Earth-Science Reviews 176 (2018) 387–394

Contents lists available at ScienceDirect

Earth-Science Reviews

journal homepage: www.elsevier.com/locate/earscirev

Discussion

Comment on “Thermal history modelling: HeFTy vs. QTQt” by Vermeesch and Tian, Earth-Science Reviews (2014), 139, 279–290

Kerry Gallagher[a,*], Richard A. Ketcham[b]

[a] Géosciences Rennes/OSUR, Université de Rennes 1, France
[b] Jackson School of Geosciences, University of Texas at Austin, USA

1. Introduction

The paper, Vermeesch and Tian (2014), henceforth referred to as VT, compares two programs (one developed by each of the current authors) available for inverse modelling of thermochronological data. VT states two goals. “First, it provides a ‘glimpse under the bonnet’ of these two ‘black boxes’ and second, it presents an objective and independent comparison of both programs”. We were both invited to review the original manuscript of this paper, but declined, in large part due to other commitments. Additionally, neither of us uses the other's software, and we believe we would not have been able to give an adequately informed opinion on both. We suggest overall that it should be more productive to have independent reviewers, ideally users familiar with both pieces of software and the principles underlying each. However, as published, VT does not consider fundamental aspects of both modelling approaches, and their comparison highlights both how not to use the software and how not to approach the inverse modelling problem. Unsupported anecdotal statements are widespread and many are unclear, selective, misleading or just wrong. Consequently, the purpose of this comment is to try to clarify, qualify, or correct some of the more important of these statements.

We start with two minor examples. The first is on the 2nd line of the abstract, “QTQt is an alternative program whose name refers to its ability to extract visually appealing (‘cute’) time–temperature paths from complex thermochronological datasets”. This is not true. The acronym refers to Quantitative Thermochronology with a user interface developed in the Qt programming platform (the latter is acknowledged in a footnote later in VT). The name has nothing to do with the nature of the graphical output. A second example is in the caption of VT Fig. 8. It is stated that differences in the projected track lengths between the two programs are due to the way the parameter Dpar is used to calculate the projected track length. Dpar does not enter into the calculation of projected length, nor does any compositional parameter. The projection is purely a geometrical adjustment based on the orientation of a measured track to the c crystallographic axis. Any difference in calculated values would have been due to a typographical error in Ketcham et al. (2007), where a factor of 2 missing in the equation on their page 793, 2nd column, end of 1st paragraph. This was originally implemented in QTQt and has since been corrected. We thank VT for helping us to identify and rectify this problem.

After a short introduction the paper contains 2 main sections (section 2: part 1 and section 3: part II). The first considers a simple regression problem (fitting a polynomial to noisy data) and the second inverse thermal history modelling using both HeFTy and QTQt on a set of real data from Tibet. We address parts of these sections in turn.

2. VT section 2: Part I

One of our most important criticisms is that their primary conclusion is due to a basic misunderstanding of the inverse modelling process and the apparent disregard of whether or not a “suitable model” is able to fit the input data. A polynomial example is developed using synthetic data (with noise added), with the data being described as linear with respect to x (first order polynomial $y = a + bx$), weakly non-linear or strongly non-linear, the last two datasets being generated from a second order polynomial ($y = a + bx + cx^2$, with different values for the coefficient c). The inverse problem is addressed using (i) the correct model, (ii) a weakly incorrect model or (iii) strongly incorrect model, such that the data in cases (ii) and (iii) were modeled with model (i), the first order polynomial. As noted in VT, this problem is linear for the model parameters (a,b,c) and solutions can be obtained with standard linear inverse methods (least squares as implemented in Excel for example).

The results are presented in VT in terms of what would be the output from a random Monte Carlo, Frequentist p-value approach implemented in HeFTy or a Markov Chain Monte Carlo (MCMC) Bayesian approach implemented in QTQt and these are different. Both approaches deal with case (i) adequately. The HeFTy approach does not identify acceptable models for case (ii) and (iii) (acceptable being defined with a user-specified p-value), while the QTQt approach produces output for (ii) and (iii), but for case (iii) the predictions from the inverse model are clearly not consistent with the input data. This leads VT to the conclusion that “the ability of a Frequentist algorithm such as HeFTy to find a suitable inverse model critically depends on the quality

[*] Corresponding author.
E-mail address: kerry.gallagher@univ-rennes1.fr (K. Gallagher).

https://doi.org/10.1016/j.earscirev.2017.11.001
Received 19 March 2017; Received in revised form 18 September 2017; Accepted 1 November 2017

and quantity of the input data; while (b) the opposite is true for a Bayesian algorithm like QTQt, which always finds a suite of suitable models, regardless of how large or bad a dataset is fed into it". VT does not define what is meant by suitable models but surely this would include fitting the data adequately.

In the words of Sambridge (2006) "inverse problems are as much about asking the right questions of a data set than building a model that fits it". The inverse modelling process involves selecting an appropriate model structure and finding values for the parameters in that model structure that can explain the observed data to varying degrees. For the moment we consider just the first condition, that of model structure.

In cases (ii) and (iii), the wrong model structure is adopted (so the wrong question is being asked), although for case (ii) the coefficient c is small enough that the model is not so wrong. This brings to mind the point made by Box and Draper (1987, page 74) "Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful". Often it will be the subsequent application of a model that determines whether it is useful or not.

Case (iii) is strongly rejected (no acceptable models found) by the HeFTy approach, while the QTQt approach produces output showing a series of results that are the most probable, given the assumed model structure (i.e. a first order polynomial). Clearly the predictions do not fit the input data. In our opinion, this is neither a suitable nor useful model. Note that using the same approach with any software that deals with such linear regression problems (e.g. Excel) would give effectively the same answer, as they will find the parameters and presumably plot the results for a first order polynomial model that will give the minimum value of a least squares misfit function (e.g. Eq. (3) of VT).

