Response to Comment SC1

Comment by van der Beek et al in red. Our response in black.

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).

Van der Beek et al. list a large number of studies with the implication that these somehow contradict the results of Herman et al. (2013). This is not the case. Many of these studies found an increase in erosion rates over the last 5 Ma. Others were modeling studies that were not parameterized in a way that was capable of establishing a change in erosion rate over the last 5 Ma. None of these studies can differentiate between tectonic and climate-change induced changes in exhumation rate. All of these studies include interpretations or models that are not inherently better than the approach taken by Herman et al. (2013). To use these results to "test" Herman et al. (2013) requires posing an objective test (See our comment AC-6). This was not done by Schildgen et al. (2018). They made a subjective selection as to which aspects of previous studies they took and which they ignored, and in every case where there was a discrepancy, they concluded that Herman et al. was wrong and the alternative was right.

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.

We addressed this point directly in AC-5 and AC-6. The approach of Schidgen et al (2018) is predicated on the idea that they could determine the "correct" answer from the literature in order to "test" Herman et al. (2013). No quantitative tests were ever applied.

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;

We addressed this in AC2 extensively.

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;

We addressed this in AC2.

3. Biased post-processing operator and resolution: We argue that Willett et al.'s criticism of the use of postprocessing 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 don't see the basis of this statement. With the replacement of Figures 2 and 3 of Herman et al. by Figures 24, 25 of esurf2020-59, there are no ratios and thus no ratio bias at all in Herman et al., so this statement is unfounded. Most of esurf2020-59 is resolution analysis, so we also see no basis for the second part of this statement. Schildgen et al. made no quantitative statistical analysis of resolution and provide no evidence that they used the Herman et al analysis other than using

one value as an analysis cut-off. Most importantly, they interpreted NR maps to infer the existence of errors, e.g. spatial correlation bias, without any knowledge of the error mapped into the NR.

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 80 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).

We will respond to these below.

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)…

We responded to this in AC2 and again in AC5. We stand by our contention that the large errors (due to their choice of a different, constant-T, boundary condition) in Schildgen et al.'s synthetic data models render them invalid. They have not corrected this error to show that they obtain the same result. They did show that their age errors produced errors of up to 28% in the inferred erosion rates of the high erosion region and did not check the low erosion rate region. (see AC2).

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.

These models have errors in the ages of 10s to over 100%. Van der Beek et al make no demonstration that the geotherm error "does not significantly affect these tests". Their argument is only that because the error is not intuitive in its second derivative (acceleration), it must not be significant. This is not a valid argument. They have not corrected the error and shown that the model produces the same result. We see no reason to respond to models whose predictions are ambiguous due to this large error. We presented an extensive suite of models equivalent to these, but without geotherm errors. We think the discussion should be restricted to these models.

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 40Ar/39Ar ages. Neither Herman et al. (2013) nor Fox et al. (2016) used any mica 40Ar/39Ar cooling ages in their inversions, for the simple reason that these do not exist. The few 40Ar/39Ar 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 40Ar/39Ar; 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

180 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.....

We responded to this comment in AC2 and again in AC5 and stand by those comments. To summarize, the difference between the models of van der Beek et al and our paper is that our models use ages that are calculated in a manner that is internally consistent with the inverse model and van der Beek et al. do not do the same. In the cases where van der Beek et al use a decimated version of our data, they have turned the high-resolution-models into low-resolution models. Our paper carefully presents a range of resolution test cases including low-resolution cases for contrast; van der Beek et al., have replaced this suite with a set of only low-resolution models, or shifted the analysis to parts of the model with low resolution, largely defeating the purpose of the models.

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.

We don't really see the point of this figure. Younger timesteps will always be better resolved. This is normal. Ages provide an integral constraint on erosion rates, therefore there are always more ages constraining younger timesteps (see AC3). We do not know where we stated that bias to the prior becomes important at resolution values of 0.4, but didn't intend to. One cannot simply take one value of resolution and state that bias to the prior is now significant. There are too many other factors involved and there is no reason to expect a one-to-one relationship between the temporal resolution parameter and bias-to-the-prior. For example, younger timesteps that have no ages in them, but do have older ages (as in Fig 1 "Low Resolution" region) will have lower values of resolution but less dependence on the prior as they must be consistent with the older ages. The strong dependence on the prior is for regions with no data or timesteps so old that there are no ages sampling the interval. This figure of van der Beek is mixing these different spatial and temporal resolution characteristics together and speculating that there is a one to one relationship between resolution and bias. This appears to be another attempt to use a fixed value of resolution to justify overly broad conclusions. There is no analysis or test provided by van der Beek et al. to show that any of this speculation is true. Herman et al. (2013) tested for this problem by running inversions with different priors and found no effect, thereby demonstrating that this argument is false.

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.

The synthetic tests (Figs 1 and 2 of van der Beek et al.) have large data errors and should not be considered valid (AC2).

