-axis, and D_{par} are measured and x–y coordinates are recorded for the measured grains. Finally,* the sample is analysed using LA-ICP-MS to obtain U, Th, Sm, U–Pb age, *and trace element (as an option)* data for the AFT age grains, ensuring that the laser spot coincides with the track count area to minimize any potential problems with inhomogeneous U distributions. *Jepson et al. (2021) discuss how U-Pb age, trace element, and AFT*

data can be used to enhance thermal history interpretations. As an *additional* option, *U-Pb age*, U, Th, Sm, *and other trace element* data can be acquired for the length grains as well."

We changed line 169-171 to read as, "*Currently,* we recommend that elemental data be acquired using electron probe microanalysis (EPMA) rather than by LA-ICP-MS even if the latter method *allows for elemental data to be acquired at the point of age measurement and it is* more convenient to integrate in the workflow." Also the paragraph starting on line 167 was split into two paragraphs to accommodate requested changes by GSC internal reviewer, Jeremy Powell (see document outlining changes to the manuscript in response to his comments).

We added the following two sentences to the beginning of the paragraph on line 187 ("EPMA is undertaken…). "*We try to obtain EPMA measurements on a smooth "clean" surface but this is not always possible for small grains with many etched tracks and other imperfections. Missing elements and track void space in the electron beam excitation volume can result in elemental totals that are less than the 97-100 wt% expected for good analyses (see Issler et al. (2021) data tables) and this can happen for samples with AFT data acquired by LA-ICP-MS or EDM. Fortunately, we have observed that suboptimal analyses still yield an elemental signature that allows for discrimination of different kinetic populations and replicate elemental analyses with good and lower elemental totals yield similar results for the study samples.*"

We added the following sentence after the text on replicate analyses on line 190. "*Results indicate that compositional zoning is not a common problem for the samples we have studied which is important because the grains are not probed at the exact point of age measurement (laser ablation precedes EMPA). This could contribute to the occasional compositional outlier in kinetic parameter space if elemental zoning is present but kinetic populations are still better resolved compared to when conventional parameters D_{par} or Cl are used (see below). Changing the order of steps so that EPMA is done before laser ablation may help alleviate this problem and reduce the number of AFT grains without elemental data but it could delay sample analysis time by up to several weeks because samples must be transferred back and forth between labs and schedules need to be coordinated. Our current method is efficient and works well for the majority of our samples but it can be modified as needed to deal with more problematic samples. For example, if compositional zoning is a significant issue, it might be better to do laser ablation after EPMA and obtain track lengths from age grains only.*"

Finally, we added an inline reference to section 2.1 on line 337 in section 2.3 and eliminated the following two sentences:

"Although elemental measurements can be made with high precision, they are the last step after the grains have been subject to multiple treatments (e.g., etching, laser ablation, Cf-irradiation and etching) and it may not be possible to find a clean spot for EPMA that is close to where the age was measured. Therefore, if compositional zoning is present, the grain may appear as an r_{mr0} outlier."

[line 269] Although Dpar imprecision is certainly responsible for a lot of the scatter in Fig. 2e, it's not clear it's the main reason; the authors might try only plotting the points within the 20% bars in Fig. 2c and seeing what the Dpar vs. eDpar scatter looks like. The even scatter might simply be an indication that the things that throw Dpar off are bidirectional; a little OH might increase resistance to annealing compared to no OH (i.e. F-apatite), and a lot of OH might decrease it (e.g., OH-apatite HS from Carlson et al. (1999), but the more OH you have the higher Dpar is.

We have given a lengthy reply in our initial author response (AC1). However, we don't see how any of this would fit well in the paper without increasing its length. We think it might best serve as an online reference as is.

[Figure 2] Maybe smaller symbols would be better to avoid some of the "solid cloud" effect; some "N =" annotations also would not hurt, and maybe correlation coefficients for d and e.