The example prompts several further comments. Firstly, it is not a bad dataset that was used. These are synthetic data, with some known noise added. It is the model that is bad, or inappropriate. It is neither allowed to be complex enough, nor have enough parameters, to capture the variation in the data. Secondly, the motor under the bonnet of QTQt is not fixed dimensional MCMC, but transdimensional MCMC, mentioned briefly in appendix B of VT, but not implemented in their example. The transdimensional MCMC approach treats the number of model parameters as an unknown (or a model parameter) and polynomial regression is a typical example of its application (e.g. Mallick, 1998; Sambridge et al., 2006). Application of this approach to data similar to that for case (iii), gives the result shown in Fig. 1. Here we see that the probability of obtaining the result given in VT (polynomial of order 1) is 0, while the probability inferred for the polynomial of order 2 is > 98%, while the probability of 3rd order or higher is < 2%. The transdimensional algorithm implemented in QTQt adapts the complexity of model to improve the fit to the data, but tries to avoid overfitting the data. Thus, it converges to a second order polynomial with high probability. Finally, it is clear that a model forced to be a 1st order polynomial cannot fit the data, based just on visual inspection of the predictions relative to the observed data. We return to this point later.

3. VT section 3: Part II

This section addresses differences in the approaches relevant to thermal history modelling, and follows directly from the previous section.

We first consider the application of HeFTy, the purported "breaking" of which was a consequence of how the software was used by VT, not of its statistical limitations. HeFTy is not quite such a black box that it does not require some common sense and thoughtful engagement in the modelling process, and the way VT set their models up had shortcomings on two fronts.

First, VT used a series of consecutive large constraint boxes that generate a large proportion of time-temperature (t-T) paths that cannot possibly work. For example, in VT Fig. 7 the penultimate constraint box from 15 to 45 Ma allows temperatures of up to 140 °C, which are clearly out of bounds given the fission-track age of 107 Ma. As a result, at least a third of the time-temperature paths generated using this box are impossible – they could never fit the age data, even before considering the details of the track length distribution. When this effect is extended across multiple constraints, the set of paths generated includes only a very small fraction of remotely plausible ones, making the inversion very inefficient. As larger numbers of confined tracks are included, and the solution space shrinks, the problem of sampling solutions that can fit the data is amplified. This is the curse of dimensionality referred to later in VT.

A common approach when using HeFTy is to run a model briefly with relatively broad constraints, and then as valid-time-temperature paths are found to shrink the constraint boxes to more tightly surround the solution space. Similarly, constraints can be added where t-T paths are naturally coming together. If done with care, this technique can greatly speed up the inversion without restricting the range of solutions (e.g. being trapped in local minima).

The second shortcoming is the indiscriminate use of constraint boxes with no recognizable motivation for their number, size, or placement. Users of HeFTy are encouraged to have a purpose for each constraint box, such as representing a piece of independent information (e.g., present-day temperature, deposition time and temperature, burial history indicated by overlying strata, etc.) or testing or exploring a hypothesis (e.g., was cooling monotonic, or was there reheating; what

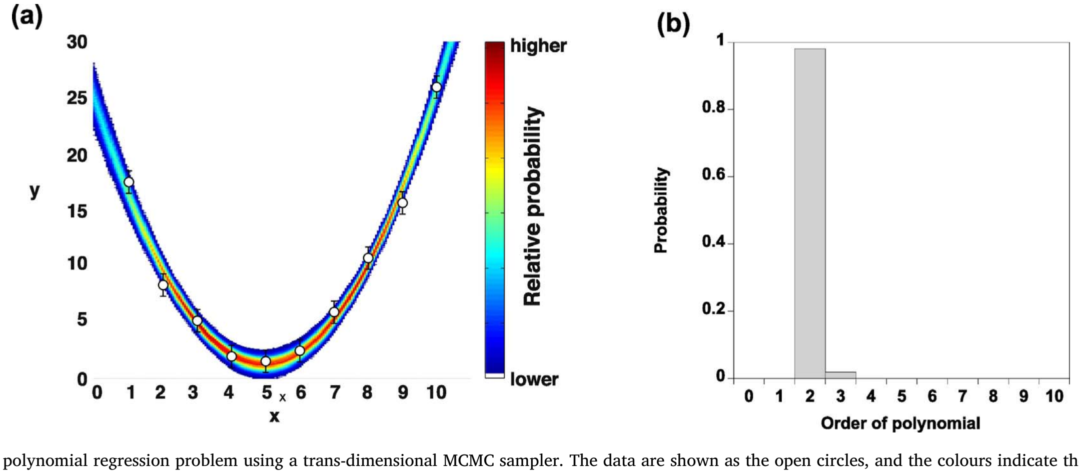

Fig. 1. (a) Result of polynomial regression problem using a trans-dimensional MCMC sampler. The data are shown as the open circles, and the colours indicate the relative (marginal) probability based on the accepted solutions.
(b) Probability distribution for order of the polynomials summarised in (a). Polynomial order 1 has zero probability given the observed data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