The final statement (also on line 247 of comment) demonstrates the stretched logic of the van der Beek et al. analysis: they consider two possibilities: (1) Herman et al. are in error due to bias to the prior, or (2) Herman et al. are in error due to spatial correlation bias. Elimination of (1) is therefore proof of (2). A third option, that Herman et al. are correct, is simply not considered, and therefore is not tested. This is not a valid analysis.

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 don't understand where van der Beek et al come up with this statement that Herman et al focussed on the magnitude of the signal, whereas Schildgen et al. focussed on the sign. Herman et al. quantified their results, but summarized them as showing an "increase in mountain erosion rates since about 6 Myr ago, by nearly a factor of two for the Pleistocene compared to the Pliocene. " (Herman et al., 2013, page 3). "Nearly a factor of two" is hardly a quantitative estimate, because we recognized that what was important was the sign, and that the area where Herman et al resolved erosion rates was small. Herman et al. did nothing further with this number ("nearly two"). Herman et al concluded that these regions experienced an increase in erosion rate, not an increase in erosion rate of precisely some value. Their conclusions would be unchanged if it were smaller or larger, provided it remains above 1, i.e. the direction of change is unchanged. We would thus argue that Herman et al. also focussed on sign. We will add a new paragraph to our paper summarizing the conclusions of Herman et al., since several reviewers also questioned what these conclusions were.

One cannot modify a quantity with a non-linear operator and not track the error through the calculation. With a linear operator, this is simple and things like thresholds map directly, but division is not a linear operator. This is basic error analysis.

Importantly, much of the Schildgen et al analysis was an interpretation of the normalized difference maps. The assessment of the Herman et al. inversion results, including all the Schildgen et al. statements regarding the presence of spatial correlation bias was based on visual inspection of these maps. Schildgen et al. contains no spatial analysis of the Herman et al. (2-13) results; it has only an interpretation and that interpretation is based on spatial patterns. There are numerous examples where the interpretation directly refers to the location of the maximum values of NR. If they used other aspects of the Herman et al analysis, they give no indication, but they did publish twenty-some normalized ratio maps and no other portrayal of the Herman et al. results such as resolution maps or posterior variance. This undoubtedly gave a distorted view that impacted interpretation. This is easily seen from the figures of our paper. Consider Figure 17 as an example. Schildgen et al. took the results of Figures 17 a-i, reduced them to Figure 17j, omitted all resolution and variance information, taking only a single contour of the continuous resolution map as a binary quality criterion and made an interpretation. As a summary of the information in Figures 17 a-i, Figure 17j is neither complete, nor accurate, and omitting the information of Figure 17k makes it worse as it leaves the impression of largely uniform values of NR across the region with no measure of relative quality. How much did this influence Schildgen et al.s interpretation? We don't know, because an interpretation is not reproducible, but the examples we gave in our paper (section 5.3) give much circumstantial evidence that they made direct interpretation of the exaggerated regions of these maps, at the expense of the well-resolved parts. We can check our text that we have not overstated this, but we don't think it is unfair to make an interpretation of their analysis. We note that their "proof" of the existence of a spatial correlation bias, as described in their paper, was given exclusively as an interpretation of these maps.

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 e2 instead of max(e1, e2), as done by Schildgen et al. (2018) (line 467 of Willett et al.; note that in their definition e1 > e2 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.

Equation 14 is reversed in definition of e1 and e2; this is a typo in the equation, not a difference in the calculation and we will correct the equation. All our figures have the correct sign. Thank you for drawing this to our attention.

The second "error" is not an error or modification of the definition. We state on line 438 that we assume that e2 is the larger quantity for the variance analysis between lines 438 and 480, because we did not want to double the number of equations we would need to write, as one needs to do with use of the max{} function. This is explained again on line 467, where we indicate that we "drop the absolute value for simplicity" This does not affect the generality of the analysis, just the brevity. For the results shown in the paper, both NR and the variance of NR use the definition as in Schildgen et al.

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.

We don't see the significance of these comments. There are no consequences of using the mean or the median as nothing is done with this number in either Herman et al. (2013) or esurf-2020-59. The difference between them is also not important for any of the issues under discussion or for the conclusions of Herman et al. (2013) or Esurf-2020-59.

The thermochronometry community of Taiwan have been discussing for years which ages are reset and which are not, with no resolution of this question. If Van der Beek et al have finally solved this problem, it would be considerate if they reported this to this community along with an explanation for how they determined it.

The last statement is undemonstrated speculation.

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.

