Simulating sedimentary burial cycles...

McDannell and Issler

This manuscript presents a series of modelling results to demonstrate that use of compositionally defined subpopulations of apatite fission track (AFT) data can provide more detailed information on reheating events in protracted (deep time, 1 Byr) thermal histories than would be intuitive based on the general understanding of annealing of fission tracks. I found the paper a bit long and wordy, even flowery at times - a bit stream of consciousness sometimes - and the wood is getting lost in the trees. Having said that, this review may suffer from similar meanderings.

The paper falls in the scope of GChron, presents new ideas and demonstrates the utility of detailed compositional data for modelling AFT data, on the assumption that the calibrated annealing models are OK.

Perhaps the authors could try to reduce and reorganise the text to help the reader. For example, 2.5 pages on rmr0 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). 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 ?). 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).

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 rmr0 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). 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 ? 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.

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.

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.

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.

Some general small points

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.

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.

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.

Specific comments

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

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.

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

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

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.

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

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.

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

$$E(x) = \int x\, p(x)\, dx$$

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.

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 ρ_s/ρ_i ratio which is then used in the likelihood function with the measured N_s and N_i values for each grain (see Gallagher 1995, EPSL).

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 ?)

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.

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).

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)

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....

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.

Line 388 - data quality is important too...

Line 391 - what do you mean by more extensive ?

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.

Line 401 - without necessarily requiring

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

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

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

Line 418 assist - use focus perhaps.

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

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.

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 indiviudal thermùal 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.

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

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.

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).

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).

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.

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.

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

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 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.

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...

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

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.

Much of section 54. 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.

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.

Line 594 - what recent publications ?

Line 595 - not misfit, but poor fits.

Line 596 - degrades in what sense ?

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

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.

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.