Author Response to SE-2017-117-RC1 (P.A. van der Beek, 2017)

We appreciate the time and energy that reviewer 1 put into the evaluation of our manuscript. The comments and questions were insightful and addressing them has improved the quality and the clarity of the presented science. We have arranged our response by 1) reiterating the comments of the reviewers (black text) 2) providing our response (dark red, indented text) and clarifying where the comment was addressed in the revised manuscript.

RC1

Gilmore et al. present a sensitivity analysis for a recently developed modelling approach in which structural restoration is combined with forward thermal-kinematic modelling to predict thermochronometer ages in fold-thrust belts, and subsequently use these ages to constrain the timing and rate of thrust-(sheet) motion in such settings. This is a promising approach, which h is being developed by several research groups separately (e.g., Almendral et al., 2015; Erdös et al., 2014; McQuarrie and Ehlers, 2015; 2017). However it still faces challenges, in particular how to take into account the topographic evolution through time and how to handle the large degree of freedom in the models. The present manuscript explores some of these challenges, in particular the effect of material properties (heat production rates), reconstructed geometry and kinematics, and the topographic history, which all influence the predicted thermal histories significantly but are very difficult to constrain. It is therefore a useful contribution to the still small but growing number of papers on this subject, and I would recommend publishing this in Solid Earth after moderate revisions.

I have two major comments and a number of smaller, more specific comments on this manuscript. The first major comment concerns the context of this study and what is exactly new in it. When I started reading this, this was not very clear for me. Long et al. (2012) presented the structural cross-section and thermochronology data used here, as well as similar data for the parallel more westerly Kuri Chu cross-section. McQuarrie and Ehlers (2015) modelled the data for the Kuri Chu cross-section in a similar manner to what is done here. What is new in this manuscript is the modelling of the (eastern) Trashigang cross-section. This is a valuable exercise in itself, and the comparison of the outcomes of the two modelling exercises in enlightening (see below), but I think it would be useful if the authors presented this context and the relationship of this study with previous work straight up in the introduction, so that readers are not left wondering what is new or different here with respect to previous work by the same group of authors.

> Introduction was revised to highlight new contributions and improve context with previous work. In particular we describe what is new in the introduction; p. 2 lines 13-21.

My second comment concerns the inferred history of shortening rates; in particular the strong variability in these rates that the analysis suggests. I have been intrigued by this outcome since the initial paper by Long et al. (2012). I reviewed that paper at the time and already queried the authors about the robustness and implications of that finding but am still struggling to understand it. Starting from what

we know (and progressing toward lesser constrained inferences): the modern convergence velocity between India and Tibet is ~20 mm/y; the total India-Asia convergence rate is about twice that. If we accept the results of Molnar & Stock (2009), India-Asia convergence rates have decreased since 20 Ma; from 54-83 mm/y before 11 Ma to 34-44 mm/y after that, for points in the NW and NE corner of the Indian subcontinent respectively. That total India- Asia convergence rate should be distributed between far-field deformation in the Tibetan plateau and its northern borders, shortening in the Himalaya, and underthrusting of the Indian plate beneath Tibet. It is interesting, and reassuring, to note that most of the tested models predict shortening rates in the order of 5-6 mm/y in the last ~10 Ma, which is consistent with estimated "overthrusting" rates in simpler thermokinematic models used to predict thermochronology ages (e.g. Brewer and Burbank, 2006; Whipp et al., 2009; Robert et al., 2009; 2011; Herman et al., 2010; Coutand et al., 2014, and others). Any increase in shortening rates up to the total India- Tibet convergence rate of ~20 mm/y could potentially be explained by temporally variable partitioning between "overthrusting" and "underthrusting"; since these concepts are really defined by a particular frame of reference only (which is in my view controlled by the erosional efficiency in the Himalaya), that could be plausible and possibly linked to temporal variations in erosional efficiency. If one wants to invoke further increases up to the India-Asia convergence rate, that would only be possible by temporally transferring far-field deformation to the Himalaya, but it remains in the realm of possibilities. The inferred rates of ~70 mm/y during building of the Upper Lesser Himalayan duplex are more problematic, because – if true – they would necessarily imply north-south extension in other parts of the Himalaya-Tibet system, for which there is very little evidence. The inferred reconstruction requires significant amounts of shortening to build this duplex (at least 150 km or ~1/3 of the total shortening since 20 Ma according to Fig. 3) and I wonder whether a more conservative structural solution would not be possible to fit the surface observations for this duplex. In any case, the preferred models with variable shortening velocities pose significant questions, which should be addressed more directly. The reader is really left wondering how well resolved these shortening histories are, given the significant number of unconstrained parameters in the models. Some of the specific comments below refer to these unknowns.

