Earth Surf. Dynam. Discuss.,
https://doi.org/10.5194/esurf-2020-59-AC5, 2020

Earth Surface Dynamics
Discussions
Open Access
EGU

Interactive comment on "Bias and error in modelling thermochronometric data: resolving a potential increase in Plio-Pleistocene erosion rate" *by* Sean D. Willett et al.

Sean D. Willett et al.

swillett@erdw.ethz.ch

Received and published: 25 October 2020

Preface to our Response: We specifically chose ESurf as the journal for submitting the Willett et al., manuscript (this paper) so that an open scientific discussion could occur around the contrasting results of Herman et al., (2013) and Schildgen et al., (2018). While this comments by van der Beek et al., and our responses, may appear as a heated exchange, we would like to emphasize for the broader scientific community reading this that it is our intent to have a healthy scientific discussion on this topic so that the strengths, weaknesses, and caveats associated with the inversion of thermochronometric data are clear. With that in mind, we respond in the following to the

latest comment by van der Beek et al.

COMMENT: Willett et al. focus their response on two aspects of our initial comment: (1) the role of the different geotherms in the forward and inverse models and (2) the resolution and data density of the synthetic models. We will provide some additional clarification on these aspects below. Before we do so, however, we feel it is important to underline once more that Schildgen et al. (2018) and Willett et al. pursue fundamentally different goals. Schildgen et al. (2018) posed a geological question: whether the "worldwide acceleration of mountain erosion under a cooling climate" since 6 Ma inferred by Herman et al. (2013) from thermochronologic datasets was robust. In contrast, Willett et al. aim to identify different sources of error in the GLIDE inversions of thermochronologic ages for erosion rates, thereby posing a methodological, model-framed question.

RESPONSE: We don't really see a difference in the goals of Schildgen et al. (2018) and Willett et al. (this paper). Schildgen et al. made no new GLIDE models of erosion rates from the data of Herman et al., so they did not test "robustness", which is defined as consistency of results in the face of data or assumption errors. Probably van der Beek et al. meant to say "accuracy", although their subsequent paragraph argues that this is impossible to do, but whatever the goals, Schildgen et al. (2018) was in effect was a comment on Herman et al., and demonstration of the accuracy of the GLIDE inversions is the centerpiece of both papers.

COMMENT: As stated in our initial comment, a detailed analysis of the potential sources of error within the GLIDE inversion procedure used by Herman et al. (2013) was outside the scope of the Schildgen et al. (2018) analysis. Schildgen et al. (2018) did not set out to test whether the model could make robust predictions of erosion history, but whether it had done so in the Herman et al. (2013) analysis. While Willett et al.'s new analysis is a useful exercise, it does not address the question of whether the model application to real-world data leads to accurate predictions of erosion histories. As Willett et al. state in their response: "we are attempting to learn something about the method, not trying to simulate the way in which the model is applied in normal circumstances". Therefore, it is unclear (and Willett et al. do not clarify) how their new analyses validate the Herman et al. (2013) results. In fact, validating the application of such models to natural systems is inherently impossible (Oreskes et al., 1994): even if Willett et al. would have shown that the inversion is capable of retrieving accurate exhumation histories in the conditions they provided (and the reanalysis of some of Willett et al.'s synthetic tests in our initial comment shows that it does not), these results do not address the analysis performed by Herman et al. (2013).

RESPONSE: This is an interesting perspective on the approach taken by both Schidgen et al.(2018) and our current paper, and we will return to this issue in a later response. However, we would like to point out that the paragraph above states that "validating the application of such models to natural systems is inherently impossible (Oreskes et al., 1994)", whereas the same paragraph states "Schildgen et al. (2018) did not set out to test whether the model could make robust predictions of erosion history, but whether it had done so in the Herman et al. (2013) analysis." In other words, Schildgen et al. set out to do what Oreskes et al. (1994) argued is inherently impossible.

In fact, Oreskes is correct, it is not possible to confirm or refute a model directly, simply because one never knows the true answer. Schildgen et al. do not compare the results of Herman et al (2013) to the true erosion rate history; they don't know the true erosion rate history. They compare the Herman et al. results to a different set of inferences that they derived by extraction from published literature. Herman et al (2013) and Schildgen et al. (2018) are essentially alternative models with differing results. To judge between them, it is necessary to assess each to determine if one or the other is more likely to come up with the correct answer. This is why we (Willett et al., this paper) have conducted a complete error analysis of Herman et al. to try to determine where errors might arise. We would do the same and make an in-depth analysis of the model of Schildgen et al. (2018), but because their paper included no reproducible

tests (other than the tests of GLIDE, which were done with the wrong geotherm), we cannot conduct an error analysis. Hence our focus remains on the Herman et al. (2018) model.