We have decreased the plot symbol size to try and reduce the point saturation. It is tricky to get around this when you are plotting thousands of points at a reduced scale. Given the very large scatter of the data in Figs. 2c, 2d and 2e, we chose not to use linear regression but instead show how the data are distributed around a 1:1 line. We prefer not to include correlation coefficients because we don't think the relationships would be very meaningful, especially for Fig. 2d where there are strong systematic differences between eCl (a function of many elements) and Cl. Coefficients would change with the addition of more data from apatite of different elemental composition because eCl can vary considerably for a given value of Cl. The plot data are included in the supplement to the paper should someone wish to analyse it in a different way. We removed the number of data points from the title for each plot and added "N=" as inset text for each panel.

[line 291] "colour-coded"

Quotation symbols have been added.

[line 295] It may be worth noting that compositional populations may also be good candidates for shared inheritance. Although eCl is one such possibility, insofar as it combines a number of compositional variables into one number, apatites with similar eCl may get there via different compositional components, and thus not constitute a good candidate for shared inheritance. This is discussed further below.

See comment AC1. We don't think it should be mentioned at this point in the paper for the reasons given. The topic of inheritance is mentioned in the discussion.

[line 333-337] Maybe here or elsewhere, discuss the choice between switching which bin a grain is in, versus leaving the grain out altogether.

See detailed reply in AC1.We added the following two sentences at the end of the paragraph to emphasize the choices we make and the reasons for doing so. "*In our experience, the LA-ICP-MS method produces consistent and reliable AFT ages and therefore we use as much of the data as possible. If a higher precision AFT age matches an existing age population but plots as an outlier in kinetic space, we prefer to reassign it to the matching population rather than omitting the grain.*"

[line 438] The claim that population 3 has retained tracks from 540 Ma, or from about 245°C (Figure 6) is eye-catching, and probably overly optimistic about the ability of AFT to retain information about such high temperatures. It appears to stem from a difference in how AFTINV evaluates total annealing versus HeFTy's "oldest track". HeFTy assumes total annealing after reduced mean length falls below 0.4095 for non-projected lengths, corresponding to a mean length of just under 7 µm, whereas AFTINV appears to have total annealing correspond to a mean track length of 2 µm (line 419). This may be based on a slight misinterpretation of what's written in Ketcham et al. (2000); the 2 µm limit mentioned there corresponds to the smallest track that can appear in a track length distribution. However, such occurrences are due to including a population of tracks with a higher mean and large standard deviation. The 0.4095 value arises in part from the observation that no annealing experiments reported by Green (1988) or Carlson et al. (1999) had a mean length below 7 µm (although there are some 6's and 5's reported by Barbarand et al.

(2003), and even an occasional 4 or 3 by their Analyst 3). Willett (1997) uses a similar value of 0.428 as the zero-density intercept of reduced length versus density reported by Green (1988). In other words, by the time a mean length falls below some limit, the track population becomes undetectable. I believe this provides a more realistic basis for evaluating total annealing and the oldest retained track. Using the revised criterion, the TA for the oldest track for an rmr0=0.491 apatite is closer to 200°C, which seems a lot more reasonable considering the closure temperature is 161°C. This is not the most crucial of issues, but it's prudent to avoid distracting claims.

See detailed reply AC1. Short tracks exist but are rarely observed, in part because they are less likely to be intersected by an etchant pathway than longer tracks. The issue of model retention ages was also raised by internal reviewer, Jeremy Powell, and we made changes in response to his comments. We changed the range of annealing temperatures from ~100-245°C to ~110-185°C (Abstract, line 16 and Introduction, line 84) based on the ability of the model to resolve temperatures rather than on predictions using retention ages for very short tracks. For figures 6 and 7, we added a dashed line and arrow to show the model temperature range that is unconstrained by the AFT data and we removed the reference to total annealing temperatures for the ~2 μm tracks used to calculate retention ages. The figure captions have been updated accordingly. We also removed reference to total annealing temperatures linked to retention ages in the text (lines 429-430 and 454-455).