> We agree that the fast rate from ~ 13- 8 Ma are unexpected, and yet this is a robust part of the model and is a function of the suite of ZHe ages that are all 8.5-10 Ma in the Kuru Chu area and 9.5 to 11 Ma in the Trashigang area. These rocks cool through the ZHe closure temperature as the Baxa duplex forms, and accommodates 155-165 km of shortening. 160 km in 2 Myr is 80 mm/yr. That is essentially the problem. We appreciate the suggestion for a more conservative structural solution to reduce the shortening expressed by the Baxa Duplex. However, this is a region where the shortening amount is remarkably well constrained. Shortening magnitude in its simplest sense identifies an area (a box), and calculates the length of a unit with thickness X necessary to fill that box. The Kuru Chu and Trashigang sections from Long et al. (2011b) show how well-constrained this box is. Unlike sections in Nepal, the Baxa duplex in this area is almost entirely exposed and fault bedding plane relationships show the hanging wall cut-offs for the Baxa faults have (almost all) been eroded (implying more shortening possible). Yet, there are erosional remnants of the Paleoproterozoic Shumar/ Daling rocks carried by the Shumar Thrust exposed in fault klippe almost all of the way to the MBT (Long et al., 2011). These fault klippe

define the top of the box as being essentially immediately above the erosion surface. There is just enough space to erode the hanging-wall cut offs of the Baxa faults. The base of the box is defined by the décollement. The décollement depth for the Long et al., (2011) cross sections in the region of the Baxa duplex is directly between the 2 permissible depths estimated by Coutand et al., (2014) and matches geophysical constraints in the region (Mitra et al., 2005; Singer et al., 2017). If anything, estimates of the décollment depth are deeper (Coutand et al., 2014) which would just exacerbate the problem. The only variable left is the thickness of the Baxa, which can be observed in the field, as can the faults that repeat it. Field observations provide several thickness estimates that all fall between 2.1 and 2.5 km and well constrained shortening estimates of 150-165 km.

Thus in both the Ehlers and McQuarrie (2015) and in this manuscript, we have tried to figure out what is an acceptable age range that does not violate the data. Shortening rates can viably increase up to the India-Asia convergence rate of 40-45 mm/yr. The expectation is that during that window of time the Himalayas are taking up the entire magnitude of convergence. Shortening rates above that (45-70 mm/yr do require coeval extension to be viable.

We have conducted more simulations looking at the sensitivity of shortening rates, particularly using the new geometry. Due to limited measured cooling ages between 70 -100 km from the MFT there is more flexibility in the Trashigang section than the Kuru Chu and rates as low as 45 mm/yr (at plate tectonic rates) are permissible. Our new thoughts are that a revised geometry for the Kuru Chu section (two ramps) may facilitate more exhumation in this region and thus lessen the need for excessively fast rates (55-75 mm/yr) for that section.

Intriguing enough (because I (McQuarrie) have never been a huge fan of extrusion or channel flow) the age of this rapid shortening in the Baxa duplex overlaps with the age of the STD in this portion of the Himalaya --12.5 Ma (Th-Pb monazite age from Kula Kangri at the border of Bhutan and Tibet) and 7 Ma (ZHe ages)(Edwards and Harrison, 1997; Coutand et al., 2014). Although shortening rates should not be faster than plate convergence rates, it is permissible if it is accompanied by fault parallel normal faulting, such as is postulated by channel flow models. To me (McQuarrie), one of the strongest arguments for channel flow/ extrusion like behavior is thrust faulting rates above plate tectonic rates. The observed thrusting rate would be the shortening rate plus the extension (extrusion) rate.

Numerous changes were made to the manuscript in section 5.3 to address this comment and our new simulations.
(1) Without relating all of the justifications for the cross-section, we have included the statement that the cross section itself is a minimum shortening estimate and that any change to the cross-section will increase the shortening. We referred again to the Long et al. (2011b) paper where the details are laid out. (p. 22 ~l. 30)
(2) The manuscript includes revisions to discussion evaluating the permissible ranges of deformation ages and rates based on our simulations (~p. 23 l. 5-8)

(3) To present these new simulations in the paper, a new figure 11 has been created and introduced in this section, and table 3 updated.
(4) Comparison of the Trashigang section to the Kuru Chu section, thermochrometer data available along the sections, and reasons for differences between rates proposed by this study and by McQuarrie and Ehlers (2015) are included. (p. 23 ~l. 19-35)
(5) Because so many of the permissible shortening rates are above plate tectonic rates we have also expanded on our discussion of modeled rates to include their relationship to convergence rates in section 5.3. (p. 24 , l. 15-25)

Overall, the paper is fairly well written and illustrated. On a number of occasions, phrases don't run because a verb is missing or because of singular/plural confusions. A certain number of typos also remain. All of these can be weeded out by some careful editing. The use of some internal "modelling jargon" like "Python topography", "Split KT" etc. does not add to the general understanding of the manuscript – the authors might want to find some more eloquent terms to describe these modelling settings.

The manuscript has been edited to correct typos and clarify wording in areas that are currently mistyped or confusing. We changed the topographic estimations from Python Topography and Template Topography to Responsive Topography and Static Topography respectively. Since "Split KT" refers to Kakhtang Thrust motion at two different periods of time (versus all early or all late), we could not find a word that was more descriptive or more accurate and that would improve the readability of the paper. If you have a suggestion, we would be more than willing to incorporate it.

Specific comments, tied to page and line number:

p. 1 l. 7-10: the first two phrases of the abstract do not really set up the problem in a very clear manner or "draw" the reader into the problem – you may want to consider rewriting these into something more clear and specific.

Abstract was revised. Comments about the first two sentences of the abstract were raised by multiple referees.

p. 2 l. 13-20: this first paragraph of the "Geologic background" section looks a bit lost on its own; it is not very informative (why is the onset of motion on the MCT important here?) and could easily be combined with the following "Tectonostratigraphy" section. The Daniel et al. (2003) and Tobgay et al. (2012) references are missing in the reference list.