Within each area with resolution greater than 0.25, there is a range of resolutions, from better to worse. Schildgen et al. report only a single interpretation per site (Figure ED2, Table 1), so if they made an assessment of well-resolved and poorly-resolved areas within each site for values > 0.25, they did not report it. They lumped them into one outcome. We gave several examples where Schildgen et al, (2018) (see quotes in section 5.3) use the values from the relatively poorly resolved periphery to characterize all the results at a site. Furthermore, We don't know what Schildgen et al. (2018) used in their interpretation as it is an irreproducible interpretation, not a quantitative analysis, so we have tried to limit our criticism to those cases where they explicitly described the signal on the periphery of the resolved region as evidence for spatial correlation bias, and where the argument would be difficult to make had they focused on the best-resolved regions. They show twenty-some maps of normalized erosion rate. None of these include the resolution information, only a threshold cut-off. All regions with resolution above 0.25 appear to be treated equally in their interpretation.

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.

We don't see what objective criterion is being used to establish what constitutes a "global scale". In fact, Herman et al. (2013) and esurf-2020-59 make no claims as the"global" significance of their results. "Global applies to the distribution of the data, and does not suggest that the increase is applicable to every point on the globe. This is stated in Herman et al. (2013), but we will add a section to ESURF 202-59 explaining the conclusions of Herman et al., as this has come up a number of times. Results are presented for what they are. If there are only seven regions, there are only seven regions. We present the results and readers can determine how significant they are. We don't see by what criteria van der Beek et al can judge and declare how many sites are enough; this is subjective.

Furthermore, the hypothesis of Schildgen et al. is not: "Herman et al included too much area in their analysis". It is "Herman et al.s method has a spatial correlation bias, such that all their results are wrong". If it were the first, we would not be writing this paper. Van der Beek et al. seem to be adding this argument as backup to their actual hypothesis.

Note that claims regarding Wasatch Mountains, Alps, Apennines, Taiwan and Fiordland are unsupported by any models or hypothesis testing (see AC6).

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).

There is no confusion *on our part* regarding isotherms or isochrons. We are drawing isochrones as we state in our paper. van der Beek et al. seem to think we are drawing isotherms because, even after we have pointed out their errors, they have not returned to the original literature to see what the original models show and how they differ from what is portrayed in Schildgen et al Figure 1. Our isochrones are consistent with the literature cited in our paper, whereas we don't know the origin of the Schildgen et al. (2018) versions. The attribution in Schildgen et al. was imprecise; they write only "discussed in the main text or supplement", and we suspect that there is no source, as we know the literature in this field very well and are not aware of examples as shown in their figure. There are two points in this comparison. First, many tectonic settings are characterized by spatially constant erosion rates, not gradients. Second, the underlying hypothesis of Schildgen et al. is that they could "test" the results of Herman et al. (2013) by a careful reading and analysis of the literature, yet they were unable to accurately summarize even the simple models of their Figure 1 without introducing modifications sympathetic to their arguments. For example, in a separate publication in the same year, Schildgen and van der Beek (Fission track thermochronology, Chapter 19, Malusa and Fitzgerald, eds., 2018) published a wedge kinematic model taken from Willett and Brandon (2002), and it is portrayed accurately. The equivalent wedge cartoon in Schildgen et al. (Nature, 2018) is represented differently, with zones of constant age shown by Willett and Brandon replaced by continuous gradients in age. If confirmation bias comes into even cartoon representations of kinematic models, it is also likely to come into the more unconstrained interpretations of natural settings.

This also serves as a good demonstration that thermo-kinematic models are not inherently better than other analysis methods. The kinematic model that (Schildgen et l. Figure 1b) envision for a thrust ramp was implemented in their test case of Figure ED5. This kinematic model includes normal sense shear across the entire thrust ramp. We have never seen observations or models suggesting such a deformation mode, primarily because it fails to balance cross-sectional area. If used to predict thermochronometric ages, it would give an answer, but an incorrect one. Yet Schildgen et al. (2018) throughout their paper treat thermo-kinematic models as unquestionably correct, thereby "testing" the Herman et al. results, whereas they are simply another model, and are subject to serious errors in cases where the practitioner has misconceptions regarding, for example, ramp kinematics or geotherms.

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 lengthscale 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.

We don't see the significance of showing these figures. The Wasatch data are exactly as we describe them; once corrected for elevation (not done by van der Beek et al) there is almost no gradient over the region with double dating. The point about the Alpine Fault data was not that there is no gradient, but that it is not a simple thrust ramp. All studies that could fit the thermochronometry data required more complex kinematics. The final statement that these are "affected by the spatial correlation bias" is hypothesis without testing (see AC6).

5 Comments on natural examples

320 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.

This comment raises a point that we did not discuss in our paper, the significance of incised valleys, but as van der Beek et al mention it at several points in their comment (lines 301, 328, 378) it is perhaps more important than we originally assessed and is worth adding a discussion to our paper.

