Earth Surf. Dynam. Discuss.,
https://doi.org/10.5194/esurf-2020-59-AC4, 2020

Earth Surface Dynamics Discussions

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: 24 October 2020

RESPONSE: This review is rather heavy on opinion and somewhat lighter on criticisms that are backed up with observation, but we will nonetheless attempt to address what points we can. We will ignore the first paragraph as this is opinion that falls outside the scope of peer review and contains no backing support, not even reference to sections in our paper to permit us to respond.

As to the factual criticisms, our general impression is that the reviewer must have been intimidated by the length of the paper, and so not read it very completely, as nearly all the requested additions are already in the paper. At a minimum, justifications and

explanations for specific reviewer-comments are already provided in the text, so we draw attention to these, but see little need for additions or major modifications to the paper.

REVIEWER: Figure 2, and the discussion surrounding it, are not well done, and actually seem to harm the case the authors are trying to make rather than setting the stage for it. To begin with, the figure itself is confusing, as depth and age axes are flipped without annotation (e.g. age rises to the right in some cases, to the left in others). Point i went from being the oldest one in 2d and 2e to about the same age as point d in 2f – it appears they also crossed themselves up by flipping an age axis.

RESPONSE: The letter labels in part (e) on some points are reversed. Points are plotted correctly. We will correct the figure, but there is no effect on the meaning.

REVIEWER: The text often refers to the wrong part of the figure (I think they mean 2d in line 161, not 2e; 2e in Line 171, not 2b; and the first 2d in line 182 should be 2e).

RESPONSE: We will correct the latter two typos in revision; line 161 is correct.

REVIEWER: In the text, the reader needs to essentially open Schildgen et al. (2018) and Willenbring and Jerolmack (2016) to follow the authors' attempts to translate those papers' approaches to the figure, and some of those translations appear selective in a pernicious way (e.g., lines 178-180).

RESPONSE: Line 178 is pointing out that Schildgen et al. plotted data points with the lowest elevation as having the largest distance above the closure isotherm. We don't know how this would occur. Without explanation as to why they did this, we cannot reproduce this figure or address it. That is all we are pointing out. We are not sure how this is regarded as pernicious on our part.

REVIEWER: The overall example itself is so oversimplified as to also be confusing, and possibly selfdefeating. The authors seem to assert that their Average 1 is best, or "unbiased", but it's hard to justify extrapolating uplift rates backward in time to before

the oldest ages in the faster-uplift regions (points c and f). One could just as easily, and certainly more conservatively, say that there is no evidence for earlier erosion in those regions, making Average 3 preferable – or, better yet, stopping the attempt to calculate a regional average uplift rate at age c and not going further back in time. Average 1 actually imparts an assumption (bias?) of spatial correlation, saying that the subregions defined by abc and def were also exhuming at time i, despite the data having no information from those subregions indicating that this was so, only proximity to the ghi region. One would need to look at the detailed geology to defend such a claim. . .

RESPONSE: We are not advocating for any average or arguing that any one is better than another. The purpose of this figure is explained on lines 224-242: it is to demonstrate that there are many ways to calculate averages or to parameterize models. To demonstrate that any specific method of averaging has bias is not to prove that ALL treatments of thermochronometric data are biased. We don't see any way in which this figure undermines this argument. We took the thought experiment directly from Willenbring and Jerolmack and find it a useful illustration. It provides a more intuitive alternative to equations and numerical models.

REVIEWER: It is also confusing to call the same model unbiased (line 171) and biased (line 395).

RESPONSE: Line 395 is not referring to the same bias. An average can be unbiased with respect to a regional average rate and still be biased towards a local rate. We will modify the text at this location to make it clearer.

REVIEWER: The difference in boundary conditions used between models for synthetic tests is certainly unfortunate, but the authors did not demonstrate that it had a large effect on the comparisons in Schildgen et al. (2018); the counter-example provided by van der Beek et al. believably indicates that it did not affect their conclusions.

RESPONSE: So, an error of 100% in ages (our figure 4) is "unfortunate" but "doesn't

have a large effect"? This will be reassuring news to many thermochronometry labs who are currently under the impression that generating data with 100% error would not be acceptable. See our response of Oct 5. There is no acceptable error in a control experiment. We contend that all error in the Schildgen/van der Beek models is due to this error and they have not provided the counterargument by calculating the ages correctly and recalculating an inverse model. Our models with comparable resolution, but no geotherm error have no false acceleration. We consider this an effective demonstration. The argument has (correctly) moved on to "what is comparable resolution?", but the Schildgen models and the initial van der Beek comment models are demonstrably wrong for reasons explained in the original Willett et al. manuscript and our response to van der Beek's comment. As such, these models should not be part of the discussion.

REVIEWER: The point of including a geotherm assuming 36 Ma of 1 mm/yr erosion in Figure 3 is not evident; are they claiming that Schildgen et al. (2018) went that far?

RESPONSE: This is exactly what Schildgen et al did.