[Figure 6, 7] I appreciate the authors' efforts to incorporate the CRS method into AFTINV, and intrigued by the result – it looks to be a powerful addition. I have long been considering doing something similar myself, having dropped the CRS method when I converted my earlier program AFTSolve to HeFTy. However, one of the reasons I did so may still be evident in the model results here. The CRS method has a tendency to quickly converge to a relatively smooth solution that does not explore the solution space as well as the Monte Carlo method, and thus map out the range of solutions that fit well. In HeFTy results, this allows the resolving power of the data to be evaluated by looking at the width of the solution envelopes.

We think our detailed response (AC1) is a useful online reference but we don't think this material would fit into the paper.

In the results here, what puzzles me for P013-12 is the relatively tight band of good solutions above 175°C from 600-450 Ma, and probably a fair bit younger/cooler than that. Given the 161°C closure temperature of the most resistant population, the idea that it would exert much constraint in the 175-250°C temperature range seems improbable, and is not reflected in the QTQt results either. This all is not necessarily a problem, but I think it should be discussed so people interpreting these results have a more complete knowledge of what they are looking at.

See detailed response AC1. We updated figures 6 and 7 to indicate the parts of the thermal history that are resolved by the AFT data and we added some explanatory text. We added the sentence, "*The dashed line (upper panel, Fig. 6) coincides with a change to a steady cooling rate (~1.2 °C/My) below 185 °C at ~440 Ma and marks the upper temperature limit that can be resolved from modelling the AFT data*" on line 437 for the P013-12 sample. On line 460 (end of paragraph), we add the sentence, "*Model results suggest that the AFT data can resolve the thermal history below 175 °C after ~480 Ma (dashed line in upper panel, Fig. 7)*" for the LHA003 sample.

Along similar lines, did both the AFTINV and QTQt models assume that all apatites in each sample had the same inherited, pre-depositional history? If so, was the fact that they did so, and their success in fitting

their models, and indication that there was shared provenance, or an indication that, for these samples, results are not terribly sensitive to the pre-depositional history?

See detailed response AC1. We added the following paragraph after line 460 at the end of section 4.1 AFTINV.

"*The P013-12 and LHA003 samples were modelled assuming a common inherited, pre-depositional history for each kinetic population in order to better resolve the post-depositional thermal history by taking advantage of relative annealing. We cannot determine whether the kinetic populations within each sample have a shared inheritance because this information has been degraded by thermal annealing of the less track-retentive populations. For LHA003, the pre-depositional thermal record for the lower retentivity population one was erased completely by thermal annealing and therefore pre-depositional cooling is only constrained by population two. For P013-12, pre-depositional thermal history has been erased for population one and population two experienced significant post-depositional annealing. Pre-depositional cooling is dominated by population three which can easily overlap with any residual cooling record for population two. Overall, the post-depositional thermal histories are not very sensitive to the pre-depositional cooling for these samples*."

Or, are the results sensitive – do the few earlier-cooling 0.5 paths for P013-12 corresponds to the earlier peaks T's at ~195 Ma and/or ~70 Ma?

We answered this specific question in our reply (AC1) but we don't see how to directly address this in the paper without it looking like we were answering a reviewer question. See response immediately above.

The manual (AFTINV) and automatic (QTQt) raising of the rmr0 values for the most resistant populations in each sample is interesting. What seems to be going on is that the different populations need greater separation in their partial annealing zones to produce their respective divergent age and length distributions. It's further interesting that the higher resistance is corroborated by the vitrinite data for sample LHA003, though less so for P013-12. The authors recommended approach of "anchoring" on low-resistance kinetic seems like a good one. Another possible "advantage" of the Ketcham et al. (1999) model over the (2007) one beyond the different rmr0 equation is that it has a much higher temperature range, which these results may imply is necessary to create these divergent populations.

Our reply (AC1) adds some information that may be useful as a reference but doesn't seem to be necessary to include in the paper.

Lastly, the comparison between AFTINV and QTQt results appears to gloss over their differences a bit. For P013-12, the first reheating peaks at ~168 Ma in AFTINV and could go as far back as 195 Ma, whereas QTQt appears to strongly say that it was at about 140 Ma. Similarly, AFTINV implies that the first peak reheating for LHA003 was at 345 Ma, compared to 300 Ma for QTQt. If you lay the models pairs on top of each other, they appear to exclude each other at these times. Is this because QTQt calculated different kinetics than the manually-shifted ones in AFTINV, or because of QTQt favoring simpler histories, or some combination of these and possibly other factors?

