Bias and error in modelling thermochronology data: A comment on Willett et al.

Peter van der Beek[1], Taylor Schildgen[1,2], Rasmus Thiede[3], Hugh Sinclair[4]
[1]Institute for Geoscience, Potsdam University, Potsdam, 14476, Germany
[2]GFZ German Research Centre for Geosciences, Telegrafenberg, Potsdam, 14473, Germany
[3]Institute for Geoscience, Christian Albrecht University Kiel, Kiel, 24118, Germany
[4]School of Geosciences, The University of Edinburgh, Edinburgh, EH8 9XP, UK

Correspondence to: Peter van der Beek (vanderbeek@uni-potsdam.de)

The manuscript submitted by Willett et al. shows a welcome appreciation by the authors of the potential biases, errors, and resolution in inversions of thermochronometric data using GLIDE. However, it also contains several misrepresentations of both the Herman et al. (2013) study and Schildgen et al.'s (2018) criticism of that study, and it overlooks some of the most important model and operator errors that affected the Herman et al. (2013) results. Based on additional analyses presented here, and our own critique of the material presented in Willett et al., we find that the spatial correlation bias remains a dominant bias in the results presented by Herman et al. (2013). Contrary to the suggestions by Willett et al., several of their synthetic tests demonstrate a clear spatial correlation bias, while others were carefully designed to avoid it. But, even in the tests designed to avoid spurious accelerations in erosion rates due to the spatial correlation bias, we show that these accelerations still occur, but just a few million years earlier, in time windows not shown by Willett et al. We also show that with only minor modifications to the input data of those tests, strong accelerations in erosion rates caused by spatial correlation biases emerge within the last few million years. Issues raised by Willett et al. concerning geotherm differences among models used to create and invert synthetic ages in Schildgen et al. (2018) are insignificant, and do not produce the spurious accelerations that Willett et al. claim they produce. Finally, the conclusion by Willett et al. that the results of Herman et al. (2013) remain valid is without support. We will discuss in detail below how we arrive at these conclusions. Before we do so, however, we feel it is useful to provide a short synthesis of these two preceding studies, to provide context for the current debate.

Herman et al. (2013) compiled global thermochronology data for four thermochronometers and inverted these data to generate erosion histories using the GLIDE code (Fox et al., 2014); the approach used globally uniform parameters, such as a correlation length scale of 30 km and an input background ("prior") erosion rate of 0.35 mm/yr. Herman et al. (2013) considered all predicted erosion rates, averaged over 2-Myr time-bins since 8 Ma, with a resolution (i.e., a measure of how well the erosion rate is resolved by the data at any given point in space and time, as defined by Fox et al., 2014) of > 0.25 to be "well resolved". Herman et al. (2013) found widespread accelerations in these "well resolved" erosion rates between the time bins 6-4 Ma and 2-0 Ma, and they interpreted these accelerations as being linked to a global signal of cooling climate during that time. The analysis by Herman et al. (2013) depended heavily on: (1) the ratio of 2-0 Ma / 6-4 Ma erosion rates, and (2) inclusion of all predicted rates with a resolution > 0.25. Fig. 1C of Herman et al. (2013) shows these erosion-rate ratios for all points that meet their resolution threshold in both time windows and is the basis for arguing the "worldwide" nature of accelerated erosion rates; their Fig. 2 shows a histogram and cumulative density distribution of erosion-rate ratios, as well as box plots of ratios split into latitude bins. Fig. 2 is the basis for the following assertions: "*The distribution of erosion rates shows that more than 80% of the regions with high-resolution values exhibit an increase, with an erosion rate ratio between >1 and 4 (Fig. 2). This increase is observed at all latitudes, but is more pronounced at latitudes outside the inter-tropical zone (inset of Fig. 2)"* (Herman et al., 2013; p. 424). These two figures and associated text thus constitute the core of their argument.

Disparities between the findings of Herman et al. (2013) and earlier interpretations of exhumation rates and patterns in many of the regions where they saw increases led Schildgen et al. (2018) to explore the robustness of the Herman et al. (2013) results. Several of these earlier studies employed thermo-kinematic modelling to quantitatively interpret the data, including some of our own studies (e.g., Bermúdez et al., 2011; Beucher et al., 2012; Glotzbach et al., 2011a; 2011b; Robert et al., 2009; Schildgen et al., 2009; Thiede and Ehlers, 2013) and some by co-authors of the Herman et al. (2013) paper (e.g., Ehlers et al., 2003; Fuller et al., 2006; Herman et al., 2007; 2010; 2009; Willett et al., 2003). Schildgen et al.'s (2018) approach was to examine the spatial patterns of the inferred erosion rates, compare these to the original data and mapped major structures, and review the literature to assess the original interpretations of these data. The approach was thus essentially *abductive*, which has been shown to be an efficient and even desirable logical approach in geomorphology (e.g., Baker, 1996). Schildgen et al. (2018) also developed synthetic tests to better understand what was likely driving the results of Herman et al. (2013). Schildgen et al. (2018) focused only on the predictions that Herman et al. (2013) deemed well resolved; only these points were shown on the maps of Schildgen et al. (2018).

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.

In the following, we will first focus on three main points made by Willett et al.:

1. **Model error due to variable geotherm calculation**: We show that, although the geotherm calculation method in GLIDE is based on a poor choice of boundary conditions, Willett et al.'s dismissal of Schildgen et al.'s (2018) synthetic tests based on this difference is a red herring;
2. **Bias to the prior versus the spatial correlation bias**: We show that bias to the prior erosion rate is an additional source of bias towards acceleration in the Herman et al. (2013) results, but new synthetic tests presented here imply it is less of a problem than the spatial correlation bias;
3. **Biased post-processing operator and resolution**: We argue that Willett et al.'s criticism of the use of post-processing operators and lack of regard for resolution is misdirected at Schildgen et al. (2018), should instead be directed at Herman et al. (2013), and has no impact on the Schildgen et al. (2018) analysis.

We then continue with (4) a few comments on the **spatial variability in thermochronometric ages** and (5) on the **additional field examples** presented by Willett et al. We finish (6) with comments on the **definition of "spurious"** as used by Schildgen et al. (2018). In an appendix, we provide some line-specific comments on the Discussion section of Willett et al., which is replete with mistakes and mischaracterizations of the work presented by themselves, the work of Schildgen et al. (2018) and that of Herman et al. (2013).

1 Model error due to variable geotherm calculation

Willett et al. claim that due to the differing boundary conditions between the model Schildgen et al. (2018) used to predict synthetic ages, Pecube (with a basal temperature boundary condition), and the model that performs the inversion, GLIDE (with a basal heat-flux boundary condition), "*the inversion will infer an increase in erosion rate with time in order to fit these ages*" (line 553). They use this argument to dismiss the synthetic tests presented in Schildgen et al. (2018) without further consideration or demonstration of the effect they argue exists. Willett et al. omit mentioning that the GLIDE input requires a basal temperature, not a flux. As the code is undocumented, it was assumed in Schildgen et al. (2018) that this was the temperature *at steady state,* but it appears to be the temperature for a stable geotherm (without advection).

This choice of a stable geotherm with a flux basal boundary in GLIDE, together with vertical exhumation paths, is unsatisfactory, as it implies that the thermal effects of exhumation are felt throughout the thermal lithosphere, leading to both an excessively strong disturbance of the geotherm and excessively long response times, equivalent to the thermal response time of lithosphere. In natural collisional orogens, which are the sites of most rapid erosional exhumation, depths of exhumation are generally limited to the upper and middle crust, above major mid-lower crustal detachments (e.g., Willett et al., 1993; Jamieson and Beaumont, 2013 and references therein). Material below this detachment is subducted and/or underthrusted, i.e., is advected *downward* instead of *upward*, with corresponding downward displacement of isotherms (e.g., Herman et al., 2010; Jamieson and Beaumont, 2013).

Regardless whether the boundary condition incorporated into GLIDE is appropriate, if the differing boundary conditions between GLIDE and Pecube were significant in the synthetic tests presented by Schildgen et al. (2018), as argued by Willett et al., they should lead to consistent underestimations of exhumation rates in these inversions and consistent accelerations toward the present, as the effect is stronger for higher-temperature than for lower-temperature thermochronometers (Willett et al., lines 552-555). We tested this supposition by using GLIDE to invert synthetic data produced with a forward-run of Pecube, using Schildgen et al.'s (2018) synthetic "western Alps" case also utilised by Willett et al., but with data only from the NW side of the fault, where exhumation rates are steady through time and space at 1 mm/yr (Fig. 1). In this test, we prevent any potential accelerations resulting from the spatial correlation bias by imposing spatially constant exhumation rates. If geotherm differences are important, GLIDE inversions of the Pecube-generated ages should produce accelerations. However, we predominantly see decreases in exhumation rates through time, or no change (Fig. 1). The underestimation of erosion rates results both from reverting to the prior erosion rate when this is set to < 1 mm/yr, and from the differing boundary conditions between the forward and inverse models. The effect of the geotherm alone is best assessed in the case where the prior erosion rate is equal to the input erosion rate (1 mm/yr; Fig. 1 g-i): in this case, median predicted erosion rates are 0.84 (± 0.09 at 1σ) mm/yr for the 6-4 Ma age bin and 0.72 (± 0.12) mm/yr for the 2-0 Ma age bin, i.e., they underestimate the "true" erosion rate by ~20%, which is within the range of the typical uncertainty associated with quantitative inferences of exhumation rates from thermochronology data. We thus conclude that the influence of differing boundary conditions on these synthetic tests is minor and, importantly, cannot be invoked to explain the accelerations in the synthetic tests including spatially variable exhumation rates presented by Schildgen et al. (2018).

The reason for the minor influence of the differing boundary conditions on the synthetic tests likely relates to several factors. First, we note that Willett et al. also changed other thermal input parameters with respect to Schildgen et al.'s (2018) synthetic models, including the basal temperature and the crustal heat production. We are unable to reproduce Willett et al.'s Figure 4, for instance, because the values of thermal parameters used in that calculation were not reported. Second, in contrast to Willett et al., we added realistic errors to our synthetic ages (see below), which are thus more scattered than what is shown in Willett et al.'s Figure 4. Third, the accelerations that we should see based on differing boundary conditions may be counteracted by the model error associated with the calculation of the geotherm in GLIDE described by Willett et al. (lines 739-741), which should lead to decelerations. Finally, while the effect is largest for deepest (highest-temperature) thermochronometers, these systems close earlier in time, while the transient geotherm is still cooler (see Fig. 3 of Willett et al.). The exaggerated transience of the geotherm in GLIDE thus has the effect of mitigating the differences between predicted ages in Pecube and GLIDE.

2 Bias to the prior versus spatial correlation bias

2.1 Bias to the prior versus spatial correlation bias in synthetic tests

Willett et al. argue that the problematic results presented by Herman et al. (2013) are affected by a Bayesian bias to the prior erosion rate rather than a "spatial correlation bias". This argument is partly semantic; the inversion interpolates both temporally and spatially between incomplete data, which when done incautiously, introduces both types of bias. The spatial correlation bias causes spurious accelerations in inferred erosion due to the combination of areas that experienced rapid exhumation (and hence have young ages) with slowly exhumed areas that yield older ages. The Bayesian bias to the prior, in contrast, returns the (input) prior erosion-rate estimate when the data do not constrain the erosion rate for a given time window. We illustrate the effects of these biases by a set of additional synthetic tests. We have run these tests in the same way as in Schildgen et al. (2018); as demonstrated in Section 1, the different calculation of the geotherm between the forward and the inverse model does not significantly affect these synthetic tests.

