Earth Surf. Dynam. Discuss.,
https://doi.org/10.5194/esurf-2020-59-AC2, 2020

Earth Surface Dynamics Discussions

Interactive comment on "Bias and error in modelling thermochronometric data: resolving a potential increase in Plio-Pleistocene erosion rate" *by* Sean D. Willett et al.

Sean D. Willett et al.

swillett@erdw.ethz.ch

Received and published: 5 October 2020

Response by the authors to the Comment by van der Beek et al.

The comment offered by van der Beek et al. is long. There are many issues in their comment that we will clarify in subsequent responses, but we begin with a couple of the more important points here. We do this in the spirit of an open discussion format offered by this journal and for the benefit of the broader ESurf community so as to avoid additional confusion in the thermochronometer community concerning the application of inverse models to thermochronometer data.

Response to comments on Model error in geotherm calculations

The most important point to address is the existence and importance of the errors made in the geotherm calculation in Schildgen et al. (2018) and demonstrated in Willett et al. (2020). In the section starting at line 84, van der Beek et al. (2020) acknowledge that they made the error, but claim that it is not a significant error. We do not see where they have demonstrated that this error is insignificant, and do not find any reason to take results of an analysis with acknowledged errors, over that conducted in our paper, for reasons that we will discuss here.

The fundamental problem is that one cannot use a steady state geotherm in a forward model to generate ages, invert them with a transient thermal model and expect to obtain a meaningful result. Regardless as to how one calibrates the two, a steady geotherm cannot approximate a transient geotherm over a significant period of time. It seems that we now agree on this point. However, van der Beek et al. argue that the steady solution based on a 30 km fixed temperature boundary condition is actually more accurate than the fixed gradient model included in Glide. This issue is not important to the error of mismatching forward and inverse models, but it is important to the thermochronometry community who need to accurately calculate geotherms, so we will digress for one (long) paragraph to discuss this issue.

We agree that there are settings where crustal thickening or underthrusting leads to downward advection of heat. There are also extensional systems, in which tectonic advection leads to upward advection of heat. Erosion leads to upward advection of heat. Any change of crustal thickness leads to changes in heat production (by crustal thickening or thinning) and thermal resistance. All of these processes lead to transience of the thermal field. However, the timescale of relaxation of this transience is determined by the thickness and heat content of the thermal lithosphere, which extends through the thermal boundary layer of the upper mantle to the asthenosphere. This has been well established since the first papers in plate tectonics (Williams and von Herzen, 1974; Sclater and Francheteau, 1970; McKenzie (1978); England and

Thompson; 1984; Furlong and Chapman, 2013). We have seen nothing in the subsequent 50 years to change this observation and tens of thousands of papers that have accepted and built upon the concept of the thermal lithosphere, so the claim that a base of lithosphere boundary condition has “exaggerated transience” (line 126) is not supported in the literature. There are studies that have chosen to approximate lithospheric transience with a thermal boundary condition at a shallower depth in the lithosphere, often because of computational limitations, but this has been (or should have been) recognized as an approximation. This approximation for a basal boundary condition can be either a constant temperature or constant gradient (also referred to as a constant flux condition based on Fourier’s Law) and can be applied at a material point, e.g. at the base of the crust, which moves up or down with time, or at a spatial point, e.g. at a constant depth. A constant gradient has the advantages that the time to a new steady state is longer than the fixed temperature approximations, and is therefore closer to that of the full physical problem. The constant gradient and material point boundaries have the advantage that advective heating or cooling through the boundary is approximated, by changing temperature in the former, and changing position in the latter. The worst approximation is a constant temperature applied at a constant depth. The response time for a temperature boundary condition applied at a fixed, and shallow, depth is much shorter than the lithosphere problem and the fixed temperature suppresses all advective heat flux across this boundary, so that energy is not conserved at the boundary. Depending on the tectonic setting, and the timescale of interest, these approximations for a basal boundary condition provide different levels of accuracy. van der Beek et al argue that because one setting (underthrusting during contraction) has a downward component to its advection, an isothermal boundary at a constant and very shallow depth (e.g. 30 km) gives a better representation of the temperature field than a model that has no approximation to the thermal timescale, but includes only the erosional component to the advection. With crustal thickening there is a downward component to advection. Neither Glide nor PECUBE nor the 1-D model of Schildgen et al. (2018) include this downward advection, so all have a positive error

