Below we address reviewers' comments. Reviewer's comments are in black text. Our replies are in blue text.

Reviewer 1: Kerry Gallagher

Perhaps the authors could try to reduce and reorganise the text to help the reader. For example, 2.5 pages on r_{mr0} calibration is a bit of a distraction - much of that could go into appendix/supplementary (perhaps keep something on effective Cl for the main text).

We agree. This material is better suited for the next paper on using natural AFT samples to illustrate the multikinetic method. We will retain minimal information to explain what we did and the remaining r_{mr0} discussion will be removed as it distracts from the flow of the paper.

There is a lot of detail from the forward model (predicted mean lengths, initial lengths) in the text that is also on the figures, so just keep the latter, or put all the numbers in a table (and the table perhaps in supplementary?).

We will make appropriate changes to reduce duplication.

Also, the results of using the wrong model (e.g., mono-compositional when it should be multi) are probably too long. I think most of us would appreciate that using the wrong model is likely to be a problem. The important point may be that we can still fit the data reasonably well (using a single sample).

We will try to reduce text where possible but we believe this is a very important point that deserves serious attention. We respectfully disagree with the last two comments. The data interpretation (data model if you like) is incorrect and therefore modelling produces a significantly different thermal history. We are not sure that most people can appreciate how different it can be without an example that illustrates this. The ability of the model algorithm to fit data closely is not a sufficient criterion for a good solution if the interpretation is erroneous. Generally, it is not difficult to fit data using a single AFT population. If there really is only one population, then it will provide useful constraints on the parts of the thermal history where it has sensitivity. However, for a multikinetic sample, separate populations can have significantly different thermal annealing behaviour. The single population interpretation means that all components are combined and treated as having the same annealing behaviour. This means that the length distribution, AFT age, and kinetics are different from the original synthetic data. Therefore, the ability of the combined data to discriminate different heating events is diminished because all grains are assumed to have the same thermal sensitivity by being lumped together. Inevitably, the model thermal history must change to accommodate this assumption. If multiple populations are unrecognized then the thermal history will be distorted in order to fit the data. The degree to which this happens is data-dependent. A very close fit to the data under these circumstances can give a false sense of confidence in the solution. Our single population example without constraints shows continuous cooling, completely missing the two separate heating and cooling events, and the geological interpretation of the results is quite different from the original history. If constraints are added, then thermal peaks appear but they are offset significantly in time and temperature with respect to the true model thermal history. Although the fit to the data is excellent for the combined sample, we do not consider this a reliable thermal history prediction. This is essentially the main point of the manuscript.

They need to state clearly up front the assumptions underlying some of the models - for example the multi-element/compositional models for the calibration to r_{mr0} or effective Cl (eCl) are not perfect, are likely to contain correlations between the fitted parameters). If they do not have access to the original data or calibrations, then perhaps some kind of resampling could done (e.g. take some elemental composition data, resample those data using typical uncertainties and recalibrate the model).

We do not think that model recalibration is warranted or appropriate here and it is well beyond the scope of this short contribution. These comments underscore why we need to move the discussion of r_{mr0} calibration to our next paper, which uses natural multikinetic samples, because it is peripheral to what we are trying to show here. First, these are synthetic data and they represent one example from an enormous range of possibilities. We have already prescribed the kinetic parameters beforehand based on values that were determined for some natural samples in order to generate the synthetic data. This is not a real dataset so it is unclear what new insight will be gained by resampling and recalibrating the model. Model recalibration will not affect the two basic conclusions of this paper: (1) multikinetic data can contain a more detailed record of the thermal history than a single population, and (2) failure to recognize multikinetic behaviour could adversely influence thermal history results. You should arrive at similar conclusions no matter what kinetic model you choose. The r_{mr0}-based kinetic scheme of Ketcham is the standard for AFT modelling and the 1999 or 2007 model will lead to the same conclusions. Any new proposed kinetic models where annealing temperatures differ between populations will also lead to the same conclusions. Discussions of model calibration are more relevant for natural samples where real elemental data are transformed into kinetic parameters and this is a topic of our next paper.

Also, their example (synthetic data) are very clean and distinct in their compositions. Do we see/expect such well separated populations often, and if so how have these been dealt with previously? When does the ability to resolve the thermal history based on compositional groups start to deteriorate if the compositional groups are less distinct?

These are all important questions and they are best addressed by reference to a suite of natural multikinetic samples exhibiting different characteristics. A substantial number of Phanerozoic detrital samples from northern Canada show good population resolution but this cannot be demonstrated in the current paper. A future goal is to get these data released in a series of publications for different study areas. Our synthetic data are meant to be ideal to show that distinct populations can lead to well resolved thermal histories. To modify a phrase used by the reviewer, I think most of us would appreciate that the ability to resolve thermal histories will deteriorate as kinetic populations become less distinct. We do not see the point in investigating this for the current paper because the results will be unique to this synthetic example. There are many factors to consider (see above comments) that make generalization of these type of model results problematic. We think a more thorough analysis of factors governing model resolution should be informed by what is observed in natural samples.

