

We appreciate the insightful comments and questions posed by the reviewers and addressing them has improved the quality and clarity of the manuscript. We reiterate the reviewer's comment, followed by our response and any references to relevant changes in the manuscript in the green text.

RC1: 'Comment on egusphere-2026-1560', Anonymous Referee #1

This is interesting contribution - where the authors reconstruct the deformation in space and time in an improved way. Also they illustrate by using the extended thermochronologic dataset - advanced thermal modelling across the range of Himalayan transects in western Nepal to help to prove or disprove two previous published cross sectional models - favor a reconstructed new cross sectional model..

Compared to other studies - they provide detailed solution, how the MHT looks like and have geologic/structural solution, what kind of processes form ramps in the MHT and how they look like (a much debated topic and first order structural features of the Himalaya - still not well constrained).

Also these models have very interesting implication - showing dramatic changes within deformation of the orogenic wedge despite a plate tectonic mostly constant underthrusting rate of the India versus Eurasia.

Fig. 6 is one of the key figures of the study - where the authors show reconstructed and favoured balanced cross section at various time steps and the proposed tectonic evolution and change of the crustal thermal field related to the deformation history.

However, it's a bit unfortunate that the authors miss to define all parameters presented in figure capture of Fig. 6 - used for obtaining the modelling results. Crucial parameters are velocity and displacement - what do they exactly mean. This needs to be introduced better and referred to the other studies of the authors - where these parameters have been studied in more detail.

As the reviewer most likely understand - displacement is total fault offset of the active fault. Velocity is more challenging to understand - but is likely the shortening rate of section studied of the orogen.

The legend within each subplot has been updated to "Shortening rate" to better clarify the meaning of the previous text of "Velocity". The caption has been updated on lines 380-382 to expand on the meaning of "Displacement" in each subplot legend.

If the reviewer understand to results correctly - the authors suggest that this compartment of the orogen has accomodated changing ammount of the total shortening of the Himalayan over time since the Miocene until today. This is fundamental constraine to obtain good agreement between model predicted data and measured thermochron results. This imply that at middle Miocene time-window - the Himalaya accomodates nearly the entire plate-tectonic shorting between India and Eurasia. The authors have studied several studies of other segments of the orogen - which come to similar results - however - there assumptions/result are so fundamental - that they need to be introduced properly and make the reader aware of these temporal change in shortening. Even if they refer to two independent parameters - flexural banding and sediment filling of the foreland basin, as well as cooling pattern of the orogen.

Changes in shortening rate along the section can be a response to slowing convergence rates between India and Asia (e.g. Molnar & Stock, 2009) and/or strain partitioning, with structures in e.g. Tibet accommodating a portion of the convergence. In Model C, shortening rates reach ~45-48 mm/yr during RMT motion at ~18.1-13.6 Ma to be able to accommodate the amount of shortening required to reproduce the surface geology over the window of time required by the MAr cooling ages in the Dadeldhura klippe. During this time the Himalayas would be accommodating approximately the entire magnitude of convergence.

We have incorporated the results for the constant velocity model with shortening rates of 24 mm/yr in the Supplementary Materials, Fig. S4. We also address the requirements for fast shortening rates during RMT motion and slower motion during the LH duplex, as well as the tectonic implications of both the initial fast rates and the subsequent slowdown in the revamped Section 5.3 'Exhumation drivers in the western Nepal high Himalaya'.

With this revised version of the manuscript they nicely show - that is potential agreement with other recent studies of the region - and they do not nicely explain - why they agree. Fig. 8 panel and based on their results, they suggest significant reorganisation of deformation across the range since the middle Miocene. Most prominent over the

While Fig. 8 (now Fig. 9 in the revised manuscript) illustrating how exhumation magnitude changes in space and time in our sequential model can be interpreted as a significant reorganization of deformation across the range, the thermokinematic models are really characterized by a foreland-ward progression of faulting that gradually affects deeper to shallower crustal levels, punctuated by periods of out-of-sequence motion (e.g. Fig. 6). The features and processes we model are fundamental to fold-thrust belts. The location of the active ramp gradually shifts southward/foreland-ward as a new fault becomes active and rocks are accreted to the overriding plate. What stands out in Figs. 6 and 8 is the out-of-sequence faults and their resulting effect on exhumation amounts. These are either incorporated to reproduce a specific exhumation signal, such as that required by the MAr cooling ages in the north, or specifically to reproduce the surface geology (e.g. faults at the MBT and at ~39 km south of the MCT in Fig. 5).

Sections 5.2 'MHT geometry in western Nepal' and 5.3 'Exhumation drivers in the western Nepal high Himalaya' have been edited to better explain Fig. 8 and to relate the model results and implications to other studies in the region.

Fig. 8 is very important - but the authors lack to discuss the implications - but rather leave the interested author making his own view - therefore the manuscript will only be appreciated by small community - that focus on this kind of research.

We have modified Section 5.3 'Exhumation drivers in the western Nepal high Himalaya' to expand the discussion of Fig. 8.

Last 5 Ma of cross section history forms super well constrained and important time window because it provides the best resolution of both geomorphic and thermal constraints. What do we learn from this - with respect to the earlier evolution? For instance the authors propose that out-of-sequence deformation from there studied section is only recognized over the last ~five million years of Himalayan evolution in western Himalaya. Also major tectonic reorganization are proposed within this time frame - due to availability of the geomorphic parameters - which are lacking for earlier time windows. In figure 6 they have to add a new ramp at ~70 km south of the MCT into the lower thrust nappe - with out any geologic evidence - just to match topographic parameters from the surface. This illustrates how much interpretation the presented section are.

While we agree that it is possible for there to be additional in- or out-of-sequence faults in the earlier evolution of the orogen that we do not have evidence to constrain, we do argue that the geometry, kinematics, and shortening rates for ~5-18 Ma are well constrained by integrating multiple thermochronometer systems that cover a range of closure temperatures (~350°C to ~70°C) and the basin accumulation record with the thermokinematic models. The complexities in the geometry, kinematic sequence, and shortening rates we propose for Model C are required components of the burial and exhumation history to be able to reproduce the surface geology, measured cooling ages, basin thickness and accumulation data, and modern topographic relief, slope, and elevation.