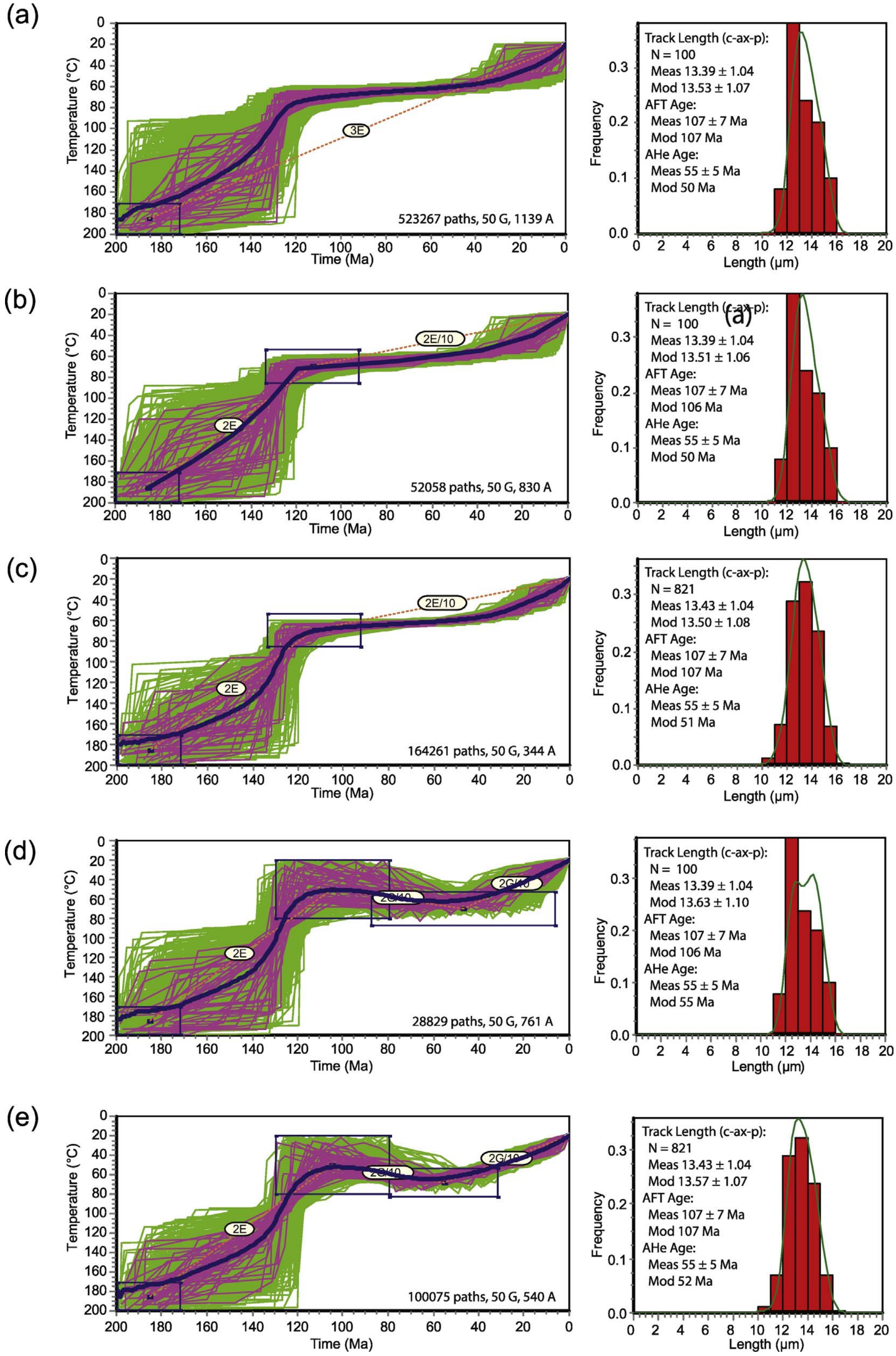

Fig. 2. HeFTy inversions of AFT data in files mm1, mm3, and mm4 downloaded from the supplementary data for VT. All models include AHe age 55 ± 5 Ma, with no alpha ejection effect included, as surmised from comparison with VT modelling results. (a) Sample with 100 lengths, modeled as cooling-only with 2 constraints. (b) Same as (a) with an additional optimizing constraint. (c) Sample with 821 lengths. (d) Sample with 100 lengths, with a history with constraints arranged to permit reheating, similar to result from VT. (e) Sample with 821 lengths allowing reheating.

was the timing and temperature of maximum burial, etc.). The output of HeFTy can then be interpreted as supporting or contradicting the hypothesis in the context of the thermochronologic data and specified constraints, and providing the range of possibilities permitted by the data within that hypothesis.

These concepts are illustrated in Fig. 2, which shows several successful runs of HeFTy on the VT data sets. In Fig. 2a, the 100-track-length data set (with AFT and AHe ages included) is successfully fitted using only two constraints: an early, high-temperature one representing an initial condition starting substantially above the closure temperatures and well before the oldest ages, and one for the present. The broad coverage of the paths before ~120 Ma indicates that the thermal history is only constrained by the data after that time, at least in this scenario. Fig. 2b shows the same model, with an additional, optimizing constraint placed around the region where the paths begin to converge. The optimization does not affect the solution, but reduces the number of paths needed to reach the ending condition chosen (finding 50 'good' paths) by an order of magnitude from over 520,000 to about 52,000. Fig. 2c shows the same inversion as 2b, in which the entire 821-track data set is used. Here HeFTy does not "break" as claimed by VT, but instead finds a tighter range of solutions; this increased tightness might be considered the "reward" VT believe there should be from having more track-length measurements. Fig. 2d shows a 100-track version of a model more similar to those shown by VT, which includes the opportunity for reheating, and Fig. 2e shows the 821-track version of the same model. Again, the increased number of track measurements does not eliminate all solutions, but instead tightens the solution space. The success of the models in Fig. 2c and e indicates that the data support both the cooling-only and reheating hypotheses, and in fact do not distinguish between them.

This example demonstrates that the concerns expressed by VT about the fragility of the statistical methods used by HeFTy are misplaced. They may be valid in some abstract theoretical limit, but as demonstrated here HeFTy is able to handle and find solutions for data sets far larger than are obtained in current standard practice, including the one gathered by VT with obvious intention of breaking it.

Turning to the use of QTQt in this section, the idea VT puts forward follows from that proposed in their earlier section on fitting polynomials: i.e. irrespective of the integrity of the data, QTQt will give output, that will be adopted as an appropriate solution. VT use some of their own data to demonstrate "Garbage in, Garbage out", but do not actually cite examples of such practice from any previously published work. The output of QTQt is conditional on the input data (and their assumed uncertainties), and the assumed range or domain of possible values for the thermal histories themselves, referred to as the prior. The results also depend on the nature of the forward models that we assume for fission track annealing, diffusion, etc. (which can also be regarded as a form of prior information). Given all this, we can obtain a posterior distribution of thermal histories. Ideally the posterior will be different to, and more informative than, the prior. If the posterior and the prior are the same, then the data have told us nothing we did not think we knew already. The important point is that the inferred posterior is conditional on the input data and model assumptions, as all Bayesian inverse solutions are.

