Comment on "Discussion: Extracting thermal history from low temperature thermochronology/A coment on the recent exchanges between Vermeesch and Tian and Gallagher and Ketcham", by Paul Green and Ian Duddy, Earth Science Reviews, https://doi.org/10.1016/j.earscirev.2020.103197

Kerry Gallagher

Journal Pre-proof

Comment on "Discussion: Extracting thermal history from low temperature thermochronology/A coment on the recent exchanges between Vermeesch and Tian and Gallagher and Ketcham", by Paul Green and Ian Duddy, Earth Science Reviews, https://doi.org/10.1016/j.earscirev.2020.103197

Kerry Gallagher

Please cite this article as: K. Gallagher, Comment on "Discussion: Extracting thermal history from low temperature thermochronology/A coment on the recent exchanges between Vermeesch and Tian and Gallagher and Ketcham", by Paul Green and Ian Duddy, Earth Science Reviews, https://doi.org/10.1016/j.earscirev.2020.103197, *Earth-Science Reviews* (2021), https://doi.org/10.1016/j.earscirev.2021.103549

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Comment on "Discussion : Extracting thermal history from low temperature thermochronology/ A coment on the recent exchanges between Vermeesch and Tian and Gallagher and Ketcham", by Paul Green and Ian Duddy, Earth Science Reviews, https://doi.org/10.1016/j.earscirev.2020.103197

 Kerry Gallagher,

Géosciences/OSUR, University of Rennes, Rennes, France

The discussion of Green and Duddy (which I will refer to as Geotrack, and use quote marks to indicate text from their discussion) aims to "highlight a key issue" concerning different approaches to thermal history modelling with thermochronological data. This key issue seems to be based around an assumption of continuous cooling in thermal histories modelling. While I and many others would agree with the statement of Geotrack that thermochronological data have limited thermal history information and the results are dependent on the nature of the assumed style of history, I would like to make one clear statement regarding the implementation of thermal history models in QTQt. This statement is : there is no assumption of continuous cooling history. Indeed, there is no default assumed style of thermal history. Perhaps then, the key issue highlighted by Geotrack represents a misunderstanding of how the QTQt software (and the algorithms implemented therein) work and how the output should be assessed. I welcome the chance to clarify these points further below.

As in most approaches, a thermal history in QTQt is parameterized in terms of discrete time-temperature points and linear interpolation between these points is used to construct a thermal history function. Note that this linear heating-cooling segment type of thermal history is not what we expect in reality. Given the nature of thermal diffusion, temperature histories should be relatively smooth, even when depth changes associated with burial or erosion are rapid and effectively linear. Parameterizations of the thermal history used with linear segments will then tend to over-estimate the maximum temperature and underestimate the duration spent at or near the maximum temperature.

Irrespective of this point, a specified thermal history function is used to make predictions which are then compared with observations through a quantitative measure known as the likelihood (or misfit). The algorithm in QTQt (reversible jump Markov

chain Monte Carlo, RJMCMC) used to sample candidate thermal histories is formulated in a Bayesian framework and has the characteristic of parsimony. This is the preference for simpler thermal history solutions over more complex ones, provided the former can provide adequate predictions relative to the observations (i.e. the model can fit the data). As shown by Gallagher (2012), taking the single thermal history which provides the best fit to the observed data but has no constraints on complexity can lead to a which is grossly over-complex, particularly when using data from just one sample. This was demonstrated using near perfect synthetic data produced from a known and simple thermal history in which spurious rapid heating-cooling events are inferred from the much simpler true thermal history (see figure 2 of Gallagher 2012).

When a simple thermal history is inferred from thermochronological data alone this implies a lack of information in the data. If a sample is on the surface at the present day, then the simplest possible model will be one of cooling from an unknown time-temperature point in the past. In QTQt the first sampled model is based on a single time-temperature point drawn from the prior (the range of time-temperature specified for sampling), although it is also possible to specify a different starting model if desired. If the present day temperature is at the low end of the prior temperature range (as we anticipate for a surface sample), there is a higher probability that the first sampled temperature point in the past will be hotter than the present day temperature, producing a cooling thermal history. The sampling algorithm progresses, adding new time-temperature points and moving, or deleting, existing ones, to produce new proposed thermal histories. Apart from the prior distribution, there are no default constraints on where the time-temperature points can be selected. If the data really need more complexity in the thermal history, the models will adapt and we will see more structure in the solutions, according to the information in the data. Given, that, there is certainly no requirement or format underlying QTQt to produce continuous cooling or episodic heating-cooling events (the latter implied by Geotrack at the start of their 2nd last paragraph).