The geologic background was removed and the critical information was included in section 2.1 on tectonostratigraphy. Daniel et al. (2003) and Tobgay et al. (2012) was added to the reference list.

p. 3 l. 20-21: how were the data exactly projected into the cross-section? This is a critical step, as the ages (in particular for the low-temperature systems) will be influenced by the local topography. See further comments below.

> In order to maintain structural context along the cross section, all of the data (including data from Coutand et al., 2014) were projected onto the cross-section along-structure (i.e. in the direction of the trend of fault while maintaining distance from structures as possible). The exceptions to this in the original manuscript were minor and have been corrected. We have corrected all figures data projected along the section to be consistent with the along-structure projection method, and text in section 2 describes this projection.
>
> Since most samples were not taken exactly along the line of section, the elevations of most samples vary from the elevations at these projected location. However, our models do not use present-day elevation in the models either (discussed below in response to RC1 comment on p. 8 l. 15-19). We have plotted all of the data with respect to elevation and limited age elevation trends emerge strongly suggesting the ages are controlled by structural uplift and minimally modified by topography – this is clarified in the manuscript in section 2.2.

p. 3 l. 30: why do the ZHe ages require "rapid" cooling? This inference can only be drawn by comparing them to other thermochronometer data, or by assessing age-elevation profiles for instance.

> There is no a priori reason to indicate rapid due to the age and the adjective has been removed.

p. 3 l. 32: three ZHe cooling ages north of the MCT are shown on the cross-section (but only two on the map?). Also, the cross-section of Fig. 2 gives the impression that the samples between ~57-65 km are from the lower Greater Himalayan sequence, while the map shows they are from the upper. Maybe you should sketch in some of the geology above the topography to make this clearer. This also brings us back to the question above of how these data were projected into the cross section. What was their imposed elevation? Simply plotting them on the topography in the cross-section puts them on a much lower structural level than where they actually are!

> As explained above, this was a plotting error and has been corrected in figures and in text where fit has changed because of the re-projection. Overall results are not impacted by this revision. Since the modeled ages are all predicted at the surface, projecting the samples in the air would have limited applicability to match modeled results. Where discrepancies between modeled and measured ages exist we do examine both the structural and topographic elevations that the samples are from. For example see new text in section 5.3, p. 23 l. 5-10

p. 4 l. 5-9: why do you take this approach? It is easy enough to model the individual data using the combined Move/Pecube approach . . .

> At the scale we are evaluating predicted versus measured ages and what is controlling the change in ages, these samples plot basically on top of each other, particularly when projected

into the cross section. In the version of Pecube we use, the ages plot as the age trend shown on figures 5, 6, etc. In our view, they represent a true variability in sample age and can be considered a clustered datum rather than several data for our purposes. The one minor caveat to this is the cluster of AFT age in structurally higher Greater Himalayan rocks. As expanded on in our responses to reviewer 3, (and included in the text at the end of section 2) there is a modest age elevation trend here. However the exhumation rate given by the age-elevation differences is 0.4 mm/yr while an average 3.5 Ma AFT age suggests more of a 1-1.7 mm/yr exhumation rate. Additional details of possible age elevation relationship are mentioned at the end of section 2.2.

p. 4 l. 17-18: the question here is obviously: "how was the new topography obtained?" this is discussed further on – you may want to refer the readers to this later discussion here.

We mention where the approach is discussed further in the first paragraph of section 3.1.

p. 4 l. 26-27: Note that a subsequent similar model by the same authors (Hammer et al., GRL 2013) comes up with much lower estimates for the elastic thickness in Bhutan (< 25 km) than in Nepal.

Yes, the very low values (in Hammer et al., and in Berthet et al., 2013) are in part a function of their approach for estimating EET that varies spatially (something that we are unable to mimic using the flexural algorithms in Move). In addition, the solution is for modern EET, which for Bhutan is strongly depending on the narrow width between the MFT and Shillong Plateau. 1) Our EET is a much longer-term average and, 2) is not meant to be viewed as a calculation of the EET in the area. However we can state with confidence that using low (25-40 km) EET values in the flexural-kinematic model will not reproduce the foreland basin thickness, the modern dip of the décollement or the geology exposed at the surface today. This section has been modified appropriately.

p. 5 l. 2: here you could reference some of the previous studies using the same approach.

Although there has been a suite of groups moving forward with linking cross-sections to advection diffusion models the details of the kinematic model are not always clear particularly if or how flexural loading and erosional unloading were accounted for. A good example of the potential influence is Erdős et al. (2014). They noted that a cooler crustal thermal structure was needed to match the measured high-temperature cooling data (than the lower temperature data) in the Pyrenees. Alternatively, their model could be restoring the rocks to a position that is too deep (thus becoming too warm) because thrust-related isostasy was not taken into account, or perhaps accurately accounted for as the section was retro-deformed backwards in time. What this paper highlights is that accounting for flexure (and erosion) in the kinematic model is a critical and necessary component.

We added text addressing this in section 3.1 as well as 3.1.1. In both sections, references to work using this approach were added. We added more detail in section 3.1 to discuss the

kinematic modeling process, in particular how different groups account for flexure, erosion, and thus paleodepths, because these decisions are going to control the estimated temperature histories and ages

p. 5 L. 27-31: a self-consistent approach would be to use a critical-taper topography in the models – it is not clear if the "Python topography" is based on such an approach, but the link between the imposed topography and a critically tapered wedge model could be outlined here.