Consequently, it is important to assess, even if only visually, how consistent predictions of the input data are with the observations given the posterior model. This may typically involve consideration of the distribution and magnitude of the residuals. So the model should not be consistently overpredicting or underpredicting relative to the observations, the residuals ideally should not have systematic trends and a prediction that lies well outside the uncertainty associated with an observation may suggest that observation or the relevant predictive model perhaps should be reassessed.

We would like to make several points concerning the approach adopted to produce the results summarised in Fig. 8 of VT. First, they grouped 5 single grain analyses, with an age range of 47 to 66 Ma, into a single datum with an age of 55 Ma (changed to 102 Ma for 8ii). However, QTQt has always been capable of dealing with the individual grain ages, with the grain specific parameters (e.g. grain size), and there is no need to use some kind of average aliquot age. We would note also that in the downloaded data files, the apatite grain is defined as a cube with sides of 62 μm in length, which implies an equivalent spherical radius of 31 μm. The data files show that no alpha ejection was applied during the modelling, but it is not clear from the text if the ages are alpha corrected values or not. However, as we show below, the AHe age data play a very minor role in the inference of the thermal histories.

Second, the fission-track length data are not the same in the two examples. The first uses 821 track length measurements, while the second uses 100 track lengths. Visually, the two track length distributions look quite different, with the former being relatively symmetrical, perhaps slightly negatively skewed, while the second is clearly positively skewed. Third, the data files available in the supplementary material are not consistent with the data presented in the figure (and it seems that the website has swapped the identifier of each file so that it refers to mmc5 as the data used in Fig. 8(ii) but the file has 821 track length measurements, whereas file mmc6 is referred to as the data used for Fig. 8(i), but it has only 100 track length measurements). Additionally, both files have what appears to be a typographical error in one of the induced counts (Ni), with a value of 38,978, which probably should be 3898 if the data in the files associated with HeFTy are correct. This outlier does not change the central age too much, the data in the files giving 102.3 Ma, while correcting the apparently anomalous Ni value, gives 104.8 Ma. Neither set of counts data yields a central age that passes the typical χ^2 test (at the $p(\chi^2) = 0.05$ level) and the problem with the bad Ni value is particularly obvious in a radial plot. Putting these problems aside for the moment, the results presented in VT Fig. 8, with quite different inferred thermal histories, were used to argue that QTQt will always give output irrespective of the data. As mentioned previously, this is because the Bayesian approach will produce a posterior distribution (of thermal histories) conditional on the input data and model assumptions.

The differences in the inferred thermal history shown in Fig. 8 are attributed to changing the AHe age from 55 Ma to 102 Ma, equivalent to the reported value for the AFT age. We ran the same types of models using the files directly downloaded from the supplementary material, and obtained the results given in Fig. 3. We have similar timing of cooling around 100 Ma, irrespective of the AHe age, in contrast to the results presented in VT Fig. 8. However, similar to the predictions from VT Fig. 8b, the fit to the observed data is poor, with both the AHe and AFT ages being considerably younger than the observed values, and the track length distribution is poorly fit for the 100 length data. If we correct the anomalous Ni value then the timing of cooling changes (Fig. 4) to around 130 Ma, the predicted AHe ages are either older or younger by 20 Ma, while the AFT ages seem reasonably well reproduced. The post 130 Ma thermal histories do show some difference in structure for temperature < 70–80 °C, with more structure evident in the dataset with 821 track length measurements. Using the original AHe age of 55 Ma in both of the datafiles with the correct Ni value, we get much the same thermal history results as Fig. 4, indicating that the anomalous AHe age has little effect on the difference on the inferred thermal histories. Removing the AHe age data leads to the results given in Fig. 5 indicating that the differences in the inferred thermal histories, and their resolution in terms of the 95% credible intervals, reflect the differences in the track length data (as the count data are identical in both cases). Looking more closely at the length data, it seems that the 821 lengths dataset has 44 measured lengths shorter than 10 μm, while 100 lengths dataset has just 5. As shorter track lengths tend to have a large influence on the inferred thermal history, this may explain part of the increased structure after the initial cooling around 130 Ma. We could examine this further by taking out these shorter tracks, but this is beyond the aim of this comment. The main point is that the comparison made in VT is based on 2 quite different data sets.

If we accept VT Fig. 8 at face value, then when assessing the quality of the predictions, there is a problem with the discrepancy between the predicted and input ages for both AFT and AHe, the predicted values being about 20 m.y. younger, or nearly 3 times the input error. Such discrepancies should not be ignored when assessing modelling results and an assessment of the agreement (or not) between the predicted and

