Reply to the Comment on the Reply to the Comment on Vermeesch and Tian (2014)

Pieter Vermeesch[1] and Yuntao Tian[2]
[1] University College London and [2] Sun Yat-Sen University Guangzhou

May 31, 2019

A useful way to test any model is to push it to its extremes. In statistics, it is desirable for models to asymptotically converge to the true solution in the limit of infinite sample size or zero noise. We have found that testing models with extreme examples is equally useful in geology. For example, Vermeesch (2012) used this approach to demonstrate that probability density plots break down when applied to large and/or high precision datasets. And Vermeesch (2018) used unrealistically large datasets to demonstrate that the youngest age peak is a poor estimator of the maximum depositional age because it drifts to younger ages with increasing sample size. Continuing in the same vein, Vermeesch and Tian (2014, 2018, hereafter referred to as VT1 and VT2) used extreme examples to highlight the fundamental differences between `HeFTy` and `QTQt`. The main thrust of the Comments by Gallagher and Ketcham (2017, 2019, hereafter referred to as GK1 and GK2) is that the case studies used by VT1 and VT2 are unrealistic. But this is exactly what they were meant to be.

VT1's main objective was to explain the algorithmic underpinnings of `HeFTy` and `QTQt` to non-expert users. We think that it is important that thermochronologists are aware of the great differences between these two software packages. It was our aim to be fair and balanced in our review, so we discussed both the strengths and the weaknesses of the two algorithms. Because of their complementary design philosophies, any strength of one program can inevitably be perceived as a weakness of the other. So it was also inevitable that our paper would meet with some resistance from the creators of `HeFTy` and `QTQt`, who invested a tremendous amount of time in their programs. This is why we invited both of them to review our paper (see VT2 for details).

The two most contentious conclusions of VT1 are that (1) `HeFTy` tends to 'break' when it is supplied with large or high precision datasets, whereas (2) `QTQt` always manages to find a solution even for nonsensical datasets. GK1 and GK2 do not deny the validity of these two points, but dispute their importance: they (1) dismiss `HeFTy`'s sample size limitations as a theoretical problem that does not affect 'real' datasets; and (2) propose that poor `QTQt` model fits can be identified by inspecting the residuals.

In their response to GK1, VT2 reiterated and reinforced the main points made by VT1, which included the importance of residuals. VT2 also elaborated on some limitations of thermal history modelling that had only been briefly discussed by VT1 for the sake of brevity. More specifically, they emphasised the nonuniqueness of time-temperature (t-T) histories, which undermines the way by which `QTQt` combines multiple 'trans-dimensional' t-T paths to form a single colour-coded graphic. GK2 felt that several of their numerous detailed comments were not adequately addressed. We have therefore provided a 9-page, line-by-line rebuttal of GK2 in the Supplementary Information. In the main body of this Reply, we will stick with the big picture and summarise the main points in this rebuttal.

GK2 label `HeFTy` as a 'non-learning Monte Carlo algorithm'. This perfectly summarises the issue at

hand. Because HeFTy does not learn, the user has to 'hold it by the hand' to find the solution space. But because the program uses p-values as a goodness-of-fit parameter, it may not find any acceptable solution at all. This is more than just a theoretical problem, in spite of claims to the opposite by GK1 and GK2.

For example, in a recent thermochronological study of Grand Canyon incision, Winn et al. (2017) "*were unable to find time-temperature paths that predict [their] observed AHe ages within error*". In order to get HeFTy to accept their data, these authors "*increased the measured uncertainty proportionally until [they] were able to find time temperature paths that could explain the data, which is equivalent to lowering the p-value and accepting more paths*". Note that this workaround was previously suggested by VT1. The Winn et al. (2017) example refutes GK2's claim that HeFTy's inability to handle large and/or precise dataset is only a theoretical possibility with no real world implications. HeFTy's sensitivity to sample size also hampers its ability to simultaneously model multiple samples. Doing so is possible in QTQt but not in HeFTy.

QTQt does not use p-values and therefore has no problems finding acceptable solutions to large and/or high precision datasets. And because QTQt is a 'learning algorithm', it converges to the solution space without much user intervention. The program therefore solves two of HeFTy's problems. However, this solution comes at a cost. It is left to the user's discretion to decide whether QTQt's model predictions are a 'good' or 'bad' fit to the data. In contrast, HeFTy makes this decision on the user's behalf. For the sake of neutrality, VT1 did not express an opinion as to whether QTQT's subjectivity is a price worth paying. But after two rounds of Comments and Replies it should be clear that we agree that it is indeed a price worth paying.