in temperature. van der Beek et al. argue that because the fixed 30-km-temperature boundary condition introduces a negative temperature error, i.e. it is much too cold, these two errors will offset each other to obtain a more accurate solution. We would like to see such a fortuitous outcome demonstrated, or supported with references. We are not aware of any previous studies documenting this in the geothermics community. In the van der Beek et al. comment, there is no model or error analysis offered and therefore no assurance that the negative error is not two or ten times larger than the positive error. Nor is there any reason to presume that spatial distributions in these errors will match. van der Beek et al. do not address extensional settings, where these errors would add constructively, or non-tectonic settings, where the fixed temperature condition provides the only error. In most convergent settings, the downward advection is more than compensated for by the increase in heat production, so orogens tend to have high lower crustal temperatures (Pope and Willett, 1998; Beaumont et al., 2004; review by Furlong and Chapman, 2013). Only subduction forearcs are characterized by lower than average crustal temperatures. Our model of a constant flux applied at a large depth will correctly represent the timescale of the thermal lithosphere and correctly includes advection due to isostatic uplift in response to erosion, but neglects changes in heat production and neglects advection due to vertical velocity gradients within the lithosphere that are negative for crustal thickening and positive for crustal thinning. Where erosion rates are large with respect to the vertical strain rate integrated over the lithospheric column, the model will be accurate.

To return to the more relevant issue, van der Beek argue that their geotherm error cannot be the sole error in the model because a model with no spatial variation in erosion rate does not result in an inferred acceleration of erosion (their Figure 1), as predicted by Figure 3 of Willett et al. (2020). However, this statement is not correct. The geotherms in Figure 3 are for a constant exhumation rate of 1 mm/yr. By including several tens of ages, incorrectly calculated, as input to the inverse model, the model they generated no longer recovers the correct erosion rates or the correct geotherm, which is coupled to the erosion rate. The geotherms in van der Beek et al., Figure 1

are not identical to the ones shown in our Figure 3 and these differences can easily account for the decceleration observed in the van der Beek et al result. We see no evidence that this model is anything other than exactly as we stated – in error because of the errors in the age calculation. To clarify the manuscript, we will change the caption to Figure 3, to express this caveat, to state: Schildgen et al. (2018) used synthetic ages calculated using a steady state geotherm similar to the blue curve, inverted them using temperatures predicted from a transient geotherm similar to the red curves, and concluded that the failure to recover the correct exhumation rate demonstrated a problem with spatial correlation in the GLIDE inversion method.

The second point made by van der Beek et al. is that the 16 to 28% errors that they do obtain are too small to be important, because these errors are "within the range of typical uncertainty associated with quantitative inferences of exhumation rates from thermochronology data" (line 114). There is no evidence that these errors are negligible, and it is unlikely that they will be. First, this is a single example and there is no assurance that another example will not give 40% or 80% error. In any case, we would argue that 20% error is over any significance level. These models are control experiments, whose sole purpose is to control for, and thus identify, error. They are not intended to simulate typical conditions with typical uncertainty (see discussion below). We remind the reader that the model includes a vertical, dip-slip fault with 36 km of exhumation, a feature that does not exist on Earth, so these are not typical conditions. The point of these models is to systematically control sources of error, in order to identify their relative importance (see our paper, Section 4.0 and 4.1). In fact, if an error cannot be eliminated entirely, the model loses its purpose, as errors in an outcome will remain ambiguous as to their source.

For this reason, we continue to see no reason to consider relevant any of the models of the van der Beek et al comment or Schildgen et al. (2018), based on PECUBE ages analysed through GLIDE. They contain errors in the ages of up to 100% and these translate into errors of tens of percent in erosion rate (at a minimum), and of unknown

magnitude and direction in acceleration. Even if there are other errors, such as spatial correlation bias, these can never be separated from the geotherm errors, so these models do not meet the intent of their construction, to identify the source and relative magnitude of errors. In contrast, the models in our paper have zero age error and have been constructed to isolate specific errors. Further support for our position is offered in the following section.