COMMENT: 1 Variable geotherm calculation

Willett et al. dismiss the synthetic models in Schildgen et al. (2018) and in our initial comment because of the difference in boundary conditions when calculating the geotherm, writing: "all error in the van der Beek et al. comment and Schildgen et al. (2018) models is due to these geotherm errors". But Willett et al. provide no support for the assertion that differences in boundary conditions produce spurious accelerations, neither in their manuscript nor in the response to our comment. Nor do they address why spurious accelerations are present in inversions using the high-density synthetic data generated by Willett et al., where there are no differences in geotherm calculation between the forward and the inverse models.

RESPONSE: There are no spurious accelerations to explain in the high density synthetic data models (Figures 5, 6, 7, 8).

COMMENT: We reiterate the observations on which we based our rebuttal of Willett et al.'s dismissal: âĂć Figure 1 in our initial comment shows that when the inferred cause of accelerations in our synthetic models (what we term the spatial correlation bias) is removed, the models do not predict accelerating erosion rates in the last few Myr, despite the difference in geotherm calculation between the forward and inverse models;

RESPONSE: The model shown in Figure 1 of the comment has large errors due to the wrong geotherm (Willett et al., Fig. 4). We don't know in detail how this affects the inverse model result – acceleration or deceleration, given the complex feedbacks on the thermal model, triggered by data errors. Adding and removing data with large errors can easily flip errors between acceleration or deceleration, simply as a function of the errors. We don't think it is our responsibility to do the error analysis on Schildgen et al.'s or van der Beek et al.'s models. We demonstrated errors of up to 100% in age. Errors in erosion rates are in the tens of percent - probably higher where erosion rates are low. This is unacceptable by any possible standard we could come up with. We are not inclined to spend additional time analysing models with such large errors, where any interpretation is tainted by speculation regarding the effect of these errors. Our Figures 5,6,7,14) show similar models with no geotherm errors and there is no false acceleration.

COMMENT: âĂć Figure 3 in our initial comment shows that Willett et al.'s models with a transient geotherm (but ages calculated with the same thermal boundary conditions in the forward and inverse model) show spurious accelerations in regions of spatially variable exhumation rates, either at earlier times than shown in Willett et al. or in the last few Myr when the high-temperature ages are removed;

RESPONSE: We did explain these errors - in detail. We explained this in our last response and stand by that statement. These are resolution errors, combined with a low value of the prior. It is always possible to keep removing data, until a model is poorly resolved. Or find some part of a model in space or time that is unresolved. In the early timesteps, there are no ages in the fast uplifting region, so there will be smoothing of the low uplift rates into the high uplift zone. These will remain in the early history until a sufficient number of high uplift-zone ages enter into the problem. Until then, the problem is defined as poorly resolved.

COMMENT: âĂć Figure 4 in our initial comment shows that Willett et al.'s models with a fixed geotherm similarly produce spurious accelerations in regions of spatially variable exhumation rates. Our conclusion from these tests is thus the opposite of what is suggested by Willett et al.: the geotherm errors cannot explain the spurious accelerations observed in these synthetic tests. We provided some tentative explanations for why this may be the case, but we did not argue that one or the other model provides a "more accurate" solution to the geotherm problem; we acknowledge that the thermal boundary conditions in both Pecube and GLIDE are imperfect.

RESPONSE: Not sure we understand this comment. The use of a steady geotherm to generate ages, inverted with a transient geotherm, explains the spurious accelerations in all models from Schildgen et al and Figures 1 and 2 from the comment by van der Beek et al. These models are, in our view, invalidated by these large errors.

If this refers rather to model errors involving advection amplification of resolution errors, it is correct that this amplification is important only in the full thermal models. However, there are still resolution errors in the constant gradient models shown in van der Beek et al., Figure 4. These also arise from removing data or showing early (under-resolved) timesteps. With higher resolution these errors vanish (van der Beek et al. comment, Figure 4a,b,c). This was our stated conclusion- these figures are consistent with our conclusions, along with the other dozen models we constructed. At high resolution, there is no error; at low resolution, there are errors, but the nature and direction of the errors are not generalizable. Again, we are not sure why we are responsible to explain every model van der Beek et al. can produce. We presented a large and consistent set of models in our paper, which provide more insight than isolated examples of poorly-defined phenomenon with guesses as to causal relationships. For example, our data set C is very similar to van der Beek's Figure 4. It did have large errors and false accelerations. These disappeared when the prior was increased (Willett et al., Figures 10-12) or when the age distribution of the data was more favorable (Figure 9).