We give a very detailed response (AC1). We added more text to the second paragraph of section 4.2 on QTQt to discuss the differences in model results and the probable factors responsible for this.

We inserted the following sentence after the second sentence in the second paragraph of section 4.2 QTQt (line 480). "*These temperatures are at the upper end of the range defined by the CRS solutions in Figure 6 (132–147 °C between 150–195 Ma and 102–110 °C between 45–70 Ma) but the 95% confidence region overlaps with the AFTINV results.*"

The following two sentences were added after the sixth sentence (line 487). "*QTQt predicts a younger time for the first thermal peak than the AFTINV CRS solutions (Fig. 7; ~320–360 Ma) but the temperatures and times for the other peaks show good agreement and both models overlap in the 95% confidence region. AFTINV uses larger eCl values and more closely fits the AFT parameters for the most retentive population in both samples.*"

We started a new paragraph beginning at the last sentence (line 487) and added the following five sentences at the end of the QTQt section to summarize reasons why the QTQT and AFTINV model results differ in some details.

"*Overall, the AFTINV and QTQt results are very similar, even with the subtle trade-offs between the different thermal minima/maxima inflection points and preferred model population kinetic parameters. Model results differ in detail for a number of reasons. Compared with QTQt, AFTINV uses more model points, constructs thermal histories differently, allows for manual fine tuning of kinetic parameters, and generates a much larger set of "acceptable" and "good" solutions. QTQt generally prefers simpler histories and there is a trade-off between the number of time-temperature points and data fit. QTQt converts LA-ICP-MS AFT data to EDM AFT data and uses Ns and Ni count data rather than ages for modelling whereas AFTINV models either EDM or LA-ICP-MS AFT data using central or pooled ages, depending on* χ^2 *and age dispersion statistics. This difference is most evident for sample P013-12 where QTQt uses younger observed population ages for model input than AFTINV (compare Figs. 6 and 8).*"

Review 2 (Karl Lang) Comments

Reviewer comments are in black font and specific author changes to the manuscript are in red font.

Use of "detrital" was a little confusing to me at first, since many applications of detrital thermochronology are now also focused on interpreting cooling histories of source rocks prior to deposition, and not simply the common cooling history of detrital minerals in a sedimentary rock after deposition. This is a semantic difference, but perhaps adding a sentence to state this explicitly at the beginning of the manuscript might clear up any confusion amongst readers.

We changed the first sentence in the Abstract to read "…cause for AFT age dispersion in sedimentary samples" rather than "…cause for age dispersion in detrital AFT samples." We changed "…detrital AFT samples with apatite of variable cation and anion composition to have significant age dispersion…" to "…sedimentary samples with apatite of variable cation and anion composition to have significant AFT age dispersion…" in the first sentence of the Conclusions. We retained detrital elsewhere in the text.

Why does the manuscript include a vigorous preference of LA ICPMS over EDM approach? This seems unrelated to the central motivation of the paper and, in my opinion, is largely unsupported (see comments by line). The authors should explain why they chose to use LA-ICPMS instead of EDM, but they should avoid generalized claims about the relative efficacy of one method over the other (e.g. "The LA-ICP-MS method has some distinct advantages compared with EDM" [117]).

We did not intend to indicate a strong preference for LA-ICP-MS with respect to EDM. We have excellent EDM AFT data that show clearly defined multikinetic AFT populations that can be modelled successfully. We anticipated that there may be critical community comments concerning the LA-ICP-MS AFT method (we surmised correctly in this case) and so we inserted some text to emphasize that the two methods are different in some respects but that these differences may have different advantages. However, if these comments are being viewed as an assault on the EDM method then we appreciate this

being pointed out. On rereading this section, we agree that the text is peripheral to the main point of the paper and, as suggested by the AE, these ideas are better left to a future study that compares both methods. Therefore, we have replaced the first two paragraphs in section 2.1 (lines 114 to 146) with the following paragraph.