Response to: Bias to the prior vs spatial correlation bias

The second important point that we need to address is the way in which the synthetic models are constructed and what they show. This is first addressed in van der Beek et al 's comment section 2. They present several models (Figs 1 and 2) with large errors that they argue show combinations of resolution and correlation bias. However, these models also contain the geotherm errors discussed above. They also add random errors to the data, which adds a fourth source of error to the problem. Adding another dimension to the error is opposite to what these control experiments are intended to do – isolate, not compound, errors. Having four potential sources of error makes it impossible to identify which is responsible for the observed error. In contrast, the models presented in our paper isolate errors, in particular by removing the geotherm error, and not adding random noise. The comparable models in our paper show none of the same errors, even when data from both sides of the fault are included (e.g. Figs. 6, 14). We have not run the analysis on data from only one side of the fault, but if there is no error in the combined data, there will be no error in the simpler case. The simplest conclusion is that all error in the van der Beek et al. comment and Schildgen et al. (2018) models is due to these geotherm errors, possibly with a component due to the added noise.

Rather than accept this conclusion, van der Beek et al. argue that we have "cleverly designed" (Line 163) the data set in order to avoid the errors produced by their models. Figure 5 of our paper shows the model data sets that we generated. We attach here a modified version of this figure that includes the data distribution of the Alps in a similar

format for comparison, and will be included in the revised paper.

The first point to make is that comparing the model data sets to the Alpine data in the detail that van der Beek et al. do in lines 172 to 180 is not appropriate. We could refer to the extensive discussion in our paper regarding the purpose of these models (Section 4.0-4.1), but it is easier to quote directly from van der Beek et al.'s comment (line 338): we "used it [the synthetic test] 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". We understood this, elaborated extensively on the purpose of the tests in our paper and agree completely with this statement. But recognizing this purpose, counting numbers of co-located points, looking at elevations and numbers of ages from specific chronometers serves no purpose. If we are not attempting a realistic simulation of the Alps, there is no reason to match details of the data characteristics, particularly while ignoring the actual values of the ages. The model includes a vertical dip-slip fault with 36 km of exhumation over 36 Million years, a feature not present in the Alps, so a model which matches the sample locations, elevations and thermochronometer type, would, and does, fail to match the ages. We also should point out there was no "design" element to the creation of these data sets, in the sense that we did not test them, modify values or in any way iterate until we found a desired result. It makes no sense that we would do so, given that we also included data sets, B, C, and E, which all have poor distributions and resolution, so were clearly not designed with some hidden intent. The purpose of these models is to contrast high-resolution and low-resolution examples, and thereby isolate resolution errors. We made no pretenses about this fact. Data sets A and D are intended to be high resolution tests and are explicitly constructed to do this.

Data sets A and D are intended to show high resolution data; the Alps have this feature in common with these synthetic data, but it is the commonality in ages, and thus resolution, that is important, not specifics of the metadata. As much of our paper argues, resolution is defined by the values of the ages and their distance from the closure

isotherm. Because we are interested in the 6 to 0 Ma time frame, this requires dense age coverage of this time interval, and if we also want to avoid geotherm advection errors in the fully coupled model, and maintain its independence from the prior, we require some age coverage of the preceding time interval. The External Alps data have very good coverage over this time window, even with only 3 thermochronometers, because AFT ages are spread across a range of 2 to 14 Ma and ZFT ages are typically between 8 and 20 Ma. However, the synthetic model with its high geothermal gradient and high erosion rate, has no ZFT ages over 8 Ma and no AFT ages over 6 Ma, even with the greater elevation range used for our synthetic data. To obtain a high-resolution model test with this uplift function, it is necessary to push back the age range, either by changing the closure temperature of the existing chronometers, increasing the elevation range, or by adding a higher temperature system. We chose to do the latter two. The fact that there are no muscovite 40Ar/39Ar data in the actual Alps is not relevant; as pointed out by van der Beek et al. themselves (see quote above); we are not trying to model the Alps, we are attempting to model a "densely sampled" example. We cannot have a dense sampling if we do not have ages in the appropriate time interval. As we discuss in section 4.2, as a high-resolution test, it would be completely appropriate to add many more data and many more systems, even approaching an infinite number of data. The point of this model is to eliminate the resolution errors, leaving only the spatial correlation error, and following standard practice in error analysis, we are justified in adding as many data as necessary to accomplish this. In fact, we have been restrained, and limited the number of data generated to something similar to the natural Alps (new Figure 5).