The predictions from each proposed thermal history model are compared to the observations (through the likelihood) and a probability-based decision is made to accept the proposed model or not. If the data require a more complex thermal history (e.g. heating/cooling events, changes in cooling rate, etc.), then new time-temperature points will tend to be accepted. A basic characteristic of any MCMC sampling algorithm is that

the initial model is forgotten so the final results are independent of the starting model. Therefore, a simple starting model (e.g. linear cooling)

has no effect on the final result. The final result is of course a distribution of thermal histories, referred to as the posterior, and this distribution is more informative than just (one) thermal history (as in the title of the comment of Geotrack).

An initial modelling exercise with no or minimal constraints on the samples is useful. Minimal constraints, which can also incorporate uncertainty, could represent present day temperatures and stratigraphic ages for samples. The rationale for modelling with no constraints is to assess explicitly what thermal history information the data alone can provide. Here I directly contradict the statement of Geotrack that the "much of VTGK debate appears to centre on methods to extract information that is not within the data". With a single sample or a single AFT analysis, the thermal history information is often limited and I think there Geotrack and most of us would agree again. Given this, there is a clear motivation for modelling multiple samples together (joint modelling - see Gallagher et al. 2005, Stephenson et al. 2006). In doing that, valid thermal history information will tend to be reinforced, while data noise will tend to be cancelled out as it is random. This also reduces the tendency to over fit the data, in that we reduce the potential of unintentionally treating data noise as real signal. However, the relevant point here is that a thermal history inferred from the data alone, and that fits the data, can then be considered as an end member model. When using a sampling algorithm that prefers simpler models to complex ones, then if all else (e.g. the data fit) is equal, we can also think of this end member model as a lower limit on the complexity of the thermal history. Thus, the true thermal history may be more complex, but the data alone do not require not justify more complexity.

In the context of QTQt, we need to remember that we have many models to choose from (which define the posterior distribution). In chossing one model, we typically consider the maximum likelihood (ML) model, maximum posterior (MP) model or the expected model. It seems to be this latter model that Geotrack focus on as the assumption of continuous cooling (and I thank Pieter Vermeesch for pointing that out to me). The expected model is not actually a sampled model and there are no asumptions about its form. It is the weighted average of the population of accepted sampled models, with the weighting being the posterior probability. In this case, the data fit for the expected model can be variable depending on the features of individual models captured

in this average model. If a range of models, or different subgroups of models, are compatible with the data, then the expected model may effectively fall between these good models. This behaviour is captured in an old joke that a statistician can have her head in an oven and her feet in ice, but on average she feels comfortable.

In contrast to the expected model, the ML and MP models are both individual thermal histories that were directly sampled during the model run. The first is the sampled model that best fits the data, is often relatively complex and can have features that do not change how well we fit the data. The second is the sampled model that tries to balance fitting the data with keeping the model relatively simple. This model will then not fit the data as well as the first, but in general it will not be too different. Sometimes, these two models exhibit characteristics common to subgroups of the overall population of accepted models, demonstrating that there is a range, or even clusters, of possible thermal history solutions, with possibly quite different forms. This can only really be assessed by looking at the population of accepted models.

An example of the differences between the various models can be seen in figure 2 of Gallagher and Ketcham (2020), reproduced here as figure 1, based on synthetic data produced from a known thermal history with 2 heating/cooling events. In this case, the ML solution is close to the true solution, while the MP solution is simpler and notably does not imply any reheating, which would clearly require more complex models (i.e. more time-temperature points). The MP model does not imply the same maximum temperatures as the ML model at the time of reheating. However, the difference between the predictions for these two models is minor, especially given the noise anticipated in real data. Figure 1 also shows a selection of the accepted models colour coded by the likelihood and posterior values, demonstrating two different forms of models. While the models coded by the likelihood show the two reheating cooling events in the true model, the inferred rates of heating/cooling are quite variable. The models coded according to the posterior suggesting the earliest reheating event is not particularly evident in, or not really required by, the data. Some models do hint at limited resolution of the later event, but these are more complex models and their scarcity suggests this information, or the evidence of reheating, is not particularly strong. It is worth noting here that the credible intervals are informative about the resolution of the temperature at different times. In this example, the credible interval becomes relatively narrow around the time of the maximum temperature for the two heating events, but broader elsewhere,

demonstrating less resolution on the heating/cooling rates and the thermal history between those times.

Returning to the expected model in figure 1, in this case it does not predicted the data nearly as well as the ML or MP models. As stated above, it is the average of the accepted sampled models, and averaging will tend to act as a smoothing process. It also reflects the fact that the posterior distribution is often not symmetrical, but skewed to lower temperatures. Thus, in this example, the expected model does not capture the maximum temperatures of the heating events, while it also has low temperatures relative to the MP model between the heating events. As a result, it, over-predicts the AFT age relative to the ML and MP models.. For a skewed posterior distribution, the expected model also does not always follow the mode (or maximum value) of the posterior distribution at each time (brighter colours in figure 1a). These points actually define the maximum mode model (MM in figure 1) and we can see this is similar to the MP model in this example. However, this is not always the case, for example if the posterior distribution is multi-modal.