The ramp at ~70 km south of the MCT is new and is required to reproduce various geomorphic indices. In addition to the young out-of-sequence motion needed to reproduce a specific exhumation signal or geomorphic signal, such as that beneath the Dadeldhura klippe, Models A-C also incorporated older out-of-sequence faults that are needed to reproduce the surface geology (e.g. at ~39 km south of the MCT in Fig. 5). Given enough time and exhumation, the out-of-sequence fault at ~70 km south of the MCT could be exposed and would likely appear similar to the fault at ~39 km south of the MCT – apparent in the surface geology, but not necessarily in the thermochronologic record. While geomorphic indicators argue it is active/recently active now, that it represents a significant reorganization or will continue to be active in the future is uncertain. A likely scenario is that the fault is helping build taper locally, so that the main decollement and southward propagation can continue. So we agree it is possible for there to be complexities in the earlier history of the orogen, but we would argue that most of

these complexities would be in the units that are now erosionally removed and thus not encompassed by our models. Complexities experienced by the rocks above the decollement and still within the model space would have to be limited in magnitude and extent as to not influence the preserved measured cooling ages or the visible and mapped surface geology.

We have edited Section 5.3 ‘Exhumation drivers in the western Nepal high Himalaya’ to better address what components of the modelled geometry and kinematics are required to meet the constraints. We have also added a new section to the Supplementary Materials, Text S6 titled ‘Assessing cross-section geometry, kinematics, and rates’ that provides additional details regarding how the geometry and displacement on the new faults (under the klippe) were determined as well as other geometric and temporal constraints.

Even if the first order topography is modelled - this will be averaged to mean topography. The dramatic changes they propose of the exhumation rates related to changes in evolving ramps (Fig. 8 upper panel) - implies that the model topography is oversimplified - and most likely orogenic wedge formed activity out-of-sequence fault or reactivated from ramps several time - however this information is lost and most likely due the lowering of resolution of the thermal history during middle and early Miocene evolution. Here only the real first order deformation processes are recognized.

The dramatic changes in exhumation rate in space and time in Fig. 8 (Fig. 9 in the revised manuscript) are a response to the increase in exhumation that occurs above a decollement ramp and the slower, more passive exhumation that occurs in the absence of active uplift as these rocks are translated away from the ramp, as well as pulses of rapid exhumation during out-of-sequence motion. Previous work (e.g. Braza & McQuarrie, 2022a; Gilmore et al., 2018; McQuarrie & Ehlers, 2017) has explored the influence of how topography is modelled in a section on the flexural and thermal responses and have shown that if the estimated topographic angle is consistently applied it does not matter which angle is selected, and to recreate the geology exposed at the surface today, the same total amount of erosion needs to occur. The selection of topographic angle, EET, and rock density is non-unique and there are multiple solutions that would predict the same cooling ages – what makes a difference is if the erosion angle is changing with time (such as systematically getting lower) to force more erosion later in the model. The only way to reduce the ‘dramatic’ changes in the proposed exhumation rates would be to have higher modelled topography directly over the ramp to lower the exhumation magnitude. For the hinterland this would entail significantly (unrealistic) higher topography than the modelled 5 kms. The point we are emphasizing is large ramps (~8-10 km) result in high magnitudes (~8-10 km) of exhumation as rocks are displaced over the ramp.

We have modified Section 5.3. ‘Exhumation drivers in the western Nepal high Himalaya’ to expand the discussion of Fig. 8 (now Fig. 9 in the revised manuscript) and better address what is driving the high exhumation rates and spatial/temporal variability in exhumation rates.

Also the mega-ramp for the onset of modelling - is simplified solution to generate useful thermal field for the stating of the modelling - however have no geologic evidences for this set-up. All these implication - need to be presented in some why or other - to help the reader to understand and appreciate the great work of the authers.

The height of the ramps in Model C are controlled by the stratigraphic thicknesses, as determined from the map distances and unit orientations at the surface. The height of the “mega ramp” at ~460 km in Fig. 6a is set by the mapped extent of GH rocks at the surface to the north of the MCT in Fig. 1. Since this ramp is large, it results in high magnitudes of exhumation (Fig. 9) and resulting advected isotherms (Fig. 6a).

We use the peak temperatures (Fig. S1) to gain insight into the conditions the rocks experienced while at temperatures hotter than the MAr closure and to ensure the cooling path at least begins at an appropriate temperature. As the reviewer points out here and in other comments, the temperature of the GH rocks has an impact on the modelled thermal field. This is important because it affects how much heat is being advected during MCT motion, which impacts the peak temperatures of the lower LH rocks in the footwall. We recognize that peak temperatures can be influenced by many processes and factors beyond burial depth, such as ductile shearing, but to a first order, the peak temperatures in western Nepal are reached when rocks are at their maximum burial depth (e.g. Braza et al., 2023). For GH rocks, the maximum burial depth is reached immediately prior to the initiation of MCT motion, which is the start of our models. We thus set the ramp height to cover the full stratigraphic extent of GH rocks.

We more clearly specify how the ramp height in the cross-section is determined on lines 549-551 and in the new Supplementary Text S6. ‘Assessing cross-section geometry, kinematics, and rates’. We have also modified Section 5.3. ‘Exhumation drivers in the western Nepal high Himalaya’ to expand the discussion of Fig. 6.

In Fig. 8 the authors present exhumation pattern of MBT hanging wall - however have poor or no data to constrain this history in more detail. After there reference frame the MBT is exposed at the surface at -120 km south of the MCT. First ZHe and AFT are at ~110 km south of MCT. In other segements of the Himalaya obtained thermochron data (ZHe and AFT) in vicinity of the MBT yield reset ZHe (< 3 Ma) and the pattern of young AFT ages imply deep ramp likely linking directly to the MHT - horizontally running about ~40 km into MBT hannging wall.

The initial exhumation of the MBT hanging wall is strongly constrained by the MAr and ZHe cooling ages in the Dadeldhura klippe, the basin depositional ages in the lower Siwalik unit, and the lower LH provenance signal depth and depositional age in the basin. The final motion on the MBT that emplaces the upper LH rocks over the Siwalik units follows the constraints from previous work, with displacement that is younger than the Siwalik units in the footwall.

We limited the cooling age data used to constrain the thermokinematic models to the region shown in Fig. 1, due to lateral variations in the surface geology, locations of elevated topography, and geomorphic metrics such as stream steepness indices. Along the Simikot transect and adjacent sections to the east and west, measured AFT ages are ~5-10 Ma and ZHe

ages are either partially reset or as old as ~12-15 Ma within 20 km of the mapped location of the MBT at the surface (e.g. Braza et al., 2023 for the Api section to the west and our unpublished data for the Juphal section to the east). These lateral variations suggest that there could be underlying changes in the geometry, kinematics, shortening rates, etc that are influencing what we see at the surface, both across the Himalaya and within Nepal alone. Projecting cooling age data from outside the area in Fig. 1 thus has the potential to mislead our interpretation of the geometry and/or timing of faulting along the Simikot transect in western Nepal. The final MBT motion proposed for Model C is internally consistent with the available data, even though (regrettably) there are no measured ages between the MBT and the RMT.