We illustrate the impact of the Bayesian prior bias alone with inversions that include data only from the NW side of the fault (Fig. 1 a-i) or only from the SE side (Fig. 1 j-l). As discussed above, when including only data from the rapidly but constantly exhuming NW side, the model returns constant rates or minor decelerations since 6 Ma (Fig. 1 a-i). The absolute rates depend on the employed prior, especially toward the edges of the model, where the resolution is lowest. A similar result (no acceleration) is obtained when only inverting data from the more slowly exhuming SE side of the fault, although in that scenario, the low resolution (< 0.25 everywhere) implies that both the 6-4 Ma and 2-0 Ma time bins largely revert to the prior erosion rate (Fig. 1 j-l).

We illustrate the combined effects of the Bayesian prior bias and the spatial correlation bias for the synthetic western Alps case by including data from both sides of the fault and varying the uniform prior erosion rate (Fig. 2). These tests show that spatially variable exhumation rates add substantial bias to the inversion results (compare Fig. 1 and. Fig. 2). The bias in Fig. 2 contains elements of the Bayesian prior bias; the two biases are impossible to disentangle in areas of spatially variable exhumation. The compounded effects of both, nevertheless, are clear: accelerations occur when data from both sides of the fault are included in the inversion, regardless of the chosen prior erosion rate, suggesting a predominance of the spatial correlation bias. Furthermore, we see that by running an inversion with data from both sides of the fault, the resolution is higher on both sides compared to the resolution found when using data from only the NW or the SE (compare resolution contours in Figs. 1 and 2). Thus, data are being combined from both sides to "better" constrain the exhumation history of each side, and that more highly resolved result produces the spurious increase we call the spatial correlation bias.

2.2 Spatial correlation bias in the synthetic tests presented by Willett et al.

At this point, we can ask why our synthetic results are so different from those of Willett et al. The answer to this question is twofold: removal of the spurious erosion-rate increase in Willett et al.'s inversions is achieved using spatially variable prior erosion rates and thus independent *a-priori* knowledge of this spatial distribution (their Fig. 8), the careful design of input data that have an idealized spatial and temporal distribution (their Figs. 6, 7, 9, 14 and 15), or both (their Fig. 13). The models of Willett et al. with a more realistic data distribution (i.e. Set C; their Figs. 10-12) show exactly the accelerations we expect (partly counteracted in Fig. 12 by the reversion to a very high prior erosion rate), and illustrate the dominance of the spatial correlation bias over the reversion to the prior bias similar to our Fig. 2. Although Willett et al. claim that their dataset A (and D, which has the same data distribution) "*roughly corresponds in pattern to the Alpine data set*" (line 581), it is in fact carefully designed to produce the desired result. In particular:

- Datasets A/D contain five thermochronological systems (and Sets B/E contain four), whereas the real Alps dataset only includes three systems and is dominated by apatite and zircon fission-track ages (290 out of 309 data, the remaining 19 being apatite (U-Th)/He ages). Most noticeably, 30% of the ages in Willett et al.'s Sets A/D (and 22% in Sets B/E) are mica $^{40}Ar/^{39}Ar$ ages. Neither Herman et al. (2013) nor Fox et al. (2016) used any mica $^{40}Ar/^{39}Ar$ cooling ages in their inversions, for the simple reason that these do not exist. The few $^{40}Ar/^{39}Ar$ dates available for this part of the Alps are crystallisation ages for minerals in fault zones (e.g., Egli et al., 2017; Rolland et al., 2008; Rossi and Rolland, 2014), and thus not representative of regional cooling related to exhumation.
- The datasets by Willett et al. are characterised by a very high number of co-located data (120 out of 176 data locations in Sets A/D combine two thermochronometers, in general associated with mica $^{40}Ar/^{39}Ar$; 15 have three or more thermochronometers). In the real dataset, in contrast, only 48 out of 251 data locations have two thermochronometers (all are apatite and zircon fission-track), and five locations have three (with additional apatite (U-Th)/He).
- All datasets by Willett et al. consist of "perfect" ages, as predicted by the forward model. In contrast, we added a random scatter of up to 10% to our synthetic ages in order to better reflect imperfect natural data.
- Sample locations in Sets A/D are more heavily weighted toward high elevations than the real data: 38% of the data from the NW zone are from locations >2000 m, whereas only 28% of the real data are from such high elevations. The data locations in these datasets also have a much wider spatial spread than the real data.

To demonstrate the dependence of Willett et al.'s synthetic tests on the idealized data input, we reran the inversion of their Fig. 14 (Set D) both with and without the hypothetical $^{40}Ar/^{39}Ar$ data (i.e., addressing only the first bullet point in the above list). Figure 3 shows our results: first, we reran the exact same inversion of dataset D using our version of GLIDE to make sure we obtained the same results; we do (compare Fig. 3 a-c with Willett et al.'s Fig. 14 d, a, g). In particular, the inversion strongly overestimates the erosion rate SE of the fault in the 6-4 Ma time bin, leading to a significant deceleration after this time, centred on the SE half of the model. However, Willett et al.'s explanation for this deceleration as being due to the sluggish transient response of the geotherm to exhumation rates that are higher than the prior rate (lines 735-745) can only apply to the rapidly exhuming NW side of the fault; that explanation cannot be applied to the slowly exhuming SE side of the model, where the decelerations are largest. Willett et al. did not show what happens in the earlier time bins of this test; Fig. 3 d-f shows that between the 10-8 Ma and 6-4 Ma time bins, predicted erosion rates show a significant spurious increase, both to the NW and to the SE of the fault. In fact, the addition of the high-temperature mica $^{40}Ar/^{39}Ar$ data (with older ages) has simply pushed the acceleration due to the spatial correlation bias back to these earlier time bins.

To demonstrate this effect, we removed the mica $^{40}Ar/^{39}Ar$ data from Set D and reran the inversion; the result shows major accelerations from 6-4 Ma to 2-0 Ma on both the NW and SE sides of the fault (Fig. 3 g-i), similar to our previous test shown in Figure 2. Interestingly, removal of the $^{40}Ar/^{39}Ar$ data from Set D results in no substantial change in the resolution for the 6-4 Ma or 2-0 Ma time bins (compare resolution contours in Fig. 3 g-h and Fig. 3 a-b). The acceleration is again independent of the chosen prior erosion rate, as shown by our third test, in which we increased the prior erosion rate to 1 mm/yr (Fig. 3 j-l). Thus, the acceleration is predominantly due to a spatial correlation bias.

To show that this behaviour is independent of the chosen geotherm model, we repeated this test using Willett et al.'s dataset A and the version of GLIDE with a stable geotherm. The results, shown in Figure 4, are similar to those discussed above for dataset D and a transient geotherm. In this case, erosion rates are constant between the 6-4 Ma and the 2-0 Ma time bins (Fig. 4 a-c, compare to Willett et al.'s Fig. 6) but there is a similar spurious increase earlier in the history (Fig. 4 d-f). Like the case with Set D, when removing the high-temperature mica $^{40}Ar/^{39}Ar$ data from Set A, this spurious acceleration occurs in the recent time bins (between 6-4 Ma and 2-0 Ma) despite no substantial loss of resolution in these time bins (Fig. 4 g-l). Thus, for both the stable (Fig. 4) and transient (Fig. 3) geotherm cases, even Willett et al.'s most highly temporally resolved dataset (A/D), with the densest data in terms of both spatial and temporal coverage, leads to spurious accelerations due to the spatial correlation bias, either somewhat earlier in the exhumation history than what Willett et al. showed, or in the most recent time bins after removal of the mica $^{40}Ar/^{39}Ar$ ages.

We agree with Willett et al. that temporal coverage of the data is important; the above tests show this dependence. However, the tests also show how, in the absence of temporal coverage in a particular time bin, information can be brought in from surrounding regions (creating a spatial correlation bias) rather than locally reverting to the prior. We acknowledge that the synthetic data in our tests have reduced temporal coverage compared to the real data. However, creating synthetic datasets that do not attempt to recreate the full age range of real datasets, but rather faithfully recreate real data distributions and the available thermochronometers, is fully justified. As Willett et al. note, "*the resolving capability of data is determined by the complexity of the erosion rate field being sampled, so resolution should be evaluated in the context of tectonic variability"* (lines 771-773). For a synthetic site with a very simplified exhumation history (i.e., a steady exhumation rate through time), one would not expect the same distribution of ages that are obtained from a real site that has experienced a far more complicated exhumation history. By maintaining the full range of ages exhibited at a real site with a more complicated history, Willett et al. are providing their inversions with far greater temporal coverage than could ever be reasonably expected to constrain a simple exhumation history. Yet, even in this case, spatial correlation biases occur, only at different times in the history.

We finally note that the western Alps dataset is one of the densest in the world, which is why we decided to explore this case in our synthetic test; if the inversion fails here (and our tests show that it does), it is likely to fail elsewhere. We also note that a relatively high resolution is no guarantee of an unbiased result, as our Figures 2 – 4 show spurious accelerations in erosion rates that have a resolution of > 0.5.

2.3 Bias in reversions to the prior in the Herman et al. (2013) results

Willett et al. argue that, when "reversion to the prior" occurs, such a result will not have a "*generalizable tendency toward acceleration*" (lines 1128-1129). Although we agree with this general statement, in the application of GLIDE by Herman et al. (2013), reversions to the prior erosion rate will likely create spurious accelerations. This is because, as originally argued by Willenbring and Jerolmack (2016), all of the "resolved" regions include sites of rapid exhumation and therefore young thermochronological ages, which will tend to increase resolution in the most recent time bin. To demonstrate this point, we illustrate the distribution of resolution values from the results of Herman et al. (2013) in two different time bins, 2-0 Ma and 6-4 Ma (Fig. 5 a, b) and the distribution of erosion rates in those time bins (Fig. 5 c, d). These plots illustrate, first, the tendency of the data to be better resolved in the 2-0 Ma time bin (median resolution of 0.48) compared to the 6-4 Ma time bin (median of 0.33). Moreover, 75% of all data points in the 6-4 Ma time bin have a resolution lower than 0.4, whereas only 32% of data points in the 2-0 Ma time bin have a resolution lower than 0.4. For co-located points, 90% show a higher resolution in the 2-0 Ma time bin compared to the 6-4 Ma time bin (Fig. 5 b). Hence, reversion to the prior, which Willett et al. suggest affects most of the results with resolution below 0.4, will much more likely affect points in the 6-4 Ma time bin than in the 2-0 Ma time bin. Moreover, most of the erosion rates in the 2-0 Ma time bin are higher than the prior of 0.35 mm/yr: the median erosion rate in the 2-0 Ma time bin is 0.48 mm/yr, and 80% of the erosion rates in the 2-0 Ma time bin have an erosion rate higher than 0.35 mm/yr (Fig. 5 c, d). Given the relatively high erosion rates that characterize the 2-0 Ma time bin, reversion to the prior in the older time bin will commonly produce spurious accelerations in exhumation.

The test by Herman et al. (2013) to explore the effect of the prior on the global compilation is an interesting counterexample to the argument by Willett et al. that most of the lower resolution results are affected by this reversion to the prior. Herman et al. (2013) reported that when choosing a prior erosion rate of 0.7 or 1.0 mm/yr (the latter of which is higher than many of the inferred erosion rates in the 2-0 Ma time bin), they still see a predominance of accelerations in their inversions. Whereas Herman et al. (2013) used this result to argue for the robustness of their conclusions, we argue instead that it illustrates that Herman et al.'s (2013) results are not dominated by a Bayesian prior bias, but rather they are predominantly affected by a spatial correlation bias, which creates spurious accelerations, but is less affected by the choice of the prior.

Our synthetic tests (Figs. 1 and 2) show that bias to the prior exists, but in most cases, it is overpowered by the spatial correlation bias. The fact that the results from Herman et al. (2013) are insensitive to the prior erosion rate is not a sign of robustness of the results, but rather the pervasiveness of the spatial correlation bias.

