Earth Surf. Dynam. Discuss.,
https://doi.org/10.5194/esurf-2020-59-AC6, 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: 12 November 2020

There is a general issue in this debate as to the objectives, goals and hypotheses of the various papers involved. In several places in comments (ESURF-2020-59-SC1, SC2) van der Beek et al. have tried to differentiate the goals of the Schildgen et al. (2018) analysis and the Willett et al. (ESURF 2020-59) goals and analysis. For example:

COMMENT (ESURF 2020-59 SC2, suppl., line 13): 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.

Later, van der Beek et al. stated:

COMMENT (Esurf-2020-59-SC1-supplement, line 57) : In other words, 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."

These comments indicate that Schildgen et al. (2018), took as a given, that they could determine what is the true erosion rate history at any given site, compare this to Herman et al. (2013) results, thereby checking not only the accuracy of the Herman et al. predictions, but also the source of the errors. Both of these are strong claims. In general, it is not possible to "know" any given erosion rate history as this is a model quantity (see discussion in ESURF-2020-59-AC5), and to assign a source or cause to error is rarely a simple process, requiring extensive error analysis, which is why this is the focus of our paper. This raises an important aspect of our paper, the critique of Schildgen et al. (2018), in which we argued that Schildgen et al. did not present a reproducible methodology, capable of testing these hypotheses (ESURF2020-59, line 1310). Van der Beek et al. suggest that we are asking a different, methodological, question, rather than a geologic question, but we would argue that we are investigating the same question, but, we are providing the hypothesis testing that is not present in Schildgen et al. (2018). We would like to elaborate on that point here to explain why this distinction is important.

The basic scientific method involves four stages. These include: (1) observation, (2) hypothesis formulation, (3) testing of the hypothesis with new observations or experiments, and, (4) conclusion, including possible hypothesis revision.

Schildgen et al. (2018), and now van der Beek et al., made a number of observations (stage 1) regarding age distributions, tectonic processes, or previous models, and from this, they formulated a number of hypotheses (stage 2). Their hypotheses include a

direct interpretation of the data, for example if ages reflect exclusively tectonic drivers or not, but mostly their hypothesis involves statements of errors in the Herman et al analysis. For example, in the comment by van der Beek et al.:

COMMENT: (Esurf-2020-59-SC1-supplement, line 57) “A spatial correlation bias, as defined by Schildgen et al. (2018), creates spurious accelerations in erosion rates when data from areas with spatially variable exhumation histories are inappropriately combined. Although Schildgen et al. (2018) concluded that the spatial correlation bias was a common problem with the Herman et al. (2013) results, it was not the only problem discussed. Other sources of error in the Herman et al. (2013) results arose from model errors, notably (1) assuming vertical rock exhumation in regions where lateral advection of rocks is important, and (2) ignoring the impacts of changes in topography on thermochronometer age patterns; as well as from operator errors, such as (3) the inclusion of samples reheated by volcanic flows or hydrothermal fluids, and (4) the inclusion of partially reset or unreset samples from sedimentary rocks in the inversions. A detailed analysis of the potential sources of error within the GLIDE inversion procedure was outside the scope of the Schildgen et al. (2018) analysis, which focused rather on the implications and robustness of the Herman et al. (2013) results. In other words, 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.”

The numbered list in this quote from van der Beek et al. defines a set of hypotheses (Stage 2). Stage (3) of the scientific method would require a set of tests to demonstrate that these potential errors are, in fact, errors, and if so, how important they are, i.e. are they significant errors, and important enough to reverse the conclusions of Herman et al. (2013)? However Schildgen et al. (2018) contains no tests. There are no tests for the importance of non-vertical rock exhumation. There are no tests for the effects of transient topography. There are no tests for the effect of inclusion of hydrothermally-influenced ages. There are no tests for the effect of including partially

reset ages. There are no tests for any specific field sites using a modified data set, having removed questionable data. There are no tests for robustness** of the Herman et al (2013) results, which would require rerunning these models with modified data or model control parameters. The complete lack of hypothesis testing means that it is impossible to either confirm or refute any of the hypotheses made above, or to judge the validity of any conclusions in Schildgen et al. (2018). One might accept that published work had already proven the existence of some of the data errors, but most of the list above are model errors (non-vertical paths, etc.) which might have published models to demonstrate process, but are not proven to be significant for specific examples. And in all cases, the impact of each of these models or data errors on the Herman et al. GLIDE models needs to be tested, and this was not done. It is completely possible that all model and data errors are critical or are negligible; without testing, one cannot say.