Furthermore in Fig. 8 south of the MCT (-40 km to the MCT) they are not able to reconstruct the tectonic evolution of early Miocene - as most likely vigorous erosion has removed cover units of the LH duplex.

While the cover units of the LH duplex have largely been removed by erosion, Fig. 8 (now Fig. 9 in the revised manuscript) shows the exhumation and burial history for only the rocks that are presently exposed at the modern surface. The white spaces at ~40-0 km south of the MCT in the original Fig. 8a indicate the rocks are not experiencing exhumation and are instead being buried. The rocks at ~40-20 km south of the MCT are in the immediate footwall of the RMT and reach their peak burial depth at the end of RMT motion. These rocks do not begin exhuming until formation of the LH duplex at ~13.6 Ma. We have updated this portion of the figure with grey shading to more clearly show that these rocks are experiencing burial.

Also the studied section in Western Nepal provided very little thermochronologic constraints for the hanging wall of the MCT (data are <10 Ma). However - other sections of the Himalaya data more data are available. Therefore the presented models of exhumation pattern from the MCT hanging wall for the early Miocene is not well constrained by the presented data and a lot of interpretation is used here. Therefore much more out-of-sequence deformation of direct MCT foot- and hanging wall (RMT-duplex and 20 km of MCT hanging wall) during Early and Middle Miocene might not be recognized by the modelling and available data.

We agree that it is possible for there to be additional deformation of the MCT hanging wall by either (or both) in- or out-of-sequence faults. For example, Carosi et al. (2010) propose the Tojiem shear zone was active in the Oligocene with top-to-the-south motion that repeated the GH section. Their observations are ~75 km to the east of the Simikot section. This motion is proposed to have occurred prior to initiation of MCT motion, which is when our models begin. This pre-MCT faulting would influence the depth of the GH rocks and the temperature conditions at the beginning of MCT motion. Incorporating faults that repeat the GH would negate the necessity of the “mega ramp” (as the reviewer calls it in other comments), as a thinner initial package of GH rocks would define the height of the ramp. However, this would require additional interpretation that we do not have the data to constrain for this section.

The constraints we do have for the deformation history of the GH rocks are the amount of shortening required to connect the lower LH and GH rocks in the Dadeldhura klippe with the respective units in the north, the requirement for a flat-on-flat relationship between the GH and lower LH rocks that is formed by emplacement of GH rocks atop the lower LH rocks by MCT motion, that the MCT was active by ~25-30 Ma, the MAr ages in the Dadeldhura klippe are ~18 Ma in the southern limb and young to ~14 Ma in the northern limb, and the MAr ages in the north are ~6 Ma. As our models highlight that the MAr ages in the Dadeldhura klippe were set during RMT motion when GH and lower LH rocks were translated over the northern edge of the upper LH rocks, all of the shortening on the MCT must occur between ~25-30 Ma and ~18 Ma, prior to initial uplift of the southern limb of the Dadeldhura klippe over the northern edge of the upper LH rocks. Exhumation of the northern GH rocks by the young out-of-sequence motion is constrained by the ~6 Ma MAr ages that indicate a significant pulse of cooling at this time. These constraints do not preclude periods of significant in- or out-of-sequence motion during MCT motion (prior to 18 Ma) or at ~18-6 Ma on faults in the north, but they also do not require this additional displacement.

While the geometry and kinematic sequence for Model C are more complex than a geometry that focuses on minimizing shortening or a strictly in-sequence progression of faulting, the complexities we introduce are required components of the burial and exhumation history to be able to reproduce the surface geology, measured cooling ages, basin thickness and accumulation data, and modern topographic relief, slope, and elevation. We have edited Section 5.3 'Exhumation drivers in the western Nepal high Himalaya' to better address what components of the modelled geometry and kinematics are required to meet the constraints. We have also added a new section to the Supplementary Materials, Text S6. 'Assessing cross-section geometry, kinematics, and rates' that provides additional details regarding the geometric and temporal constraints.

Detrital Ar/Ar data from Siwaliks from Szulc et al., 2006 a study the author refer to - shows that the maximum exhumation of Greater Himalaya during early Middle Miocene is 3.6 mm/yr - based to short lag times (<4 Myrs) between depositional and cooling age between 16-13 Ma. This do not match with exhumation proposed from the modelled sections - see Fig. 6. However - this is not discussed. All these are limitation the authors do not or partly miss to discuss or present in any way. The reviewer gets to some degree the impression the authors do a kind of cherry picking - just choosing studies that support obtained results - neglecting studies contradicting obtained results.

Our intention was not to cherry-pick among studies but were instead trying to keep the discussion as concise as possible. The detrital MAr data of Szulc et al. (2006) has been incorporated into Section 2.2 'Basin record' and Fig. 2. The predicted detrital MAr distribution is compared to the measured distribution in the new Fig. 7 and in Section 5.3 'Exhumation drivers in the western Nepal high Himalaya'. The predicted detrital MAr distribution for Model C does fully reproduce the distribution of measured ages. The Szulc et al. results (3.6 Myr lag time and

2.6 mm/yr exhumation rate from 16 Ma to 10 Ma), is very similar to the 3 mm/yr the long-term average of the more detailed exhumation presented in Fig. 9.

Tectinal suggestion and comments:

Line 68-90 - 22. Basin Record:

Introduce also other parts of the study by Szulc et al.: Detrital Ar/Ar data from Siwaliks from Szulc et al., 2006 and discuss later the mismatch between your results and their results.,

The measured detrital MAr data has been incorporated into Section 2.2 'Basin record' and Fig. 2. The predicted detrital MAr distribution is compared to the measured distribution in the new Fig. 7. As detrital MAr ages are a function of both where rocks are exhuming as well as the magnitude and size of mica in those rocks, our goal is to reproduce the distribution of detrital MAr ages and not focus on the magnitude of the detrital MAr ages or specific peak locations in the distribution. Model C does sufficiently reproduce the distribution of measured detrital MAr ages.

Line 135 - how can you publish thermokinematic modelling results without having published new thermochron-data ???