"This section discusses the type of data required for multikinetic AFT thermochronology; more details on sample analysis are in Issler et al. (2021). Our AFT data were acquired using the LA-ICP-MS method (Chew and Donelick, 2012; Cogné et al., 2020; Donelick et al., 2005; Hasebe et al., 2004) although the technique works equally well using the older external detector method (EDM; Hurford and Green, 1982). A key difference is that the U data needed for AFT age determination are acquired for the spontaneous track count area after counting is completed (LA-ICP-MS method) whereas counting of spontaneous and induced (proxy measure for U from sample irradiation) tracks are done at the same time (EDM). Typically, 40 single grain AFT ages and 100–200 track lengths are obtained per sample, depending on apatite yield. Generally, this amount of data is sufficient for most multikinetic samples with two or three kinetic populations, but more data may be required for samples with unevenly distributed populations or with more than three populations. In contrast, many EDM AFT studies have used a lower number of age grains per sample (usually ≤ 20) for thermal history studies. The greater number of counted age grains naturally increases the statistical probability of χ^2 failure that may complicate mixture model interpretation (McDannell, 2020; Vermeesch, 2019)."

118. It has not been my experience that analytical costs are lower for LA-ICPMS than for the EDM when measured on a per grain basis. If you can measure 100 grains per mount and 50 mounts fit in a $1000 irradiation package, that's $5/grain. By comparison, LaserChron (probably cheapest option in US, at least) charges $9-16 per grain for 100 grain samples, not including costs for CL imaging. Also, throughput is not necessarily higher for LA if you have to wait several months for lab time to become available. In my experience the analytical time to produce a complete fission-track dataset is comparable regardless of the analytical approach. I worry that comments like this will gradually discourage scientists from using the EDM, which is a well established and data-rich method.

Our reply (AC2) provides some background reference pertaining to cost. We removed "and costs are lower" from line 118. Although costs are lower for us, it is difficult to generalize when different labs and different methods are being used.

137-138. Wouldn't observer bias have a greater impact on age determination when it is only accounted for in spontaneous track counting? It seems to me that observer bias may actually be reduced when it is accounted for in both the spontaneous and induced track counts, rather than just in spontaneous counts. Either way, I don't consider it fair to say that LA-ICPMS is "more objective" if it still relies on user interpretation and collection of spontaneous track data.

We removed the text but we believe that the issue of objectivity and the potential for grain selection bias should not be dismissed categorically. We maintain that LA-ICP-MS is more "objective" in that analysts must count track areas without advance knowledge of U concentration. They may be able get some assessment of potential U zonation from the distribution of spontaneous tracks which can be helpful for avoiding areas with strong zonation but all track counting is done prior to U measurement. With EDM, there is the potential for track counting to be influenced by accumulating Ns/Ni count data, especially if there is a desire to have samples with minimum age dispersion. We know in the real world that outcomes can change when people have additional information and that the "Monty Hall effect" is a real phenomenon. We do not know how significant a factor this is for real measurements but we point out that

the potential is there for it to be a factor that may lead to underrepresentation of multikinetic populations. Nevertheless, this discussion belongs elsewhere because we are not dealing with EDM data in this study.

140. This is not an inherent limitation of the EDM, simply a choice by the operator to count fewer grains. Many detrital studies regularly count more than 100 grains per sample with the EDM.

Yes, but in practical applications of EDM for thermal history analysis, most studies have counted a lower number of grains compared to what we measure using LA-ICP-MS.

142-143. It is convenient to make this argument here, but one could also make an alternative argument that the induced track print actually allows for more robust data collection because you can avoid the zonation issues you mention to be a problem on line 128-130. I don't understand why this is cast as an example of making the EDM less objective.

This point is generally overlooked as a possible factor when comparing differences in LA-ICP-MS and EDM AFT data so we think it is worth mentioning in future work that compares the two methods. Zonation is a possible reason for data discrepancies, but if lab work is done in a careful and consistent manner, we believe that it is not a widespread problem based on the success we have had with analyzing Phanerozoic sedimentary rocks in northern Canada.