Going to the extreme, the conclusion that we might draw from this study is that we should model each grain with its own specific compositionally defined annealing model (and model parameters). I agree with the authors that we often need to consider sub-populations of data from a given sample both for AFT and AHe, and that averaging the data prior to modelling is probably not a good idea (or at least we need to acknowledge that we will obtain some kind of average, perhaps unrepresentative, solution and that we are potentially throwing out information). However, the other side of the argument would propose that the predictive models are not that sophisticated, not free of uncertainty, and not even really based on a well developed understanding of the physical processes and how they operate on geological scales.

We do not agree that you can draw this extreme conclusion from the ideal synthetic data we are presenting in this paper. This sounds more like an expectation based on other experience. We think people have been preconditioned to expect poorly resolved AFT populations because proxy kinetic parameters that are in common use have low resolving power, and in addition, simply due to the classically low precision of the AFT method in general. Therefore, population overlap is the normal situation if multikinetic data are present in a sample, which is probably due to random geologic noise as well as imperfect kinetic model calibration. From our own experience, many natural multikinetic samples display distinct statistical populations when plotted using the elemental-based r_{mr0} parameter and analysed using conventional approaches (age mixture modelling and radial plots). There are already published examples of natural multikinetic samples, so their existence in nature has been demonstrated. It is pointless to model single grains if, statistically, they fall into one of the discrete populations. The same criticism could be made of conventional modelling. Why not assign an individual Dpar value to every grain and model them that way. You would not do it because the modelling depends on the assumption that they are part of a single population. What we are doing is not so radical. Instead of assuming one overdispersed population, we are saying that elemental variation can result in multiple populations being present and that exploiting this can result in improved t–T modeling results.

Again, none of this can be demonstrated in the current paper that deals only with one synthetic example. These topics are the subject of our next paper, which uses natural examples to illustrate multikinetic interpretation methods and modelling. We agree that empirical models have uncertainty and are a simplification of complicated underlying physical processes. However, it is unclear to us why a multikinetic approach to modelling would be more adversely affected than current modelling applications. It has been demonstrated clearly by laboratory annealing experiments that AFT annealing temperatures vary with changes in apatite composition (and this has been mostly dealt with unsystematically for decades). Why would ignoring this fact produce a better model?

Overall I think that main premise could be demonstrated more efficiently. The idea is that chemical composition (and perhaps associated mineral structure changes) has a major effect on annealing and diffusion in apatite, and this effect is multi-element, rather than just Cl/F as sometimes assumed for fission track annealing. It is a good idea to promote an analytical protocol of measuring a wide range of elements, rather than just Cl, or a proxy such as Dpar, as these data may be useful in future for annealing model recalibration and/or provenance (e.g. O'Sullivan et al. Earth Science Reviews 201, 2020). However, the available model calibrations are not based on a lot of data, and as stated in Carlson et al. (1999) "*in the absence of any physical understanding of why compositional variations impede or enhance annealing, we have little confidence that it can be used meaningfully to predict the annealing behavior of apatites not included in the experiments*". The concern is that these preliminary calibrations are assumed to give us the definitive model, free of uncertainty, and the rather strong conclusions about the inference of thermal histories are based on that assumption.

We agree that we need to focus this paper better to avoid distractions concerning model calibration, which is discussed in our next paper involving natural AFT samples. In this paper, the ability to convert real elemental data into kinetic parameters is not relevant. We have already predetermined that the synthetic sample has three multikinetic populations with different relative annealing behaviour based on what we have observed in natural samples. The question is, “What are the consequences for modelling the thermal history if you have populations with different annealing kinetics?” In principle, can more thermal history information be retained in a multikinetic sample than a monokinetic sample? It must be kept in mind that the quote from Carlson et al. (1999) is an expectation that the empirical model would not be reliable because of limited calibration. This is an inference that was not tested with follow up studies. Unfortunately, this statement may have deterred people from attempting to use the method. It is also worth noting that in the absence of a perfect understanding of *why* composition alters FT annealing behavior does not negate the fact that empirical evidence indicates it is a real phenomenon and does not ultimately prevent it from being useful.

For example, the results presented are based on more or less ideal data with well separated (kinetic/composition and age) populations (as described in 236 to 249). In this case, we can pretty much recover what we started with including heating out needing to specify near surface constraints on the thermal history. Furthermore, the modelling approach implemented in QTQt tends to prefer simple models, conditional on fitting the observed data adequately relative to more complex models. I think this should mean that individual models making up the credible interval range figure 3a will tend to look like the ML model shown in that figure, and we do not fully recover the thermal history of 20°C for about 500 m.y. prior to the first heating event) - I would guess it is just the timing of reheating that changes, and probably the same for the second heating event.