The necessity for publishing new thermochronologic data lies with what is available from previous studies for a given location and the distribution of the existing data. New thermochronologic data is published for this cross-section to fill in the location gaps between the extensive published cooling ages. For example, our new ZHe ages are the only ZHe samples at ~40-0 km south of the MCT. With sufficient location coverage of multiple chronometers from existing datasets, it can be possible to re-evaluate the cooling ages using thermokinematic forward modelling without needing to analyze and publish new samples.

Line 175: OK, very good. What are and where in manuscript are these these predictions tested and confirmed?. Is this also confirmed and varified for the "RMT duplex and MCT-hanging wall ? What are the measured peak-temperatures in "RMT duplex" and MCT-hanging wall of Western Nepal???

The text has been updated to include a reference to the Supplementary Materials where the peak temperatures are shown, line 198.

Line 322-324: This is most tricky part of the reconstruction - because authors have very little independent information to constaine this. Entire new faults and not reactivation are suggested by the authors - to improve the fit between Modell and topographic parameters.

Thus are we at a transistion to major reorganisation of the deformation with Western Nepal over the last million year - as others have suggested? Even if the location of ramps of MHT are

different in this solution - other (e.g. Harvey et al., 2015) have concluded that Western Nepal might in process of tectonic transition???

Only the fault proposed beneath the Dadeldhura klippe is new and is required to reproduce various geomorphic indices. We initially tried to reproduce the modern topography in the landscape evolution models without the new out-of-sequence fault and it quickly became apparent that a mechanism for young uplift is required in this region. We tested fault locations within ~80-60 km south of the MCT, with varying ramp angles and amounts of shortening to evaluate the influence of the location and angle of the proposed fault on the AHe ages measured in the Dadeldhura klippe in the thermal model and on the location, magnitude, and slope of the elevated topography in the klippe in the landscape evolution model. The location of the new fault is consistent with the region of “intermittent/recent duplexing” proposed by Harvey & Burbank (2024) though it occurs at a shallower structural level, in part due to the overall shallower decollement depth in Model C (as constrained by the measured basin depth) compared to the geometry shown in the Harvey & Burbank figure (their Figure 6). The position and geometry of the new out-of-sequence fault allows slip to be fed to the RMT splays in the southern limb of the Dadeldhura klippe, consistent with the distribution of observed microseismicity in the region.

We have added a new section to the Supplementary Materials, Text S6. ‘Assessing cross-section geometry, kinematics, and rates’, that provides additional details regarding how the geometry and displacement on the new faults (under the klippe) were determined as well as other geometric and temporal constraint.

Line: 325-329: Results of Model C shown in 'Figure 5. Is the revised geometry and kinematic sequence new - or is it based on earlier work?

The geometry and kinematic sequence tested with Model C are new. The improvements in Model C are based on what we learned from the results and misfits of Model A (Robinson et al., 2006 geometry) and Model B (Olsen et al., 2019), as well as from interim models that tested various ramp locations, sizes, and angles. The geometry in Model C is also based on our updated geologic mapping. We have clarified this by adding lines to Section 4.4.3 ‘Model C: Revised geometry’ to remind readers that this is the new cross-section shown in Fig. 3.

Line: 335: Figure 6: Please define - Displacement and Velocity in figure capture - so interested reader better understand - how to read the figure. The result are very interesting - and potentially have a lot of implication - for a overall understand of the Himalayan wedge deformation - not only for western Nepal Himalaya - but rather Himalayan evolution in general.

The legend within each subplot has been updated to “Shortening rate” to better clarify the meaning of the previous text of “Velocity”. The caption has been updated on lines 380-382 to expand on the meaning of “Displacement” in each subplot legend.

These changes in velocity of the model - which relates to shortening rate within this segment of the orogen - is some degree different to the results presented of models of the best results yield in parallel but earlier published study in Tectonics 2023 by the authors - why is that?

The text in Section 5.3 'Exhumation drivers of the western Nepal high Himalaya' has been updated on lines 568-570 to incorporate how the timing and magnitude of out-of-sequence shortening compares for far western Nepal in Braza et al. (2023).

Each cross-section geometry is independently tested using the thermokinematic forward modeling approach to determine what is required to reproduce the measured datasets that are available (basin data, cooling ages, peak temperatures, geomorphic indices, etc) for a given section. While independently tested, the thermokinematic models for the Simikot section in this study and for the Api section in Braza et al. (2023) both highlight similar requirements for the geometries and kinematic sequences, including the northerly ramp location that is near or north of the MCT in the hinterland, the faster shortening rates during MCT and RMT motion, the slowdown in shortening rates beginning ~15-11 Ma and continuing until ~1 Ma, the initiation age of the LH duplex is ~13.6 Ma, and out-of-sequence motion on the Ramgarh-Munsiari thrust is younger than 6 Ma. Model C of the Simikot section is also consistent with the temperature-dependence of deformation style examined by Braza et al., with long, low-angle high-displacement thrusts active while temperatures are above ~375°C (MCT, RMT) and a transition to non-temperature dependent processes below ~375°C.

In detail though, there are distinct along-strike changes in the surface geology and location of units and major structures at the surface, locations of elevated topography and geomorphic indices, and distribution of young cooling ages in between the Simikot and Api transects, at ~100 km distance apart. For example, the MAr ages in the RMT and MCT hanging walls in the north of the Simikot section are very young at ~6 Ma, while those in the equivalent location in the Api section are ~13 Ma. What is driving the cooling and exhumation of these rocks in the north through the MAr closure temperature must be different. As the thermokinematic model results for Model C highlight, a significant pulse of out-of-sequence motion in the hinterland is necessary to exhume GH rocks through the MAr closure temperature at ~5-6 Ma. Models for the Api section also highlight out-of-sequence motion in the north as a required component of the deformation history, but this is to influence the ZHe and AFT ages in the RMT and MCT hanging walls, not the MAr ages. As such, the magnitude of out-of-sequence motion is less than required by the Simikot section. The timing of out-of-sequence motion is similar to that proposed here for the Simikot section. We have edited the text in Section 5.3 'Exhumation drivers of the western Nepal high Himalaya' to address this.

In (a) the authors present the Initial model configuration. Try introduce here a kind of "Mega Ramp" in the MHT at 25 Ma - however do not define, if the have field observation or other constraints justify this. What is motivation - or is it way to try set up crustal thermal field for the sections modelled later??? This has strong impacts of the modeled thermal field for the entire

Miocene - therefore the reader needs more information about this and motivation - explanation why this is justified.

The height of the ramps in Model C are controlled by the stratigraphic thicknesses, as determined from the map distances and unit orientations at the surface. The height of the ramp at ~460 km in Fig. 6a is set by the extent of GH rocks mapped to the north of the MCT in Fig. 1.