145. Again, this is not an inherent problem with EDM it is a choice by the EDM user.

We agree that user choices can influence results and that experienced users who are aware of these issues can mitigate these problems.

Jeremy Powell (GSC Internal Review) Comments

Reviewer comments are in black font and our replies are in red font.

Abstract, line 12. Consider “we present an interpretation and modelling strategy that exploits multikinetic AFT annealing kinetics of samples with compositional variability. These multikinetic thermal histories provide more detail and better resolution compared to conventional methods”

We inserted extra text (in italics) in the sentence so it reads as, “We present an interpretation and modelling strategy *for samples with variable apatite composition* that exploits multikinetic AFT annealing to obtain thermal histories that can provide more detail and better resolution compared to conventional methods”.

Abstract, line 52, p2. Substitute “regions” for “areas”?

We retained areas because both words are synonymous. Seems to be a matter of personal preference.

Line 84, p. 3. Should specify that these are based on the thermal history models and not independent annealing studies. That’s not clear in this sentence.

We changed the total annealing temperature range from ~110 °C to 245 °C to ~110 °C to 185 °C based on the range over which the model can resolve thermal histories. Sentence has been modified to say “The AFT kinetic populations of this study have a wider range of total annealing temperatures (~110 °C to 185 °C *based on model thermal histories*) than typical fluorapatite (~110 °C)…”

Table 1 – should include which NAD is referenced.

NAD83 datum was added to the bottom of Table 1.

Multikinetic AFT methodology section, line 90-100. I agree with not discussing the geologic setting or providing a map in this manuscript, but consider referencing other publications or geologic maps that the readers can look up if interested.

We updated a sentence (new text in italics) to read "Table 1 summarizes basic sample location and stratigraphic information *and a geological map with plotted sample locations is available in Issler et al. (2021).*"

Multikinetic AFT methodology section, line 90-100. Available thermal maturity data could be discussed here as well – both for the samples, and for the overlying Cretaceous section.

We added the sentence, "*Percent vitrinite reflectance (%Ro) data (Issler et al., 2021) indicate that both samples experienced paleotemperatures that were high enough (~135 – 175 °C) to cause substantial AFT annealing.*" We also modified a sentence (new text in italics) to say, "Approximately 1 km of Cretaceous strata *(≤ 0.6 %Ro)* overlie the unconformity…"

Figure 1, Step 4: shouldn't the referenced equation be Carlson et al., 1999?

Yes, Ketcham et al. (1999) was changed to Carlson et al. (1999) for step 4 of flowchart.

Figure 1, Step 6: the grey kinetic boundary is hard to see. Might not show up online. Maybe make it darker?

The light grey dashed lined was replaced with a dark grey.

Line 135. Rework this sentence. It's a bit of a mouthful, which makes it hard to understand the point you're making.

We have divided the sentence into two sentences. It now reads as, "*We think that both the LA-ICP-MS and EDM methods can yield equally good results for the age grains that are measured. This has been demonstrated repeatedly in publications where EDM and ICP-MS methods were compared or where ICP-MS AFT dates were referenced to samples with well-determined absolute ages (e.g., Ansberque et al., 2021; Cogné et al., 2020; Hasebe et al., 2004; Iwano et al., 2019; Seiler et al., 2013; Soares et al., 2014).*"

Line 150-151. This section feels a bit redundant given the description in lines 103-112. Is one of these sections the figure caption?

We removed the first two sentences, "Figure 1 illustrates the key steps in our multikinetic workflow. Step 1 summarizes the procedures needed for acquiring AFT and related data using the LA-ICP-MS method" to avoid duplication of the text above. We modified the next sentence by inserting "*(Step 1, Figure 1)*" at the end of the sentence. This should convey the essential information.

Paragraph on lines 150-164. Consider expanding the utility of AFT + UPb by referencing Jepson et al 2020 and their work on double dating.