Yes, using an ideal data set, we can recover much of the input thermal history except for the low temperature part where the model lacks sensitivity. Although this may appear to be a self-evident conclusion – modelling of multikinetic data is not a straightforward exercise. The ability to recover good solutions depends on the choice of modelling method and on how it is applied. Here, we demonstrate that good quality multikinetic data may preserve a record of multiple heating events under the right conditions.

We do not think that this is widely appreciated because people are used to dealing with single populations that are sensitive to a narrower range of the thermal history. This in part explains the fixation by many workers with whether or not AFT samples pass or fail the X^2 test — where passing does not absolutely ensure the absence of multiple sub-populations, nor does X^2 failure imply a 'poor quality' dataset or offer a direct means for understanding failure of the statistical test, which could occur due to many reasons such as high precision single-grain ages, large analytical sample size, or compositional heterogeneity (i.e., differential annealing). If none of these variables are assessed, dealt with, or ruled out – or the necessary data collected to do so – then how can the AFT community make any real progress towards advancing data interpretation and modeling practices?

The authors then demonstrate that combining these sub-populations and assuming an average composition (but generally fixed) leads to lack or resolution and/or spurious results for the inferred thermal history. The authors often imply that the latter is common practice, but do not really give any concrete examples. I think many, if not most, people working with fission track data are aware of the potential for over dispersed age data and hopefully would deal appropriately with an over dispersed population (using subpopulations based on composition or age, or perhaps remove egregious outliers - also this issue seems to have different significance depending on whether the data are collected with the traditional EDM method or LA-ICPMS, the latter method tends to have greater dispersion but similar central ages.

We think that modelling of mixed AFT populations is more common than realized but that it is unintended. We could cite specific examples but that could be viewed poorly, as we do not think the practice of 'multikinetic misidentification' is carried out on purpose, but instead with this paper and others in the future, we would hopefully draw attention to deeper investigation of complicated datasets by the AFT community. Awareness of overdispersed data and taking appropriate actions to understand it are two different things, with the latter likely yielding to the former in the majority of cases. If this were not so, there would be more examples to point to in the literature. We believe that multikinetic populations are best-resolved using elemental data (which are rarely collected) whereas population overlap is normal when using low-resolution kinetic parameters such as Dpar. If you cannot discriminate between populations, then modelling mixed populations is unavoidable. These issues cannot be addressed here with synthetic data but are the subject of the next paper. We will choose our words carefully here. We want to make the cautionary point that modelling multikinetic data as a single population can distort the results and it is something to take into consideration. In our experience, multikinetic populations are evident in both EDM and LA-ICP-MS AFT data. The reported higher dispersion with the LA-ICP-MS method is not a problem and both analytical approaches and derived data have been shown to provide similar results. The key is to have enough age and length data to characterize the different populations properly. This topic will be discussed in a subsequent paper that uses natural AFT samples.

The idea of AFT ages increasing then decreasing with Cl content was mentioned some time ago, I think by Barry Kohn, who had some data from Canada implying that the age decreases at high Cl (6% ??) which was associated with a change in crystallographic system from hexagonal to monoclinic. Not sure he ever published that though.

We are not sure what the reviewer wants here. We do not include real elemental data so this extra detail would not add significant information to our paper.

Not clear if the sensitivity of the data to composition as proposed is enhanced due to the long timescales, or this is a general feature.

This is a general feature of the data independent of timescale. We chose a deep time problem because it is harder to deal with than a Phanerozoic situation that may have more geological constraints available, and in general, deep-time problems are afflicted by greater uncertainty, where more data would be even more beneficial.

Also you switch between different combinations of AFT and AHe data, but the latter are fairly minor to the main proposition (and their sensitivity is based on the radiation damage model based on a FT annealing model). You could remove the AHe aspect totally and not change the message...expect perhaps in section 5.4, which comes pretty late and out of the blue.

We will try to rework the AHe part. It is true that the most important part of the paper concerns multikinetic FT data. However, we know that it can be difficult to reconcile AFT and AHe data sets for different reasons, especially for cratonic histories. We wanted to present one case where AFT composition may influence AHe modelling. Compositional data are not routinely used for AHe, but the commonly used radiation damage models are based on AFT annealing. If AFT annealing is truly affected by the composition then the radiation damage model should reflect the change in AFT annealing related to composition. We wanted to investigate this end member situation where all of the AHe age variation is controlled by radiation damage just to make a point that apatite composition may affect how alpha radiation damage accumulates. We will add a section upfront in the introduction to clearly outline why the AHe data are included and the reasoning behind the modeling. This should alleviate concerns that those data are a distraction, yet offer seeds for discussion for future work regarding if and how apatite composition affects He diffusion.