The issue of how high and low resolution are defined, we take up below.

COMMENT: 2 Resolution and Data Density

In their response, Willett et al. frequently invoke "high" versus "low" resolution and data density, but without quantifying these classifications. The first argument in Willett et al.'s response is that the "commonality of ages" between the Alpine dataset and their "high-resolution" synthetic datasets A and D makes the results from these synthetic tests applicable to the Alps. We have already refuted this argument in our original comment (lines 210-222).

RESPONSE: There is a simple definition. High resolution is obtained when resolution errors vanish. If one can continue to add data and the answer changes, the initial case was not high resolution. By removing data until errors appear, van der Beek et al. have found the limit between high resolution and low(er) resolution.

In fact, “high” and “low” resolution are relative and context specific. The purpose of the models in Figures 6, 7, 8, 13, 14, is to differentiate between model and resolution errors. This is a standard first test in error analysis and is absolutely necessary. If one does not know if errors are model or data based (resolution), it is impossible to understand any additional analysis. Justification and explanation for the model test is given in Lines 583 to 605 of our paper. To complete this test, it is fully justified to use an infinite number of data, so high resolution is defined to be anything that fully eliminates resolution errors up to and including an infinite number of data.

We stand by our original statement, by removing data or looking at earlier timesteps van der Beek et al have sidestepped the purpose of the model test by changing a high resolution data set into a low resolution data set.

We don’t see relevance of comparison to the Alpine data. This was not our justification.

COMMENT: Second, Willett et al. argue that looking at earlier time-steps of the model is unwarranted, writing: “Because these timesteps are not the target of the experiment, we did not generate ages over these intervals, they are poorly resolved and there is a correspondingly larger resolution error”. This statement is incorrect: Dataset D includes ∼30 mica 40Ar/39Ar ages within the 8-10 Ma time-window (Fig. 5 of the Willett et al. manuscript, or Fig. 1 of the response). Moreover, we could have taken the 10-12 Ma time-window as our starting point, which contains >50 mica 40Ar/39Ar ages in both datasets A and D. The predicted erosion rates are lower in the 10-12 Ma time-window than in the 8-10 Ma time-window (leading to larger accelerations when compared to later timewindows) but the resolution is similar (Figure 1).

RESPONSE: See point above. Age counts are not an exclusive measure for resolution.

First, timesteps must be bracketed by ages, not simply sampled. Second, there are many other factors such as spatial position relative to other ages. If we could just count ages, why would we go through the numerous exercises revolving around resolution determination?

We also point out that the spatial distribution of data in these models based on Data set A, particularly with the Argon data removed, is very poor in the sense that data are clustered with the majority of ages on either side positioned near to the fault, meaning they are separated by less than 1 correlation length. Also given that the fault displacement and the erosion rate contrast is greater than anywhere observed on Earth, we would not speculate as to how many ages define “well” or “poorly” resolved. We accepted these model conditions from Schildgen et al.’s test, but it is not appropriate to contend that these are “typical” of any natural examples. The only extrapolatable conclusion from these tests is that there exists a threshold for data density, above this there are no errors, below this, there are errors. Where this threshold is for the model is easy to establish; where this threshold is for a natural example, such as the Alps, cannot be determined by direct comparison with the models because the erosion rate function and data/age distribution are different.

We also note that we discussed this potential situation in our paper (lines 1179-1186) and acknowledge that it is possible to find a situation where a false increase in erosion rate derived from the prior or nearby data might not be recognized. But this requires a very particular set of circumstances. The fact that van der Beek et al. have modified our model until these circumstances are met is nothing we did not already acknowledge; the important point is how likely is that these circumstances would appear in nature without recognition.

COMMENT: Third, Willett et al. argue that by removing the hypothetical mica 40Ar/39Ar data from the datasets A and D, we have “converted a high-density model into a low-density model and thereby reintroduced resolution errors”. We note, however, that (1) since all the synthetic mica 40Ar/39Ar ages are >8 Ma, the synthetic tests in which they are removed contain exactly the same number of data within the critical 6-0 Ma time-windows as the original tests;

RESPONSE: Same point as above. If errors appear by removing data, these are resolution errors by definition. If there is a model where errors disappear by adding data, the first model was not high resolution. All ages older than 6 Ma contribute to the estimate and resolution of the 6 to 0 Ma window, not just the ages that fall into the particular timestep.