Contrast this to the way that van der Beek et al. (ESURF-2020-59-SC1) responded to the realization that they made a geotherm error in Schildgen et al. Figures ED 3-6, as identified in our manuscript (Willett et al., ESURF-2020-59). They immediately made a model test and announced that even though all their ages were wrong by up to 100%, it was a negligible error given the low precision of most inferred erosion rates, and therefore all their models were still valid. If a 10 to 100% error in every individual age of a data set results in a negligible effect on a Herman et al. (2013) GLIDE model result, doesn’t it seem important to check the significance of all the data or model errors listed above? Schildgen et al. (2018) did not think so; it was sufficient to hypothesize them, assume that they were critical errors, i.e. fatal to the Herman et al. (2013) conclusions, and move directly to step (4), their own conclusions.

As we discussed (Esurf 2020-59-AC1), the test van der Beek et al. conducted for the models of Schildgen et al. Figs ED 3, 5, 6) was not valid. No error is valid for a control test, certainly not 30%, but the important point, methodologically, is that discovery of model errors or data errors (what van der Beek et al. call operator errors*) should be tested. In fact, we have no problem with the first two steps of the Schidgen et

al. (2018) analysis. It is completely appropriate to identify potential errors and model inadequacies of Herman et al. (2013). We welcome critical assessments of all our work. We would welcome future work in which the importance of these errors is tested. What we object to is a set of very strong conclusions based on hypotheses with no testing.

When we state that Schildgen et al. (2018) have no reproducible methodology, we are referring to this lack of reproducible hypothesis testing (stage 3). The three models that were presented in Schildgen et al. Figs ED 3, 5, 6) were a test, but were conducted incorrectly as described in Willett et al.(ESURF2020-59). Once these tests are removed, there is no remaining hypothesis testing and therefore no basis for any of their conclusions.

The question of reproducibility also applies to the Schildgen et al.'s (2018) assessment of a "spurious" result in Herman et al. (2013). We argued that this term is undefined and van der Beek et al. attempted to clarify this with the following comment:

COMMENT (esurf-2020-59-SC1-supplement lines 409-419): "Schildgen et al. (2018) used the term "spurious" simply in its generally accepted meaning of "false" or "fake", describing in detail why any given acceleration was deemed "spurious" for each region. In addition to inappropriate combination of data, the reasons also included models errors and operator errors that are not considered by Willett et al., such as (1) inappropriate assumptions of purely vertical exhumation in regions where lateral rock advection plays an important role (e.g., Southern Alps of New Zealand, Olympics, Apennines, Taiwan); (2) inappropriate assumptions of no change in surface morphology where such changes were shown to be critical for understanding thermochronometer age patterns (e.g., Aconquija, Fiordland, western European Alps, southern Peru, Bolivia, Coast Range); (3) inappropriate inclusion of samples that were reported to have been reheated by volcanic flows (e.g., San Juan Mountains, southern Peru) or hydrothermal fluids (Eritrea); and (4) inappropriate inclusion of partially reset or unreset data from sedimentary rocks (e.g., Taiwan, New Zealand). Some of the spurious increases may

have arisen due to a reversion to the prior in some cases, but in reality, the reason for the spurious acceleration does not matter so much as the fact that it is fake. By strictly limiting the definition of "spurious", the authors have sidestepped addressing the true extent of problems in the Herman et al. (2013) inversion results. Uniform application of a model that takes no account of changes in surface morphology, rock-exhumation pathways, or tectonic features to a global dataset that includes many data points unrelated to exhumation is bound to fail in some places. For the results presented by Herman et al. (2013), we conclude that it has failed in the majority of cases."