Lines 197 (and 213) There seems to be the assumption that AHe ages depend on composition, but is this just because the assumed model for radiation damage is one incorporating a fission track annealing model

The reviewer's statement is outside the scope of this paper but — yes, some published work has suggested composition may influence AHe ages directly but some work has suggested the opposite. No studies show a strong direct connection, but logic and the *ab initio* modeling suggest it should be so. The reviewer's point is a good one in that an additional issue could be secondhand via the kinetic model calibration. But annealing must involve diffusion, right? So the diffusion that heals tracks depends on small substitutions? The jury is still out. In our example, composition indirectly influences AHe ages because the radiation damage model is tied to AFT annealing, which can vary strongly as a function of composition.

Line 214 - QTQt will generate thermal histories regardless of ..data.... This is true of any sampling based modelling approach. The important point is how the generated thermal histories are accepted or not - QTQt effectively uses the ratio of data fit (likelihood) between a current and proposed model, while HeFTy, another piece of software for modelling thermal histories from thermochronological data, uses an absolute approach (p-value as a measure of data fit) for each thermal history. Perhaps accept is a better word than generate (also on 218), but keep the statement about the user needing to assess the output, particularly how well any particular model predicts the observed data.

Okay we will change generate to accept.

Line 249 - 3% seems small for AHe age data ??

We are assuming good quality, homogeneous AHe grains of suitable grain size and 3% is a normal analytical uncertainty without factoring in the Ft correction. We are trying to reproduce an observed date produced from a forward model, we are not modeling real data where data uncertainty may be cause for concern, when we in fact are trying to recover an unknown history. Ultimately the uncertainty should not be a major concern regardless of the value applied — especially in QTQt where better data quality is rewarded.

Line 256 - what happens if you do not constraint the heating/cooling rates ?

It is appropriate to determine suitable boundary conditions that are commensurate with the scale of the problem. The 5°C/Myr rate limit is more than ample for a timescale of 2 Ga. Given the unimodal and relatively broad length distributions and the old AFT ages, it is not possible to resolve extreme heating rates over this timescale. The main effect of not imposing a rate limit is to have longer model run times with the possibility of introducing rapid heating/cooling artifacts. This is especially deleterious for nondirected Monte Carlo schemes such as HeFTy and AFTINV. We will show results for the AFTINV model and discuss required boundary conditions that are needed to focus the model in promising areas of solution space. We chose QTQt for modelling this deep time problem because it has a learning algorithm that refines solution space and model boundary conditions as it evolves. For nondirected Monte Carlo schemes, the model may generate millions of trial solutions and only converge on a small number of acceptable solutions if boundary conditions are too broad. Even with perfect data, it may be hard for nonlearning models to find answers over such large timescales.

Line 257 - "t-T points were only *accepted* if they provided.."..the points are added, but the important step is whether they were accepted or not.

We will change “added” to “accepted.”

Line 262 - put the constraint definitions in section 3.1, and perhaps explain what they represent geologically

We can elaborate on the geological meaning of the constraint boxes. However, discussion of constraints does not belong in section 3.1 that deals with forward modelling (no constraint boxes are used here).

Line 264 - the ML model is potentially more complex than the MP model, but not always, and similarly the MP is not always simpler than the ML model...they can be the same.

We will reword to reflect this. The ML model is commonly more complex. The MP model is usually simpler.

Line 285 - not sure what you mean by a simple temperature weighting....the expected model is defined as

$$E(x) = \int xp(x)dx$$

and in this case the p(x) is the posterior. Given that the distribution of accepted models is the posterior then we just take the arithmetic mean of all accepted models to do that integral. The lower temperatures are because the distributions on temperature around the time of maximum temperature are often skewed (to lower temperatures) and so that leads to lower values for the expected (mean) maximum temperature. However, we often see that the duration of time close to the maximum temperature is greater for the expected model than say the ML model. This may tend to compensate a little in terms of fitting the data, but often not enough...you need to look at the predicted values relative to the observations. Note that we do not generally expect sharp V shaped thermal histories anyway (due to diffusion), but that is another issue.

This was a minor point and could be removed outright, but due to poor wording this should have more clearly stated that the averaging effect effectively pulls temperature downward. We will reword. The main point is the bias toward cooler temperatures.

Line 314 - QTQt does not use the central age directly (or even indirectly) as a model constraint. The predicted age for a given kinetic parameter and thermal history is used to infer the equivalent predicted ρs/ρi ratio which is then used in the likelihood function with the measured Ns and Ni values for each grain (see Gallagher 1995, EPSL).