Modified line 153-155 (new text in italics) to read, "The sample is analysed using LA-ICP-MS to obtain U, Th, Sm, U–Pb age, *and trace element (as an option)* data for the AFT age grains, ensuring that the laser spot coincides with the track count area to minimize any potential problems with inhomogeneous U distributions." Added the following sentence, "*Jepson et al. (2021) discuss how U-Pb age, trace element, and AFT data can be used to enhance thermal history interpretations.*" Modified line 158-159 to read,

"As an option, *U-Pb age,* U, Th, Sm, *and other trace element* data can be acquired for the length grains as well."

Line 170-175. Provide a comment on the challenges of measuring F using EPMA as well. It would be useful if you recommended analytical conditions for measuring apatite composition via EPMA – are you doing two runs and changing the operating conditions for halogens vs majors? Is there a processing software that the lab uses to mitigate beam migration, etc.

We split the paragraph into two paragraphs and added the following two sentences at the end of the first paragraph. *Analytical conditions for elemental analysis are summarized in the elemental data files included with the Issler et al. (2021) sample report. For the sake of efficient sample processing, a single setup was used for elemental analysis and time-dependent corrections were used to deal with halogen migration (similar to Nielsen and Sigurdsson, 1981), with the knowledge that crystals oriented with their c-axis parallel to the electron beam could yield some inaccurate results.*

Line 224, equation 2. I think eCl is a really confusing concept to many readers, so it's good to be crystal clear in your description of it. For example, It might be a bit tough for readers to connect that Cl* of Ketcham et al 1999 is the eCl of this paper, and that Abs(Cl-1) is the measured Cl component.

We have chosen to use the same equations as given in Ketcham et al. (1999) and all parameters are defined. For those less comfortable with the mathematical aspects, we cited equation 1 of McDannell and Issler (2021) which shows the final transformed equation for eCl. The sentence on lines 229 and 230 now reads as (new text in italics), " Equations (2) and (3) can be used to transform measured kinetic parameters (i.e., Dpar and Cl) to r_{mr0} values or vice versa by rearranging the equations in terms of Cl and Dpar *(see equation (1) of McDannell and Issler (2021) for eCl)*."

Line 239, change "eCl" to "The eCl parameter.

Done.

Line 253. Can't quite tell, but are their two ranges indicated around the 1:1 line? The figure says 5-10% but I only see one set of dashed lines.

No. The dashed lines represents ±0.03 apfu. This value is typically 5-10% of the range of eCl and Cl values present in a sample.

Line 253. Also, consider adding a figure showing replicate analyses of F (maybe in the supporting information document). I think that the EPMA composition could be a sticking point for some reviewers. The description of EPMA data acquisition is limited, and some reviewers might suspect that the Cl vs eCl difference could partially be due to poor F measurements resulting in changes in stoichiometric OH. I know that's almost certainly not the case, but it might help this manuscript to be a bit more descriptive in how you acquire F and how replicate measurements look.

We added F data to table S1 in the supplement and included a plot showing the results of replicate F analyses with the table. We added the following two sentences after the discussion of the replicate results for Cl and eCl. *A similar plot of 322 replicate results for F is included in the supplement with Table S1 and it shows that 92% of the measurements are within ± 0.2 apfu with larger variations associated with zoning and nonstoichiometric F values. Overall, the accuracy of the halogen measurements is sufficient for estimating OH contents and calculating r_{mr0} values.*

Line 258. Of major and minor elements – might be worthwhile specifying, as readers could relate this to your discussion of U-Th-Sm zonation and LAICPMS vs EDM above.

We think this may not be necessary as it should be implicit that chemical zoning refers to major and minor elements.

Line 298, Fig 3 doesn't show chi square values for individual population.

These are listed in Table 2 for the kinetic populations. DensityPlotter only shows the χ^2 probability value for the whole sample. The χ^2 numbers are more relevant for the interpreted kinetic populations which are used for thermal modelling.

Also, Fig. 3 caption needs to mention what the values in brackets next to the peak ages are.

Added the following to the caption, "*(estimated percentage of grains per population in brackets).*"

Line 373-374. Maybe mention that dynamic limits are determined iteratively and only applied to make the modelling more efficient? Some might misinterpret this as forcing the model.