This point is important because van der Beek et al. and Schildgen et al (2018) have the relevance of this observation effectively backwards. Mountainous regions experiencing an increase in erosion rate often respond by initial incision of primary river valleys as the largest rivers respond to the increase in local base level fall or faster incision with increased river discharge. If erosion rates increase due to valley glacier incision, this can be even more pronounced. Deeply incised valleys are thus evidence for a recent increase in erosion rate. The surrounding hillslopes, ridges, or lower-order river valleys might not have the same increase in erosion rate, given that most landscape response times are millions to tens of millions of years, but eventually they catch up to valley incision. Thermochronometric data from the valleys will detect the increase in erosion rate and, depending on topographic form and data distribution, will measure the valley incision rate, including its increase. However, they may or may not detect the lack of change on surrounding high elevation regions that have not yet had time to respond, which is what Glotzbach et al. (2011) demonstrated. The regional mean erosion rate is the average of the erosion rate in valleys and surrounding ridges, so if the ridge erosion rate remains constant and the valleys have accelerated erosion, the average erosion rate of the region has also increased. The thermochronometric ages reflect this increase in erosion rate and accurately measure the increase in rate of the valleys, although the estimate of the regional mean change (including valleys and ridges) might be overestimated. The Glide analysis of Herman et al. (2013) accurately shows this increase in erosion rate, and the conclusions of Herman et al (2013) that this is a region that has had accelerated erosion rate is correct. If there is a potential overestimation of the regional rate, this would still not change the sign of the change, which is an increase in erosion rate.

In spite of this, van der Beek et al. and Schildgen et al. (2018) argue that the existence of deep valleys refutes the Herman et al. (2013) results and that the existence of these valleys implies that the thermochron-derived increase is “spurious” in areas including the Alps, the Olympics, Patagonia and Fjordland, New Zealand (Schildgen et al., 2018, Table 1). Their argument seems to be that because the volume of sediment removed through valley incision is smaller than for spatially uniform erosion, the result is somehow invalid. We see no justification for this argument. Herman et al. made only a qualitative comparison to sediment flux, but made no argument that the raw numbers coming out of the inversion were representative of either global or regional sediment fluxes (AC3). The important conclusions of Herman et al. were that erosion rates in mountainous regions increased and that the increase was measurable by thermochronometric data. With deep valley incision, we not only have the high erosion rates and high relief needed for thermochronometry to accurately measure changes in valley erosion rate, we also have the incidental geomorphic evidence for a recent increase in erosion rate (i.e. deep valleys, less eroded high topography). In other words, we have the needed conditions to resolve an erosion rate increase and independent evidence that such an increase has occurred. Schildgen et al.’s (2018) assessment of “spurious” is based entirely on a *hypothetical* misuse of the Herman et al. results by some future studies that *might* interpret these results in terms of sediment volume. The thermochronometric data, and the geomorphology support an increase in erosion rate. Schildgen et al’s (2018) analysis finds support for an increase in erosion rate (recall that they claim their analysis is focussed only on sign). Herman et al. (2013) find an increase in erosion rate. And yet these cases are classified as “spurious”. This highlights another case where Schildgen et al.'s (2018) summary Table 1 and Supplement Figure 2 have attached the label of “spurious” to sites and analyses for such disparate reasons as to render the categorization meaningless.

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.

The part about the Pennenic Line as a normal fault is partially correct; we were in part responding to conference talks and even the pre-review versions of the Schildgen et al (2018) paper which we saw and not only suggested, but explicitly stated that the Penninic Line was a normal fault. This material seems to have been omitted from the final published version of Schildgen et al., so we will check that we are not commenting on statements no longer made by the authors. However, the main reason this issue is important is because Schildgen et al. in both model and interpretation of the data assume that there is a sharp boundary between the external and internal Alps. There is no support for this, with multiple sub-5 Ma ages south of the fault, and geochronologic evidence that the Penninic Line has not been active since the mid-Miocene. Observed ages can easily be fit by a broad doming uplift of the Alps, not a vertical fault. This is one reason why the ages in their “Alps model” is very different from the real Alps.

The aim of the synthetic test has been discussed in AC2 and AC5. van der Beek et al (lines 166-182) go into great detail as to how our model was constructed incorrectly because the synthetic data differ from that of the Alps, and now it is claimed this was not meant to be a realistic simulation of the Alps. Both cannot be true. In any case, we understand perfectly the point of any synthetic data models and have extensive discussion in our paper in the introduction to section 4. The long introduction to our modeling section is in part because Schildgen et al. did not define the goals for their models, and so did not differentiate between model errors, resolution errors, bias to the prior or spatial correlation bias so their models were were not designed, nor sufficient, to test for any specific errors, and we want to be careful not to repeat this error.

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 40Ar/39Ar 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 40Ar/39Ar 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 40Ar/39Ar ages outside the massif are > 10 Ma without exception, and they are mostly > 20 Ma. Moreover, careful interpretation of mica 40Ar/39Ar 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).