We use the peak temperatures to gain insight into the conditions the rocks experienced while at temperatures hotter than the MAr closure and to ensure the cooling path at least begins at an appropriate temperature. The temperature the GH rocks reach affects how much heat is being advected during MCT motion, which impacts the peak temperatures of the lower LH rocks in the footwall. The available data that constrain the thermal field for the early portions of Model C is the GH rocks reached peak temperatures of ~700°C (e.g., Braden et al. 2020; Soucy La Roche et al. 2016, 2018) and then cooled through the MAr closure temperature at ~14-19 Ma in the Dadeldhura klippe and at ~6 Ma in the north. Lower LH rocks in the footwall of the MCT preserved in the north reached peak temperatures of ~450-650°C. Lower LH rocks preserved in the southern limb of the Dadeldhura klippe cooled through the ZHe closure temperature at ~13-15 Ma. This indicates that thrust burial of lower LH rocks in the MCT footwall ceased prior to ~13-15 Ma and lower LH rocks had already cooled from their peak temperature conditions by more than 200°C, meaning MCT motion ceased long prior to 15 Ma.

We recognize that others may have similar questions and thus added a new section, Text S6, to Supplementary Materials titled 'Assessing cross-section geometry, kinematics, and rates' to address these comments and others regarding the sensitivity of the model results to these parameters.

Line 432: Exhumation drivers in the western Himalaya - Please explain better what the effect of the shortening rate changes shown in Fig. 6 between ~15 to 13.5 Ma - Model step 32 - 39. This is indeed very interesting - if you have good reasons. Is this change fundamental to get a good fit between model and measured data???? But also discuss the results other thermochronologic approaches such as Szulc et al., 2006 or DeCelles et al., 2020. Aren't the models fits very strongly supporting work by Wobus et al., 2006 - - Model step 56 - Out-of-sequence deformation in the vicinity of the MCT with Late Miocene reset Ar/Ar-Mica data???

We expand on these points in the revisions to Section 5.3. 'Exhumation drivers in the western Nepal high Himalaya', including discussing the relationship between shortening rate and exhumation rate. We have added a citation for Wobus et al. (2006) on line 568. In detail, there are a series of constraints that are used to determine shortening rates. We specify a few below and have included these in our new Supplementary Materials section, Text S6, titled 'Assessing cross-section geometry, kinematics, and rates'.

The possible range of constant velocities are rates that maintain an MCT initiation age of ~25-30 Ma (e.g. Soucy La Roche et al., 2018). At the lower bounds of the MCT initiation age of 25-30 Ma and similar to the 25 Ma age in Model C, a velocity of 24 mm/yr provides an initiation age of ~25.6 Ma for the amount of shortening outlined in Table S3. This faster constant rate

predicts basin depositional ages that are much younger than the measured data, only reproducing the depositional ages for ~3-3.5 km basin depth. At the upper bounds of the MCT age range, applying the modern GPS shortening rate of ~20.5 mm/yr in west Nepal (Ader et al. 2012) to the duration of Model C has an MCT initiation age ~30 Ma. This slower constant rate provides an older basin age than 24 mm/yr rate, but still predicts a basin that is generally younger than the measured data and only reproduces the depositional ages for ~3-3.8 km basin depth. While the timing of RMT motion generally falls within the required window with constant rates of 20.5-24 mm/yr, the beginning of LH duplexing occurs at ~9.8 Ma with the 24 mm/yr rate and ~11.5 Ma with the 20.5 mm/yr rate and continues until ~5-6 Ma, too young to reproduce the ZHe and AFT ages measured in the Dadeldhura klippe. Additionally, the out-of-sequence motion on RMT faults required to produce the ~6 Ma MAr ages in the north occurs at 4.34-3.3 Ma with 24 mm/yr and 5-3.87 Ma with 20.5 mm/yr, both too young to reproduce the measured ages. We have incorporated the results for the constant velocity model with shortening rates of 24 mm/yr in the Supplementary Materials, Table S3, Text S6, and Fig. S4.

Line 455-460: This the most interesting part of the study - however much to surfically and short - this needs to be significantly extended - also to really appreciate the lovely findings of the study and all the implications - if the setting favour by the authors are getting close to how the tectonic evolution of the frontal Himalaya have evolved since the Middle Miocene - Pleiocene - until today.

This part the of the discussion is the most important part - to better understand the results - and implication the authors have worked out.

It is good choice to focus on the last five Ma - show here the entire strength of the thermochronologic approach in active orogen such as the Himalaya - vigerouse erosion and exhumation rates. Combining both geomorphic and other constaints and setting to the results in context to the model results but also to other studies illustrates its strength. It is most likely here - where the models the authors present are best justified - because the input parameters have their highest resolution.

At the same time these results have a lot of implication - which are lost and not mentions by the authors.

They illustrate the spatial deformation pattern has changes on million to two million year time scales in different compartments of the orogen. Some of the parameters here available are much less well constraint for the earlier time steps of the modell. Also detailed resolution of the thermochronolic dataset lowers significantly at earlier time windows studied here. This implies that the presented results might show nicely the first order cooling pattern - however lack to exactly reconstract earlier evolution in the same detail - as this apporach is able to do of the last

five Ma. The reviewer is convinced that the authors are aware of the this - however not readers will be aware of this. 'This has also strong implication, how reliable the results presented in Figure 8 are justified - especially of the earlier history or the orogen. Compare the again differences ob the obtained model results with Szulc et al., 2006 or DeCelles et al., 2020. Not only with Harvey and Burbank, 2024)

Understanding the last five million in this detail means - for instance also earlier time windows were affected most likely by significant deformation switches forth and back in million to two million year timescales - but this likely beyond the resolution to the presented datasets for earlier timescales. Here plays the critiacal collomb wedge models an important role - how the Himalayan orogenic wedge deformed during shortening and erosion through time.

Therefore beside the first order discussion on the architecture of the MHT and this part should be discussed in much greater detail - to better appreciate the strength of this approach and what we can learn from it. This is sometimes a bit lost in just listing the first order aspects of the study.

Our original intention was to keep this section as concise as possible, but we see how this ultimately resulted in a more surface-level discussion. We have reworked Section 5.3 'Exhumation drivers in the western Nepal high Himalaya' to expand our evaluation of the drivers of exhumation and to better incorporate Figs. 6 and 8 (Fig. 9 in the revised manuscript) into the discussion.

Also please consider a small paragraph on temporal resolution of the study and which details (parameters) can be considered.