The analysis shown by van der Beek et al. in their Figure 3 has removed the point of the resolution test. van der Beek et al. make two analyses of our high-resolution model of Figure 14. First, they look at the earlier timesteps of the model. Because these timesteps are not the target of the experiment, we did not generate ages over these intervals, they are poorly resolved and there is a correspondingly larger resolution error. Second, they remove all the age data from the 8 Ma to 14 Ma time window, finding that

there are subsequently errors in this timeframe and some of the younger timesteps. There is no revelation in this result. By removing all the key ages, they have converted a high data-density model into a low data-density model and thereby reintroduced the resolution errors that this model was intended to remove. This has nothing to do with spatial correlation bias or any other error. The model response to increasing resolution error is complex because of the coupling of resolution errors to the model errors associated with the geotherm. As we explain in our paper (section 4.5), with the thermal coupling, resolution errors are compounded, which is why it is important to have age coverage prior to the timeframe of interest, so as to minimize advection errors. Also, as discussed at several places in our paper, the temporal resolution metrics do not reflect these geotherm errors. We also note that these geotherm errors are much higher than any natural settings because of the extreme advection due to 36 Ma of erosion at 1 mm/yr. This synthetic test is very challenging to our model as it would be to any other. We could go back and redesign a different synthetic uplift model with lower erosion rates, closer to what is actually observed in the Alps, so it would not be necessary to introduce 40Ar/39Ar data to obtain high resolution, but this could be perceived as designing the test in our favour, so we have retained the uplift function of Schildgen et al. (2018).

As a final point, we strongly object to the casual dismissal of models using the true solution as a prior model (line 158). van der Beek et al. state that because the true solution is not known, it cannot be used as the prior in practice, and so these models are not relevant. However, much like the discussion above, that does not acknowledge their own point, that this is "not to attempt a realistic simulation of the European Alps", or to demonstrate how the model performs under typical real-world conditions. We are attempting to learn something about the method, not trying to simulate the way in which the model is applied in normal circumstances. We remind the reader that the model includes a vertical, dip-slip fault with 36 km of exhumation, a feature that does not exist on Earth, so there is nothing typical about this exercise. These experiment results (our Figs. 8 and 13) are actually quite remarkable. We have taken a model

result with large errors (take any figure with errors in our paper – they all respond the same way), changed one number in the inversion parameters (the prior erosion rate in the NW corner of the model), and the errors are gone – not reduced, but effectively gone. This requires an explanation, not casual dismissal. Suppose for a moment errors are a mix of resolution errors and spatial correlation bias, which is a model error. There is no theoretical reason for model errors to be related to the prior model in a Bayesian inversion. In fact, it goes against the very principle of model errors, that they are built into the parameterization and thus unavoidable through the addition of data or changing of inversion details, including parameter priors. In explaining spatial correlation bias, neither Schildgen et al. (2018) nor Willenbring and Jerolmack (2016) made any statement that the bias applies only to Bayesian models with an incorrect prior. Nor should they; this would make no sense. An averaging error should apply to all models that include that averaging. On the other hand, theory (Eqn. 9 in Willett et la., 2020) shows that resolution errors in a Bayesian model are eliminated if the prior and true solution are equal. In this experiment, model (spatial averaging) errors should be retained, but resolution errors eliminated. The outcome is: all errors are eliminated. The conclusion reached from this is that all errors in the GLIDE inversion are resolution errors. The fact that this is also the conclusion of the analysis comparing high and low resolution models provides additional confirmation - theory, and two control experiments align with a common outcome: there is no spatial correlation bias as defined by Schildgen et al. (2018). van der Beek et al. provide no explanation for this experiment. We anticipate a response that this is a rhetorical difference and that spatial correlation bias is a type of resolution error. However, Schildgen et al. (2018) never state this and all their treatment and analysis imply that they regard it as a model error. If they now suggest that spatial correlation bias is a resolution error, they must accept the implications. In particular, resolution errors go to zero with high data density. Many examples, e.g. Alps, are argued by Schildgen et al. (2018) to have both high data density (van der Beek comment, line 338) and to be dominated by spatial correlation bias (Schildgen et al., 2018). These cannot both be true if spatial correlation bias is a resolution error.

The main point of this discussion is that we see no evidence of anything wrong in our synthetic tests. Nor anything incorrect in the methodology of generating the data. We see no need to modify our paper, nor even how this would be done without removing the essential point of the error analysis.