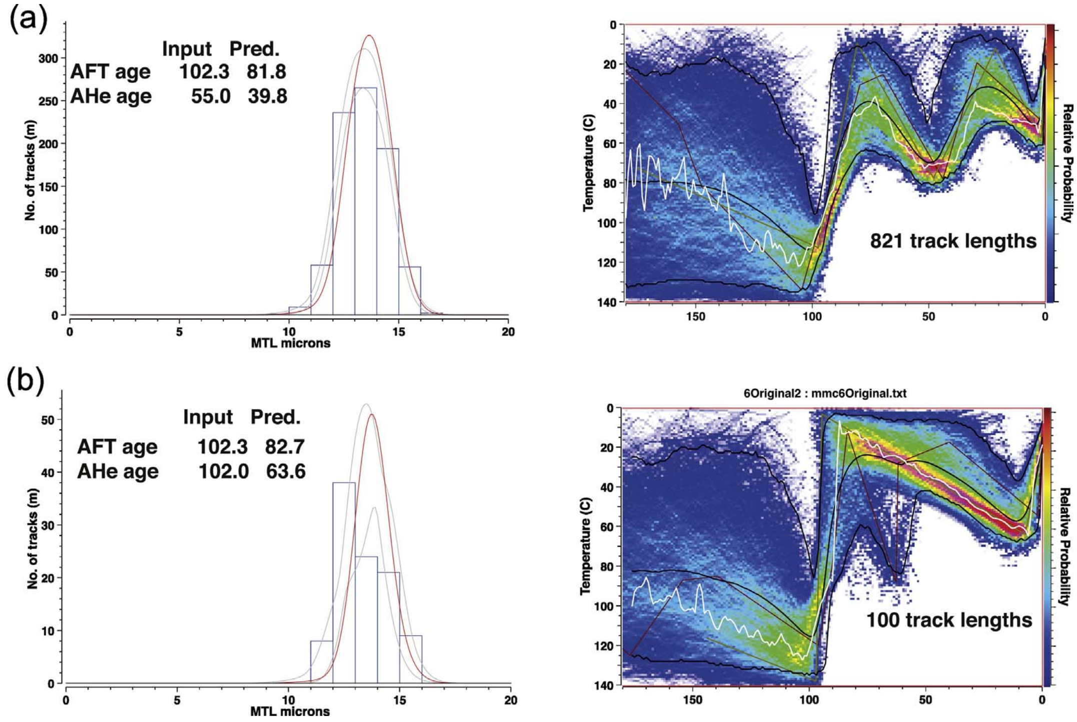

Fig. 3. Inferred thermal histories and predictions for the 2 data files mmc5 and mmc6 downloaded from the supplementary material for VT, but both seem to have an anomalous Ni value. (a) Sample with 821 lengths and AHe age = 55 Ma. (b) Sample with 199 lengths and AHe age = 102 Ma. In contrast to the result presented in VT, these two samples have the initial cooling period around the same time about 100 Ma.

input data is an important aspect of any inverse modelling study. It may be that better data fitting models cannot be found, and then the data, models and assumptions may need to be re-examined. However, the main conclusions here are that the differences in the thermal histories in VT Fig. 8 are due to a problem with one outlier value of Ni for Fig. 8i (but apparently not for 8ii), and the fact that quite different track length distributions were used for each case. This example does serve to demonstrate a practical strategy to assess the fidelity and consistency of different data sets and data types. We can break the total set into subgroups (e.g. AFT data and AHe data) and undertake modelling on

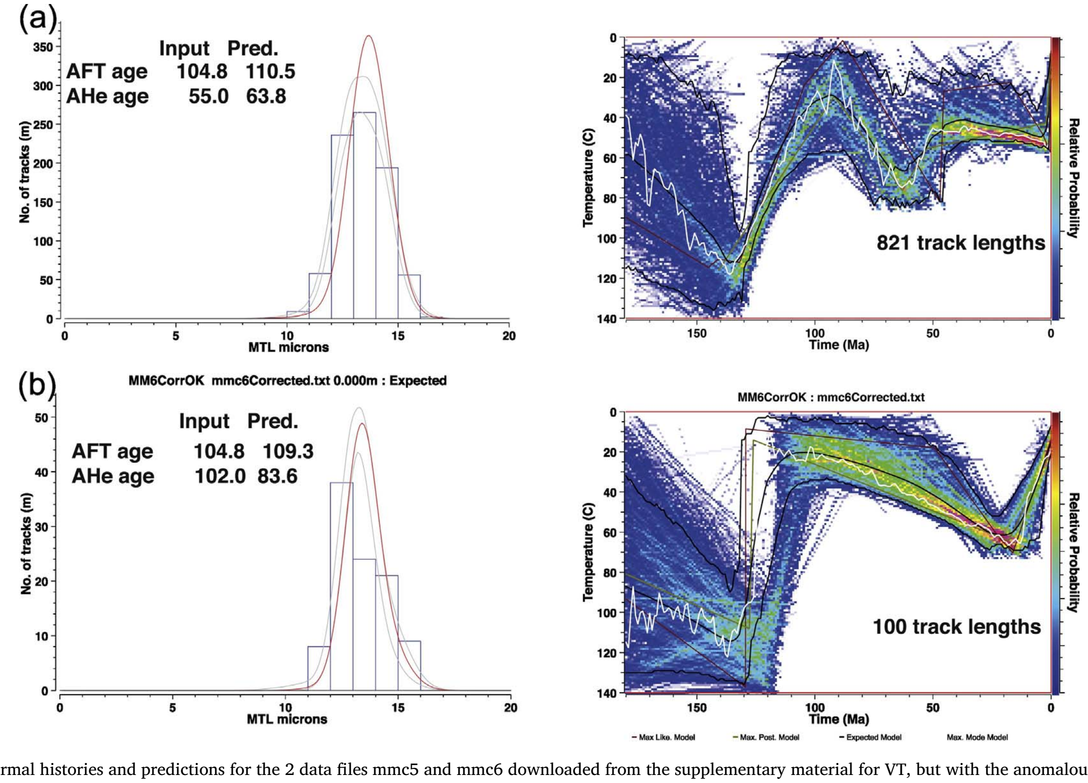

Fig. 4. Inferred thermal histories and predictions for the 2 data files mmc5 and mmc6 downloaded from the supplementary material for VT, but with the anomalous Ni value corrected. (a) Sample with 821 lengths and AHe age = 55 Ma. (b) Sample with 100 lengths and AHe age = 102 Ma. In contrast to the results presented in VT, these two samples have the initial cooling period around the same time (about 130 Ma).