The "Python Topography" (now Responsive Topography) may be viewed as a simplified critical taper approach, with the first order angle of topography estimated from modern topographic angles in the Himalayas. A key difference is that we do not systematically vary the topography angle based on the décollement angle. Please see further discussion response to p. 8 l. 22-24 comment below.

p. 7 l. 6-17: see general comment on variable shortening rates above. More justification and discussion of these rates is needed.

As mentioned in the general comment above, a whole range of velocities were tested and we acknowledge that a full suite of parameters tested (including velocities) was not reflected in the previous version of this manuscript. We have addressed this in section 3.2.2 and Table 3. This is also more fully addressed in the discussion section 5.3.

p. 7 l. 16: it seems that this is the first time the Kuru Chu section is mentioned; it hasn't been introduced previously (but should be).

The Kuru Chu section and corresponding studies are now mentioned in section 2.2, (multiple locations), and earlier in section 3.2.2, and quite a lot in section 5.3.

p. 7 l. 19 (and numerous other occurrences): why do you call the reconstructions "flexural models"? This is surprising and confusing, as flexure is only one component of these models; the structural reconstruction is at the heart of them. You could call them "kinematic models" or something like that.

The decision to call the models "flexural models" stems from the multi-step process of achieving a viable "kinematic model" in Move – and from our suspicion that the flexural component is missing from most thermo-kinematic modeling approaches that use cross section kinematics (clarified in the revised end of section 3.1). Without accounting for flexure in the kinematic solution, the evolution of the décollement cannot be determined and thus the estimated depth history (and resulting thermal history) of a given rock becomes a complete guess. Thus a forward model taking into account flexure is critical. We are weighting the flexural component with the term 'flexural'. The work flow for any given kinematic model is to first find a pure kinematic solution (the "kinematic model") with only fault motion accounted for, the second round of iterations is the flexural component that requires an evolution of topography, erosion, foreland basin development, and décollement flexure.

We have revised the name to include both adjectives, Flexural-kinematic model, to make the model name more intuitively descriptive. We also clarify the reasoning for this in the revised end of section 3.1.

p. 7 l. 30: the INDEPTH lines were shot in the Yadong rift, which overlies the Yadong cross-structure – a probably important lateral ramp in the Main Himalayan décollement. Is the 4° dip you cite here relevant for the décollement west or east of the Yadong structure? In any case, this would be valid for western Bhutan and not necessarily for eastern Bhutan. It is not obvious that comparing the décollement dips with data that are not from the same region is very informative, given the probable lateral segmentation of the MHT.

We have removed this reference and added Singer et al. (2017), which has estimates for both the décollement and Moho for this region of eastern Bhutan.

p. 8 l. 1-4: this is counter-intuitive. The flexural response should be driven by the topographic loading, not by the kinematic scenario. Therefore, if the different kinematic models lead to differences in flexural loading profiles, it must be because the (imposed) topographic response to the kinematics is different between these models.

Yes, this is correct. We have rephrased this to make it much more clear and more accurate. See revised section 4.1

p. 8 l. 15-19: why do you not simply use the present-day topography as the final topography in the model? This is a known entity, and at least that would help in comparing kinematic and thermal histories at the right structural and topographic levels for the data points.

While at first impression it seems that using present-day topography as the final topography would improve the integrity of the models, that is only true of a model that can 'predict' a topographic evolution where the next to final topography is almost identical to the modern topography. If there is significant discrepancy between the penultimate predicted topography and the present-day topography (if inputted as the final step) the result would be unrealistic "deposition" of material in areas that are modeled in the prior step with a lower topographic elevations than actual topography. Simultaneously, in areas that have lower actual topography than modeled, using present-day topography could simulate several km of unexplainable erosion.

We recognize that topography of the Earth's surface is altered by more processes than are accounted for in our simplified, first-order estimation of topography such as river incision, the geometry of interfluves, and the effect of axial or transverse drainages. Our approach to modeling topography is outlined in McQuarrie and Ehlers (2017): "the more simplified critical taper model that responds to regions of uplift or subsidence will account for the longest-wavelength, and most significant, topographic effect (i.e., valley and ridge topography) in the thermal calculation."

Each kinematic scenario prescribes a different evolution of topography because as Reviewer 1 stated in the p.8, l. 1-4 comment, "topographic response to the kinematics is different between these models." Our goal is to determine if the estimation of modern topography using the python script can successfully replicate the first-order patterns of present-day topography. This is why we compare where and how the modeled topography deviates from the actual topography.

p. 8 l. 22-24: this phrase is hard to read and also appears counter-intuitive. In the critical-wedge model, the surface topography (α) and décollement dip (β) are linked through the critical taper angle (which itself depends, among other things, on β). Therefore, it might be more self-consistent to try to find a surface topography angle that corresponds to the critical taper for each time step (and degree of topographic loading). This would be an iterative approach, but I'm sure it can be done. See comment on p. 5 l. 27-31 above.

