Earth Surf. Dynam. Discuss.,
https://doi.org/10.5194/esurf-2020-59-RC4, 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.

Anonymous Referee #3

Received and published: 22 October 2020

Revised review of Willet et al., including responses by van der Beek et al.

The contribution by Willett et al. is perplexing and disappointing. The author list is a who's who of accomplished thermal modelers, but the result is long yet incomplete, pedantic yet sloppy, and preoccupied. The paper is essentially a protracted comment on Schildgen et al. (2018), whom they want to counter on seemingly every point, down to the colors of circles on a schematic diagram. It's plainly overdone. At 43 pages of text without contributing much that is new, it comes across as an attempt to overwhelm the reader into submission. I am hoping the authors recognized, or even intended, the irony of calling van der Beek et al.'s initial 15-text-page response "long." In terms of

style, Willett et al. has much that is unacceptable. In its frequent digressions, it guesses at thought processes and motives by Schildgen et al. (2018) and Willenbring and Jerolmack (2016) instead of just sticking to the facts. It puts words into their mouths, refuting hypotheses they did not pose. It is hard to avoid the conclusion that the entire affair has become partly personal, and that this is tainting the thinking and judgement of the authors.

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. 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). 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). The overall example itself is so oversimplified as to also be confusing, and possibly self-defeating. 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. . .

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

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.

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.

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? 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.

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.

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.

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.

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. If the authors cut down the paper by »50% 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.

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