REVIEW #1

In the manuscript, titled "U and Th zonation in apatite observed by synchrotron X–ray fluorescence tomography and implications for the (U–Th)/He system", the authors describe a clear and well-documented novel method to investigate (U-Th)/He thermochronology. Great value is found in this manuscript, as the authors do a great job of making clear every step of the process they followed, in a way that allows the reader to reproduce it. Furthermore, the proposed method makes use of non-destructive analytical methods, which pose a great complementarity to the currently existing mainly destructive methods. The obtained results are fairly similar to the results obtained using the previous benchmark methods, and the authors discuss what and how these differences are caused.

All in all, this manuscript is very nice to read, the science within is sound and the story is well-supported by data. There are, however, some parts that could do with some additional elaboration. Please find my comments and suggestions below:

Abstract: in the first sentence, the "(…) in whole apatite crystals at the scale of one µm³." is found to be misleading. Perhaps it is better to specify that this is the resolution scale, instead of the crystal size scale.

We thank the reviewer for this clarifying comment, and will modify the text to improve clarity.

P7, line 168: The authors state the focal spot size of the beam was 1×2µm² (V×H). Can the authors provide a measure of the beam's divergency, as this will affect the voxel size of the tomographic reconstruction, particularly at the edges of the sample.

In the horizontal direction, the beam converges from 300 microns to 2 microns from a focusing mirror that is 200 mm away from the sample position. The beam will grow to 3 microns about 750 microns away from the ideal focal position. The vertical is a little better: although the size is 1.5 microns, the mirror is 400 mm away.   So, the beam is collimated enough that the beam size will not double in size over the depth of the sample.

P7, line 180: please provide the density of apatite that was used for these calculations.

We used 3.2 g/cm3, and will add this detail to the text. See https://xraydb.xrayabsorption.org/atten/Ca5P3O12F/3.20/0.100/3000.0/21000.0/10 for an example.

P10, line 207: "to determine an overall scaling factor between abundance for each element of interest and the full intensity of its fluorescence. These elemental scaling factors are then calibrated to give reasonable total abundances and used to produce sinograms of abundances": This statement is rather vague and unclear.

We thank the reviewer for this clarifying comment, and we agree that adding more detail about XRF data processing would strengthen the manuscript. Rather than adding this method detail to the main text of this section, we suggest adding an appendix to the manuscript that would explain calculation of the elemental scaling factor in more detail.

What do the authors mean by 'the abundance for each element of interest'? Is this the average abundance within the full apatite crystal, or some locally determined abundance? How did the authors determine either? This should be stated in the manuscript.

We thank the reviewer for this clarifying comment, and add clarifying explanation to the text to explain that we use "abundance" to mean the concentration (values in ppm) for each element of interest. That is locally determined from the analysis XRF spectra that uses one calibration scaling parameter.

In a similar way, it is unclear whether the approached method can indeed be followed to generate a sinogram of 'abundances', as of course every pixel in a sinogram in fact represents the total yield of a line through the sample (at a specific sample orientation).

Yes, this is how computed tomography works. Every pixel in a sinogram has an intensity that is a line integral through the sample. By measuring multiple angles, the intensity for each pixel can be determined. As above, we converted sinograms of XRF spectra to sinograms of abundance. Using computed tomography methods, we then converted these to virtual slices and volumes of abundance.

For instance, reconstructing said sinogram will not result in a virtual cross-section displaying the spatially resolved abundancies/concentrations (at least not without further scaling of said cross-section, using pre-existing data). Given how careful the authors are in describing the used methodologies, it would be opportune to elaborate on this (important) step as well.

Reconstructions of sinograms do result in a virtual cross-section with spatially resolved abundancies/concentrations. As explained in the text, this does require a scaling factor from "weight of the elemental abundances" that is selected to give reasonable total abundances (in ppm) from the spectra of total grains. Again, we emphasize that these absolute values are best viewed as orders of magnitudes, while ratios of abundances of different elements will be more accurate, and variations of abundances for each element even more accurate.

Later in the publication mention is made of ppm quantified results. Unless the above quoted sentence entails this, please provide (additional) information on how the quantification procedure was performed. In addition, a method validation (or reference thereof) would be greatly appreciated.