This is an intriguing point and one that we have thought about. As elaborated on in our reply to Reviewer 3, *Move* is a purely kinematic model and thus not governed by mechanical responses. Critical taper is a mechanical response that is dependent on a ratio of internal rock strength to décollement strength (i.e. resistance to sliding) (Dahlen 1990; Suppe, 2007). Thus assuming constant critical angle (one in which the topography angle becomes smaller over time as the décollement angle becomes steeper) would most likely misrepresent the topography evolution of the fold-thrust belt because décollement strength changes as lithologies change. As pointed out by Stockmal et al. (2007), pure critical wedge solutions become more limited when evaluating the effect of material differences, particularly ones with original horizontal geometries, and the ways in which those initial planes of weakness impact the internal structural geometry, strain history patterns, etc. This non-uniform behavior alters the predicted erosional response. An example may be the front of the fold-thrust belt dramatically propagating forward (on a weak décollement) before the development of a duplex system. The jumping forward would dramatically reduce the taper angle and the duplex response would be to increase structural and topographic elevation to regain "critical" taper (so the system can move forward). Using a constant (say 2° topography angle) in a model suggests that the taper angle is increasing through time. A true self-similar response would argue that the initial topography angle of the cross sections presented here would be 2.5°- 3.5° with an initial décollement angle of 1.5° to produce a final critical taper of 6°- 7° (broadly similar to the modern 4-5° décollement and a 2° topographic slope).

What we do know is the geology that is at the surface today, the modern dip of the décollement, and the cooling ages of a suite of minerals. What we can test is a topographic evolution that best matches all of those constraints because the ability of the model to predict older and deeper thermochronometer ages reflects its ability to accurately estimate the relationship of those rocks to the evolving surface of the earth. Critical taper theory gives us broad bounds for what may be a realistic topographic evolution though time. And that is an evolution that can get tested (using a range of permissible topographic angles) to see how

accurately it reproduces the first order features in the modern topography.

Regardless, a taper angle is topography plus décollement, and defines an area that is filled with folded and faulted rocks. If the area does not change, (because the taper angle des not change, then a lower topographic angle would require a steeper décollement. We have rephrased this in section 4.1 to make this clearer.

p. 9 l. 11: you may have modified your version of Pecube, but in the "standard" model, heat production is constant with depth, so that "surface heat production" is a bit of a confusing term in this context.

Following the approach and rationale summarized in McQuarrie and Ehlers (2017), we prescribe an exponential decrease in heat production with depth, as opposed to assuming a constant crustal heat production. An exponential decrease in heat production with depth requires definition of a surface heat production (Ao) and an e-folding depth. One caveat of this approach is that material properties are not exhumed during the simulations to modify the surface heat production value. However, an exponential decrease in heat production with depth has the advantage of honoring observations that heat production diminishes with depth through the crust and that this decline is not monotonic (Chapman, 1986; Ketcham, 1996; Brady et al. 2006). This approach not only allows honoring measured surface values of heat production in the Himalaya (e.g. see Whipp et al. 2007), but also produces reasonable mid and lower crustal temperatures that would not produce partial melts. This text has been added to section 3.2.1.

p. 9 l. 15: this seems a fairly obvious result, since the kinematics of the models do not change, only the thermal field. The samples have the same "normalized" thermal histories; the temperatures are simply somewhat higher throughout for the models with higher heat production.

Yes, we agree. We have added this phrase when we first talk about the differences in the predicted ages. i.e. "The most apparent trend among all three thermochronometer systems is that predicted cooling ages become younger as the radiogenic heat production increases from 1.0 to 3.0 $\mu W/m^3$ due to the higher temperatures throughout the model." In addition we now talk about how changing values of heat production effects the three thermochronometer systems differently.

Specifying the changes in predicted cooling ages as Ao values change is necessary to fully address the concern raised in p. 13 l. 31-32, when we altered both heat production AND geometry , Reviewer 1 was left wondering "OK, but how much of this improved fit can be ascribed to the new structure and how much to the increased heat production." The background we have expanded upon here is needed to emphasize what signals are a function of changing geometry and what signals are a function of changing heat production when both change later in the manuscript (sections 4.3.1, 5.1.2 and 5.2).

p. 9 l. 19: "ages" not "rocks", I think.

Corrected

p. 10 l. 3-5: a bit of a rambling phrase that is difficult to read/understand.

We revised this paragraph to make it easier to read.

p. 10 l. 24: “later” not “earlier” I think?

Revised to “more recent”

p. 10 l. 32-33: there are many free parameters in these models: not only an infinite number of shortening-rate histories, but also significant degrees of freedom in the imposed structure and the topographic evolution. I fully understand and appreciate the difficulties in exploring this complex parameter space, but how robust are the inferred rates really? This is not obvious, and given the important implications of the shortening-rate history, this should be discussed. An alternative approach would be to not allow shortening rates that are greater than the plate-scale convergence rates at any time (i.e. use the plate-convergence rates as a constraint) and try to find models that can explain the data using this constraint.

We agree that the sensitivity of the model to the prescribed rates needs to be more fully discussed. The questions that Reviewer 1 raises on how-well constrained shortening magnitudes are, helps to elucidate what additional information is needed.

To address this comment, we removed much of the last paragraph in section 4.2.2 that emphasized the variations in shortening rates. Instead we ended with the very important observation that even with dramatic changes in shortening rate, the model still can not accurately predict cooling ages through the greater Himalayan section. We return to the discussion of shortening rates and the sensitivity of the predicted ages to these rates in section 5.3. We discuss the sensitivity of the model to rates that are at plate convergence rates (~45 mm/yr) versus faster than plate convergence rates when we present the revised geometry. In the end, there is limited usefulness in evaluating rates with a geometry that will never reproduce the measured ages.

p. 11 l. 10-11: why is this your expectation? The erosional history would depend on the topographic history through time, rather than the final topography. In the no-topography scenario, if I understand well, there is no topographic change through time. If in the other topographic scenarios topography diminishes locally in the final timesteps, this will predict younger ages.