COMMENT: (2) as already noted in our initial comment, the resolution values for the recent time-windows are nearly identical between the two tests (compare panels a-c and g-i in Figs. 3 and 4 of our initial comment). Given the similarity of data and resolution values within the critical time-windows, we do not understand on what grounds Willett et al. characterize one model as “high-density” and the other as “low-density”.

RESPONSE: Based on the existence of resolution errors. This is not circular because we recognize them as resolution errors because they vanish as more data are added or as the prior goes nearer to the true value (compare Willett et al., Figs 10 and 12). This is why one model has little value – it is only through multiple and contrasting models that one understands the sources of errors and behavior of a method. It would be great to have a binary criterion, so one can black-box an analysis, but this does not exist for most estimation problems. The values of posterior resolution and how well they reflect the situation are a separate issue that we will address later.

Most of these comments by van der Beek et al seem to be conflating the multiple purposes of the model tests. There are two purposes, carefully laid out by the organization of our paper:

Establish whether or not there are model errors or only resolution errors. Section 4.3. Figures 6, 7, 8, 13,14. There are no significant errors (including no false accelerations) showing that errors are indeed resolution errors.

Having established that errors are resolution errors, the next steps are:

to determine how many data, in what configuration and over what age range are adequate to resolve an erosion rate history. This is the subject of section 4.4 or our paper and the models in figures 6, 7, 9, 10,11, 12, 14, and 15. to determine how well posterior metrics predict accurate solutions. See line 646 in our paper.

van der Beek et al seem to not recognize the dual purpose and so argue that there is something artificial in the data construct of data set A, when in fact, we probably should have included many hundred more data and more thermochronometric systems to properly address objective (1). We tried to include the minimum necessary number of data in the first experiments, so as to address objective (1), but retain their relevance to Objective (2) (see line 651). It is therefore no surprise that as soon as data are removed the model falls into the “lower” resolution state.

Objectives (2) and (3) are much more complicated questions and it is nearly impossible to generalize a simple answer, which is why we have spent much of our paper discussing how one does this in practice (e.g. section 6.1). We have constructed a range of data sets (Fig. 5), some of which are clearly at the low end of resolution, in order to make a sincere effort to show the range of possible outcomes, including demonstrating failure of our model (e.g. Figure 10). We are happy to continue and expand this discussion, but it should be done on the basis that the extensive suite of models we have presented, which are only a subset of the hundreds of models that we have run over the years, are a legitimate portrayal of the behaviour of our models. Rather than engage with this suite of models and outcomes, van der Beek et al. continue to try to argue that we are covering up some universal failure of our models by perturbing the successful models until they fail. We don’t find that this approach is productive.

We are preparing another response in which we can go into more depth as to how one determines resolution in practice and as to why counting ages in specific timesteps is inadequate, although we thought we covered this rather thoroughly in section 6.1.

COMMENT: As a final point, we provide some clarification of why Willett et al.'s models that use the true solution as the prior (Figs. 8 and 13 of the Willett et al. manuscript) are irrelevant. In Bayesian theory, if the posterior probability distribution of the parameter values (i.e., in this case, the predicted spatial-temporal distribution of erosion rates) is equal to the prior probability distribution (i.e., the input prior erosion rates), then the available data have not contributed to constraining the solution. In this case, the prior erosion rates are equal to the true erosion rates, which have been used to predict the synthetic data. Therefore, for each point in space and time, the model's "initial guess" (the prior), is correct and perfectly reproduces the data. The data cannot influence this problem, and the only thing we learn from this test is that there are no obvious errors in the code. This is the explanation for why the errors disappear. From a practical viewpoint, this test is irrelevant because we cannot know a-priori what the exhumation history of a particular region is (i.e., there is no way to know the prior, it is necessarily a guess) and if we did, there would be no point collecting data in such a region because we wouldn't learn anything new from that data.

RESPONSE: Thanks for the clarification. However, this is not quite correct. Glide is a linear method; there is no initial guess that is to be improved upon. What there is, is a joint probability function for each parameter that reflects a balance between (1) staying near the prior; (2) being consistent with the data (ages); and (3) remaining within the constraints of the model (in this case, the geotherm and spatial correlation structure). By setting the prior equal to the true solution, we eliminate the need to balance fitting the ages and staying close to the prior, but not the third constraint. We still need to balance (1) and (2) against (3). This trade-off is expressed in Willett et al., current paper, Eqn. 12. There is a spatial correlation structure and it does impose averaging; we don't question this. The question is how large is this effect. This is not a test for errors in the code; this is a test for errors in the model, for example, excess smoothness.

Interactive comment on Earth Surf. Dynam. Discuss., https://doi.org/10.5194/esurf-2020-59, 2020.