See the new Supplementary Materials, Text S6. 'Assessing cross-section geometry, kinematics, and rates'.

RC2: 'Comment on egusphere-2026-1560', Anonymous Referee #2

General comments

Overall, this is a well-executed study that compiles an impressive thermochronological dataset, (new and published) with basin accumulation rates, thermokinematic and landscape evolution models to investigate structural controls on exhumation in western Nepal. The integration of balanced cross-sections with thermal constraints is a methodologically sound approach, and the effort required to synthesize and iterate across this volume of data is commendable. In general, the work makes a valuable contribution to understanding Himalayan thrust belt tectonics and is relevant to EGU Solid Earth. The paper is well structured and written, the title and abstract accurately represent the content, and the figures are clear and easy to follow. The central conclusion of the work is that the structure, especially MHT ramp geometry, is a primary driver of observed cooling dates, basin accumulation rates and topography from Simikot transect in

western Nepal Himalaya. The authors evaluate this with thermokinematic models, which are in turn compared with landscape evolution models to support the main conclusion. However, I have several questions and comments below.

1. Averaging and treatment of thermochronology data

A study of this scope inevitably requires averaging and generalization of data, especially when compilations span a large spatial extent and datasets from multiple published work. However, my biggest concern is the treatment (mean/averaging) of ZHe and AHe ages (as mentioned in Line 96). If all samples rapidly cooled, then the ages can be justifiably averaged but if there is inter-grain variability, averaging cooling ages can yield geologically meaningless dates. These are important for ensuring that the ages input into the thermokinematic models genuinely reflect cooling rather than grain-scale artifacts. Additional information on the treatment of thermochronology data (screening of outliers? age-eU correlations?) would help strengthen the authors use of averaged datasets.

The grain ages presented in Table 1 are averaged with an error that encompasses the analytical error and the grain distributions. We agree that care must be taken when evaluating individual grain ages and averaging their distributions, especially when this data is being used as an input to constrain an inverse thermal model. The thermokinematic models we present are forward models, with the Pecube model predicting the thermal response of the crust to the specific geometry, kinematic sequence, shortening rates, topography, and flexural solution of the Move flexural-kinematic model. The measured cooling ages are not used as an input constraint on the model, but are instead separately compared against the cooling ages predicted by Pecube. We also note that in many locations where the data themselves have larger error bars due to the averaging of ages, that there is also a pronounced spread in age amongst samples showing a true uncertainty in any given age (e.g. 105 km to 65 km south of the MCT). As such, we can still use the large error bar data to understand the broad, first order trends of cooling. The models are also sufficiently constrained by additional datasets, so we are not only relying on the large error bar ages. We have edited section 3.4 Thermokinematic forward modelling to more clearly specify that these are forward models rather than inverse models. We have also edited Section 2.3 ‘Published thermochronologic record’ to address that the datasets have been screened for outliers and age-eU relationships (lines 105-106).

2. Pecube modeling details

As I (and many readers) are not deeply familiar with Pecube, can the authors clarify the use of constant and /or variable shortening rates? Is this routine or deliberate methodological choice? For variable shortening, the authors mention that it was based on cooling ages (depth, depo ages) from the foreland which is reasonable, but what were the range/uncertainty of the variable shortening? This context would strengthen the reasoning. Broadly, please add more information

about the iterative process that produced Model C. Were Models A and B evaluated and adjusted until Model C yields a satisfactory fit? Or was Model C independently assessed and yielded a better fit? Additional details about what changes were made between iterations, how many iterations were attempted, and why >80% fit threshold was adopted to select models for CASCADE will be good for clarification. Also, in Figure 6, both displacement and velocity varies between each model slip. How was the required displacement evaluated? How sensitive are the models are sensitive to these parameters, especially the velocity (varies 24, 48, 45, 16, 15, 18, 20 mm/yr in Figure 6)?

The Pecube methods have been edited to better explain how the variable velocities are determined (lines 167-170) and to address the various model iterations that led to Model C (lines 209-213). A constant shortening rate model has also been added to the Supplementary Materials, Fig. S4. We have also added a new section to the Supplementary Materials, Text S6. ‘Assessing cross section geometry, kinematics, and rates’. As we are following procedures established in previous studies, the details of the flexural-kinematic modelling and thermal modelling are included in the Supplementary Materials and we limit the methods in the main text to the specifics of the Simikot section/western Nepal.

The required displacements are constrained by the cross-section geometry and begins with the minimum amount of shortening required to reproduce the surface geology, and in the case of Model C, still maintain a more northern ramp location. The total displacement on the fault is broken down into ~10 km steps and each step is isostatically accommodated and eroded in Move. The shortening rate is determined by the amount of displacement and the increment of time this displacement occurs over. The age for each increment of shortening is determined by the cooling ages and which modelled cooling ages are sensitive to a given fault motion and age. The modeled velocities are then adjusted to better reproduce the measured ages and/or the measured basin depositional ages. For example, sediments at ~4 km depth in the modelled basin accumulate in Step 52 of Model C, during LHD-1 motion in the LH duplex. The depositional age of strata measured at an equivalent depth in the basin is ~6.8-12.7 Ma, so the timing of Step 52 must fall within that window. Multiple different combinations of shortening rates are tested to identify the best fit to the cooling ages and basin depositional ages, and to assess the sensitivity of the ages to the rates. In general, thermokinematic models of the Nepal Himalaya show a stronger sensitivity to the age of fault motion, rather than the shortening rate (Braza et al., 2023).

Each cross-section geometry is independently modeled and evaluated against the measured cooling ages. The first model for any cross-section is an in-sequence (foreland-ward progression of faulting), constant velocity model, as this provides the least complex exhumation pathway. From the fit/misfit to the measured data, the shortening rates are varied to see if the timing of uplift prescribed by a specific kinematic sequence can be adjusted to better reproduce the cooling ages and basin depositional ages. The number of velocity combinations tested depends on how poorly the measured ages are reproduced and whether any geologically viable changes in shortening rate are sufficiently improving the model fit. From there, the geometry (ramp location) and/or kinematic sequence (in- or out-of-sequence order of faulting) is modified

and input into a flexural-kinematic model and then thermal model. This iterative process is completed until a viable exhumation pathway is identified.