3 Biased post-processing operator and resolution

Willett et al.'s insistence that it is the post-processing operator, i.e. plotting normalized erosion-rate differences, rather than the inversion that creates the bias, is difficult to understand. Increased erosion rates through time appear as such, whether they are visualized from direct comparison of time-bin maps or plotted as differences, ratios, or normalized differences. The magnitude of the metric changes in each case, but the sign (positive or negative) does not. In contrast to the analysis of Herman et al. (2013), the interpretations of Schildgen et al. (2018) focused on the sign of the change, not its magnitude, and on whether that change is reasonable considering (1) the spatial and temporal (age) distribution of the data, and (2) previous, more detailed analyses that include independent geologic data and more appropriate modelling of the data (e.g., with changing topography and/or lateral components of rock advection, and exclusion of data unrelated to exhumation). The ratios of erosion rates used by Herman et al. (2013), namely the erosion rate from the more recent time bin divided by that of the earlier time bin, are problematic, as the values tend to blow up when the earlier erosion rate is small. Although the normalized difference used by Schildgen et al. (2018) is arguably also imperfect, it has the benefit of tracking fractional changes in exhumation rates (i.e., a change from 0.5 to 1.0 mm/yr or from 0.05 to 0.1 mm/yr both result in a value of 0.5), and being symmetric (i.e., changes in erosion rates from 1.0 to 0.5 mm/yr or from 0.5 to 1.0 mm/yr yield respective normalized differences of -0.5 and 0.5). We note that Willett et al. (1) make a mistake in their description of the normalised difference NR (their Eq. 14 returns a negative number in case of an acceleration); and (2), more critically, in their analysis of the resolution of NR, they modify the definition of NR to maximise its value and minimise its resolution by dividing by e_2 instead of max(e_1, e_2), as done by Schildgen et al. (2018) (line 467 of Willett et al.; note that in their definition $e_1 > e_2$ in case of an acceleration). By altering the definition of the NR in their error analysis, Willett et al. are analysing a metric that mixes Herman et al.'s (2013) original ratio and Schildgen et al.'s (2018) normalised difference.

The absolute differences now advocated by Willett et al. are problematic in their own right, as areas with high erosion rates tend to dominate any global "signal". Moreover, Willett et al. report mean values from their histograms of erosion-rate differences, neglecting to take into account how the dominance of extreme values is exacerbated by the use of the mean instead of the median as a measure of central tendency in these positively skewed distributions (their Fig. 24). For example, the median difference for a resolution cutoff of 0.5 is 0.42 mm/yr, whereas the mean is 0.65 mm/yr (Fig. 24c in Willett et al.). We can further illustrate the problem with absolute differences by considering the impact of removing one region of rapid erosion rates from the compilation. Again using a resolution cut-off of 0.5, after excluding results from Taiwan, where extreme reported increases in erosion rates are related to the inclusion of unreset thermochronometer data in the inversion (Schildgen et al., 2018), the mean drops from 0.65 to 0.37 mm/yr and the median drops from 0.42 to 0.39 mm/yr. However, these numbers still cannot be taken at face value, as the results still suffer from spatial correlation biases, reversion-to-prior biases, model error, and operator error.

Oddly, Willett et al. claim that the analysis of Schildgen et al. (2018) was focused on areas of poorly resolved results and that Schildgen et al. (2018) "never address resolution". We reiterate from Schildgen et al. (2018) that the analysed results were those that passed the threshold defined by Herman et al. (2013) as "well resolved", i.e., with a resolution > 0.25, and were used by these authors to support a "worldwide increase in erosion rates". Willett et al. show a welcome new appreciation for the importance of better resolved results (lines 1000-1003), but the repeated suggestion that Schildgen et al. (2018) misdirected attention to areas of poorly resolved results is unfounded. The spurious increases documented in Schildgen et al. (2018) and here (Figs. 2 – 4) comprise from best to least resolved areas, and all are above the "well resolved" threshold of 0.25 used in Herman et al. (2013). Nevertheless, we agree that focusing on more highly resolved results is better practice, although doing so does not eliminate spurious increases (Schildgen et al., 2018 and Section 2 above). The implications of the selected cut-off resolution value are substantial: as noted in Schildgen et al. (2018), increasing the resolution threshold to 0.5, which characterizes the "well resolved" region in the Alps that Willett et al. prefer to focus on, would eliminate 90% of the "resolved" erosion ratios reported by Herman et al. (2013), and would comprise data from only seven distinct regions, as shown in Willett et al.'s Fig. 25. One of those regions (central Himalaya) comprises a single resolved point that shows a decrease in exhumation rates through time, not an increase; five regions (Wasatch Mountains, Western Alps, Northern Apennines, Taiwan, Fiordland) suffer from spatial correlation bias, model or operator errors and sometimes a combination of these, as discussed in the Supplementary Information of Schildgen et al. (2018); in three regions (Coast Mountains of British Columbia, External Massifs of Western Alps, Fiordland) glacial valley incision has been previously inferred (Shuster et al., 2005; 2011; Valla et al., 2011; see below). Restriction of the results to these few locations precludes any attempts at generalization to a global scale.

4 Spatial variability of thermochronometer ages

Willett et al. expend considerable effort in critiquing cartoons by Schildgen et al. (2018) that have no vertical or age scale, and only a rudimentary horizontal scale (lines 793-831; 1320-1322 in Willett et al.). We do not see much merit in this discussion, but note that Willett et al. appear to confuse *isotherms* (surfaces of equal temperature, which are sketched into the cartoons of their Figure 16 a-c) and *isochrones* (surfaces of equal thermochronologic age, which were drawn in the cartoons of Schildgen et al. (2018), reproduced in Willett et al.'s Figure 16 d-f). The point that Willett et al. appear to be trying to make is that thermochronologic ages will be spatially constant over length scales larger than the correlation length-scale used in the GLIDE inversions in most tectonic settings. To assess this point, we illustrate some original data that inspired the cartoons (Fig. 6). For the Wasatch Mountains (Fig. 6a), the AHe and ZFT ages shown in Ehlers et al. (2003) increase steadily with distance from the range-bounding Wasatch Fault, whereas AFT ages show only a slight increase until 17 km from the fault, where they start increasing more rapidly with distance. Likewise, in the Southern Alps of New Zealand (Fig. 6b), all thermochronometer ages increase rapidly with distance from the Alpine Fault, starting from distances of 10-20 km from the fault (e.g., Herman et al., 2009). Importantly, in both these cases, the ages show significant variation over length scales of ~30 km, which was the correlation length scale used in Herman et al. (2013), making these settings prime examples of where that inversion was affected by the spatial correlation bias.

5 Comments on natural examples

5.1 European Alps

Willett et al. focus much of their analysis and discussion on the European Alps, presumably because, as they claim, the Alps "*play an important role in the study of Schildgen et al. (2018)*" (line 863). We consider the European Alps to be no more important than any of the other 32 locations where Herman et al. (2013) reported resolved changes in late-Cenozoic erosion rates. Schildgen et al. (2018) opted to highlight the Alps as one of three examples in the main text due to the very high density of available data, which gives GLIDE the best chance of performing well. Although Willett et al. have gone to great lengths to demonstrate the reality of the erosion-rate increase in the western external Alps, they neglect to consider that increased erosion in the External Crystalline Massifs, to the north and west of the Penninic thrust front, is limited to localized valley incision, as has been demonstrated with both apatite $^4He/^3He$ studies (Valla et al., 2011) and detailed thermo-kinematic modelling that includes a temporally evolving topography (Glotzbach et al., 2011b). The latter study also demonstrated how, when steady-state topography was assumed, the pattern could be mistaken for a generalized increase in exhumation rate, as happens in the GLIDE inversions. This difference is not trivial, as presuming a regional increase in exhumation rather than localized valley incision has major implications for sediment flux to the oceans, carbon cycle impacts, and landscape evolution.

Incidentally, on lines 939-940, Willett et al. state "*Finally, we address the geologic evidence of Schildgen et al. (2018)'s hypothesis that the external and internal Alps are separated by an active normal fault along the Penninic Line.*" Although Willett et al. delve into considerable detail in the following paragraph to argue against this hypothesis, Schildgen et al. (2018) never suggested this. Willett et al. appear to have misunderstood the aim of the synthetic test presented in Schildgen et al. (2018): it was used simply to test for the occurrence of a spatial correlation bias across a densely sampled and strong gradient in thermochronologic ages, not to attempt a realistic simulation of the European Alps.

5.2 Nanga Parbat

It is difficult to assess the inversions Willett et al. present for Nanga Parbat, because they do not discriminate between ages inside and outside the massif. All apatite and zircon fission-track ages within the massif are < 3.4 Ma (Treloar et al., 2000; Zeitler, 1985; Zeitler et al., 1982), and all such ages outside the massif are > 3 Ma (see the Supplementary Information of Schildgen et al., 2018). Willett et al. now add mica $^{40}Ar/^{39}Ar$ data, but it is unclear if Herman et al. (2013) used those data in their inversions. Herman et al. (2013) did not report any such data, and the resolution values they report for Nanga Parbat are considerably lower than the resolution values shown by Willett et al. (e.g., maximum resolution in the 6-4 Ma time bin of 0.4, rather than ca. 0.6 shown in Willett et al.; and maximum values in the 4-2 Ma time bin of 0.5, rather than ca. 0.7 shown in Willet et al.), implying that the inversion presented in Herman et al. (2013) used less data. Mica $^{40}Ar/^{39}Ar$ data from the core of the massif are ≤ 4 Ma, with one exception along the Indus River (Treloar et al., 2000; Zeitler et al., 2001); older ages are only encountered within the massif-bounding shear zones. Mica $^{40}Ar/^{39}Ar$ ages outside the massif are > 10 Ma without exception, and they are mostly > 20 Ma. Moreover, careful interpretation of mica $^{40}Ar/^{39}Ar$ data from Nanga Parbat is required, as excess Ar is a commonly reported problem, and several reported ages are crystallisation ages rather than cooling ages (Schneider et al., 2001). If any of these complications in the data were considered by Willet et al., they are not reported, raising the possibility of operator error.

Despite there being no information from in-situ data within the massif prior to 4 Ma, and there being no ages < 4 Ma outside the massif, the GLIDE solution shows reasonably resolved moderate erosion rates within the massif (< 0.8 km/Myr; resolution ~0.5) prior to 4 Ma and rapid recent rates "bleeding" outside the massif since 2 Ma (Willet et al.'s Fig. 20). Both inside and outside the massif, the inversion predicts large increases in erosion rate with time, which were included in the "worldwide pattern" of Herman et al. (2013). This example provides one of the clearest instances of the spatial correlation bias, as data are combined across major massif-bounding faults that are generally considered to be active (see Butler, 2019 for a recent review).

It can be easily shown that the inferred exhumation history inside the massif presented by Willett et al. is erroneous, as U-Pb ages as young as < 2 Ma on metamorphic monazite and granite dikes imply much greater exhumation within the last 2 Myr than the < 8 km predicted by the GLIDE inversion (Zeitler et al., 2001 and references therein; Crowley et al., 2009; Butler, 2019). Moreover, < 4 Ma granites currently outcropping within the massif solidified at ~700 °C and 350-500 MPa (Crowley et al., 2009), or 13.0-18.5 km depth (assuming a crustal density of 2750 kg m^{-3}), which is significantly higher than the < 10 km exhumation predicted by the GLIDE inversion since 4 Ma.