Yes, a topographic scenario where topography diminished with time would produce younger ages, and the expected exhumation difference would be approximately the change in topographic elevation (maximum 2-3 km). Our expectation that the No Topography scenario would produce younger ages is because these models always produced higher total exhumation where the final cross section was over eroded by 1-2.3 km. The age in which this exhumation happens is a function of the age that a given structural relief was being generated. As an

example, some component of over-erosion happened as the upper Lesser Himalayan duplex moved up and over the pronounced ramp at 65 km. Thus our expectation is that predicted AFT ages that show this exhumation would be younger. The conclusion is that since the magnitude of erosion that happens during this displacement in each topographic scenario is significant, the additional 1-1.5 km of extra erosion in the No Topography scenario is not significant – particularly when viewed incrementally (e.g. Valla et al., 2010).

p. 11 l. 14-15: a list of 6 adjectives ("Python topography model fully reset Mar ages") followed by another of 4 . . . Maybe rewrite?

This was revised.

p. 11 l. 20-23: this is an important point but it also seems fairly obvious. It clearly points to the need of a self-consistent treatment of topographic evolution. The best way forward may be to combine these models with simple surface-process models to erode the topography through time.

We agree, a self-consistent treatment of topographic evolution where the modeled topography is a function of the deformation is a key result from this work. Although this seems like an obvious result, it is also a common approach to use a DEM of modern topography in models and assume that topography is in steady state and not changing – this result highlights that assumption is not valid either (and may also cause burial of material where particle points are subsiding and topography is not, and produces over-erosion of material where rock uplift occurs but topography remains static.

Also, while it is obvious to Reviewer 1, how topography is estimated particularly over long time windows is still a rather new item of discussion and application for thermokinematic modeling in compressional orogens. As outlined in the introduction, several other studies that have used Pecube have not used a method of applying topographic evolution that account for localized structural uplift and isostatic subsidence. Rather, they apply a muted topography similar to present-day elevations, infer topographic changes that seem appropriate or increase/decrease topographic slope over time. Yes, a self-consistent way to estimate topography is critical.

While we see the value of using other surface-process models such as Cascade to erode topography over time, the Python code (or an equivalent Matlab code) we use in this study, which approximates the first order topographic slope and specifically accounts for increasing topography in regions of active uplift and subsiding topography, provides a critical first step for estimating topographic change particularly in the isostasy calculations in *Move*. What may not have come through in the paper was the iterative process of finding a flexural solution (which is why we referred to it as flexural modeling). The kinematic displacements are known, and we are searching for a solution where the sequential kinematic restoration in *Move* (using flexure) can reproduce the depth of the foreland basin, geology at the surface and the dip of the décollement. This may take 20+ iterations to achieve using 20 km shortening increments. Thus whatever mechanism is being used to generate an initial topographic estimate needs to

evaluate the magnitude of topography change and predict a new topography in <1 minute to be viable in the iterative process. In addition, 1D erosion models require an estimate of time (which would have to be approximated for the initial reconstructions). Of course if the velocity were to change then the flexural-kinematic reconstruction in move would need to be redone. 1D erosion models also do not account for sedimentation (in a growing foreland basin). Our thought process is that the thrust loading and erosional unloading are much more sensitive to the first-order component of topography and thus using the responsive topographic taper approach is the best approach for Move. Once the displacement field has been determined (and then the resulting velocity fields), Pecube can run in conjunction with Cascade, to predict a more realistic and variable topography. As a double check --this Cascade Topography can be imported again in Move – just to make sure the resulting isostatic load is the same. We are currently working on fully integrating our modified version of Pecube and Cascade.

p. 12 l. 2: I think you are discussing MAr ages specifically here? May be useful to state this.

Revised

p. 12 l. 30: “older ages” seems more correct than “earlier ages” in this context.

Corrected

p. 12 l. 31: you have been calling this the MHT throughout the manuscript. Better stick to this acronym so as not to confuse the readers.

Changed

p. 13 l. 1-2: another somewhat rambling phrase . . .

This has been revised

p. 13 l. 5: this ramp is rather located at _90 km in the present-day geometry (Fig. 2)?

No, this early ramp is no longer visible in the cross section. See figure 3 C.2a for ramp location. We have clarified this in the Manuscript by revising the first 2 paragraphs of section 4.3 and referring to the appropriate figure location and ramp locations in the text.

p. 13 l. 12-15: is the cross-section of Fig. 9 still balanced? There is all of a sudden 35 km more Baxa group in this cross-section, while the rest of it has not been modified. Could these additional 35 km be found by reducing shortening in the upper LHS duplex? In that manner you might also be able to reduce the problematic shortening rates necessary to produce this (and the associated ZHe ages).

The new cross-section in figure 9 is balanced. Forward modeling the kinematics of a cross section ensures that it is balanced. But Reviewer 1 is correct in that the distribution of shortening has changed. All ramps north of the new Baxa footwall cutoff were shifted 35 km

north, and thus 35 km of shortening was added. Yes, we agree this does not reduce the problem of the fast rates (it can make the rates higher).