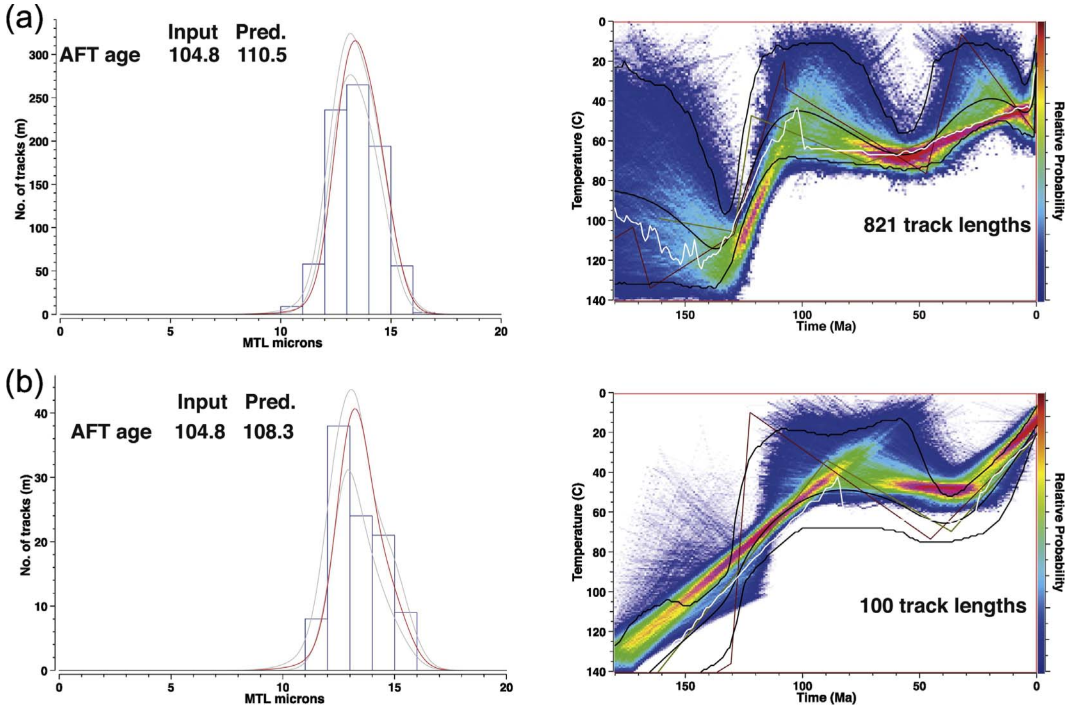

Fig. 5. Inferred thermal histories and predictions for the 2 data files mmc5 and mmc6 downloaded from the supplementary material for VT, but with the anomalous Ni value corrected and no AHe age data used. (a) Sample with 821 lengths (b) Sample with 100 lengths The differences in the thermal histories are attributable to the different track length data rather.

each subgroup in turn. In this way it can be possible to identify which features of the overall thermal history solutions are attributable to different data and so provide different information on the thermal history and/or the integrity of the predictive models for annealing and diffusion, for example.

Finally, we acknowledge the good practice demonstrated by VT in posting their data as an electronic supplement, which enabled us to conduct these re-evaluations. However, the data files reveal further irregularities beyond the single spurious Ni value discussed previously. A number of other counts are off by 1 here and there between the HeFTy and QTQt files. The zeta values are also different (380 vs. 386), as are the Nd′s (4000 vs. 5000), affecting the age uncertainties. None of these would be expected to have a large effect on the inversions, but such inconsistencies are unfortunate when the intent of the study is to compare program output given equivalent data input Additionally, the length data are not calibrated against a standard (e.g., Ketcham et al., 2015), which in one of the present authors' experience has been a cause of failing to fit a large number of lengths, but was not in this case. The point has been underappreciated in the AFT community, but it makes little sense to think one should be rewarded for measuring 821 track lengths in an unknown if these are not calibrated against one or two hundred measurements from a standard. Finally, as mentioned above, the AFT single-grain ages do not pass the chi-squared test. Strictly, such a fission track age dataset should perhaps be considered suspect for thermal history modelling that presumes the ages constitute a single population.

4. VT section 4

The comment on p. 287 that "the problem is that HeFTy is very flexible in accepting many different types of data and it is unclear how these can be normalised in a common reference frame" is bizarre; normalising different data types to a common reference frame is a requirement for doing multi-thermochronometer modelling in the first place, whether in HeFTy or QTQt. In HeFTy, for example, this is done by calculating a probability (whether it is called *r* or a GOF) that the data could be drawn from the model for track lengths, or the model result from the data and estimated uncertainties for ages. It also bears pointing out that HeFTy already uses a Bonferroni-type correction in its 'good' model criteria, specifying that the minimum goodness-of-fit value be at least 1/(N + 1), where N is the number of data sets, and that the mean be at least 0.5. This change was reported in Ketcham et al. (2009), and is described in the program documentation. HeFTy also allows the user to change the GOF value used to define 'acceptable' solutions.

There is another statement, also on p. 287, stating "the MCMC algorithm ... does not care 'how bad' the data fit is". While MCMC does not explicitly maximise the likelihood, this statement is misleading as the likelihood ratio appears in the acceptance criterion. Better data fitting models tend to be preferred, unless the prior weights strongly against them. If it was not the case then, when using a uniform prior, all proposed models would be accepted, and the posterior will look the same as the prior for the thermal histories. Furthermore, it is the interplay between the likelihood and the prior that provides the natural parsimony implicit in the transdimensional version of MCMC. More complex models will tend to fit the data better (higher likelihood), but these will tend to be penalised by having lower prior probability. The combination of the two can then lead to a lower posterior probability. As stated earlier, the results from the Bayesian approach are always conditional on the input data and prior assumptions regarding the thermal history and the forward model parameters. If one, or all of these are inappropriate, it is not the fault of the MCMC algorithm, as it produces a posterior based on the information given to it.