We will reword this statement. The input is the central age. How the model uses that information is another issue. We are trying to avoid going into too many fine details of the QTQt modeling because this information has been published elsewhere.

Section 4.2 -line 300 perhaps just discuss using the correct kinetics (here you choose rmr0, but this could be another parameter, even if less sensitive ?)

The point is that the kinetics are known in advance for this synthetic data set. The Ketcham et al. (1999, 2007) annealing model uses r_{mr0} values. If users specify Cl content or Dpar, they are converted to r_{mr0} values using empirical correlations for the purpose of the kinetic model. We chose r_{mr0} based on our experience that elemental-derived r_{mr0} values provide better kinetic population resolution than Dpar or Cl content alone. We cannot show that in this paper because we do not include real data for a natural sample.

Line 317 - when using the average eCl 0.213±0.373 in QTQt, this implies that you let the kinetic parameter vary as part of the inevrsion...what was the distribution of the accepted values ? If skewed to a higher or lower value, that may be informative concerning the sensitivity of the different subgroups of data to an average/common kinetic parameter.

We can show this information in the supplement if necessary. Yes, the kinetic parameter was allowed to vary during the inversion, however regardless of skew in either direction, the point is that you could assign any kinetic value here within reason, *but more importantly***,** if a multikinetic sample is misinterpreted as a single population and the true history is complex (like we see in our example), then the fit to the observed data may be perfect, but one would never recover a history that closely resembles the true history. Therefore, as the reviewer mentioned previously the model assumption would be incorrect, but in this case both the t–T model AND the assumed single population are both incorrect. There seems to be a misperception in the reviewer's statement because the sensitivity of the different subgroups are lost entirely when combined into a single, monokinetic population — hence the point of this exercise. Combining the data yields some 'average' AFT age and some average kinetic value that removes t–T sensitivity. It is worth noting that ***not all AFT samples are multikinetic*** and we are not claiming this. Each natural sample is unique, and some may be multikinetic and others may not be, but if compositional data are not collected in the first place, how can this be known or addressed?

Line 318 - perhaps some examples of overdispersed data treated as a single population. The impact of this is likely to depend on how overdispersed and why...failing the chi-sq pvale = 5% test is not necessarily the definitive indication (e.g. we can pass at a level of 5.0001, but fail at 4.9999). Leaving aside analytical problems, dispersion that is a real if sometimes unwelcome signal could be due to compositional effects and discrete provenance related age populations (for which compositional ranges may be similar).

We will reduce some of this section. Based on the reviews, we think part of this discussion is best left to the next paper that includes real data for natural samples. Without including real data, we cannot show why we think unresolved, mixed AFT populations may be a larger issue than recognized. However, we do provide an example of overdispersed data being treated as a single population in our model examples. See our comment above regarding the X^2 test.

Line 320 - we do not necessarily have to formally identify discrete groups with mixure modelling, but just divide compositional range into subgroups and use the appropriate values (e.g. as Geotrack seem to do for Cl binned at 0.1wt % intervals)

Yes, but how many people do this other than Geotrack? That isn't clear in the literature overall and it isn't very clear at all how Geotrack carries out thermal history analysis. We think discrete models work better than a more continuous model for the natural multikinetic samples we have observed. However, we cannot demonstrate this here with a synthetic example, so it is beyond the scope of this paper. Our synthetic example is based on features we have seen in natural samples. Dealing with the distribution and allocation of track lengths in either a discrete or continuous approach is the most important factor to consider, but is outside the scope here.

Line 330 - it is not that QTQt failed to reproduce the true AHe dates...it is because the wrong choice of model prevented QTQt from doing so....

We will reword to explain that the true AHe dates were unable to be reproduced due to incorrect data treatment, i.e., incorrect model choice (see 'line 317' comment above as well)

Line 359 - I think that Geotrack do use compositionally discrete modelling for their routine AFT studies, but we rarely get to see the predictions for their preferred models.

That may well be but unless it is clearly documented in the literature then we are not privy to the details and can only speculate on how things were done. We still think it is fair to say this is underutilized in the scientific literature.

Line 388 - data quality is important too...

Sure, we did not list everything. Data quality is a factor but it is a function of many things. You need enough measurements to define populations, you need representative measurements for each grain and this could be affected by zoning, stressed grains with dislocations, problems with analytical procedures, etc.

Line 391 - what do you mean by more extensive ?

Greater time-temperature ranges. We can state this more clearly.

Line 396 - as I said above, perhaps we should model each grain with a specific set of kinetics ? This does no necessarily require running N annealing models for N grains, but perhaps 4-5 and we can interpolate the results (e.g. predicted ages and length distributions) for intermediate compositions.