Yes the previous quoted sentence does entail concentrations in ppm, as explained in the response to earlier comment from this reviewer.

P10, line212: The authors state that XRF-CT attenuation correction is difficult and error prone. Although I agree with the former, and am sceptical about the latter statement, it may be opportune to cite here a few recent manuscripts that pertain this attenuation correcting, for readers who might be inclined to attempt to do so. Additionally, some methods exist that seemingly do not impose much errors in the final reconstruction, are fairly straightforward to use and do not require any chemical information input by making use of a neural network. These methods may be worthwhile to investigate.

We thank the review for this comment, and we will change the language from 'difficult and error prone' to instead read "difficult to do with sufficient accuracy"

Because the absorption by these samples is dominated by the host apatite mineral, and not the trace metals, we could make a correction for each "X" line in each sinogram, assuming we can identify "the edge" of the apatite mineral, and assuming that the spatial heterogeneity of the trace elements does not significantly affect the self-absorption. When we have tried such processes, we find that it does not greatly improve the sinograms and does add some uncertainties.

The fluorescence energies for the elements of the interest are relatively high energy and have a fairly narrow range (ranging from 12.97 keV for Th La1 to 15.0 keV for Y Ka1), As shown, the attenuation by the sample will vary from 0 for X-rays emitted from the surface to a factor of 2 for X-rays emitted from the center of the sample. Thus, the fluorescence sinograms and the derived abundance

sinograms do not show significant self-absorption. In contrast, sinograms for Ca Ka (3.7 keV), and even Fe Ka (6.4 keV) would be severely attenuated.

There are some published works on correcting self-absorption in X-ray fluorescence tomography, especially for lower-energy lines. Since we do not use any such method, we do not cite these works.

P 10, line 220: It could be noteworthy to mention that in principle U and Th signals can also be detected by their K-line excitation (in contrast to the L-line excitation used here), which corresponds to significantly higher energies (>90keV) and thus is virtually non-affected by attenuation effects in these samples. The drawback is of course that one has to obtain measurement time at high energy beamlines that allow microscopic focussing.

The K-edge of U is 115 keV and its main emission line, U Ka is at 98.4 keV. There are very few X-ray fluorescence beamlines that work at these energies. The beamline used does not work above 30 keV. The silicon detectors used in this study are more than 50% efficient for U La, but below 2% efficient for U Ka1 X-rays. A Ge detector would be significantly better at 13% efficient. If an X-ray energies above 115 keV was used to illuminate the sample, seeing Sr Ka and Y Ka would be very difficult, as these elements would have extremely low absorption cross-section.

It is true that self-absorption effects would be virtually non-existent for U Ka, these other effects would greatly compromise the data quality. And, to re-emphasize: self-absorption in apatite crystals that are a few hundred microns in size is not a significant effect for U La1 or any of the other elements of interest in this study.

P11, line 244: "Because strontium follows calcium and does not have an issue with self-absorption": It is more prudent to state that Sr has less of an issue with self-absorption, as also Sr characteristic fluorescence photons will be absorbed within the apatite matrix (~43% in 100μm apatite, per your calculations stated at p7, line 181). Additionally, would it not be possible to use Ca in any case, as the outer shell of the Ca should be visible/reconstructed and thus a total volume can be defined (assuming that there are no cavities within the crystal, perhaps this is what the authors fear/expect?).

We thank the reviewer for this comment and will change the language to be clearer, stating that strontium is... "minimally impacted by self-absorption"

yes, Ca can be reconstructed. It beautifully illustrates self-absorption, but since we do not use it in the analysis we have not shown that reconstruction.

In any case, an intricate issue with such approaches is always where/how the difference between background and sample is determined. In the images provided in the manuscript, the background is cut away (white). Which threshold value did the authors set? How was this determined?

The XRF fitting method does (try to) account for backgrounds in the XRF spectra, which comes largely from the tails of the elastic and Compton-scattering peaks. These scattering components are included in the set of eigenvectors used to linearly decompose the sinograms.

As a result, the sinograms of elemental abundances do fall off dramatically (to 1% of average intensity) within a few microns of the edges of the crystal.

The image reconstructtion algorithms used do automatically impose some constraints on the spatial extent of the data, but no additional constraint to spatially identify the edges of the crystal are imposed. The intensities for the resulting virtual slices also drop off dramatically (again, to a few percent of the average intensity) within a few pixels of the edges of the crystal.