Modified sentence to read (new text in italics), "Static limits define the entire model search space whereas dynamic limits are applied only at model inflection points to focus calculations into favourable regions of solution space *to improve model efficiency.*"

Line 388. As well as dynamic temperature limits?

We consider determining the style of thermal history to encompass any adjustment to rate/temperature limits.

Line 417. This description of retention ages could be more clear. The retention age concept is something that many people seem to struggle to wrap their head around, so consider adding an extra sentence or two describing it.

Replaced the existing sentence with the following two sentences. *Model retention ages represent a theoretical age for the oldest (shortest) track in each population (assumed to be ~ 2 µm based on the shortest track ever measured; Ketcham et al., 2000) and provide an uppermost temperature and time limit for track survival. However, very short tracks are rarely observed and maximum temperatures constrained by the AFT data may be significantly lower.*

Line 531. 'may yield significantly more thermal history information'

Changed "can contain" to "may yield."

Line 557. This is an important paragraph of the discussion, as it develops an argument as to why r_{mr0} is a valid kinetic parameter despite the warning of Ketcham et al 1999. I think that the last sentence loses the point a little, and could be reworded to better emphasize these points. In my opinion, the important stuff here isn't that you're 'advancing the field of thermochronology', it's that you're revisiting existing experimental data in a novel way to improve how we apply thermochronology methods to complex geological problems. Maybe that's pedantic, but some might read that sentence and think "advancing the field of thermochronology would be fully understanding these complex annealing behaviours that aren't predicted by Ketcham/Carlson."

Changed the last sentence of the paragraph to, "*Our method pursues the logical consequences of the annealing experiments and shows that it is possible to use existing techniques in a novel way to improve how we apply thermochronology methods to complex geological problems.*"

Line 566. In between these sentences might be a good place to mention that the <0.73 r_{mr0} range also corresponds with where we have no constraints in the annealing datasets, so the discrepancy b/w measured r_{mr0} values and those required by the modelling is expected.

Added the sentence, "*There is a dearth of experimental annealing data for r_{mr0} values < 0.73 so the discrepancy between r_{mr0} values calculated using elemental data and those required for modelling is expected.*"

Line 598. Just a thought as I review – but did you mention that you used Basin%Ro to model vitrinite? It would be good to reference that model and why you use it instead of EasyRo (or reference our Mackenzie Plain paper), given that this paper is emphasizing how we can extract complex thermal histories from sedimentary rocks with multiple burial cycles.

Yes, the basin%Ro model is cited in the section, "3. Thermal history modelling of multikinetic AFT data." I modified the second last sentence of the second paragraph (lines 379-380) to be (new text in italics), "Vitrinite reflectance (%Ro) values are calculated for the entire post-depositional thermal history and for the last phase of heating and cooling using the basin%Ro model (Nielsen et al., 2017) *which provides better fits to observed maturity profiles in northern Canada than the Sweeney and Burnham (1990) EASY%Ro model (e.g., Issler et al., 2016; Powell et al., 2020).*"

Line 625. Another point is that multikinetic samples may produce similar thermal histories when modelled, but have very different age populations. It would be a useful thought experiment (maybe a future paper?) to test what increasing or decreasing the temperature of one thermal event does to the pooled age and track length distribution of these populations – presumably, lower Cretaceous temperatures = more effect of pre-Cretaceous thermal history on the youngest population and a partially annealed age. Whereas, higher Cretaceous temperatures might lower the age of the Mesozoic and Paleozoic population. Judging by the samples that we've worked on, I suspect that it doesn't take much of an increase/decrease in temperature to produce noticeable effects. I mention this, because as people begin to work with multi-kinetic samples they will inevitably be frustrated/sceptical when proximal samples yield different age populations. Many will skim over your point on "value of the approach judged on the ability to generate spatially coherent thermal histories", which is a very important conclusion of your multikinetic work.

We added the following two sentences to the last paragraph of the Discussion section. "*Furthermore, multikinetic samples with different age populations may produce similar thermal histories, depending on how differences in provenance and composition interact with the thermal history. Therefore, proximal samples may not necessarily have the same age populations."*