Our modeling (and others) have highlighted the strong relationship between ramps and young cooling ages. We can use this relationship and what is required by the geology to figure out how far south we can place the southern ramp (through the Diuri) initially this was placed at its location because the pervasive northward dips in structurally higher units (the northward dipping boundary of the Shumar-Daling on Baxa, GH on Shumar-Daling and the northward limb of the STD all suggest a northward dipping ramp ~ in the location shown on both cross section). What we did was turn this large ramp in Long et al.'s (2011b) original section into two ramps to better match the cooling signal. We know that the Baxa formation *has* to be under the anticline of Shumar-Daling because of the along strike relationship shown in the Kuru Chu section of the map (figure 1 – the anticline shown in the cross section is underlain by the Baxa Group rocks repeated by faults). So, even though we could move this ramp farther south to 50 km (location of the youngest AFT age in this region -- figure 9), we can't remove either of the Baxa horses. In addition, moving the ramp farther south would make each of these horses longer, adding more shortening back into the geometry. The cross sections were constructed to minimize shortening while matching surface constraints – thus any modification to the cross section that also matches surface constraints will tend to increase shortening estimates.

This last point was added in the discussion section on rates, and we have included the restored modified cross-section below the deformed section in figure 9b.

p. 13 l. 25-29: this is problematic. First of all, you change two major inputs to the model (structural geometry and heat production) at the same time here, while previously you have carefully only changed one parameter at a time. Second, you introduce spatially variable heat production here, which you did not do previously and which could have led to better fits in the previous models. This is a large change in the thermal structure and it should be justified. Although I am sympathetic to the fact that heat production could be significantly higher in the GHS than in the LHS, to really model this properly you should ascribe heat-production values to the different units, and advect these with the units.

Yes, we agree that the jump was too large to independently see the effect of both, but our goal was to show the best fit and a reasonable number of models and iterations. Numerous kinematic and thermal model iterations were performed in addition to the specific model results presented in this paper. Most of these iterations were performed with the goal of obtaining an improved fit using the cross-section geometry published by Long et al. (2011b). Several new models (changing flexural and topographic parameters were run using the new geometry to produce several models with slightly different exhumational histories to test the sensitivity of the model results to changing these different parameters. All models run in Pecube were evaluated using heat production values ranging from 2.0 to 5.0. (in steps of 0.5) and a range of different velocity combinations. In all, nearly 100 forward modeling combinations of the Long et al. (2011b) geometry were run for this study, and over 100 for the new geometry. None of the

models from the original Long et al., (2011) cross section could reproduce the AFT age trend seen across the GH (younger ages farther north), even with significantly higher heat production values in Pecube. This unsuccessful result of not being able to match the cooling ages with the original section led to the decision to strategically explore new geometry options, beginning with the replacement of the Baxa footwall cutoff. After evaluating a range of velocities and heat production values, we concluded that it would be best to ascribe different heat production values for different units in the model – even though we agree the most accurate approach would be to characterize each unit with distinct heat-production values in a single model. However we were limited by the current capabilities of our model. Thus the simplest way forward was to combine the results of the two models at the surface location of the MCT. Using Supplementary Figure 2 and 3, one can infer the range of potential cooling ages that would be predicted if it were possible to implement unit-prescribed heat production in Pecube. This seems most important for units in the immediate hanging wall and footwall of the MCT (~52 km north of MFT) where GH rocks that are known to be hotter with higher radiogenic heat production are spatially juxtaposed with the cooler Daling-Shumar units. This area is also where the greatest amount of cooling data are available.

We have worked to make this as transparent as possible in the revised manuscript. These include figure revisions to figure 9 and supplementary figure 2 and 3, and clarifications throughout sections 4 and 5. Examples include (1) the final paragraph of section 4.2.4 which highlights what is controlling predicted AFT ages in the immediate footwall of the KT, (2) the fourth paragraph of 4.3 detailing the rationale and method for combining models with different Ao values, and (3) section 5.2 which re-emphasizes the point that, even with higher heat production, fit of AFT ages remain poor in the immediate footwall of the KT using the original geometry proposed by Long et al. (2011b).

p. 13 l. 31-32: OK, but we are left wondering how much of this improved fit can be ascribed to the new structure and how much to the increased heat production.

Supplementary figure 2 graphically presents the best results from the using the updated cross-section geometry with 2.0 and 4.0 $\mu W/m^3$ heat production values applied along the entire line of section. New supplementary figure 3 does the same with the original geometry. As mentioned in the response to the comment directly before this one, we have clarified this in the text.

p. 14 l. 5-6: following up on the previous comment; can the data really tell the difference between the improved structural geometry and the increased heat production? There is very little data in the "bump" region. You use a simple visual comparison of predicted and observed ages; it would be useful to provide a more objective and quantitative comparison to back up inferences such as this.

We have added discussion in the manuscript that quantitatively compares the predicted AFT ages from the Long et al. (2011b) geometry to the predicted AFT ages from new geometry presented in this paper in order to support our conclusions of improved fit.

p. 16 l. 10-12: this is introducing yet another unconstrained parameter. I am not sure it is the best strategy to further complexify the models to improve the fit; this seems like a bit of a "flight forward". A more complete sensitivity and resolution analysis might be a more productive way forward.