Herman et al used hornblende and muscovite 40Ar/39Ar data from Treloar et al. 2000, and used in the compilation of Thiede and Ehlers (2013). These data were inadvertently omitted from the data compilation published by Herman et al. (2013), but this information was provided to the Nature editors and to Schildgen et al. prior to their publication in 2018. Treloar et al. included data from Zeitler (1986) and report 5 Muscovite ages between 4 Ma and 8 Ma from the core or cover sequence of the Nanga Parbat massif in the Indus valley and 2 muscovite ages in the same range from the Astor valley. In addition, there are 4 Hb argon ages less than 20 Ma within the basement core. These are all internal to the Main Mantle Thrust bounding the syntaxis core. In addition, Treloar et al. report a large number of biotite Ar/Ar ages from the same area that were roughly consistent with the muscovite ages, but showed some evidence of excess argon, so we did not include these ages. Similarly, we did not include the biotite ages of Schneider et al. that were distributed across the core of the Nanga Parbat massif to the south, even though they do not appear to suffer from excess argon.

Even though Schildgen et al. (2018) were aware that the argon ages were included in the Herman study, they make no mention of these data or the importance of these to the result or resolution in the Herman et al. (2013). Instead, they present the case as one of only young ages in the core and only old ages outside the core. van der Beek et al now present this information as if seeing it for the first time, and still describe the data incorrectly as “no in situ data within the massif prior to 4 Ma”. This is not consistent with the age data of Treloar et al. (2001). If they have some argument why these data are not usable, they should submit those arguments for peer review. In any case, the result of Herman et al. (2013) is based on these data, so the van der Beek et al. argument that this is one of “clearest instances of the spatial correlation bias” is incorrect. The Herman et al. result comes from differences in age between the muscovite argon ages and the zircon and apatite fission track ages, all located within the metamorphic core. Even if one did argue to remove the argon data, there is a well resolved increase in erosion rate from 4 Ma to the present, given the extraordinary number of young (<4 Ma) fission track ages from the core region, so a real acceleration is also supported by the local fission track data, just for a slightly younger timeframe.

The Schildgen et al. (2018) description of the situation at Nanga Parbat is a good example of their tendency to equate a conclusion to a hypothesis without hypothesis testing (AC 6). They stated (and van der Beek et al. confirm) that “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” A test of this hypothesis is easy to apply. We removed all data from outside the syntaxis and reran the inversion. Results are shown in Figure 1 below. There is no significant difference in the erosion rates over the last 6 Ma in the core region. The hypothesis is disproven. We could include this model in our paper, but don’t see that it is necessary to keep showing identical results.