We think the discrete model is a better way to go. You have well defined discrete populations. Why pick something in the middle and subdivide these well-defined populations? What is to be gained by further subdivision? If coherent age and length populations appear on plots of AFT parameters versus kinetic parameter, why resolve populations to a finer scale than necessary? In any case, this is a synthetic sample with predetermined properties. We also stress that arbitrary division into populations may be problematic depending on how this is carried out. With real samples we use radial plot mixture modeling to first identify discrete populations and in nearly all cases these populations coincidentally or naturally align with breaks or divisions within the AFT age data with respect to apatite composition. This agreement suggests to us that there is real kinetic/differential annealing giving rise to variability in single-grain ages and that those ages in turn correlate with changes in composition. After all, if single-grain dates correlate with composition and the t–T path of the sample caused differential annealing, we should expect a X^2 test failure and mixture models should reveal discrete subgroups.

Line 401 - without necessarily requiring

We can add this qualification.

Line 407 - what was the first hand ? I had forgotten by the time I got here.

We will tighten up the text so you remember.

Line 413-415...not sure I understand this...explicit condition on the model prior ? Perhaps you mean on the model sampling.

Yes, will reword.

Line 415 - not correct...it is the posterior that will be lower I think, rather than the likelihood. At least the acceptance of models is based on the posterior

We will rephrase this to be correct in our Bayesian language

Line 418 assist - use focus perhaps.

We will change this.

Line 421 - proper data interpretation is vague...you mean some of sub-population classification based on composition, age dispersion ?

We will replace proper data interpretation with kinetic population interpretation.

The discussion of precision v accuracy based on the credible intervals does not really take account of the form of the individual thermal histories...although they may define an envelope that seems consistent with the known true answer, the actual thermal histories may not really capture the true thermal history as well.

Yes, we know that QTQt calculations differ from models such as HeFTy and AFTINV where the plotted individual solutions are required to fit the data at a specified level of significance. Depending on how plots look we may add all individual paths to show how and where the Expected ± 95% envelope falls.

Line 429 - again, it is difficult to assess if we really resolve the true thermal history from the credible intervals..they do not tell us about the variation of individual thermal histories...so the earlier reheating event may start at different times, but we do not necessarily resolve the period sitting at low temperatures for a long time.

Yes, we know the issue with the individual solutions. The important point is the main features of the forward model thermal history are recovered by the Maximum Likelihood and Expected models. See point above.

Line 434 - more complex thermal histories were accepted or retained rather than added.

We will change added to accepted.

The rest of this paragraph to line 439...when you allow more complex models that do not improve the data fit to be accepted, you start to sample the prior more extensively, i.e. you tend to fill up the prior box in those parts of the time-temperature space where the data do not care what happens. For example, prior to a reheating event, the temperature can be pretty much anywhere from the reheating maximum value to the lower limit of the temperature prior. This does not mean we resolve the thermal history there. Almost the opposite. All we resolve is that the temperature has to be lower than the subsequent maximum.

Makes sense. The model lacks sensitivity at low temperature and it is allowed to fill up the space. We will modify the text.

Line 443 - Apfu Cl ?..if you add normally distributed noise, or even uniformly distributed, with the same mean value as the true model, it is not a great surprise that the expected model and credible intervals do not change much as we will tend to fit the average values of the data (which will reflect average values of the kinetics, which will be sampled on average to give the correct value...if that makes sense).

We will add eCl after apfu in the brackets. Kinetic parameters are relative. If the mean is similar to the "true" kinetic value used in the forward model, the inverse model will converge to be close to the input values. If mean values are not close to the input value, the model may converge on a different value but it will need to offset the remaining kinetic parameter to satisfy the observations. The result is that model temperatures can shift upward or downward but still converge to the same relative temperature history. Depending on how input ranges are specified, final kinetic parameters can be anywhere within the input range. The model will have trouble converging on good solutions if kinetic parameter ranges are too narrow to accommodate the required adjustment to kinetic parameters. None of this changes the conclusion of this paper that multikinetic data may contain a record of multiple heating events.

Line 448 - this acknowledgement of imperfect kinetic models should be stated much earlier, and ideally quantified somehow..(i.e. the recalibration exercise I mentioned above).

Recalibration is not important for this paper because we are using synthetic data. We assume the kinetic parameters are known and we investigate whether we can recover important thermal history information from a multikinetic sample using inverse modelling. This drives home the point that our current kinetic models do have some validity. All AFT modelling suffers from the same uncertainty on kinetic parameters. The most important aspect of a multikinetic model is the relative behaviour of the system. Calibration affects the absolute model temperatures. Relative kinetic behaviour preserves important details of the thermal history and this is important distinction will be discussed in the next paper.