5.3 Olympic Mountains

The Olympic Mountains are a prime example of an orogenic wedge with curved particle paths (Willett et al.'s Fig. 16c). Any 1D analysis of such a system will infer a recent acceleration in exhumation; this acceleration is real, because for a constant particle velocity, as the particle path becomes more vertical closer to the surface, exhumation rates increase. However, this increased exhumation rate is not associated with an increased erosion rate at the surface, which is an important distinction when considering the possibility of climatically triggered erosion-rate increases. Thus, when considering the implications for surface erosion rates through time, we consider this increase to be spurious in the analysis of Herman et al. (2013), because it assumes vertical exhumation pathways. Only models that incorporate curved particle pathways will potentially be able to distinguish between changes in exhumation rates related to the exhumation pathway versus those related to changes in surface erosion rates. The western flank of the Olympic Mountains has been glaciated and deep glacial valleys have been carved into the landscape (Montgomery, 2002; Adams and Ehlers, 2017). Thus, it is possible that samples from valley bottoms on the western flank show an influence of valley incision, but this effect can only be assessed with models that incorporate both realistic kinematics and changes in landscape morphology.

5.4 Marlborough region of New Zealand

The interpretation of thermochronological data in the Marlborough region by Herman et al. (2013) is even more problematic than described in Schildgen et al. (2018). Although Willett et al. have pointed to the co-located thermochronometers close to the Alpine Fault as evidence of increasing exhumation rates for a few data locations, they dismiss the increases inferred by the model elsewhere as "non-resolved", even though these were included in the analysis of Herman et al. (2013). The zircon fission-track ages for the points highlighted by Willett et al. in close proximity to the Alpine Fault are < 6 Ma, whereas just a few km away, having crossed no major intervening structure, such ages jump to > 70 Ma (see Fig. S16 in Schildgen et al., 2018). In addition, several of the co-located apatite fission-track ages are reported as 0 Ma (Tippett and Kamp, 1993), and Herman et al. (2013) did not explain how they addressed such ages in their inversion. The clear implication is that most of the zircon fission-track ages from the sedimentary rocks in this region are unreset or only partially reset, and the young ages found only in close proximity to major mapped structures imply strong tilting of the individual fault blocks and/or local reheating due to hydrothermal fluid flow along the Alpine Fault. Although increasing the resolution threshold in this region would eliminate many of the clearly spurious erosion-rate increases illustrated in Fig. S16 of Schildgen et al. (2018), it will not eliminate the operator error associated with the inclusion of reheated or unreset samples.

5.5 Fiordland

Schildgen et al. (2018) argued that some of the well-resolved erosion-rate increases in Fiordland reported in Herman et al. (2013) are probably real, and could be linked to glacial valley incision (Shuster et al., 2011). However, like in other glaciated terrains, mistaking local valley incision for a regional increase in exhumation rates is a recurring issue with models like GLIDE, which do not consider modifications in surface morphology, and subsequently vastly overestimate the regional impact of thermochronometer age patterns controlled by localized valley incision. Apart from the clear localized influence of glaciers on valley incision, regional spatio-temporal patterns of exhumation have been argued to be linked to the evolving subduction zone (Sutherland et al., 2009; Jiao et al., 2017), an argument that was detailed in the Supplementary Information of Schildgen et al. (2018). Nevertheless, the largest increases reported by Herman et al. (2013) are spurious increases to the SE of the main range, which Willett et al. now consider insufficiently resolved.

6 On the definition of "spurious"

On lines 964-966, Willett et al. state "*According to the re-analysis of Schildgen et al. (2018), of the 32 sites identified in the Herman et al. (2013) study as showing sufficient thermochronometric data to resolve an erosion rate history over the past 6 Ma, 23 of them were what they called "spurious", meaning that they arose as a result of inappropriate spatial averaging of age data.*" In contrast, 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.

Conclusions

We have shown that the issues raised by Willett et al. in their criticism of the work by Schildgen et al. (2018) are either insignificant or unfounded. By reproducing the inversions that Willett et al. reported, and plotting results from time windows that they did not include, we have shown that (1) the spatial correlation bias is a common problem in the inversions shown by Willett et al., even in those designed in an attempt to avoid it; and (2) the effects of differing boundary conditions between Pecube and GLIDE, and consequent differences in the assumed geotherm, are insignificant when comparing predicted ages from the former and inversion results from the latter. Our use of the synthetic data produced by Willett et al. in new synthetic tests presented here show a spatial correlation bias similar to that shown previously in Schildgen et al. (2018), demonstrating that Willet et al.'s dismissal of those earlier synthetic tests is unfounded. Other issues raised by Willett et al. concerning post-processing operators and critiques of cartoons are irrelevant with regards to the Schildgen et al. (2018) analysis.

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 (e.g., inclusion of data unrelated to exhumation, such as from samples reheated by hydrothermal fluids or volcanic flows, and inclusion of unreset or partially reset ages from sedimentary rocks). The small number of remaining regions where results from Herman et al. (2013) may be reliable are inadequate for any conclusions regarding the impact of late-Cenozoic cooling on worldwide erosion rates. We are in full agreement with Willett et al. that a resolution cut-off value much higher than the 0.25 value used by Herman et al. (2013) will lead to better results, but we have also shown that even the cut-off resolution value of 0.5 suggested by Willett et al. is insufficient to avoid the biases that we demonstrate.

Appendix: Inaccuracies in the Discussion by Willett et al.

The Discussion section by Willett et al. distorts much of what was presented in Schildgen et al. (2018) and in Herman et al. (2013). Although we consider the following comments somewhat minor relative to the main issues we raise earlier, we highlight below several inaccurate points for the sake of completeness.

Willett et al. claim to have identified all sources of bias and error in their model, and conclude that they are either unimportant or do no create a tendency toward acceleration (lines 1120-1129). However, Willett et al. the authors have neglected to discuss the implications of the most important model errors as applied to several field settings, namely the assumption of vertical rock-exhumation pathways and no change in topography in the inversions. They also appear unconcerned with operator error, which takes the form of the inclusion of inappropriate data (e.g., samples reheated by volcanic flows or hydrothermal fluids, and unreset data from sedimentary rocks) for several field sites in the Herman et al. (2013) analysis. Willett et al.'s analysis instead misdirects readers toward trivial issues like geotherm differences, metrics used to illustrate erosion-rate changes, and cartoons. Willett et al. also neglected to consider how a reversion to the prior may constitute an additional bias, particularly when considering the analysis of Herman et al. (2013), as we illustrate in section 2.2, and they failed to recognize the spatial correlation bias in their own synthetic tests.

A.1 Do spatial correlation biases occur?

Willett et al. note that "*The idea that spatial differences in age, i.e., a combination of old and young ages from distinct regions, will always, or even frequently, combine to produce an apparent increase in erosion rate is false. Models in this paper were consistent in demonstrating this point*" (lines 1147-1149). These statements are odd for several reasons. First, it is certainly possible to mistakenly combine data from regions with distinct exhumation histories to produce a spurious acceleration and, as we have pointed out, many of the synthetic tests presented by Willett et al. show a spatial correlation bias. The tests that do not show the spatial correlation bias were specifically designed to avoid it, at least in the time bins Willett et al. chose to report, and/or they were run in modes that are unrelated to the application in Herman et al. (2013), such as by setting spatially variable prior erosion rates. But even in the tests designed to avoid the spatial correlation bias through highly temporally resolved datasets, spatial correlation biases occur; they simply occur earlier in time (Figs. 3, 4). The synthetic tests presented by Schildgen et al. (2018) and here (Figs. 2 – 4) further illustrate the common occurrence of the spatial correlation bias when the inversion is applied to realistic datasets with a setup equivalent to that applied by Herman et al. (2013). A model that combines real data based on a predefined correlation length, without regard for tectonic structure, will suffer from this bias whenever (1) there are insufficient data from a single tectonic block to fully constrain the exhumation history, and data from an adjacent block, exhuming at a different rate, are available, or where (2) blocks are tilted such that exhumation rates and/or depths vary across them.

Willett et al. continue to claim: "*The argument that spatial variation maps into temporal variation was based on an intuitive argument (Figure 2) that was never tested. The reason why this argument fails is that there is no temperature history that can fit multiple data that have the same closure temperature, but different ages*" (lines 1149-1151). Both statements are false. Regarding the first statement, Schildgen et al. (2018) tested and demonstrated the spatial correlation bias for several realistic field scenarios, but Willett et al. dismissed those tests because they inferred that differences in the boundary conditions of the thermal model used to predict the ages (Pecube) and the model used to invert the ages (GLIDE) creates the spurious accelerations. However, this predicted effect is insignificant, based on the tests we present here (Figs. 1, 3 and 4). Therefore, Willett et al.'s dismissal of Schildgen et al.'s (2018) synthetic tests that demonstrate the spatial correlation bias is unwarranted, as is the statement that Schildgen et al.'s (2018) argument was "never tested". If the second statement were true, then the use of age-elevation profiles to infer exhumation histories would be impossible. In reality, it is impossible to fit multiple ages with the same closure temperature only if those samples are found at identical elevations (or more precisely, identical distances travelled since closure). Given that samples are rarely reported from identical elevations, and that elevation uncertainties are typically of the order of several tens of meters in any case, this limit to the feasibility of finding solutions from regions experiencing differing exhumation histories is not nearly as restrictive as Willett et al. imply.

Willett et al. argue that with a single thermochronometer, it can be very difficult to resolve a temperature history. They state "*This is why all sites identified by Herman et al. (2013) as having sufficient resolution, have ages from more than one thermochronometer*" (lines 1193-1194). This statement is incorrect. Five out of the 32 sites deemed to show sufficient resolution by Herman et al. (2013) included data from only a single thermochronometer. Sites that comprised only apatite fission-track data include Aconquija, the Mérida Andes, the Kyrgyz Tien Shan and the western Pamir, and only apatite (U-Th)/He data were included in the inversion for southern Peru.

A.2 The "Chicken or Egg" debate

Willett et al. state: "*we have established that there are no spurious accelerations in erosion, only genuine ones*" (line 1227). While intriguing, this statement is wholly unsupported, and also contradicted by the many times that the authors emphasize how the resolution cut-off value used by Herman et al. (2013) was inappropriate. Willett et al.'s conclusion that a resolution cut-off of 0.5 is more appropriate would eliminate 90% of the results reported by Herman et al. (2013), which includes 25 out of the 32 "resolved" locations of exhumation-rate changes in the late Cenozoic. Even this higher cut-off does not address the spatial correlation biases, model errors, and operator errors that compromise the results presented in Herman et al. (2013). Thus, "genuine" accelerations in erosion rate characterise only a small minority of the cases put forward by Herman et al. (2013); to be precise, Schildgen et al. (2018) argued that "genuine" accelerations in erosion were identified in seven out of 32 regions (which are not the same as those with resolution > 0.5).

Willett et al. note that "*Given the target timeframe of the last 6 Myr, young ages are needed and this gives a bias … toward high erosion rates, but it does not follow that this leads to a bias towards recent acceleration*" (lines 1249-1251). As we argued in section 2.3, a bias does follow, if resolution is better in the most recent time bin, and a prior erosion-rate value is selected that is lower than the median erosion rate in that most recent time bin. We believe that such a reversion-to-the-prior bias is still dwarfed by the spatial correlation bias based on our analyses and those of Herman et al. (2013), but without being able to examine the results of the tests that Herman et al. (2013) performed with alternative prior values in detail, it is difficult to make a conclusive statement in this regard, as the two types of biases are intertwined.

Willett et al. suggest on lines 1253-1254 that in Schildgen et al. (2018), "*complicating factors including the difficulty of establishing cause and effect in a system with feedback were not discussed*." In fact, Schildgen et al. (2018) discussed this difficulty in detail regarding the St. Elias range in Alaska and the Kyrgyz Tien Shan, both of which were inferred to show accelerations that can be linked to changes in climate and/or tectonics. For several of the other sites where Schildgen et al. (2018) concluded tectonics was the main driver for increases in erosion rates, the rationale behind that interpretation, which is largely based on the detailed studies of the authors who originally published the data, was explained. These original studies often included more sophisticated 2D or 3D thermo-kinematic modelling of the data and independent geological support. But, in many of the locations with purported increases in exhumation according to Herman et al. (2013), there is no climate versus tectonics debate either because the accelerations noted by Herman et al. (2013) are erroneous, or because the accelerations appear real, but are very localized and limited to individual fault blocks that were exhumed due to local fault-geometry and stress-field configurations.