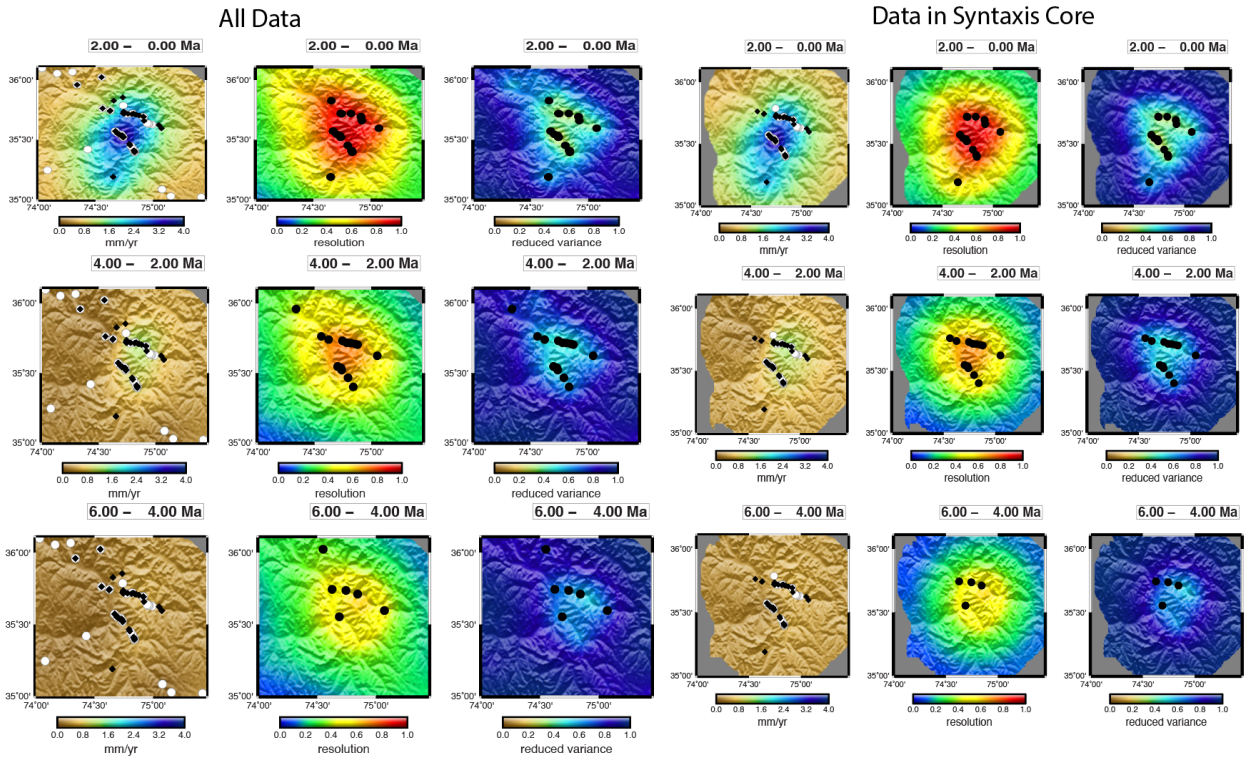

FIgure 1: Glide result for thermochronometric inversion of data from Nanga Parbat region. Left includes all data from the region and is shown in ESURF2020-59 as figure 20. On the right is the same inversion, but removing all data outside the core as defined by the MMT on east and west of the syntaxis. See Figure 20 caption for additional information on data format.

The “bleeding” into regions outside the massif is just smoothing into regions of low resolution and has no importance. The smoothing is greatly exaggerated by the NR calculation (Fig. 20j), but with no statistical significance outside the immediate region of the data (Fig. 20k).

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, 365 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.

We are not sure where these U-Pb ages are, but they are not consistent with any of the thermochronometry data within the massif. One cannot be consistent with the tens of AFT, ZFT and Ar-Ar data from the massif, if one takes these U-Pb ages as representative of the regional exhumation rate. This does not indicate a problem with the Herman et al analysis, but the kind of typical contradictions that appear in complex geologic settings where one attempts simplistic interpretations. We prefer to take the thermochronometric ages as representative of the regional erosion rates, particularly over the last 5 Ma. We also note that the resolved solutions are limited to the northern (Haramosh) massif and Indus valley. The highest rates of exhumation in the Nanga Parbat region are to the south as are most of the U-Pb ages.

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.

This comment is incorrect in every particular. There is no proof of curved particle paths. There are some models for the kinematics of the Olympic wedge that include a horizontal component to the particle paths, but these are not necessarily curved during the ascent through the relevant closure temperatures to the surface. A particle path that is linear even with a horizontal component has no false acceleration. We cited the recent work of Michel et al., 2018, 2019 (esurf2020-59 Line 1054) that did detailed thermo-kinematic modeling and found that 1-D models do not differ significantly from 2-D models that include the effects van der Beek et al. argue would be important. This is another example where van der Beek et al. have made a hypothesis with no testing (AC6). The tests were made by others (Michel et al. 2018) and the hypothesis was proven wrong. These papers were published several years ago and cited in our paper, so we don't see why van der Beek et al choose to ignore the results. The deep glacial valley incision cited by van der Beek et al. as a problem is, on the contrary, verification of the recent increase in erosion rate as we discussed above.

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.

We don't see the problems here. This is all conjecture with no testing or demonstration that any of these potential problems are important. If van der Beek et al think these are important effects, test them and publish the tests. Pointing out that there are complications that van der Beek et al can't explain is not a scientific demonstration that the Herman et al results are wrong. See AC6. Our main point of this section is that Schildgen et al. incorrectly attributed the Herman et al result to spatial averaging, when there was a good explanation of the result in terms of local data. Similar to the Nanga Parbat example, this is one of the many examples of how Schildgen et al.'s lack of hypothesis testing results in a false identification of "spatial correlation bias". As to what the true exhumation history of the region is, there are many potential data errors and we look forward to a demonstration that they are important. There is potential for results to all be a tectonic signal, as is true for many other sites, but Herman et al. did not claim otherwise. In any case, it is not proven and for the most part, the methods for differentiating tectonic and climate signals do not exist. Even single fault blocks can experience a change in exhumation rate or boundary fault slip rate through unloading stresses, all driven by climate change induced erosion. As in other cases, neither we nor Herman et al. venture an opinion, because in our view, it is unanswerable.

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.

Fiordland appears in Schildgen et al. Table 1 and Fig. ED2 as Spurious/Tectonic. If the authors now think that the increase in erosion rate is real and due to glacial erosion, they have misplotted it in their summary figures and tables. Although we do now recognize that "spurious" is defined so broadly that it can also include regions where the erosion rate change is real, but sediment volume could be misconstrued by some future study.

Local valley incision is confirmation of an increase in erosion rate, not a mistake (See discussion above). Influence of the subduction zone is a hypothesis with no testing, but in any case this is not relevant to questions of fidelity of the analysis method or the existence of a change in erosion rate. The "largest increases...SE of the main range" is an artifact of the normalized erosion rate change of Schildgen et al. as discussed in our paper (line 1100). This is one of the examples where Schildgen et al. focus on the region at the edge of resolution, not appreciating that 10 plus or minus 20 is not necessarily larger than 2 plus or minus 0.5. We have not changed our estimate of appropriate resolution, we have simply pointed out that it is

better to interpret the well resolved parts of the estimate, not the poorly resolved parts, and keep in mind that it is a relative quantity.