REVIEWER: The subsequent examples from GLIDE are also deceptive, though I imagine unintentionally so. As van der Beek et al. point out, the examples demonstrating the ability of GLIDE to correctly reproduce erosion rates across a sharp interface used far more advantageous data than Herman et al. (2013) used to derive their conclusions for the Alps. When the data better match what Herman et al. (2013) actually used (Fig. 10-12), the model produced spurious accelerations, at the resolution levels used by Herman et al. (2013), and without boundary condition mismatch.

RESPONSE: This misrepresents the intent and outcome of the modeling. The modeling is broken out into a series of tests. One of the series of tests is to investigate resolution by running a series of models from high resolution to low resolution. None of the tests is intended to model the Alps or to mimic the precise characteristics of the Alpine data, an impossible task in any case. This point is acknowledged by van der

Beek et al (comment Sept 17 line 338). The resolution models contain a spectrum of data sets that likely span the resolution of the actual Alps data; only the lowest resolution models fail, and these models have less than a third of the number of data that are available in the Alps. Figure 9 shows a case where a model with even fewer data, far fewer than used by Herman et al. (2013), accurately reproduces the input parameters, because the distribution in age is advantageous. The models and current text are explicit in pointing out this intent.

Our paper contains an introductory section (4.1) explaining this point, but the reviewer has not acknowledged or referred to this explanation, but rather is repeating the content of the van der Beek et al comment of Sept 17, including the false statement that the Alpine data look more like our synthetic data set C, rather than A (see our modified figure given in comment of Oct 5). Fig 10-12 have fewer data than the Alps and are distributed systematically younger. This is largely irrelevant in any case as the model erosion rate function is not that of the Alps and resolution is determined in part by the erosion rate function.

REVIEWER: There is also an odd tendency to proclaim victory and move on, when the data don't appear to match the words. The authors try to construct a chain of QEDs but instead leave a trail of question marks. As one example, in the Fig. 7 test, the estimated erosion rate in the SE is about twice the true one (to the best of my ability to read their color bar), but they call it a good match. This is also the only model where the SE region is resolved according to both "resolution" and reduced variance metrics. Shouldn't this be worrisome?

RESPONSE: There is a point to every model. Success is determined by the intent and outcome of a model. Not incidental characteristics. The model of Fig 7 has a 100 km correlation length to make the point that there is little sensitivity to the correlation length. Solution is too smooth, so the SE erosion rate is too high, but there is no false acceleration, which is the subject of the paper. As discussed in our paper (e.g. section 6.1), the resolution scales with correlation length, so if one used a 100 km correlation

length, a different cut-off value should be used. So, no, this is not worrisome.

REVIEWER: Similarly, the "perfect prior" tests (Fig 8, 13) do indeed seem trivial. By equation 3, if zc = A e(prior), then e(post)=e(prior). It's not clear if the equality is strictly true – that would depend on whether any noise was added to the synthetic data. Willett et al. don't mention adding noise, so I'm assuming they did not. If that's indeed the case, then in fact none of the tests they present had to withstand routine and inevitable data scatter, making them all suspect for purposes of verifying the robustness of the method. Evaluating the resolution of a statistical method using only noiseless synthetic data would be oddly incomplete.

RESPONSE: We state explicitly (line 485) that we will not consider noise in the data. We could do so, but the reviewer has already made the point that the paper is long as is. We have, of course, run models with noise in the data – it decreases resolution, but can be compensated for by data density, and plays no important role in questions of model bias. A well-designed error analysis does not mix sources of error; it isolates them.

REVIEWER: As a result of this all, I came away still not knowing when GLIDE results are robust, which promised to be a primary contribution of the paper. This became particularly evident at line 1170: "Resolution remains a relative measure, and determining a precise confidence level a priori is not possible, but can be estimated based on spatial patterns, relationship to sample locations, fit to the age data and sensitivity to the prior." In other words, run a big complex computation, and then manually inspect the results to see if you think they actually fit and are justified, nudging thresholds on a case-by-case basis as necessary. This blurs the boundary between quantification and interpretation, and opens the door for arguments about motivated reasoning.

RESPONSE: This is the reality of all estimation problems. Every field that deals with data analysis has had to deal with similar problems. What surprises us most is that so many in thermochronometry don't seem aware that this is normal in both forward and

inverse modeling. Note that reviewer 1 questioned the need for a long, pedagogical introduction, but subsequent comments and reviews have demonstrated that there is need.

REVIEWER: Even as a review and interrogation of GLIDE fidelity, the paper is a bit disappointing. The GLIDE topographic correction assumes that topography does not change through time, but the authors mention that in two of their high-resolution cases (Taiwan, southern New Zealand) all topography developed recently. What is the effect of presuming pre-existing topography that wasn't there?

RESPONSE: Agreed -topographic evolution is problematic, but this is an unfaced problem in all modeling of thermochron data. All thermo-kinematic models make similar assumptions. The justification is that the lowest T systems are the most sensitive to topography, and are the youngest ages, so topography at time of closure should be close to the modern for the most sensitive ages. We note that there is discussion of this in Fox et al., (2014). We can add reference to this paper at the location the reviewer mentions.