5. VT section 5

Section 5 addresses the selection of time-temperature constraints. When discussing HeFTy, VT discuss the "curse of dimensionality" apparently while at the same time dealing with it adequately in their sample, as discussed previously (Fig. 2). Even with only two constraint boxes and all of their length and AHe data, HeFTy starts finding acceptable solutions within the first 10,000 paths, or less than ~30 s. Concerning box size, each box ought to be the size appropriate for the data or hypothesis behind it; for example, there is absolutely nothing

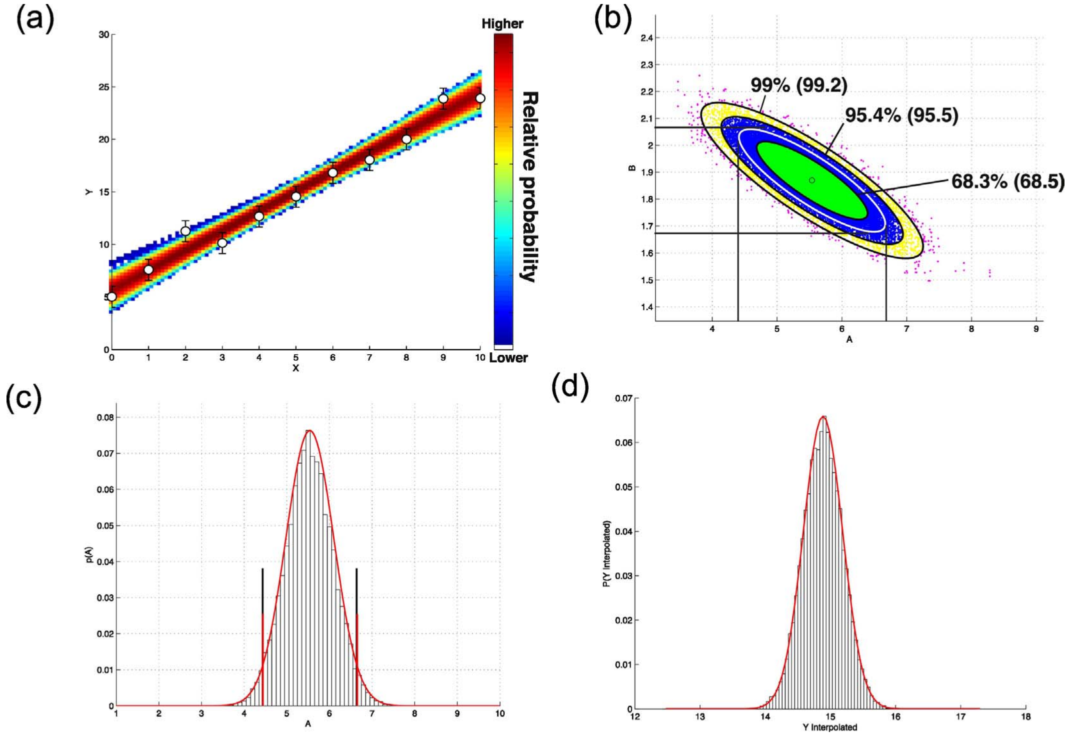

Fig. 6. (a) MCMC estimation for the parameters defining straight line or first order polynomial, y = A + Bx, given the data shown as open circles. The colours, show the marginal posterior probability for y = A + Bx over the range of X using all the accepted combinations of the parameters A and B (see text for details).
(b) The joint distribution, p(A,B|d), of the two parameters a and b. The labelled black ellipses are the analytical solutions for different cumulative probabilities (given as the percentages, and calculated from a χ^2 distribution with 2 degrees of freedom) and in brackets the proportion of the total accepted sample pairs (A,B) obtained using MCMC. The MCMC samples within each probability band are distinguished by different colours for better visibility. The white ellipse is the 95% region for 1 degree of freedom and is used to estimate the 95% confidence intervals on each parameter, shown by the black lines projected onto each axis.
(c) The red line is the analytical solution for marginal distribution on parameter A, and the histogram is the distribution estimated from the MCMC samples for the same parameter. The red and black vertical lines indicate the 95% credible intervals calculated analytically or using the MCMC samples respectively.
(d) The marginal posterior distribution of the predicted value for y = A + Bx, at x = 5. The red curve is the analytical solution, and the histogram represents the histogram estimated for y(x = 5), using linear interpolation between the predicted values for y(x = 0) or A, and y(x = 10). This is the same form of interpolation used in QTQt. We obtain the same result if we sampled A and B from the joint distribution p(A,B|d) and use predicted values for the function A + Bx, at x = 5.

wrong with deposition being represented by as small a box as independent stratigraphic data indicate.

Subsequently, VT urges users not to use geological constraints with QTQt. While the philosophy underpinning QTQt is primarily to let the data determine the thermal history and/or to avoid over-structuring the thermal history with many constraints, some constraints are obviously more geologically valid than others. For example, we often have an idea of the possible range in stratigraphic age for sediments, so we know the sample should be at surface temperatures some time in a stratigraphic age range. The duration of an erosional unconformity in a well section can potentially be estimated within a (possible uncertain) range and implies the samples immediately below the unconformity were at or near the surface during some part of that range (they could of course have been buried and re-exposed during that time too). The presence of surficial lava flows in contact with basement rocks implies the latter were at or near the surface at the time of eruption. We may also have some constraints on the present day temperature, or its range, for a given sample, either at the surface or in well. This is additional prior information that can be incorporated as a constraint defined as a box (similar to HeFTy), defined by ranges on time and temperature. As the inferred thermal history models will depend on the prior information, it is straightforward to assess the impact of such constraints by removing them (e.g. Bernard et al., 2016). In this context, it is then important to report what prior information has been used when reporting any model results.

It is appropriate here to address another erroneous statement concerning output from QTQt made in this section at the bottom of page 287, "The crudeness of these models is masked by averaging, either through the graphical trick of colour-coding the number of intersecting t–T paths…". This refers firstly to the fact that individual thermal histories are relatively simple, being constructed as a series of discrete time-temperature points as model parameters and secondly to the representation of thermal multiple history models with QTQt in the form shown in Fig. 8 of VT. In this representation, a thermal history composed of a finite number of t-T points is interpolated between these points. We can then count how many thermal histories pass through a given temperature interval at a given time interval (typically defined as 1 °C and 1 m.y.) and normalise these values by the sum over all temperatures for that time interval (so that the sum is 1). This is then plotted using colour-coding to indicate the appropriate value over the range of sampled temperatures for the given time interval.