The geometry in Model C is based on the updated geologic mapping and the fit/misfit to the cooling ages in Models A and B. The results for Models A and B indicate that the modern active ramp cannot be placed south of the young measured AFT and AHe ages in the north and certainly cannot be placed beneath the Dadeldhura klippe. We tested multiple ramp locations in the northern region of the cross-section, to the south and to the north of the MCT, to constrain the specific location of the active ramp presented in Model C. The final geometry, kinematics, and shortening rates presented for Model C are the result of more than 30 flexural-kinematic models and more than 60 thermal models (not including Pecube models with only variations in surface heat production), in addition to the variations of Models A and B that were tested. The location and angle of the ramp, magnitude of displacement, and the timing of the young out-of-sequence fault beneath the Dadeldhura klippe in Model C is further constrained by 20 landscape evolution models that tested the sensitivity of fault kinematics and geometry on topography.

The >80% fit threshold for Cascade modelling was an arbitrarily selected value. This cutoff was intended to balance the need to understand how (un)successfully the topography is generated and maintained in the landscape evolution models for a given geometry and kinematic sequence, with the time that is needed to fully complete the Move, Pecube, and Cascade modelling processes. If the geometry and kinematic sequence of a given thermokinematic model cannot reproduce even the broad trends in the cooling ages, then it becomes a question of what insights can be gained by testing the unviable exhumation pathway in the landscape evolution model? At the same time, this thermokinematic-landscape evolution modeling is an iterative process, where the results of the Move, Pecube, and Cascade models each inform the changes that will be made in the next iteration. Testing a geometry and kinematic sequence that reproduces >80% of the measured cooling ages allows us to gain insight into various aspects of the exhumation pathways and the location, timing, rate, and magnitude of uplift with the landscape evolution models to make a more informed decision for what will be tested in the subsequent model.

On a different note, the authors should move the discussion of the South Tibetan Detachment System and Tethyan Himalayan rocks from supplementary in the main text. Given that Soucy La Roche et al. (2016, 2018) document STDS and THS rocks in western Nepal in the Dadeldhura klippe, the authors should briefly clarify it in the main text.

The discussion of the South Tibetan Detachment System and Tethyan rocks has been moved to Section 3.4 ‘Thermokinematic forward modelling’, lines 151-155.

3. Discussion of alternate drivers of cooling and exhumation

Firstly, the authors do an intensive interrogation of existing structural models (Model A-B) to test with observed independent thermochronology and basin accumulation rates, but the authors do

not account for an erosional signal that could be independent of ‘tectonics’. For instance, other work adjacent to the study area in western Nepal (e.g., Sherpa et al., 2022) suggests, based on preservation of an anomalous high-elevation low-relief surface and cooling ages, that low-T thermochronology results highlights an erosional signal due to a northward propagating fluvial system. Do the authors think that the ramp is the strongest control or there is no erosional control at all? I recognize that the authors do briefly talk about fluvial incision required to replicate topography in Lines 425-427 but I encourage the authors to discuss alternate explanations.

We have incorporated discussion of the fluvial incision into a high elevation, low relief surface as a mechanism to generate the elevated k_{sn} values and raised topography in the Dadeldhura klippe (Sherpa et al., 2023) to lines 479-498.

We do assert that the first-order control on where exhumation is occurring over geologic time scales (0.2-2 Myr) is governed by structural uplift, and surface processes are responding to that uplift. Note that the HeFTy cooling models in Sherpa et al. argue for rapid $\sim >200^{\circ}\text{C}$ cooling (using ZHe, AFT, and AHe data), but argue that the AFT ages record the river incision. Note that Harvey & Burbank (2024) also argue for rapid cooling between ~ 13 Ma and 9 Ma from potentially as high as 400°C (QTQt inversion). We agree with Sherpa and co-authors that northward younging ages in the ZHe and AFT systems argue for a northward younging process. The challenge with river incision is that it cannot explain the ZHe ages, and in our opinion (looking at the impact of 3+ km in river incision in Peru that only effects AHe ages e.g. Buford Parks et al., 2023) river incision would be hard-pressed to account for the AFT ages. However, showing that is beyond the scope of this paper. Thus, we argue that the pulse of rapid cooling shown in Sherpa et al.’s inverse HeFTy modelling (consistent with 8 km of exhumation) is the expected response of exhumation over a ramp. We then argue in lines 494-498 that these high elevations are difficult to maintain due to continued flexural loading (at frontal and hinterland ramps) and erosion. Where they are maintained, it is because uplift at a frontal fault outpaces a river’s ability to incise, protecting the elevated regions in the hinterland. When that uplift slows down or steps forward, rivers take advantage of the steep now not uplifting slope and incise into the hinterland rapidly over 1-3 million years (e.g. Plaster et al., 2025, and work of ours in central Nepal that is currently in review). It is recorded by young AHe ages in the Andes but has no effect on thermochronologic ages in central Nepal.

Specific comments

Line 15. Are there any ZHe, AFT, ZFT, AHe data from the northern zone? Apatite, zircon fission track and (U-Th)/He thermochronometers have vastly different closure temperatures from muscovite Ar-Ar thermochronometer so the cooling dates may be documenting something different, unless they are all within same age range (within uncertainty).

There are ZHe, AFT, and AHe ages that are <3 Ma in the northern zone (Figs. 1, 3). With the word limit of the abstract, there is not sufficient space to clarify this.

Line 17. Add 'in western Nepal' after Simikot transect.

Corrected on line 17.

Line 42. Please add citation for the AHe, AFT study.

Added on line 42-43.

Line 46. Please add citation to the thermochronology studies or figure in the manuscript.

Added on lines 46-47.

Line 110-113. Please rethink this line. It is unclear how prograde path documented for rocks in NW Indian Himalaya (Corrie) and Bhutan Himalaya (Long et al., 2016; Long and Kohn, 2020) can be meaningfully related to this sector of the western Nepal Himalaya. Metamorphic histories are often sample specific, depending on when the rock equilibrated, and need to be carefully evaluated given the lateral heterogeneity in deformation and metamorphism across the orogen.

We agree that metamorphic histories must be evaluated in the context of the specific location and conditions the rock experienced, especially given the lateral variations in surface geology, peak temperatures, and distributions of young cooling ages that suggest differences in the timing and conditions of metamorphism across western Nepal alone. The interpretation of what these samples represent and the history they record can still provide insight for how to interpret similar features elsewhere in the orogen. We have edited the text on lines 127-128 to better clarify that the interpretations of granular distortions are for other parts of the Himalaya.

Line 155. Were the input ZHe and AHe ages raw or corrected ages? Please add.

The thermal modeling in Pecube is forward modelling that uses the input displacement grids from Move to calculate an evolving thermal field. The measured cooling ages are not used as an input in Pecube, but are compared to the output predicted cooling ages. We have clarified this on lines 178-179.