Willett et al.'s claim that in the literature review of Schildgen et al. (2018), "*the approach used was to search recent literature for evidence of active tectonics and if found, they attributed not just young ages, but also recent acceleration, to tectonics*" (lines 1255-1256) is an oversimplification and mischaracterization of the detailed analysis presented in the Supplementary Information of Schildgen et al. (2018). Willett et al. accuse Schildgen et al. (2018) of a "*confirmation bias*" (lines 1310-1339) that takes the form of neglecting to discuss the difficulty of distinguishing between tectonic and climatic forcing of exhumation in a landscape, and neglecting to discuss a number of papers that purportedly contradict the interpretations of the causes of changes in exhumation rates. However, the examples the authors give for this bias (from comments we refer to above and next) are inaccurate portrayals of what is in Schildgen et al. (2018).

Willett et al. state: "*In addition, although many previous studies using a variety of other interpretation methods found results that support Herman et al. (2013) (e.g., Zeitler et al., 1982; Ehlers et al., 2006; Thiede and Ehlers, 2013; Michel et al., 2018; Vernon et al., 2008; Shuster et al., 2005; Sutherland et al., 2009; Thomson et al., 2010a,b; Avdeev et al., 2011; Shuster et al., 2011; Ballato et al., 2015; Bracciali et al., 2016), none of these studies swayed an interpretation away from their "spurious" assessment*" (lines 1326-1330). Although we would not claim that Schildgen et al. (2018) cited every relevant paper in a world-spanning, but abbreviated review of 195 papers, several of these papers mentioned above by Willett et al. were indeed cited and discussed by Schildgen et al. (2018), and those interpretations were used to infer spurious, tectonic or glacial causes of increases. To give just a few examples, Ballato et al. (2015) was cited as evidence for a tectonically driven increase in exhumation in the Alborz mountains; Avdeev and Niemi (2011) was cited to support the interpretation of a tectonic driver for uplift in the Greater Caucasus; Shuster et al. (2011) was cited as evidence for localized glacial incision in Fiordland; and Thomson et al. (2010) was cited in support of some of the erosion-rate increases in the Apennines being real and related to tectonics.

References

Adams, B. A. and Ehlers, T. A.: Deciphering topographic signals of glaciation and rock uplift in an active orogen: a case study from the Olympic Mountains, USA, Earth Surf. Proc. Landf., 42, 1680–1692, doi: 10.1002/esp.4120, 2017.

Avdeev, B. and Niemi, N. A.: Rapid Pliocene exhumation of the central Greater Caucasus constrained by low-temperature thermochronometry, Tectonics, 30, TC2009, doi: 10.1029/2010TC002808, 2011.

Baker, V.R.: Hypotheses and geomorphological reasoning. In Rhoads, B.L. and Thorn, C.E. (eds.): The Scientific Nature of Geomorphology: Proceedings of the 27th Binghampton Symposium in Geomorphology, Wiley, Chichester, p. 57-85, 1996.

Ballato, P., Landgraf, A., Schildgen, T. F., Stockli, D. F., Fox, M., Ghassemi, M. R., Kirby, E. and Strecker, M. R.: The growth of a mountain belt forced by base-level fall: Tectonics and surface processes during the evolution of the Alborz Mountains, N Iran, Earth Planet. Sci. Lett., 425, 204–218, doi: 10.1016/j.epsl.2015.05.051, 2015.

Bermúdez, M. A., van der Beek, P. and Bernet, M.: Asynchronous Miocene-Pliocene exhumation of the central Venezuelan Andes, Geology, 39, 139–142, doi: 10.1130/G31582.1, 2011.

Beucher, R., van der Beek, P., Braun, J. and Batt, G. E.: Exhumation and relief development in the Pelvoux and Dora-Maira massifs (western Alps) assessed by spectral analysis and inversion of thermochronological age transects, J. Geophys. Res., 117, F03030, doi: 10.1029/2011JF002240, 2012.

Butler, R. W. H.: Tectonic evolution of the Himalayan syntaxes: the view from Nanga Parbat, Geol. Soc. London Spec. Publ., 483, 215–254, doi:10.1144/SP483.5, 2019.

Crowley, J. L., Waters, D. J., Searle, M. P. and Bowring, S. A.: Pleistocene melting and rapid exhumation of the Nanga Parbat massif, Pakistan: Age and P–T conditions of accessory mineral growth in migmatite and leucogranite, Earth Planet. Sci. Lett., 288, 408–420, doi:10.1016/j.epsl.2009.09.044, 2009.

Egli, D., Mancktelow, N. and Spikings, R.: Constraints from ^{40}Ar/ ^{39}Ar geochronology on the timing of Alpine shear zones in the Mont Blanc-Aiguilles Rouges region of the European Alps, Tectonics, 36, 730–748, doi: 10.1002/2016TC004450, 2017.

Ehlers, T. A., Willett, S. D., Armstrong, P. A. and Chapman, D. S.: Exhumation of the central Wasatch Mountains, Utah: 2. Thermokinematic model of exhumation, erosion, and thermochronometer interpretation, J. Geophys. Res., 108, 2173, doi: 10.1029/2001JB001723, 2003.

Fox, M., Herman, F., Willett, S. D. and May, D. A.: A linear inversion method to infer exhumation rates in space and time from thermochronometric data, Earth Surf. Dynam., 2, 47–65, doi: 10.5194/esurf-2-47-2014, 2014.

Fox, M., Herman, F., Willett, S. D. and Schmid, S. M.: The exhumation history of the European Alps inferred from linear inversion of thermochronometric data, Am. J. Sci., 316, 505–541, doi: 10.2475/06.2016.01, 2016.

Fuller, C. W., Willett, S. D., Fisher, D. and Lu, C. Y.: A thermomechanical wedge model of Taiwan constrained by fission-track thermochronometry, Tectonophysics, 425, 1–24, doi: 10.1016/j.tecto.2006.05.018, 2006.

Glotzbach, C., Bernet, M. and van der Beek, P.: Detrital thermochronology records changing source areas and steady exhumation in the Western European Alps, Geology, 39, 239–242, doi: 10.1130/G31757.1, 2011a.

Glotzbach, C., van der Beek, P. A. and Spiegel, C.: Episodic exhumation and relief growth in the Mont Blanc massif, Western Alps from numerical modelling of thermochronology data, Earth Planet. Sci. Lett., 304, 417–430, doi: 10.1016/j.epsl.2011.02.020, 2011b.

Herman, F., Braun, J. and Dunlap, W. J.: Tectonomorphic scenarios in the Southern Alps of New Zealand, J. Geophys. Res., 112, B04201, doi: 10.1029/2004JB003472, 2007.

Herman, F., Copeland, P., Avouac, J.-P., Bollinger, L., Mahéo, G., Le Fort, P., Rai, S., Foster, D., Pêcher, A., Stüwe, K. and Henry, P.: Exhumation, crustal deformation, and thermal structure of the Nepal Himalaya derived from the inversion of thermochronological and thermobarometric data and modeling of the topography, J. Geophys. Res., 115, B06407, doi: 10.1029/2008JB006126, 2010.

Herman, F., Cox, S. C. and Kamp, P. J. J.: Low-temperature thermochronology and thermokinematic modeling of deformation, exhumation, and development of topography in the central Southern Alps, New Zealand, Tectonics, 28, TC5011, doi: 10.1029/2008TC002367, 2009.

Herman, F., Seward, D., Valla, P. G., Carter, A., Kohn, B., Willett, S. D. and Ehlers, T. A.: Worldwide acceleration of mountain erosion under a cooling climate, Nature, 504, 423–426, doi: 10.1038/nature12877, 2013.

Jamieson, R. A. and Beaumont, C.: On the origin of orogens, Geol. Soc. Am. Bull., 125, 1671–1702, doi: 10.1130/B30855.1, 2013.

Jiao, R., Herman, F. and Seward, D.: Late Cenozoic exhumation model of New Zealand: Impacts from tectonics and climate, Earth Sci. Rev., 166, 286–298, doi: 10.1016/j.earscirev.2017.01.003, 2017.

Montgomery, D. R.: Valley formation by fluvial and glacial erosion, Geology, 30, 1047–1050, doi: 10.1130/0091-7613(2002)030<1047:VFBFAG>2.0.CO;2, 2002.

Robert, X., van der Beek, P., Braun, J., Perry, C., Dubille, M. and Mugnier, J. L.: Assessing Quaternary reactivation of the Main Central thrust zone (central Nepal Himalaya): New thermochronologic data and numerical modeling, Geology, 37, 731–734, doi: 10.1130/G25736A.1, 2009.

Rolland, Y., Rossi, M., Cox, S. F., Corsini, M., Mancktelow, N., Pennacchioni, G., Fornari, M. and Boullier, A. M.: ^{40}Ar/^{39}Ar dating of synkinematic white mica: insights from fluid-rock reaction in low-grade shear zones (Mont Blanc Massif) and constraints on timing of deformation in the NW external Alps, Geol. Soc. London Spec. Publ., 299, 293–315, doi: 10.1144/SP299.18, 2008.

Rossi, M. and Rolland, Y.: Stable isotope and Ar/Ar evidence of prolonged multiscale fluid flow during exhumation of orogenic crust: Example from the Mont Blanc and Aar Massifs (NW Alps), Tectonics, 33, 1681–1709, doi: 10.1002/2013TC003438, 2014.

Schildgen, T. F., Ehlers, T. A., Whipp, D. M., Jr., van Soest, M. C., Whipple, K. X. and Hodges, K. V.: Quantifying canyon incision and Andean Plateau surface uplift, southwest Peru: A thermochronometer and numerical modeling approach, J. Geophys. Res., 114, F04014, doi: 10.1029/2009JF001305, 2009.

Schildgen, T. F., van der Beek, P. A., Sinclair, H. D. and Thiede, R. C.: Spatial correlation bias in late-Cenozoic erosion histories derived from thermochronology, Nature, 559, 89–93, doi: 10.1038/s41586-018-0260-6, 2018.

Schneider, D. A., Zeitler, P. K., Kidd, W. and Edwards, M. A.: Geochronologic constraints on the tectonic evolution and exhumation of Nanga Parbat, western Himalaya syntaxis, revisited, J. Geol., 109, 563–583, doi: 10.1086/322764, 2001.

Shuster, D. L., Cuffey, K. M., Sanders, J. W. and Balco, G.: Thermochronometry reveals headward propagation of erosion in an alpine landscape, Science, 332, 84–88, doi: 10.1126/science.1202357, 2011.

Shuster, D. L., Ehlers, T. A., Rusmore, M. E. and Farley, K. A.: Rapid glacial erosion at 1.8 Ma revealed by ^{4}He/^{3}He thermochronometry, Science, 310, 1668–1670, doi: 10.1126/science.1118519, 2005.

Sutherland, R., Gurnis, M., Kamp, P. J. J. and House, M. A.: Regional exhumation history of brittle crust during subduction initiation, Fiordland, southwest New Zealand, and implications for thermochronologic sampling and analysis strategies, Geosphere, 5, 409–425, doi: 10.1130/GES00225.1, 2009.

Thiede, R. C. and Ehlers, T. A.: Large spatial and temporal variations in Himalayan denudation, Earth Planet. Sci. Lett., 371-372, 278–293, doi: 10.1016/j.epsl.2013.03.004, 2013.

Thomson, S. N., Brandon, M. T., Reiners, P. W., Zattin, M., Isaacson, P. J. and Balestrieri, M. L.: Thermochronologic evidence for orogen-parallel variability in wedge kinematics during extending convergent orogenesis of the northern Apennines, Italy, Geol. Soc. Am. Bull., 122, 1160–1179, doi: 10.1130/B26573.1, 2010.