Line 456-58 - the data quality (and the associated information on the thermal history) is first order in terms of what we can resolve. If the data are poor, scarce, we want more uncertainty.

Yes, data quality is critical. Uncertainty increases as data quality degrades. There are many factors influencing model results. These should be explored in future papers after as we learn more about natural multikinetic samples.

Section 5.3 - this deals with a separate issue to the main thrust of the paper, and adds to the feeling of stream of consciousness. I think it is not just with QTQt that we need to assess the effect of constraints on the final results and this has been said before in the exchanges in Earth Science Reviews over the last few years, and still ongoing. Any model result is conditional on the assumptions made to obtain the result. Thus a no constraint model can often be relatively boring (i.e. linearish cooling) but if the data can be adequately explained that way, then it is useful to know and can be considered as an end member model. Adding constraints is not a problem, but these need to be justified and this needs to be clearly stated in any study, preferably with some assessment of the confidence in a constraint. Additionally, when a thermal history is composed of linear segments joining up constraint boxes and we fit the data, this does not mean the constraints are justified...just that the data do not contradict these constraints and they do not require more complexity than imposed by the constraints.

A combination of inverse models and targeted forward models, often based on the inversion results, can be useful to deal with many of the problems discussed in this section and you do say that ..but you could just say that more concisely.

Yes, we will streamline this. There are many different situations and it is hard to generalize based on modelling one synthetic data set. There are different modelling strategies. Nondirected Monte Carlo models such as HeFTy and AFTINV usually require constraints in order to restrict model search space or they may have great difficulty converging. We just wanted to point out that constraint boxes might not be helpful under certain conditions. If multikinetic behaviour is unrecognized due to a lack of compositional data, for example, then thermal histories may be distorted to fit the data (see above). Use of constraint boxes will not lead to a better solution under these conditions.

Lines 474-476...not clear what this means...especially linearizing bias and Bayesian treatment of user constraints ..?

We will remove this text as it is not entirely necessary and as we can see from the comment that it can be confusing or misleading. What we were attempting to point out is that if ‘more complex’ models are prevented from being accepted (regardless of likelihood) then if there are two boxes placed within QTQt, the tendency will be to connect the two boxes with a simple ‘linear’ history segment, **IF** the data do not require a more complex path to provide a better fit. This does not mean that linear segment is legitimate or ‘real’ — which is where the point regarding a basement nonconformity being subaerial 100s of Myr before the geologic evidence (and a small constraint box) would suggest. This is likely an artifact of thermal history construction. Without boxes, the model fills the low sensitivity, low temperature space. Surface temperature can occur anywhere in the interval. When all solutions are forced to be at surface early on then temperatures ramp up because of the way temperature histories are constructed. The input history is very difficult to fit at low temperature because it is hard to have an effective heating rate of zero persisting over such a long time interval unless you force the model to do so. You can fit the data without this requirement.

Line 494..I agree with the sentiment...heralded perhaps better stated as imposed under the guise of geological evidence. This goes back to what I said above - there is not a problem using constraints/forcing a model, but the results are conditional on these and there are generally other models that can fit the data without, or with different constraints. Again, it is how valid the constraints are that is the big question.

Yes, we agree constraints have their place. We will rework the section to make things clearer.

Line 504...again Geotrack seem to use 0.1 Cl wt% bins when modelling, but we never see their predictions. You may add here that limitations arise when modelling samples independently, especially in boreholes. It is clear that jointly modelling multiple samples (plus the mutli-kinetic approach for each sample) is better -

data noise will tend to cancel out (if it is random) while real signal (the thermal history) should be reinforced. See Gallagher et al EPSL 2005 for an example with synthetic data too...

We expect models that use 0.1 wt% Cl bins will give different results than those that model discrete multikinetic populations. In the former case, kinetic parameters are expected to vary smoothly. We do not see this in most of our multikinetic samples. Instead, there are significant differences in kinetic behaviour between discrete populations. We also do not think a single parameter like wt% Cl will properly account multikinetic behaviour. This topic cannot be discussed in the current paper. It needs to be mentioned in the context of naturally occurring multikinetic populations that is the subject of a future paper.

Liner 509-510...what if the true thermal history is simple cooling ?

If that is the situation then inverse multikinetic modelling should yield linear cooling histories. We have unpublished natural examples where simple cooling can explain observed multikinetic data. In other cases, complicated histories are required (consistent with the geological complexity of the regions from which the samples were obtained).

Line 513- not sure what you mean by universal slow cooling suppositions...there is nothing forcing cooling in QTQt, apart from perhaps have a sample at surface temperature today and letting it start hotter...if the data are happy with that, then why not ? If the data need more complexity the thermal history should adapt.