Line 156. Please reconsider these lines. The model excludes distributed ductile shearing and the justification is that the model reproduces peak temperature gradients well but thermal fit alone cannot confirm the absence of a deformation mechanism.

Our intention was not to assert that ductile processes were not an active deformation mechanism in the high temperature GH rocks, but rather that our ability to reproduce the thermal field suggests that distributed ductile shearing was not the dominant driver of exhumation in our study area. Outcrops of the GH rocks certainly show evidence of ductile processes and high temperature conditions.

We use the peak temperatures to gain insight into the conditions the rocks experienced while at temperatures hotter than the MAr closure and to ensure the cooling path begins at an appropriate temperature. This is important because the initial temperature of GH rocks affects how much heat is being advected during MCT motion, which impacts the temperatures of the

lower LH rocks in the footwall. We recognize that peak temperatures can be influenced by many processes and factors beyond burial depth, such as ductile shearing, but to a first order, the peak temperatures in western Nepal are reached when rocks are at their maximum burial depth (e.g. Braza et al., 2023). For GH rocks, the maximum burial depth is reached immediately prior to the initiation of MCT motion, which is the start of our models. The peak temperatures and cooling ages are sufficiently reproduced by thermokinematic forward models. We have clarified that ductile processes are not absent on line 199.

Line 159. Some samples deviate spatially laterally from the transect significantly, why is a lateral error of ± 0.5 km applied?

The measured data in Fig. 1 is projected onto the cross-section following the strike of the major structures, to ensure the sample is evaluated at an equivalent structural location on the cross-section as the sample location. While all efforts are taken to carefully project the data, as the reviewer points out there are spatial variations in the sample locations. The ± 0.5 km lateral error allows us to recognize this spatial variability and any uncertainties introduced by projecting the samples.

Line 163. For the MAr ages, are the ages used as input in the models plateau ages, integrated or isochron ages?

The thermal modeling in Pecube is forward modelling that uses the input displacement grids from Move to calculate an evolving thermal field. The measured cooling ages are not used as an input in Pecube, but are compared to the output predicted cooling ages. We have clarified this on lines 178-179. Predicted MAr ages are required to overlap with the measured MAr age and 2σ error. The fits/misfits of the predicted to the measured data informed the changes to the geometry and kinematics tested in the subsequent models that led to Model C.

Line 238. What is considered a best-fit variability velocity and how is it determined? Please add more details for readers to follow.

Clarified on lines 265-266.

Line 433. Please replace dates with ages for consistency.

Updated on line 448.

Line 449. Please specify U-Th or Th-Pb ages reported for monazite by Braden et al., 2020.

Clarified on lines 566-567.

Line 472. A sentence about why the conclusions of this work is scientifically relevant can be valuable for the general audience to understand the novelty and importance of this work.

We appreciate the suggestion. We have added three new sentences to the start of the conclusions that state why this work is scientifically relevant to a general audience (lines 592-595).

Figure comments

Figure 1. Please remove coordinates so that it's not repeated. There is a lot of information in the figure and it adds to visual load. Maybe increase the size or bold the outline of the new sample locations so it's immediately visible. It was hard to spot which of the samples are new datapoints.

The repeated coordinates were removed from Fig. 1. The size of the new data points is slightly increased and the width of the outline thickened to improve the visibility.

Figure 3. Please add briefly in the label or add citations (TopoToolbox etc) to show how the swath profile, mean, max/min and median K_{sn} were plotted in panel 3b.

The figure caption has been updated to include the citation for TopoToolbox (line 109).

Figure 4. The Fit % within the panels are different from Total Fit%. Please describe why they are different or have brief additional information in the figure label.

The Fit %s within the panels represent the fit of the modelled to measured data for the specified dataset. The figures have been updated to indicate “Basin Fit” and “Age Fit” in addition to the “Total Model Fit”. The total model fit is calculated from the fits to the basin and cooling ages.

Figure 5. Same comment as for Figure 4.

The figure annotation has been updated to specify “Basin Fit” and “Age Fit” for clarity.

Figure 6. Regarding the thermal field, as deformation propagates toward the frontal thrust belt, what is the assumed effect on the geothermal gradient, or is there any? The modeling appears to treat heat flow from the basal fault as the primary thermal control, with deformation propagation on individual having minimal influence. Is there model sensitive to geothermal gradient variations in the individual sheets?

One of the benefits of this modelling approach is that we do not have to make any assumptions regarding the geothermal gradient, beyond the parameters applied to the crust in the thermal model listed in Table 2 and detailed in Supplementary Text S3. The thermal field in each panel of Fig. 6 is the cumulative result of each ~10 km increment of displacement and the resulting advection of the thermal field. With no deformation in panel a, the geothermal gradient is unperturbed and the isotherms mimic the shape of the topography. As our models demonstrate, there is a focusing of exhumation above the active ramp with the ongoing displacement, as rocks are uplifted and eroded. This focused exhumation drives advection of heat and produces an elevated geothermal gradient above the active ramp, as time is needed for this heat to diffuse to

the surrounding rocks and for the thermal field to return to an unperturbed state/a standard geothermal gradient. The sensitivity of the thermal field to individual thrust sheets is dependent on the size of the ramp, the magnitude of displacement over the ramp, and the velocity. Fast motion over a very large ramp can produce a very steep or even overturned geothermal gradient in the hanging wall, while rocks in the footwall maintain a lower geothermal gradient. While there is minimal heat flow from the base of the model, most of the heat is due to radiogenic heat production within the crust which is given as a value at the surface and decays with depth to a background value at the e-folding depth. There is fault heating applied in the model. Additional details for the thermal modeling process and the parameters applied to the crust are included in the Supplementary Materials.

Figure 8. How is the exhumation rate calculated? Please add brief details for the reader either in text or label.

The methods for calculating the exhumation rates have been added to lines 508-511.

Table 1. What is the reasoning for averaging a mean cooling age for samples with inter-grain variability (eg., 98)? These can be geologically meaningless age.

The grain ages presented in Table 1 are averaged with an error that encompasses the analytical error and the grain distributions. We agree that care must be taken when evaluating individual grain ages and averaging their distributions. We note that in many locations where the data themselves have larger error bars due to the averaging of ages, that there is also a pronounced spread in age amongst samples showing a true uncertainty in any given age (e.g. 105 km to 65 km south of the MCT). As such, we can still use the large error bar data to understand the broad, first order trends of cooling. The models are also sufficiently constrained by additional datasets, so we are not only relying on the large error bar ages.