Tippett, J. M. and Kamp, P. J. J.: Fission track analysis of the Late Cenozoic vertical kinematics of continental Pacific crust, South Island, New Zealand, J. Geophys. Res., 98, 16119–16148, doi: 10.1029/92JB02115, 1993.

Treloar, P. J., Rex, D. C., Guise, P. G., Wheeler, J., Hurford, A. J. and Carter, A.: Geochronological constraints on the evolution of the Nanga Parbat syntaxis, Pakistan Himalaya, Geol. Soc. London Spec. Publ., 170, 137–162, doi: 10.1144/GSL.SP.2000.170.01.08, 2000.

Valla, P. G., Shuster, D. L. and van der Beek, P. A.: Significant increase in relief of the European Alps during mid-Pleistocene glaciations, Nature Geosci., 4, 688–692, doi: 10.1038/ngeo1242, 2011.

Willenbring, J. K. and Jerolmack, D. J.: The null hypothesis: globally steady rates of erosion, weathering fluxes and shelf sediment accumulation during Late Cenozoic mountain uplift and glaciation, Terra Nova, 28, 11–18, doi: 10.1111/ter.12185, 2016.

Willett, S., Beaumont, C. and Fullsack, P.: Mechanical model for the tectonics of doubly vergent compressional orogens, Geology, 21, 371–374, doi: 10.1130/0091-7613(1993)021<0371:MMFTTO>2.3.CO;2, 1993.

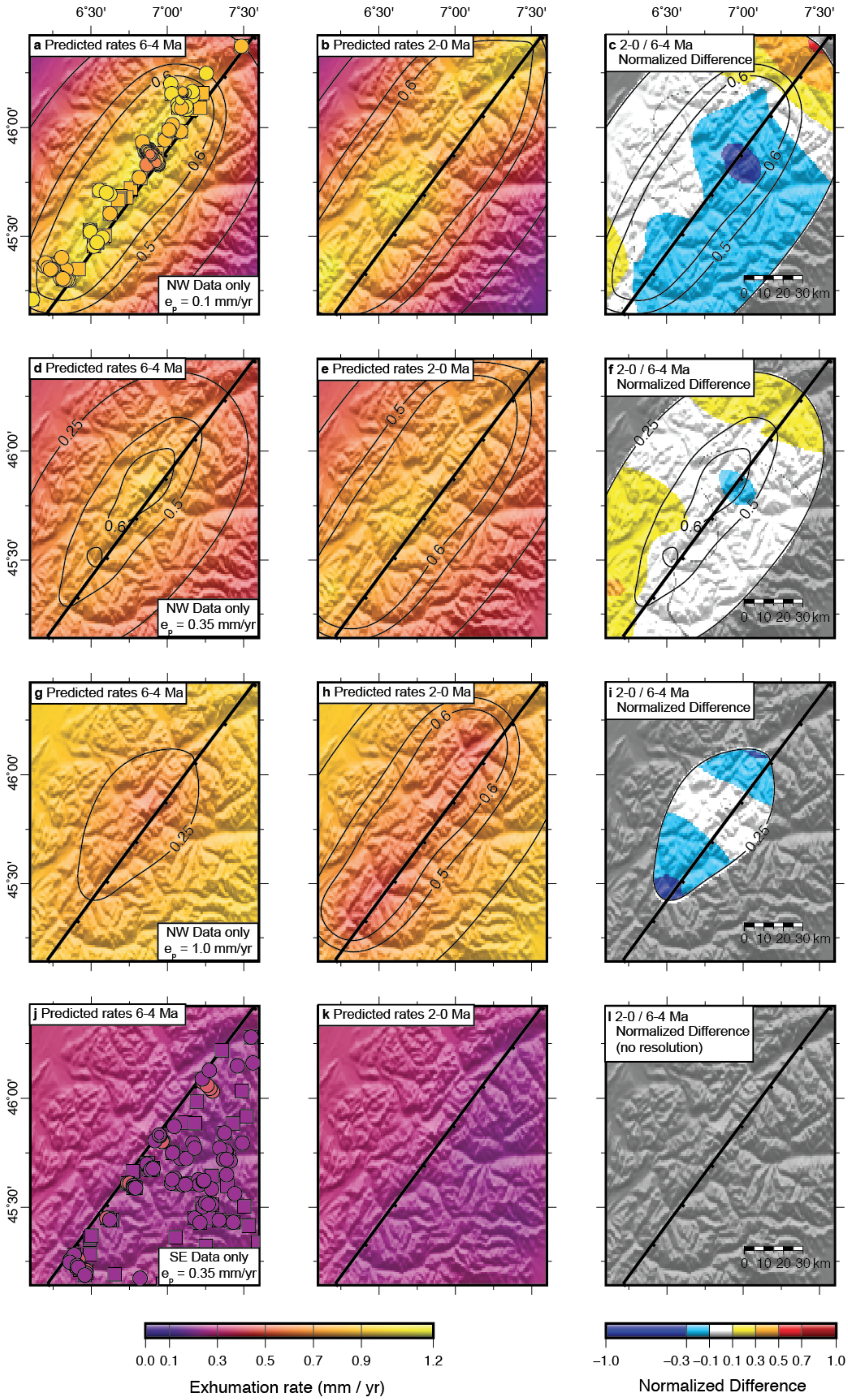
Willett, S. D., Fisher, D., Fuller, C., Yeh, E.C. and Lu, C.Y.: Erosion rates and orogenic-wedge kinematics in Taiwan inferred from fission-track thermochronometry, Geology, 31, 945–948, doi: 10.1130/G19702.1, 2003.

Willett, S.D., Herman, F., Fox, M., Stalder, N., Ehlers, T.A., Jiao, R., Yang, R.: Bias and error in modelling thermochronometric data: resolving a potential increase in Plio-Pleistocene erosion rate. Earth Surf. Dyn. Discussions, doi: 10.5194/esurf-2020-59.

Zeitler, P. K.: Cooling history of the NW Himalaya, Pakistan, Tectonics, 4, 127–151, doi: 10.1029/TC004i001p00127, 1985.

Zeitler, P. K., Johnson, N. M., Naeser, C. W. and Tahirkheli, R.: Fission-track evidence for Quaternary uplift of the Nanga Parbat region, Pakistan, Nature, 298, 255–257, doi: 10.1038/298255a0, 1982.

Zeitler, P. K., Koons, P. O., Bishop, M. P., Chamberlain, C. P., Craw, D., Edwards, M. A., Hamidulah, S., Qasim Jan, M., Khan, M. A., Khattak, M. U. K., Kidd, W. S. F., Lackie, R. L., Meltzer, A. S., Park, S. K., Pêcher, A., Poage, M. A., Sarker, G., Schneider, D. A., Seeber, L. and Shroder, J. F.: Crustal reworking at Nanga Parbat, Pakistan: Metamorphic consequences of thermal-mechanical coupling facilitated by erosion, Tectonics, 20, 712–728, doi: 10.1029/2000TC001243, 2001.



a Predicted rates 6-4 Ma
NW Data only
e_p = 0.1 mm/yr
b Predicted rates 2-0 Ma
c 2-0 / 6-4 Ma
Normalized Difference
d Predicted rates 6-4 Ma
NW Data only
e_p = 0.35 mm/yr
e Predicted rates 2-0 Ma
f 2-0 / 6-4 Ma
Normalized Difference
g Predicted rates 6-4 Ma
NW Data only
e_p = 1.0 mm/yr
h Predicted rates 2-0 Ma
i 2-0 / 6-4 Ma
Normalized Difference
j Predicted rates 6-4 Ma
SE Data only
e_p = 0.35 mm/yr
k Predicted rates 2-0 Ma
l 2-0 / 6-4 Ma
Normalized Difference
(no resolution)
6°30' 7°00' 7°30'
46°00' 45°30'
0 10 20 30 km
0.0 0.1 0.3 0.5 0.7 0.9 1.2
Exhumation rate (mm / yr)
−1.0 −0.3 −0.1 0.1 0.3 0.5 0.7 1.0
Normalized Difference

655

Figure 1. Synthetic test of the Bayesian bias to the prior erosion rate, and the effect of different geotherm calculations between Pecube and GLIDE, for the "western Alps" case. To eliminate the spatial correlation bias, this test only uses synthetic ages NW of the fault (shown in a; large squares = predicted zircon fission-track ages; large circles = predicted apatite fission-track ages; small circles = predicted apatite (U-Th)/He ages; symbols are coloured according to age/exhumation rate as in Schildgen et al. (2018); ages are predicted at the real data locations). Input exhumation rate is 1 mm/yr NW of the fault (thick black line) and is held constant for 20 Myr. All other model input parameters are identical to those used in Schildgen et al. (2018). (a-c) Predicted temporal evolution in erosion rates for a prior erosion rate (e_p) of 0.1 mm/yr: (a) predicted erosion rates for the 6-4 Ma time bin; (b) predicted erosion rates for the 2-0 Ma time bin; (c) normalized difference in erosion rates, only shown where the resolution in each time bin is > 0.25. Contours in plots (a-c) show the predicted resolution. (d-f), as plots (a-c) but for a prior erosion rate (e_p) of 0.35 mm/yr; (g-i), as plots (a-c) but for a prior erosion rate (e_p) of 1.0 mm/yr; (j-l), as plots (d-f) but using only data from SE of the fault (shown in j; symbols and colours as in a), which has a constant input exhumation rate of 0.25 mm/yr. The normalised differences in erosion rates are generally small and the inversion tends to predict decreases in erosion rates. When using data from SE of the fault only, the resolution is nowhere > 0.25.