So sometimes the expected model will not be able to make suitable predictions as it may not appropriately capture the particular features or temperatures in the sampled thermal histories that lead to better data fitting. In that case, this particular model should not be presented as representative of the good data fitting solutions. Then we might prefer to present the MP and/or ML models, although the discussion above demonstrates that each should be considered in the context of the population as a whole. We also need to consider also how these models fit or explain the observations when considering model choice. Referring back to the original article of Vermeesch and Tian (2014) an example of poor data fitting related to poor model choice is the linear regression example (their figure 6). As stated in the first comment of Gallagher and Ketcham (2017), the best fit straight line model presented just does not fit the data and the approach implemented could not adapt to the information in the data produced from a more complex (quadratic) function. The key point here is that we need to assess the quality of the data fit as part of the modelling-interpretation process, as I have gently suggested previously (Gallagher 2016).

If we want to narrow the range of possible models, then adding constraints clearly becomes an important aspect of thermal history modelling. Any inferred thermal history model will be conditional on the constraints and the issue then is how robust any

imposed constraint is. Constraints as I define them here may be parametric (e.g. rates of temperature change over time, number of time-temperature points defining the thermal history) or more geological in nature. In practice, if we understand the geological history of a region and have total confidence in a geological constraint, based on independent geological evidence, this should be specified in advance of any thermal history modelling. An example of a such a constraint was presented by Bernard et al (2016) when modelling a vertical profile from a fjord in East Greenland. The uppermost sample was collected a few tens of metres under a well dated basalt flow. It was reasonable to impose a constraint that that sample was near the surface at the time the basalt flow formed. When an imposed constraint is less sure, the scientific method would suggest we assess how data predictions change with and without such a constraint. Indeed, Bernard et al. (2016) did that also and presented results demonstrating that a post-basalt heating event can be partially recovered for some samples without the near surface constraint.

Another way of using geological information is to consider it after modelling, justifying the relevance or significance a posteriori. For example, we can force heating/cooling events in the model set up, and then argue that we can find geologically reasonable reasons to justify the resulting thermal history model. Again this may be that there is sediment or volcanic of a certain age found nearby, and then we assume that the sample was at or close to the surface at that time. In general, these arguments will have uncertainty and subjectivity and should presented in that context. For example, what is nearby, how well defined is the age of preserved sediment, can we be sure that extrapolating restricted outcrop over several tens km or more is valid ? The recent replies of Braun (2019) to the comment of Japsen et al (2019) and Jess et al. (2020) to the comment of Green et al. (2019) highlight some issues concerning geological constraints used in thermochronology. Indeed, we could rephrase the statement of Japsen et al. (2019) quoted in Braun (2019) - "...models can prove anything with the appropriate choice of parameters" to *model results can be forced to prove anything with the appropriate choice of constraints*,. The point here is that constraints are not necessarily facts bit results are strongly conditional on the constraints and their validity. Consequently, resulting models do not provide apodictic evidence in support of a given hypothesis.

In their discussion, Geotrack refer more than once to episodic heating-cooling events, e.g. "episodic heating cooling events may be much more common that generally accepted", and that even without geological constraints, "thermal history constraints can still be obtained within a framework of episodic heating and cooling, using assumed heating and cooling rates". The latter is clearly true, but the former statement seems to be circular reasoning in that if we impose heating-cooling events, then that is what we will get. When geological constraints on the surface are imposed the effect is the same as we are effectively forcing cooling to the surface, then typically reheating. If modelling data does not produce heating-cooling events without imposing them and/or the data can be explained without such events, it is difficult to agree that "a signal is actually present in the data". Rather, we can say that the data are consistent with these events, but do not necessarily require them.

I do not know if Geotrack have ever used QTQt but I can state that I have never had access to the software used by Geotrack. This software is not obviously available (nor indeed the annealing model implemented in it) for independent assessment nor have the details been published. Therefore, based on the Geotrack discussion, their publications and available commercial reports, I can only speculate on how their modelling is implemented. A thermal history seems to be based on, or parameterized, in terms of heating/cooling events, defined within a specified rate of temperature change. Geotrack state that "specific constraints can only be obtained by fixing heating and cooling rates (essential for comparing results between samples)". Here constraints implies inference on the nature of the thermal history, although clearly such inference depends on the imposed constraints (limits on the heating/cooling rates for example). Presumably, some kind of starting thermal history is specified (perhaps continuous cooling), and, perhaps depending on how this initial model can explain the data, a first heating cooling event is searched for. The likelihood for that thermal history is calculated, then an additional, more recent, event is added, presumably under the constraint that the maximum temperature during heating is lower than that of the first event. More events are added until the fit to the observed data is deemed to not improve sufficiently to justify the last event. The thermal history that best fits the data (the maximum likelihood model) is then chosen and a form of resolution analysis is made. This step seems to be a form of the likelihood ratio test to estimate an uncertainty about

given time-temperature point and the results are presented as range in time-and temperature about each point.