The phrase "graphical trick" implies this is some kind of sleight of hand with no mathematical foundation. The plots present what is known more formally as the marginal posterior distribution of temperature. This is not simply the probability distribution of temperature at say 100 Ma. Rather it is a conditional probability distribution, that is, the probability of temperature at 100 Ma, given the variation of accepted temperature values at all other times. When correctly implemented, Bayesian MCMC sampling produces a set of parameters (time temperature points) distributed according to their joint posterior distribution. The marginal posterior distribution for a parameter, or a function of parameters, is given by all the accepted values of that particular parameter or function values. The hard work of dealing with the conditional dependence on other parameters is dealt with for us by the MCMC algorithm in the way it is constructed.

To demonstrate this in practice, we return to the linear regression problem, with a 1st order polynomial, $y = A + Bx$. There are two parameters A (intercept) and B (slope) and, given normally distributed errors on the data (d), we can obtain analytical solutions for the joint posterior distribution of A and B conditional on the data, $p(A,B|d)$. We can also obtain analytical solutions for the marginal distributions, $p(A|d)$ and $p(B|d)$ which are the probability distributions of parameter A, given the data and allowing for the inferred variation in parameter B, and the same for parameter B, given the data and the inferred variation in parameter A. The marginal distribution for A, is obtained by integrating the joint posterior over all values of B, so it is defined as $p(A|d) = \int p(A,B|d)dB$. We can also obtain an analytical solution for the distribution on the predicted value of y_i at a position x_i, known as the posterior predictive distribution, $p(y_i|x) = \iint p(y_i|x_i, A, B)p(A,B)\, dA\, dB$. This last operation is analogous to interpolating between 2 discrete model time-temperature points in QTQt, where A and B would be the two time-temperature points, x_i would be the time we want to calculate the marginal, and y_i is equivalent to the temperature.

Here we compare the results from the analytical solutions (e.g. Lee, 1989; Denison et al., 2002) to those obtained using the same approach implemented in QTQt. These are given in Fig. 6, and we can see that MCMC samples and the analytical solutions agree well for the joint distribution (Fig. 6b) and the marginal on the intercept, parameter A (Fig. 6c). We could show a similar plot to 6c for parameter B, showing similar concordance between the analytical and MCMC solutions. Finally, we show the analytical and MCMC posterior predictive distributions for y at $x = 5$ (Fig. 6d) and we can do this for any value x. The MCMC approach is the same as the method used in QTQt to produce the marginal distribution on the temperature at any time by interpolating between discrete time-temperature points. Fig. 6a shows the result of the same calculation applied to the regression problem and a vertical slice at $x = 5$ gives us the result shown in Fig. 6d, which agrees well with the analytical solution.

So there is no graphical trick involved in producing the marginal posterior distribution for the thermal history results from QTQt, However, there may be some confusion about what the marginal distribution and its scale represent. It is not the probability of a particular thermal history being correct. It is the distribution of temperature over a chosen (usually small) interval of time, given the variation in the accepted thermal histories (or perhaps more easily considered as the average temperature at different times). The accepted thermal histories themselves depend on the data, priors, and forward models as stated above.

6. Conclusions

There are two overriding issues with the critique of Vermeesch and Tian (2014). The first is that it is uncalibrated. Although the points they make are reasonable and even obvious – that the predictive models for annealing and diffusion are not perfect, geological factors can make systems perform non-ideally, reproducibility of some data can be poor – they say nothing quantitative about the scale of the effects that these non-idealities have on thermal history inversion. All the points made in the final paragraph are indicative of confirmation bias, as they are unsupported by their analysis.

The second is that their critique is overly simplistic about what thermal history inversion software should do. It is not intended merely to find an answer or reward the user, but as a tool to help conduct science. “Garbage in, Garbage out” does not only refer to faulty data or incompletely known kinetics, but also to the process of setting up models, posing and testing hypotheses, and appropriately interpreting the results. We hope that this comment helps to clarify some of the misunderstandings in VT and how, in our view, inverse modelling of thermochronometry data might be approached in general.

Acknowledgments

We thank R. Brown and M. Sambridge for their helpful remarks on earlier versions of this manuscript.

References

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

Box, G.E.P., Draper, N.R., 1987. Empirical Model Building and Response Surfaces. John Wiley & Sons, New York, NY.

Denison, D.G.T., Holmes, C.C., Mallick, B.K., Smith, A.F.M., 2002. Bayesian Methods for Nonlinear Classification and Regression. Wiley, Chichester.

Ketcham, R.A., Carter, A., Donelick, R.A., Hurford, A.J., 2007. Improved measurement of fission-track annealing in apatite using C-axis projection. Am. Mineral. 92, 789–798.

Ketcham, R.A., Donelick, R.A., Balestrieri, M.L., Zattin, M., 2009. Reproducibility of apatite fission-track length data and thermal history reconstruction. Earth Planet. Sci. Lett. 284, 504–515.

Ketcham, R.A., Carter, A., Hurford, A.J., 2015. Inter-laboratory comparison of fission track confined length and etch figure measurements in apatite. Am. Mineral. 100, 1452–1468.

Lee, P.M., 1989. Bayesian Statistics: An Introduction. Edward Arnold.

Mallick, B.K., 1998. Bayesian curve estimation by polynomial of random order. J. Stat. Plan. Inference 70, 91–109.

Sambridge, M., 2006. Inverse Problems in a Nutshell, Abstract From “Mastering the Data Explosion in the Earth and Environmental Sciences”, Aust. Acad. Science Elizabeth and Frederick White Conference. (rses.anu.edu.au/cadi/Whiteconference/papers/s06-efw.pdf).

Sambridge, M., Gallagher, K., Jackson, A., Rickwood, P., 2006. Trans-dimensional inverse problems, model comparison and the evidence. Geophys. J. Int. 167, 528–542.

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