> There are no new parameters. The parameters being discussed are EET and topography (section 3.1.1 and 3.1.3), and any given solution presented in this manuscript is a function of both parameters that combine to affect the exhumation of rocks. Is the added complexity you mention changing the value of EET or topography with time? There are strong arguments that can be made that both may have changed with time- and reflects your point made previously (p. 11, l 10-11). A forward model where multiple parameters have to be evaluated, and it is impossible to see if the model is a match to present day conditions until the last step, will always be a "flight (fight) forward". Not all questions or problems can be addressed through inverse solutions.
>
> The reality (which is why this section is important) is that subtle changes in EET have a larger effect on the modeled cooling ages than subtle changes in topography (such as using a process based estimation of topography or a simplified critical taper relationship). The reason why, is that a 5 to 10 km change in EET can impart a 1-3 km difference in magnitude of exhumation. Unfortunately, the flexural response to fault motion and associated topographic displacement (solved in the kinematic model) is something that is not included in many models attempting to link cross section to thermokinematic models, yet it has a significantly larger control on the predicted cooling ages than topographic estimations. We have clarified this section to emphasize this point.

p. 17 l. 9-10: "the amount of exhumation in this model is just at the amount necessary to reset AFT ages" is strange and apparently incorrect. The ages record cooling through the closure temperature at a certain time in the past. The thermal structure is going to affect that time, but the total amount of exhumation is much larger than the AFT closure depth it would seem.

> We have rewritten and clarified this point in section 5.1.2

p. 18 l. 10-15: A bunch of hard-to-read phrases that are in need of a few commas. Also, "after 13 Ma" would be better than "longer than" and replace the colloquial "till" by "until".

> We have edited this text for clarity and grammar.

p. 18 l. 15-20: another potential issue that is not discussed concerns the diffusion kinetics of He in zircon. Recent work has shown that the effective closure temperature of the ZHe system can vary from as low as _120 _C to as high as _240 _C as a complex function of the degree of _-damage (e.g. Guenthner et al., 2013). If you have underestimated the ZHe closure temperature (I suppose you are using the "standard" ZHe diffusion parameters built into Pecube) you could significantly underestimate the duration of shortening on the upper LH duplex, and thereby overestimate the shortening rates.

The reviewer raises a very good point, and we have modified the manuscript to state this as a potential caveat, although we do not think this is important for our samples because of the high cooling rate. The text now added in Section 2.2 is as follows:

The predicted ZHe ages in this study do not account for the effects of radiation damage on the closure temperature (e.g. Guenthner et al., 2013). The potential effect of this could be to underestimate the ZHe closure temperature. However, the effects of radiation damage on ZHe (or AHe) closure temperatures are most pronounced for long durations at relatively low (~220°C) temperatures (Guenthner et al., 2013). The Lesser Himalayan samples evaluated here experienced temperatures greater than 350° (Long et al., 2011c, Long et al., 2012), have young ages (typically ~7-11 Ma), highly reproducible ages (for individual samples) and underwent extremely rapid cooling (e.g., or around 16.3-22.5 C /Myr cooling rate since closure at ~180 C), thereby leading us to infer that radiation damage effects are minimal.

p. 18 l. 25-28: the first part of this argument is somewhat circular, since the McQuarrie and Ehlers (2015) scenario was input in the models here, without extensively testing all other potential scenarios. So the fact that the model predicts these variations in rates should not come as a surprise. In contrast, the dissimilar timing between the two sections that are only _25 km apart should be worrying. How can the same structure be active at time intervals that are several million years different between two adjacent locations? Again, the reader is left wondering how much of this difference could be due to variable diffusion kinetics?

We agree that many more rates need to be evaluated and presented, and we have clarified that in the updated version of the manuscript (see section 5.3, figure 11 and Table 3). We do not think that variable diffusion kinetics play a significant role (see response to previous comment) but elevation differences might. In addition, a revised geometry for the Kuru Chu section (two ramp scenario) may allow for an older age of transition from lower to upper LH duplexing which would decrease the fast rates.

p. 19 l. 2: given the numerous unexplored degrees of freedom in the models, it appears risky to assess the validity of the data based on the modelling outcomes.

That was not quite our point—thus we have revised and removed this sentence.

p. 20 l. 1: not sure what is meant with this phrase; what is “the spatial nature of thermochronometry”?

Wording was edited to clarify this point. The second part of the sentence is the important part: “the importance of considering the aerial distribution of cooling ages in the direction of transport and their relationship to the structural evolution of a landscape.”

Figures

Fig. 1: the inset geological map of Bhutan (panel B) is very small and not very readable. You should either increase its size or decrease the amount of detail on it. Also, in the legend of the main panel (C),

the Chekha Formation should be above the Greater Himalaya to keep all units in their structural order. Finally, it would help the reader if the colours used for the different thermochronometers were consistent between this figure and the following.

> Figure 1 has been revised. The colors of data points on the map are assigned based on the original studies due to overlap in sampling (e.g. ZHe and AFT data collected at same location). Colors used to label ages from thermochronometers at each sampling location do match colors used in subsequent figures.

Figs. 5-10: much more data appears to be plotted in these figures than in Figs. 1 and 2. What do the lighter-coloured data points refer to? For clarity it would be better to take them out. In Fig. 7, why does the "template topography" model not predict AFT ages everywhere?

> Figures 9, 10, and 11 include data from the Kuru Chu region (50% transparent) as well to help evaluate similarities and differences between the two sections. This has been clarified on the figure captions and expanded on in the text. Published data are presented in Figures 1 and 2. Are plotted on figure 5-8.
>
> Template Topography in Figure 7 does predict ages along the cross-section as completely as the other two models' output shown. In some areas, there is significant overlap with the other modeling results. In the AFT output plot, the Template Topography output lines are discontinuous because predicted ages were more scattered.