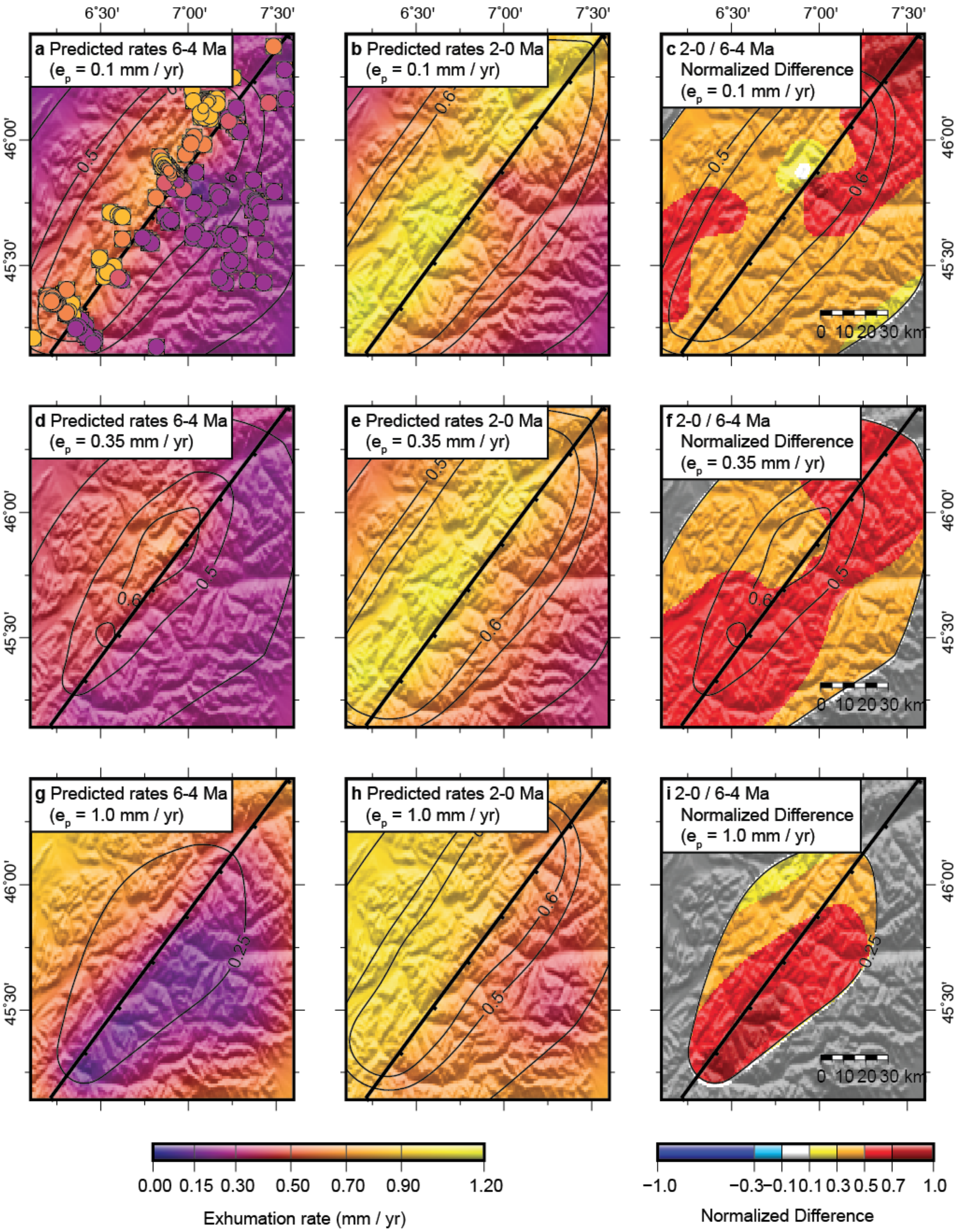
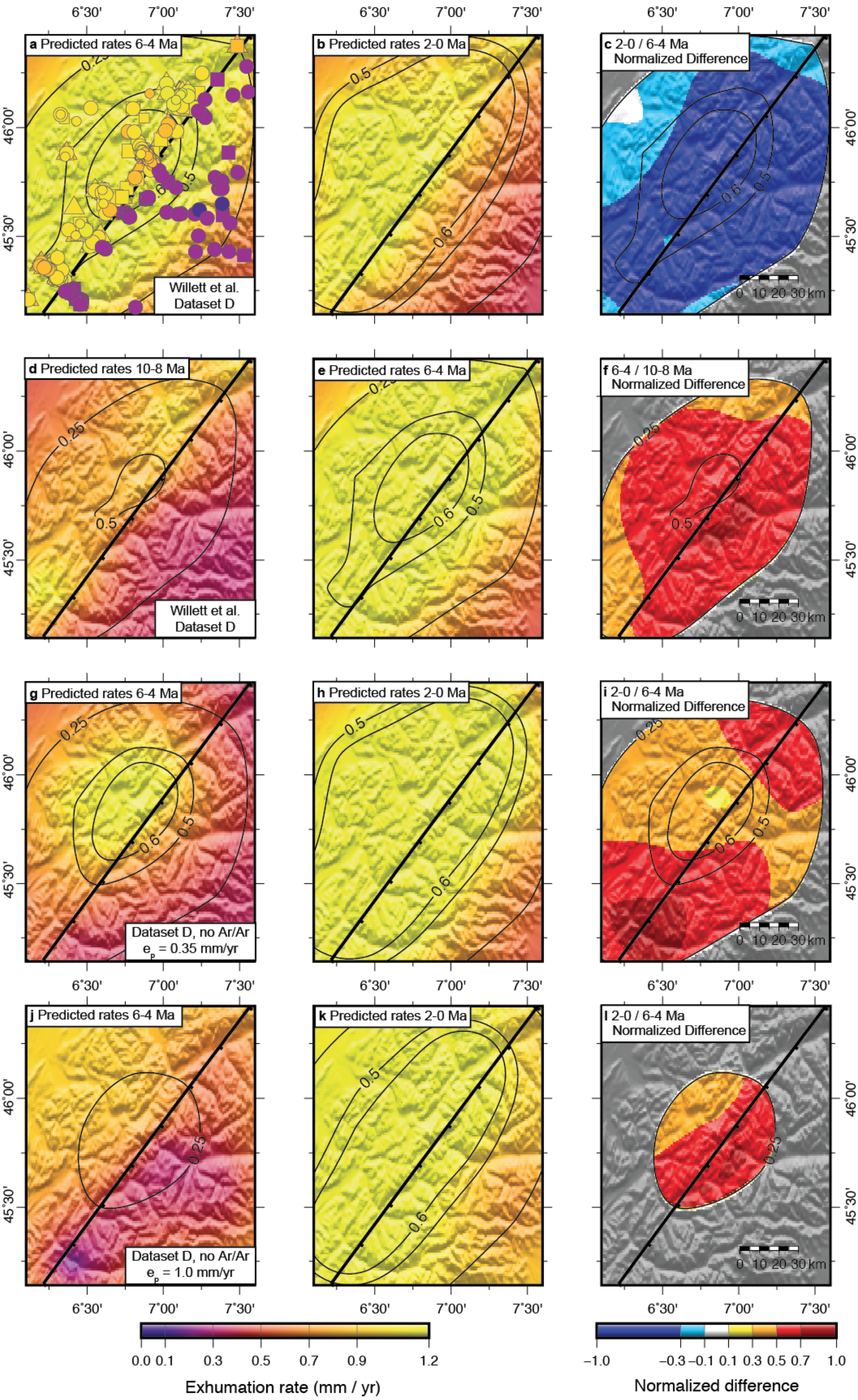


Figure 2. Synthetic test of the combined spatial correlation bias and Bayesian bias to the prior for the "western Alps" case. Input exhumation rates are 1 mm/yr NW of the fault (thick black line) and 0.25 mm/yr SE of the fault; they are held constant for 20 Myr. All other model input parameters are identical to those used in Schildgen et al. (2018). Synthetic ages both to the NW and to the SE of the fault were used (shown in a; symbols and colours as in Fig. 1). (a-c) Predicted temporal evolution in erosion rates for a prior erosion rate (e_p) of 0.1 mm/yr: (a) predicted erosion rates for the 6-4 Ma time bin; (b) predicted erosion rates for the 2-0 Ma time bin; (c) normalized difference in erosion rates, only shown where the resolution in each time bin is > 0.25. Contours in plots (a-c) show the predicted resolution. (d-f), as plots (a-c) but for a prior erosion rate (e_p) of 0.35 mm/yr; (g-i), as plots (a-c) but

for a prior erosion rate (e_p) of 1.0 mm/yr. Note that the normalised differences in erosion rates are large and positive in all cases: the model predicts significant spurious increases in erosion rates through time, although their distribution and resolution depend on e_p.



a Predicted rates 6-4 Ma
Willett et al.
Dataset D
b Predicted rates 2-0 Ma
c 2-0 / 6-4 Ma
Normalized Difference
d Predicted rates 10-8 Ma
Willett et al.
Dataset D
e Predicted rates 6-4 Ma
f 6-4 / 10-8 Ma
Normalized Difference
g Predicted rates 6-4 Ma
Dataset D, no Ar/Ar
e_p = 0.35 mm/yr
h Predicted rates 2-0 Ma
i 2-0 / 6-4 Ma
Normalized Difference
j Predicted rates 6-4 Ma
Dataset D, no Ar/Ar
e_p = 1.0 mm/yr
k Predicted rates 2-0 Ma
l 2-0 / 6-4 Ma
Normalized Difference
0 10 20 30 km
6°30' 7°00' 7°30'
46°00' 45°30'
0.0 0.1 0.3 0.5 0.7 0.9 1.2
Exhumation rate (mm / yr)
−1.0 −0.3 −0.1 0.1 0.3 0.5 0.7 1.0
Normalized difference

680

Figure 3. Synthetic test showing the effect of including higher-temperature mica $^{40}Ar/^{39}Ar$ data in the inversion. Input exhumation rates are 1 mm/yr NW of the fault (thick black line) and 0.25 mm/yr SE of the fault. Dataset D of Willett et al. (shown in a; symbols and colours as in Fig. 1, with the addition of zircon (U-Th)/He data – small squares; and mica $^{40}Ar/^{39}Ar$ data – large triangles) and a prior erosion rate (e_p) of 0.35 mm/yr were used. (a-c) Predicted temporal evolution in erosion rates using all the data (equivalent to Willett et al.'s Figure 14): (a) predicted erosion rates for the 6-4 Ma time bin; (b) predicted erosion rates for the 2-0 Ma time bin; (c) normalized difference in erosion rates, only shown where the resolution in each time bin is > 0.25. Contours in plots (a-c) show the predicted resolution. (d-f) Predicted temporal evolution during earlier time bins for this inversion: (d) predicted erosion rates for the 10-8 Ma time bin; (e) predicted erosion rates for the 6-4 Ma time bin (same as a); (f) normalized difference in erosion rates, only shown where the resolution in each time bin is > 0.25. Note the significant increase in erosion rates that precedes the decrease shown by Willett et al. (g-i), as plots (a-c) but with the mica $^{40}Ar/^{39}Ar$ data taken out; note that the removal of mica $^{40}Ar/^{39}Ar$ ages results in no substantial loss in resolution, based on comparing resolution contours in a-b and g-h; (k-l), as plots (g-j) but for a prior erosion rate (e_p) of 1.0 mm/yr. Inversions of Willett et al.'s dataset D without the mica $^{40}Ar/^{39}Ar$ data predict large spurious increases in the most recent time bin, similar to our models in Fig. 2.