In general, the publications of Green, Duddy and their various collaborators do not seem to present the predictions from their preferred model, nor the predictions from alternative models (e.g. with less heating-cooling events). It is therefore not possible to assess the variations in the data fit for different models and make appropriately informed comparisons to thermal histories inferred from other approaches. Given the general availability of modelling software such as QTQt and HeFTy, perhaps this comment will encourage Geotrack to provide access to their software to allow the research community to make their own judgments concerning the validity and generality of the results it produces.

References

Bernard, T., Steer, P. Gallagher ,K., Szulc, A., Whitham, A., and Johnson, C., (2016) Evidence for Eocene–Oligocene glaciation in the landscape of the East Greenland margin, Geology, 44, 895-898**.**

Braun. J. (2019) Response to comment by Japsen et al. on "A review of numerical modeling studies of passive margin escarpments leading to a new analytical expression for the rate of escarpment migration velocity" Gondwana Research 65, 174–176

Gallagher, K., (2012) Transdimensional inverse thermal history modelling for quantitative thermochronology, J. Geophys Res 117, B02408, doi:10.1029/2011JB00882.

Gallagher, K. (2016) Comment on 'A reporting protocol for thermochronologic modeling illustrated with data from the Grand Canyon' by Flowers, Farley and Ketcham, Earth Plant. Sci. Letts. 441, 211-212.

Gallagher, K. and Ketcham, R.A. (2018) Comment on "Thermal history modelling: HeFTy vs. QTQt" by Vermeesch and Tian, Earth-Science Reviews (2014), 139, 279–290, Earth Science Reviews, 176, 387-394

Gallagher, K. and Ketcham, R.A. (2020) Comment on the reply to the Comment on "Thermal history modelling: HeFTy vs. QTQt" by Vermeesch and Tian", Earth Science Reviews, 203, 10.1016/j.earscirev.2019.102878

Green, P.F., Japsen, P., Bonow, J.M., Chalmers, J.A., and Duddy, I.R., (2019). Thermal History Solutions From Thermochronology Must be Governed by Geological Relationships: A Comment on Jess et al. (2019). Geomorphology, p. 106848.

Japsen,P., Green, P.F., Chalmers, J.A., Duddy, I.R. and Bonow, J.M. (2019) Elevated passive continental margins: Numerical modeling vs observations. A comment on Braun (2018) Gondwana Research, 172-173

Jess, S., Stephenson, R., Roberts, D.H. and Brown, R. (2020) Reply to: Thermal history solutions from thermochronology must be governed by geological relationships: A comment on Jess et al. (2019), Geomorphology 360 (2020) 106971

Vermeesch, P., and Tian, Y., 2014. Thermal history modelling: HeFTy vs. QTQt. Earth Sci. Rev. 139, 279–290 2014.

Figure Captions

Figure 1

This figure is taken from Gallagher and Ketcham (2020), which has a more detailed figure caption.

(a) Output from QTQt showing the maximum likelihood (ML), maximum posterior (MP), maximum mode (MM) and expected (Exp) models, together with the marginal posterior distribution and the 95% credible interval (the region between thinner black lines). Note how the credible intervals narrow around the times of peak heating. . The values of the observed fission track age and mean track length and their predictions for each of the 4 individual models. SP represents sampled predicted, which summarises the mean and standard deviation of the predicted ages and mean track lengths for all accepted modelsLL is the log likelihood for each model .

(b) The predicted track length distributions for the 4 models highlighted in fig. 1a.

(c) Representative thermal histories from the posterior distribution, colour coded by the relative likelihood, calibrated against the maximum likelihood with a value of 100%. The black line is the expected model with the 95% credible interval. The higher likelihood heating-cooling thermal histories are relatively complex and the timings of reheating events are dispersed, as are the pre-reheating temperatures. Simpler models similar to the maximum posterior model in 1a, also have relatively high likelihoods. Most of the accepted models fall into one of these two groups, effectively defining two local maxima in the likelihood

(d) Representative thermal histories from the posterior distribution, colour coded by the relative posterior, calibrated against the maximum posterior probability with a value of 100%. In comparison with figure 1c, we see the simpler models define the higher posterior probability, while the complex models have moderate to low posterior probability. This reflects limited information in the data on the form of the true thermal history.

There is no conflict of interest