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.

We discuss this in AC6. This paragraph does not provide a usable definition. It is not possible to reproduce the reported findings of Schildgen et al. Table 1 or Figure ED2 using this definition. “Spurious” includes a mix of unrelated and untested factors including data errors, model errors, and interpretations where the authors believe that even real increases in erosion rate should be classified as spurious because the erosion is concentrated in valleys. The significance of data and model errors were untested. No examples of operator error were provided, although we expect this follows from a misunderstanding as to what operator error is. The inclusion of valley incision as a criterion for a “spurious increase”, when it is in fact an independent confirmation of an erosion rate increase, is particularly problematic. The terms “false” and “fake” are superfluous derogatory descriptors, also unreproducible, with an insinuation of intentional fraud, and whose libelous meaning we will simply ignore.

We agree that we should make clear how Schildgen et al. used this term and we will change the section where we describe their use of the term to better describe how they have used it. We will rewrite this section of our paper to explain that we use a simplified definition in order to have a testable hypothesis, and that we are not addressing those parts of their interpretation that are untested (data errors) or untestable (model errors where the kinematics are unknown), or simply interpretations that do not correctly describe thermochronometry outcomes (valley incision).

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.

This is addressed in comments AC2 and AC5. To summarize, the Schildgen et al. geotherm error is far above significant. As they have not recalculated ages correctly and rerun GLIDE models to test this error, they have no basis for making any statement to the contrary. The bias in our inversion models exists only when the resolution of the data is poor as we indicated in our paper. The fact that van der Beek et al can alter our data until resolution is poor is neither a test nor a proof of anything except our original conclusions: data resolution is key, and poor resolution data will have errors. The real problem in these and any estimation methods is to effectively estimate resolution to recognize the nature, source, magnitude and direction of these errors. Faults with sufficient data of the appropriate age on both sides can be resolved by our method, even with a large correlation length subject to resolution characteristics of the data.

Van der Beek et al make no new analysis regarding post-processing operators or “cartoons”, nor have they responded to our error analysis of the NR or to the errors in their kinematic models, so we do not see where they have any basis for conclusions, nor do we find these points irrelevant.

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).

We addressed this in several places, but primarily in AC6. Neither Schildgen et al. (2018) nor van der Beek et al present any tests, so the errors listed here are untested hypotheses. They conducted no correct models using GLIDE, reproduced none of the Herman et al. (2013) results with and without questionable data (robustness testing), and presented no alternative models. There is no basis for these conclusions. These are valid hypotheses and perhaps future work will test them and demonstrate that many are legitimate, but until then, they remain untested hypotheses, not demonstrated conclusions.

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.

Neither Herman et al. (2013) nor esurf2020-59 proposed worldwide erosion rates, so the first statement is misdirected. See AC3.

van der Beek et al constructed no alternative models, provided no reproducible tests and do not “know” the correct erosion rate history for the Earth, so there is no basis for the final statement.

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.

As we addressed in comment AC6, Schildgen et al. (2018) did no testing of any of the data or model errors (incorrectly referred to as operator error) listed above. We, the authors of this paper, have extensive experience in modeling thermochronometric data and disagree that we have selected the trivial aspects of the models. If van der Beek et al. think the factors they list are important, they should provide the tests and demonstrate this. Until then, these are unsubstantiated hypotheses. We focussed on the factors that we felt were most important - the geotherm errors of the Schildgen paper, which were serious enough to invalidate their only tests, and the spatial correlation bias, because this was presented by Schildgen et al. as the fundamental problem with the Glide methods. This priority was set by Schildgen et al. through the title of their paper, the focus of nearly every figure of their paper, and was reiterated by van der Beek et al in their comment where they referred to the spatial correlation bias as the dominant error (e.g. SC2, Line 249). Reversion to the prior is an important issue in large part because Schildgen et al. seemed unaware that it even existed and so attributed many cases of this to spatial correlation bias.

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.

Addressed in AC2 and AC5.

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.

As to the first statement, we stand by our original position that the models presented by Schildgen et al. are not valid as discussed in AC2 and AC5. The tests were invalid.

As for the second statement, van der Beek et al. are mixing two independent processes. Age variations due to elevation and age variations due to spatial variability in erosion rate are two completely different processes and variables. Age variations due to elevation are a consequence of different path lengths, but sample a common exhumation history; a thermal model-based analysis will have no problem fitting all ages with no bias. Age-elevation profiles ONLY work where there is a common exhumation history. Age variations due to spatial variations in erosion rate, even with additional variations in elevation, will produce a widely distributed spread of age across age-elevation space, as shown in our Figure 2d. Only some fortuitous combination of erosion rate, elevation and closure temperature will produce a monotonic function across this space. This is, in fact, what Schildgen et al. (2018) did in their Figure 2b - they covaried the closure temperature, elevation and age until they obtained a monotonic function. The probability of this happening in nature is vanishingly small.