References Cited:

McKenzie, D., 1978, Some remarks on the development of sedimentary basins: Earth And Planetary Science Letters, v. 40, no. 1, p. 25–32, doi: 10.1016/0012-821X(78)90071-7. Beaumont, C., Jamieson, R., Nguyen, M., and Medvedev, S., 2004, Crustal channel flows: 1. Numerical models with applications to the tectonics of the Himalayan-Tibetan orogen: Journal Of Geophysical Research-Solid Earth, v. 109, no. B6, p. B06406, doi: 10.1029/2003JB002809. Furlong, K.P., and Chapman, D.S., 2013, Heat Flow, Heat Generation, and the Thermal State of the Lithosphere: Annual Review of Earth and Planetary Sciences, v. 41, no. 1, p. 385–410, doi: 10.1146/annurev.earth.031208.100051. Pope, D., and Willett, S.D., 1998, Thermal-mechanical model for crustal thickening in the central Andes driven by ablative subduction: Geology, v. 26, no. 6, p. 511–514. Sclater, J.G., and Francheteau, J., 1970, The Implications of Terrestrial Heat Flow Observations on Current Tectonic and Geochemical Models of the Crust and Upper Mantle of the Earth: Geophysical Journal of The Royal Astronomical Society, v. 20, no. 5, p. 509–542, doi: 10.1111/j.1365-246X.1970.tb06089.x. Williams, D.L., and Herzen Geology, Von, R.P. Heat loss from the Earth: new estimate, doi: 10.1130/0091-7613(1974)2<327:HLFTEN>2.0.CO;2.

Figure 5: Synthetic age data sets for model bias and resolution tests, comprising two zones with differing erosion rate. Colors represent ages from different thermochronometer systems corresponding in closure temperature to: green: AHe, blue: AFT, yellow: ZHe, red: ZFT and black: muscovite 40Ar/39Ar. Data Sets A, B, C were generated using a constant geothermal gradient. Data sets D, E were generated using the GLIDE transient thermal model with a flux boundary condition. For reference, (f) shows the measured distribution for the full Alps and for the external Alps.

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

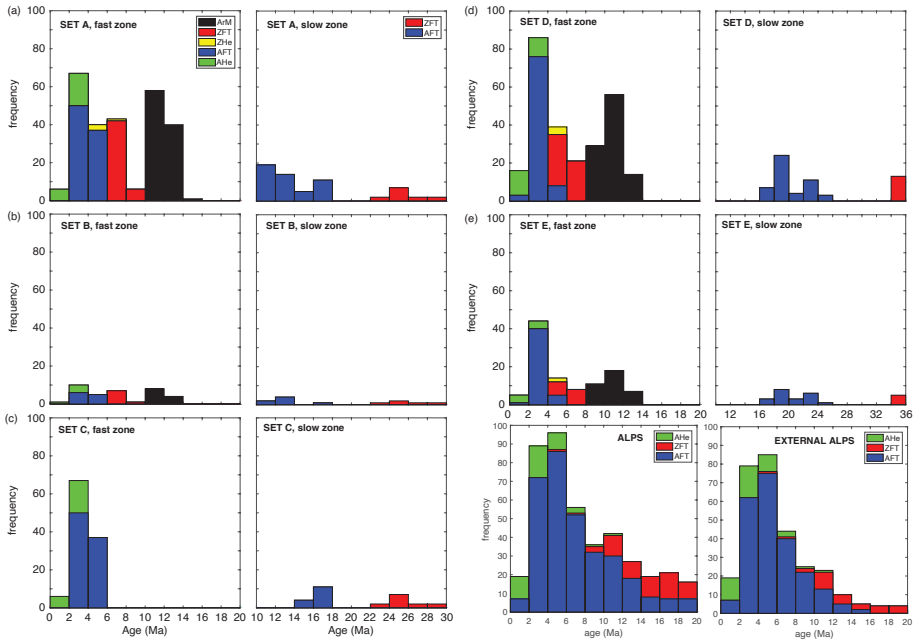
(a)
SET A, fast zone
SET A, slow zone
A+M
ZFT
ZHe
AFT
AHe
(b)
SET B, fast zone
SET B, slow zone
(c)
SET C, fast zone
SET C, slow zone
(d)
SET D, fast zone
SET D, slow zone
(e)
SET E, fast zone
SET E, slow zone
ALPS
EXTERNAL ALPS
frequency
Age (Ma)
age (Ma)

Fig. 1.