So the clarification states that spurious means whatever the authors want it to mean, again selecting from their list of untested hypotheses, reinforcing our argument that this is an irreproducible designation. Furthermore, by mixing data errors with model errors with, what are really model interpretations (tectonic vs climate changes or topographic transients), they render this assessment meaningless as a testable quantity. We will ignore the statements regarding "false" or "fake", as these are meaningless words in a scientific context.

Willett et al. (ESURF-2020-59) is focussed on the methodological aspects of GLIDE and its response to spatial gradients in age because this is the only aspect of the Schildgen et al. (2018) study that even attempted any hypothesis testing. It is not because we are "sidestepping" the other criticisms; it is because these criticisms are hypotheses with no testing, either in Schildgen et al. (2018) or the van der Beek et al. comments. If they are ever tested and shown to be important, we will address them, or, given valid tests, we would accept the results.

In addition to not testing the hypotheses regarding errors in the Herman et al. (2013) analysis, Schildgen et al.'s (2018) own assessment of individual site erosion rates and their cause is neither tested, nor reproducible. We challenge anyone to reproduce Figure ED 2 or Table 1. We are not able to, and assert that it is not possible. The results of Schildgen et al, (2018) are qualitative, subjective interpretations. One might agree with their interpretation or not agree with it, but it does not constitute a reproducible

result.

The conclusions of van der Beek et al.'s comment provide an appropriate summary of the logic of both Schildgen et al. (2018) and their comment

COMMENT (ESURF-2020-59-SC1, line 435) : "We also reaffirm the conclusions from Schildgen et al. (2018) that a great majority of the results reported by Herman et al. (2013) are unreliable due to a combination of spatial correlation biases, model error, and operator* error. Reversion to the prior erosion rate may have also led to spurious results in some sites, but the spatial correlation bias is likely the most common issue in areas that were not significantly affected by model errors in GLIDE (e.g., assumption of vertical exhumation pathways and assumption of no changes in topography through time) or operator errors".

We note that these conclusions are identical to their hypotheses, but without a single test in the interim between hypothesis and conclusion. Note that, even if the Schildgen et al. (Figures ED 3, 5, 6) tests were valid, they would have provided only an example that a spatial correlation bias might exist for their one selected data set and one selected erosion rate function, not that it does exist in any or "most" of the Herman et al results. The tests were designed as existence tests, not for significance in any of the natural examples. As for the other types of errors in their list, no tests were attempted. Spatial correlation bias is concluded, not only to be important, but to be "the most common" error. Even if they had established error, how can one conclude that a specific error is the most common, when there are no tests presented even for its existence, let alone frequency, and no tests for the existence, significance, or frequency of the alternative errors?

Based on the above concerns we raised over the Schildgen et al., (2018) manuscript and the comment by van der Beek et al., we stand by the results of Herman et al., (2013) and supporting analysis provided in our manuscript (ESURF2020-59). We stand by our original paper, that identified the lack of hypothesis testing and do not

see that we have misrepresented anything in our critique of Schildgen et al. (2018). We explained this argument in our paper at various points (e.g. paragraph at line 1310); we could expand this argument as we do above, but the paper is already long on criticism and we think the point is made effectively. We agree we have paid less attention to some of the secondary errors hypothesized by Schildgen et al., and we are willing to consider them further, once there are proper tests of the significance of these errors. Until then, we think the focus on the spatial correlation bias is appropriate.

*"operator errors" refer to failure to follow documented procedure, like landing an airplane in a field, when procedure calls for an airport. What van der Beek et al. describe are "data errors".

**"robust" refers to model insensitivity to small errors in the data, i.e. adding or removing some data gives the same result. What van der Beek et al. refer to is "accuracy".

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