Another point of contention is the issue of model resolution. VT2 introduced a synthetic dataset that was inspired by Green and Duddy (2012) to illustrate the non-uniqueness of thermal history inversions based on fission track data. This example showed that QTQt successfully recovers both the true 'sawtooth-like' history and a simpler 'hockey stick' history. Because the simple t-T path was sampled more frequently than the complex one, the 'average history' shown in QTQt's graphical output resembles the simple history. GK2 discuss our synthetic example in great detail. They show that fission track data lack the resolution to differentiate between the sawtooth and the hockey stick scenarios. GK2 show that, in such situations, reversible jump Monte Carlo methods such as QTQt will always prefer the simplest model. However it is important to note that the simplest model is not necessarily the correct model. This was the main point of VT2's synthetic example and GK2 just elaborate on this point.

The problems of non-uniqueness and model resolution are well known in the context of seismic tomography (Tarantola, 2005). In tomography it is common practice to quantify these limitations with checkerboard tests and resolution matrices. This is not the case in thermochronology although recent contributions are moving in this direction (e.g. Fox et al., 2014). We are also happy to note that version 5.6.0 of QTQt addresses the concerns raised by VT1 and VT2 about its raster visualisation by offering the possibility to plot all acceptable solutions and colour code them by likelihood.

The non-uniqueness of thermal history inversions diminishes their scientific value. So in their final paragraph, VT2 argue against the need to estimate continuous t-T histories. Here we would like to advocate for a different type of study design, in which thermochronological data are used to answer specific geological questions or to test specific geologic hypotheses, rather than to recover an entire t-T history. For example, one could constrain the exhumation rate under the explicit assumption of linear cooling. Or one might estimate the maximum temperature reached during a reheating event that occurred at a pre-specified time. Or as yet another example, the proprietary inverse modelling software developed by GeoTrack® International constrains the timing and temperature of a limited number of thermal 'events'. This software produces sawtooth-like thermal histories that more faithfully display the resolution of the solutions than QTQt's average histories do (see examples in Green and Duddy, 2012). Whether episodic thermal histories are geologically sensible is, of course, an entirely different issue.

VT1's original goal was to review `HeFTy` and `QTQt`. We did not anticipate the spirited debate (Kohn and Gleadow, 2019) that followed. But we are happy with this outcome because it resulted in a fundamental discussion about the limitations of thermochronology. Jointly considering VT1, GK1, VT2, and GK2 together with this Reply should give the reader a fair and balanced overview of this important issue. We hope that the thermochronology community will benefit from the exchange.

References

Fox, M., Herman, F., Willett, S., and May, D. A linear inversion method to infer exhumation rates in space and time from thermochronometric data. *Earth Surface Dynamics*, 2(1):47, 2014.

Gallagher, K. and Ketcham, R. Comment on "Thermal history modelling: `HeFTy` vs. `QTQt`" by Vermeesch and Tian. *Earth-Science Reviews*, 176:387–394, 2017.

Gallagher, K. and Ketcham, R. Comment on the Reply to the Comment on "Thermal history modelling: HeFTy vs. QTQt" by Gallagher and Ketcham. *Earth-Science Reviews*, [this issue], 2019.

Green, P. F. and Duddy, I. Thermal history reconstruction in sedimentary basins using apatite fission-track analysis and related techniques. *Analyzing the thermal history of sedimentary basins: Methods and case studies: SEPM Special Publication*, 103:65–104, 2012.

Kohn, B. and Gleadow, A. Application of low-temperature thermochronology to craton evolution. In *Fission-Track Thermochronology and its Application to Geology*, pages 373–393. Springer, 2019.

Tarantola, A. *Inverse problem theory and methods for model parameter estimation*, volume 89. SIAM, 2005.

Vermeesch, P. On the visualisation of detrital age distributions. *Chemical Geology*, 312-313:190–194, 2012. doi: 10.1016/j.chemgeo.2012.04.021.

Vermeesch, P. Statistics for fission tracks. In Malusá, M. and Fitzgerald, P., editors, *Fission track thermochronology and its application to geology*. Springer, 2018.

Vermeesch, P. and Tian, Y. Thermal history modelling: HeFTy vs. QTQt. *Earth-Science Reviews*, 139: 279–290, 2014.

Vermeesch, P. and Tian, Y. Reply to comment on "Thermal history modelling: HeFTy vs. QTQt" by Gallagher and Ketcham. *Earth-Science Reviews*, 176:395–396, 2018.

Winn, C., Karlstrom, K. E., Shuster, D. L., Kelley, S., and Fox, M. 6 Ma age of carving Westernmost Grand Canyon: Reconciling geologic data with combined AFT,(U–Th)/He, and ^{4}He/^{3}He thermochronologic data. *Earth and Planetary Science Letters*, 474:257–271, 2017.