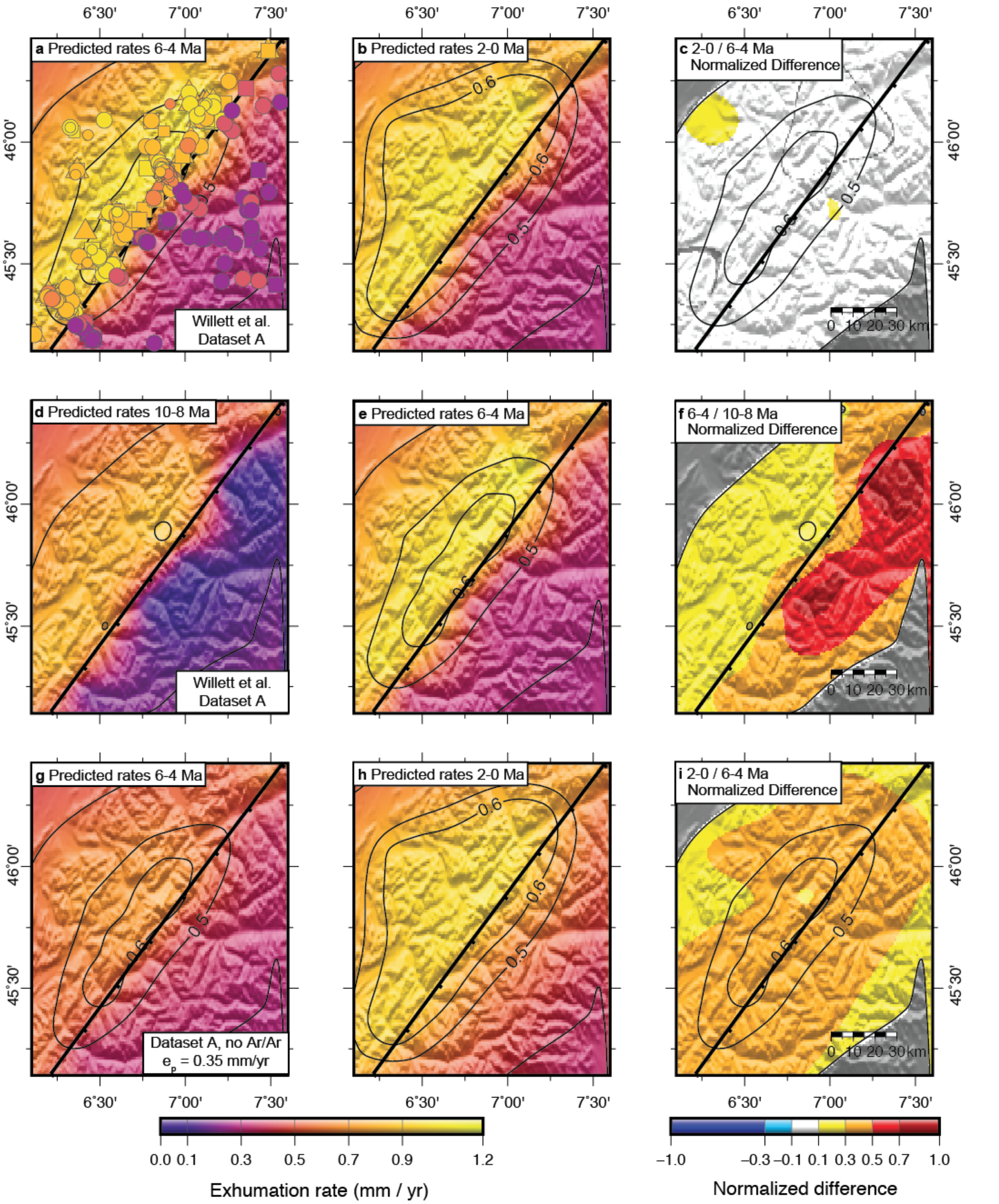


Figure 4. Synthetic test showing the effect of including higher-temperature mica $^{40}Ar/^{39}Ar$ data in the inversion for the stable isotherm case. Input exhumation rates are 1 mm/yr NW of the fault (thick black line) and 0.25 mm/yr SE of the fault. Dataset A of Willett et al. (shown in a; symbols and colours as in Fig. 3) and a prior erosion rate (e_p) of 0.35 mm/yr were used. (a-c) Predicted temporal evolution in erosion rates using all the data (equivalent to Willett et al.'s Figure 6): (a) predicted erosion rates for the 6-4 Ma time bin; (b) predicted erosion rates for the 2-0 Ma time bin; (c) normalized difference in erosion rates, only shown where the resolution in each time bin is > 0.25. Contours in plots (a-c) show the predicted resolution. (d-f) Predicted temporal evolution during earlier time bins for this inversion: (d) predicted erosion rates for the 10-8 Ma time bin; (e) predicted erosion rates for the 6-4 Ma time bin (same as a); (f) normalized difference in erosion rates, only shown where the resolution in each time bin is > 0.25. Note the significant increase in erosion rates that precedes the time window shown by Willett et al. (g-i), as plots (a-c) but with the

mica $^{40}Ar/^{39}Ar$ data taken out. This inversion predicts large spurious increases in the most recent time bin, similar to our models in Figs. 2 and 3 g-l. Also note that the removal of mica $^{40}Ar/^{39}Ar$ ages results in no substantial loss in resolution, based on comparing resolution contours in a-b and g-h.

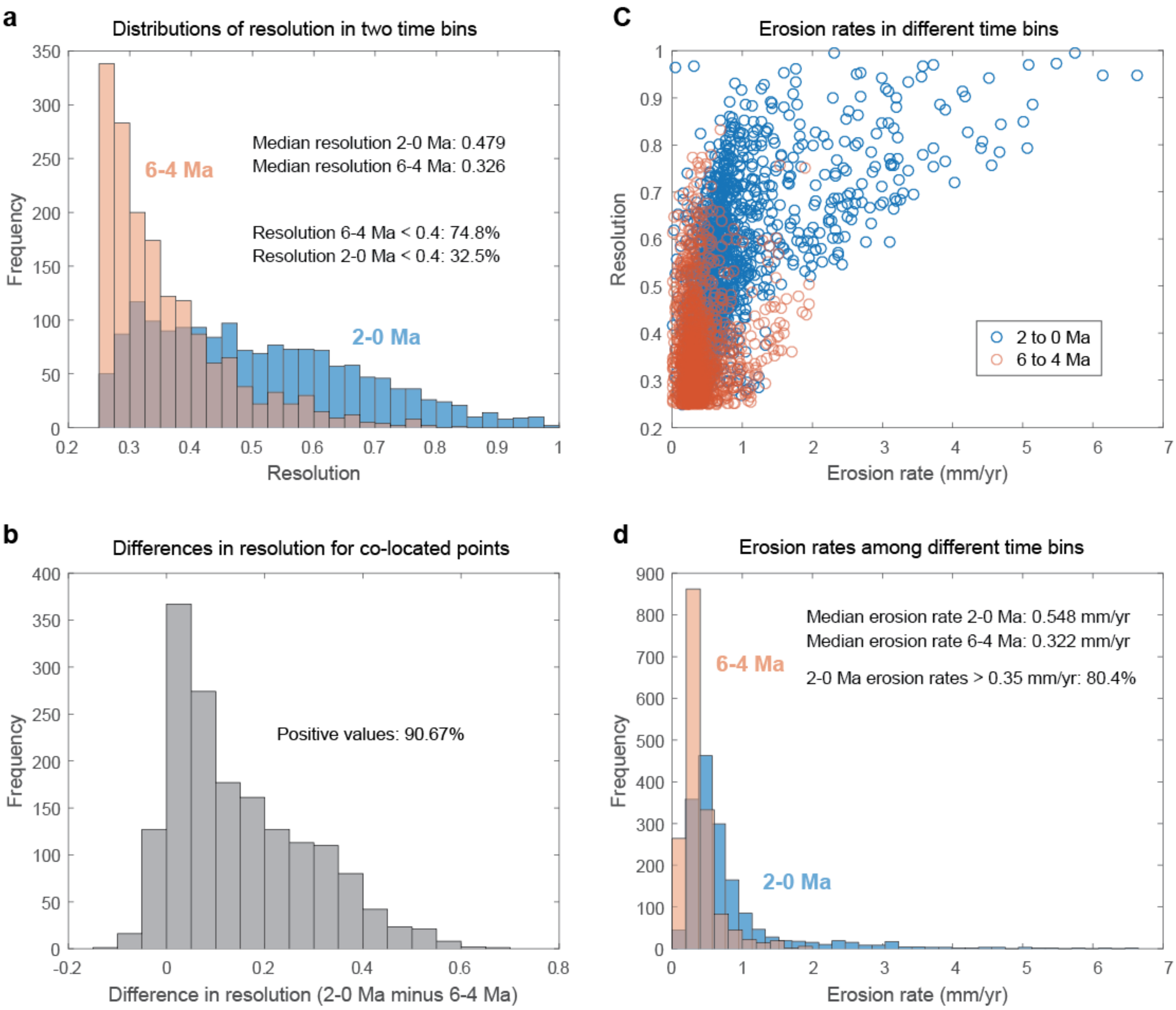


Figure 5. (a) Distribution of resolution values for all predicted erosion rates in Herman et al. (2013) with a resolution > 0.25 in the 6-4 Ma time bin (orange) and 2-0 Ma time bin (blue). Median rates and proportion of points with resolution < 0.4, which Willett et al. argue will revert to the prior, are indicated. (b) Differences in resolution for co-located points, illustrating that in over 90% of the "well resolved" values reported by Herman et al. (2013), the resolution is higher in the 2-0 Ma time bin compared to the 6-4 Ma time bin. (c) Resolution versus inferred erosion rate for the 2-0 Ma (blue) and 6-4 Ma (orange) time bins. (d) Distribution of erosion rates in the 6-4 Ma time bin (orange) and the 2-0 Ma time bin (blue).

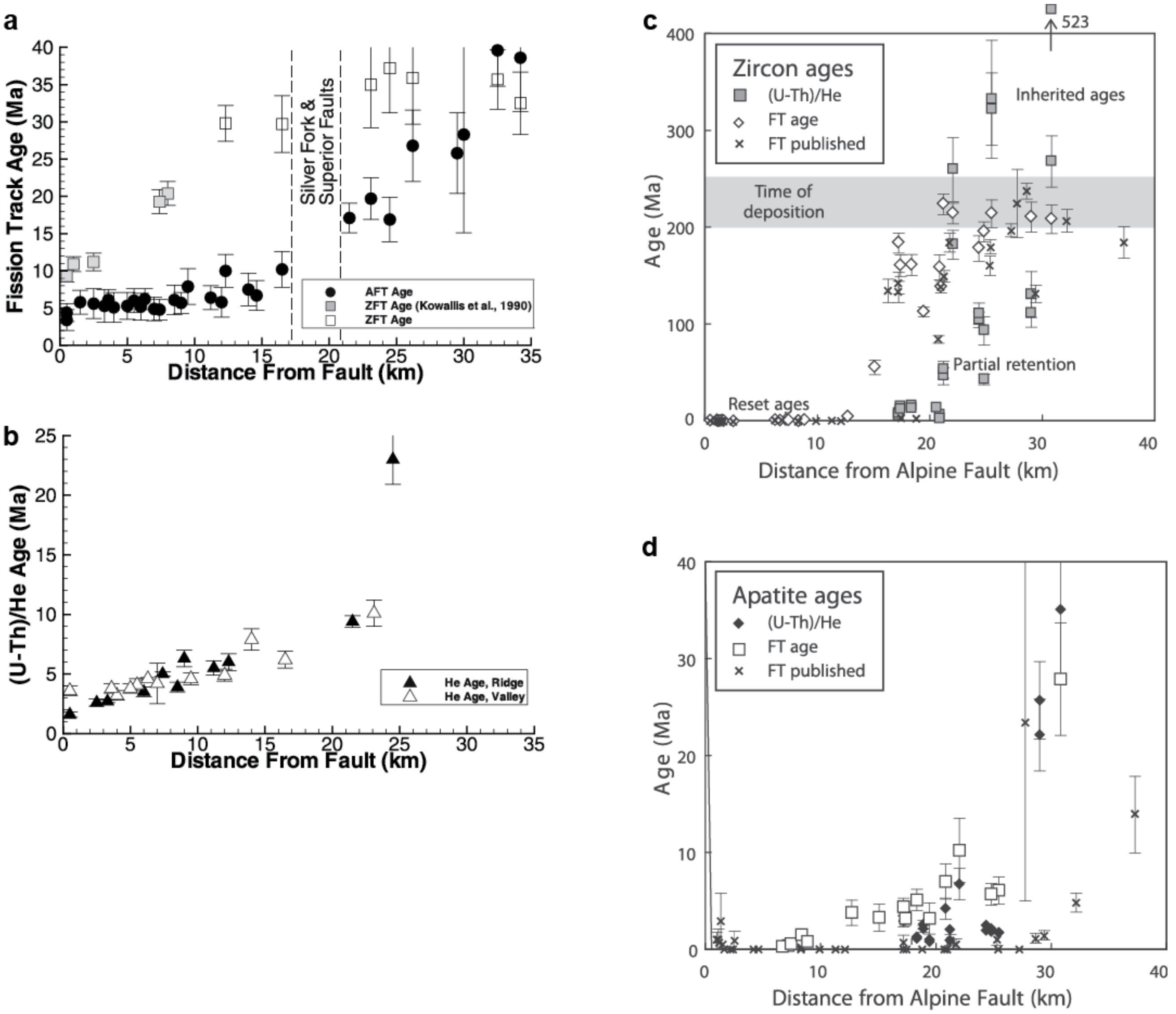


Figure 6. Thermochronometer age patterns versus distance from the Wasatch Fault in the Wasatch Mountains (a: apatite and zircon fission-track ages; b: apatite (U-Th)/He ages; from Ehlers et al., 2003, reproduced with permission from John Wiley and Sons) and versus distance from the Alpine Fault in the Southern Alps of New Zealand (c: zircon (U-Th)/He and fission-track ages; d: apatite (U-Th)/He and fission-track ages; from Herman et al., 2009, reproduced with permission from John Wiley and Sons).