REVIEWER: Other potential model errors are discussed in passing (line 401-407) in a somewhat oversimplified way. One omitted assumption is that all thermochronometers of the same name have the same closure temperature, which is incorrect. Could their model be artificially accelerating late cooling by assuming that all apatite loses helium at the rate of Durango apatite (Tc = 70C for dT/dt = 10C/Myr; Farley 2000), as opposed to low-radiation-damage apatite (Tc = 55C; Shuster et al., 2006)? This is unexplored.

RESPONSE: Agreed – it is unexplored. We acknowledge this point and state it explicitly on line 565. It was discussed in Fox et al. (2014) and Herman et al. (2018). In fact, it is likely to be much more important than any spatial correlation effects. This could be explored at the expense of lengthening the paper, which raises other objections.

REVIEWER: A puzzling part of the back and forth is the failure of anyone to do a simple test, which is to model the data NW and SE of the Penninic Line independently. Van der

Beek et al. sort of do this in their Figure 1, albeit with synthetic data for the purpose of testing boundary condition changes. If the isolated models show no acceleration, and the combined model does show it, then that's a pretty good indication that the acceleration signal came from combining data across a major structure, presumably a red flag. If one or both of the isolated models do show acceleration, then that's an indication that the signal is at least partially independent of spatial correlation across a suspect inter- face. Fox et al. (2014) sensibly state, with appropriate caution, that their method is best used "for regional studies where the: : : exhumation rates are smooth in space and are not strongly affected by surface-breaking faults. This latter complication can be easily accounted for, where these are well-identified, by building them into the correlation structure. In such cases, samples from either side of a fault could follow independent exhumation histories." Even easier than customizing the correlation structure is just running separate models.

RESPONSE: This is an excellent idea. Which is why we did it. This experiment is described on lines 928-933 with results shown in Figure 19. It gives a conclusive result; it is a shame the reviewer seems not to have read this part of the paper.

REVIEWER: The authors are essentially claiming that there is no need to follow their own advice if they use more demanding (yet fungible) resolution limits. Even after 10 figures with synthetic model results, their case is not convincing.

RESPONSE: There is a justification given (section 4.1); we took the challenge of attempting to model the most difficult case possible: a vertical fault with 13 km of relative displacement, a feature that does not exist on Earth. If we can model this successfully, and many of the models we showed do model this successfully, the kind of smooth variation in erosion rate that is typical of most places on Earth (including the Alps) would be far easier. More importantly, the point of any modeling exercise is not to show success or failure; it is to understand why a model succeeds or fails and thereby better understand the model and better predict its behaviour elsewhere, under easier conditions.

REVIEWER: One gets the feeling that the authors simply do not want to say in so many words that Schildgen et al. (2018) are basically correct that most of the data do not support the conclusions by Herman et al. (2013). Insofar as the Herman et al. (2013) claim was based on the overwhelming weight of a global data set showing the same signal everywhere, and given that the authors now admit that the majority of those data in fact do not have the necessary resolving power, one is left to wonder why the argument needs to be so ferocious.

RESPONSE: We do not see where we admit that the majority of the data do not have the resolving power. Perhaps the reviewer could point this out and reference line numbers in our manuscript.

REVIEWER: There is potentially useful information in this paper, where the authors reassess resolution and how it applies to their original data sets, although it would be better if the tests used more realistic synthetic data in terms of both noise and comparability to the actual available data.

RESPONSE: We do not see where the reviewer has engaged with (or even read) the arguments as to why noise is omitted (to isolate errors) or how it is impossible to match original data sets (if you don't know and implement the "real" erosion rates, you cannot match the ages and the locations at the same time), so we do not find this advice particularly helpful. We have made a serious effort to make a systematic error analysis, rather than to generate a few random models with "realistic" data, which we find an unproductive methodology. This point has come up in a number of comments, so we are currently preparing a response on the general issue of the proper use of models as hypothesis tests.

REVIEWER: If the authors cut down the paper by Âż50% by getting rid of the argumentation against Schildgen et al. (2018), concentrate on a more thorough exploration of possible biases and errors in GLIDE modeling of the higher-data-density areas analyzed by Herman et al. (2013), that could help clarify how much the remaining data say.

The paper is far from that point, though. Alternatively, the paper could be published almost as-is (the errors surrounding Figure 2 really should be fixed, tests in Fig 8&13 clarified or redone, etc.) and appear together with van der Beek et al.'s responses, and everyone can just move on. I don't think the authors would be well served by this, but it would provide some closure.

RESPONSE: We have selected a long-format, open discussion, and open access journal for this manuscript to provide a comprehensive analysis of bias and uncertainties associated with the result of Herman et al. and the critique of it posed by Schildgen et al. The topic addressed in both these papers is of high prominence in the thermochronology and surface processes communities and warrants a thorough evaluation. While our manuscript is long, the length of the manuscript is allowed by this journal. Furthermore, we maintain that a manuscript of this length is needed to comprehensively advance the science on this topic. If readers are vested in this topic they will read it. We feel it is important to cover both the contrasting results of Schilgen et al and Herman et al. and present a systemic analysis of the GLIDE modeling approach. Based on the above arguments, we respectfully disagree with this reviewer's request to reduce manuscript length by 50%.

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