This may be true for QTQt. However, other models such as HeFTy and AFTINV can be used to enforce continuous cooling, or enforce anything for that matter. These models may fit at the threshold of acceptability (and therefore cannot be rejected as a possibility) but may not be able to fit the data closely. The question is, “should we prefer the simple model that barely passes and has trouble finding solutions or a somewhat more complicated model that fits the data closely and converges quickly?” Complexity often yields better fits to observed data. As sample quality increases, it usually becomes easier to answer this question. With enough data, the linear cooling model may not reach the acceptance threshold. In any case, treating multikinetic data as a single population may lead to a significantly different and more simplified thermal history. The comment about *“universal slow cooling…then why not?”* highlights the issue cropping up in the literature with the use of QTQt as a black box where people dump in lots of data and hit “go” and especially over long timescales, these data can often be adequately reproduced under linear cooling assumptions. This doesn’t mean linear cooling is valid. This is a difficult problem to address.

Much of section 5.4 is relatively speculative but more importantly deviates from the main message of the paper concerning FT annealing and left me a little confused. The last paragraph is OK...but I think you could drop much of this and perhaps save it for another paper....but that is up to the authors. While I agree that allow an effective parameter to vary in the modelling, as QTQt allows, can be useful - the paper by Ricanati et al. demonstrated that yes we can improve the fit to the data by having a wiggle factor for each grain diffusivity. As Samuel Karlin said, 'The purpose of models is not to fit the data but to sharpen the question'...so demonstrating we can fit the data with an additional factor is not really the solution, but suggests we should look for a physical control on that factor. Given the arguments here, I would say go and measure apatite chemistry to demonstrate that there is some correlation of age and chemistry or effective diffusivity then you are on to something (perhaps this will be possible with the method, Pickering et al. mentioned). For me, as it is, the sceptic will take this section as special pleading for a control on AHe date dispersion that is neither understood nor constrained. Those adopting the averaging strategy for AHe data will just say we do not understand enough to do anything more sophisticated and carry on. I would add a caveat to the Karlin quote too....not fitting the data means we need to ask different questions.

We agree that we need to scale this back and save it for the next paper where we can have more in depth discussions around multikinetic interpretation and modelling. We also agree that we are only showing one possible reason why AHe data may be difficult to model and that much more work is needed to resolve the problem. We will focus the paper more on the key points we want to convey. We would also add that if the thermochronology community does not adequately understand the problems surrounding AHe data and workers have to manipulate or massage dates through averaging, ‘culling outliers’, or ‘eU binning’ should papers be published utilizing these data or should more work be done to address these issues head on? This aligns with the reviewer’s sentiments or caveat to the Karlin quote.

Line 582...Steve Bergmann was often insistent on the importance of OH as a control on fission track data....not sure if he ever published anything though.

We think this is well established in the literature and that OH is an important factor and it should be included when obtaining detailed elemental data for constraining kinetic parameters. That is why electron microprobe analysis is preferred over LAICPMS elemental data. Carlson et al. (1999) and Ketcham et al. (1999) noted that it was important enough to be a separate term in their empirical r_{mr0} equation. They also noted that r_{mr0} correlated better with OH than Cl or Dpar. This topic is peripheral and better discussed elsewhere.

Line 594 - what recent publications ?

The section on AHe will either be removed entirely or greatly reduced and reworked so these points may no longer be relevant. Also the AHe data will be downplayed in terms of significance and discussion. The relevant citations being mentioned in the previous section being Gautheron et al. 2013 or Gerin et al. 2017, Recanati et al 2017 and Powell et al. 2017 and 2020 for looking at both AFT and AHe composition/kinetic variability together.

Line 595 - not misfit, but poor fits.

We will change this.

Line 596 - degrades in what sense ?

This is a sample specific result. It is hard to generalize. For our case, it results in a simpler thermal history or a change in the timing and magnitude of thermal events relative to the input history (if constraints are enforced).

It may be the paper may be better concluded by adding a series of recommendations on analytical practice/protocol and then modelling strategies.

We cannot say too much here because we are not providing a detailed description of how to interpret multikinetic data for natural samples. Much of this will be dealt with in a future paper.

Fig 1 - perhaps put the 2 constraint boxes on the forward thermal history. They are a little strange as constraints as we might expect the constraint to be the stratigraphic age of sediment deposited at the time of the start of the heating events, rather than some time prior to the heating event.

The constraint boxes do not belong on this figure. They were not used to generate the forward model. They are considered as something inferred from fragmentary geological evidence. We can elaborate on what they represent geologically.

Fig 3 - the credible interval ranges are fine as presented in this figure, but it may be useful to put the sampling of individual models and/or the marginal distribution (the coloured plot from QTQt for a given sample thermal history) in the supplementary - then we can see how many models actually start reheating at 1200 Ma for example.

We will add more QTQt output to the supplement to assess overall model behavior.