We could clarify our paper in a number of places to avoid this technical difference explaining that when we state "same closure temperature", we imply "same elevation", because variations in elevation are equivalent to variations in closure temperature. We admit that it became tedious to state this every time we discussed closure temperatures and multiple thermochronometric systems, so in many places we omitted to say explicitly "neglecting variation in elevation". We did say it many times including at lines: 132, 136, 159, 231, 508, 641, 768, 1131, 1190, 1323, 1395, and probably more. When discussing the resolution models, we were explicit and stated: (line 641): "Age variations with elevation increase the time range and are just as valuable as multiple thermochronometers in this regard, but we will not extensively investigate this aspect of the problem." At some point, we assumed that readers would understand this principle and we would not need to add the caveat "or exhibit an advantageous elevation distribution". Apparently, we were wrong, and we will expand the text to include this in more places.

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 (UTh)/He data were included in the inversion for southern Peru.

The same clarification can also be made for this statement at line 1194 and elsewhere if we discover more places where this would be helpful. We will add " or have a fortuitous distribution in elevation", at which point this statement is correct.

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.

Our paper goes through the arguments extensively. We demonstrated that the Schildgen et al. synthetic data tests were invalid, and the remainder of their paper was untested conjecture. This does not prove that Herman et al. (2013) is correct, simply that the conjecture by Schildgen et al. is untested, so there is no reason to reject the Herman et al. results at this time.

We do not understand the obsession with numerical values of the resolution parameter that we have repeatedly explained is a relative metric. Nor do we ever state a cut-off of 0.5 is universally appropriate or that the value used in Herman et al. is universally inappropriate. There is a ranking of sites with different confidence levels from 100% of the Herman et al sites with a lower confidence to a smaller number of sites with a higher confidence. We could take a different approach and simply take the best resolved point in each geographic area, regardless of its resolution value. There are many ways to use statistical criteria; black-boxing an analysis with a fixed set of rules or cutoffs is one of the worst.

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).

None of the errors identified in Schildgen et al. were tested or demonstrated to be significant. We repeat that Schildgen et al. present no tests or new analysis, so they have no basis for commenting on the robustness of Herman et al. results. We do not accept that their interpretation of the literature qualifies as a reproducible test.

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.

We have run hundreds of Glide models and would never venture to predict how resolution correlates to bias-to-prior; van der Beek et al have no experience with GLIDE and yet are willing to speculate on this relationship. Bold, but it remains speculation until a test is provided (AC6). Their statement that “reversion-to-the-prior bias is still dwarfed by the spatial correlation bias” is even bolder given that they made no quantitative estimate of either error, so they are comparing two quantities for which they have no experimental information and yet are willing to make a conclusion as to which is the larger.

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.

Published models are just models. Thermo-kinematic models are not unquestionably correct. If they are not parameterized in a way that permits changes in climate-driven erosion rate, they will not find it. If they are parameterized in a way that permits variations in both tectonic uplift rates and climate-driven erosion rates, they still need to determine cause and effect between the tectonics, climate and erosion. In general, thermo-kinematic models cannot address the cause and effect problem because they are kinematic, not dynamic - they cannot predict erosion-stress-isostasy feedback, because they do not include stress. There is no solution to this problem, which is why it is referred to as a paradox and has been for 30 years. Schildgen et al. have not engaged with this problem and simply take model results at face value, or worse, selectively. Even individual fault blocks will have displacements and motions affected by erosion and changes in erosion rate, so scale does not avoid this problem. Where Schildgen et al. discuss that both tectonics or climate could be important (e.g. St. Elias), they make no mention of the feedback between tectonics and climate-driven erosion changes, they simply acknowledge that other authors have recognized this problem. As we write in our paper, they made a specific determination for cause (tectonic or glacial), something that is widely regarded as impossible.

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., 535 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.

Our quote above does not say that the papers were not cited. It states that these papers were cited incorrectly as supporting the authors hypothesis, when in fact they did not. Each of those papers argued for an increase in erosion rate, yet most of those sites appear in Schildgen et al. Table 1 as “spurious” or “spurious/tectonic”. Much can be hidden behind the authors catch-all term “spurious”, which includes a wide range of speculation and interpretation, but the new definition van der Beek et al. give above states the spurious is synonymous with “false”. Each of the papers listed here concludes that an increase in erosion rate is true. True is the opposite of false. Schildgen et al. were very selective on which part of these papers they would accept and which they would not, which is the definition of confirmation bias.

The main point of this section was that Schildgen et al. did not differentiate between tectonic activity and accelerated tectonic activity, although the latter is needed to explain an erosion rate increase as was identified by most of those cited papers. They simply used tectonic activity as validation of their causal interpretation.