Several automatic algorithms were coupled to manual checking and modification procedures to delineate the physical grain edge in 3D slicer as explained in the text section 2.6.

P12, line265: please explain the acronym SA/V equivalence

We will modify the text here to not use an acronym for surface area to volume ratio.

P13, line285: In the process, carefully described by the authors, usage is made of mean alpha stopping distances. At a later stage (line 295) these mean distances then essentially have to be rounded to the nearest integer value.

This rounding does not have much of a biasing effect at all because a very small percent of total voxels are bisected by the segmentation boundary of the crystal. The voxels that are cut by the crystal edge can be expected to be roughly symmetrically distributed, resulting in an equal number of roundings in both

directions, and therefore not having much of an effect on the overall FT value. At lower spatial resolutions, this would become more of a problem.

However, one could argue that using the mean alpha stopping distance is already an oversimplification of the physics that occur? Perhaps it would be more accurate to provide a sigma-deviation on this stopping distance, probed by another random number for each of the randomly oriented alpha decay vectors? Did the authors consider this, and is there a grounded reason why it was elected to forego this?

In this case using mean stopping distances is not an oversimplification. The mean values are the averages of the well known stopping distances of each of the alpha decays in the three decay chaings (8 in U238 chain, 7 in U235 chain, and 6 in Th232 chain). These average values are the same used as standard practice for (U-Th)/He chronometry widely and is sufficiently accurate to correctly predict the FT values in this study.

P14, Figure6: It would be interesting to provide the reader an estimate of the calculation time (e.g. s/CPU) required to perform these calculations. Additionally, the time required to complete these calculations will depend severely on the size of the investigated grains. Can the authors provide some insights on this size-dependency? Is this calculation time a bottleneck in the described method?

The non-linear fit needed for an XRF spectrum can take 10s of seconds to a few minutes with a single processor of a typical modern workstation or laptop. While a single fit cannot be readily parallelized, fits can be parallelized across processors. As described above, we perform a few XRF fits per grain to check that the fitting model is well-behaved and the results sensible. The decomposition of XRF sinogram maps for a single slice through the sample from these fit results also take 10s of seconds. The iterative reconstruction methods take a few seconds per sinogram, which we did for each of the abundances found for each element of interest. Thus, the computation requirements are not too onerous, but many of these steps are best done initially with "partial supervision" using an interactive GUI application. All the needed analysis steps can also be fully scripted (using Python and the Larch package), so that the analysis for a grain can be set up and run in 10s of minutes to a few hours on a typical workstation or laptop.

P14, line310: How do the authors explain that the outer dimensions of the (alpha decay) 3D array is >30μm bigger than the crystal in all dimensions, while the mean alpha stopping distance is (depending on the isotope) at most 22.25μm? Since the authors do not appear to use a probabilistic deviation on this stopping distance, one

could imagine that the alpha decay volume can maximally extend 22/23 (depending on rounding perhaps) µm beyond the outermost apatite grain voxels?

The size of the alpha accumulation array needs to be sufficiently larger than the crystal dimensions so it has space to capture the ejected alphas from decays near the edge of the crystal. The array starts entirely full of zeros and accumulates alphas only in those voxels where decays terminate, so having it extra large has no effect on the calculation.

P18, Figure8: is there a reason why the intensity scale maximum for the Sr image has been set to a relatively low threshold, thus rendering the entire image (at least apart from the Zr inclusion) yellow? A more appropriate scale may be selected, providing a more convincing (I expect noisy or heavily fluctuating in intensity?) image, that would still support the same conclusion i.e. that a Zr inclusion is present. Alternatively, a square root or log scale could be applied to the intensities for the Sr image, thus providing a more general overview of the collected data.

We chose the color scale for Sr in figure 8 to maximize the visualization of the zircon inclusion in the virtual slice. For more clear visualization of Sr distribution in apatite see figure 5.

General:

- The authors specify quite regularly throughout the manuscript that a 1µm^3 resolution/voxel size was attained. This clearly results in a good spatial resolution, but as seen from figure 2B also a 2µm^2 resolution provides quite clear spatially resolved information. This makes one wonder how the spatial resolution impacts the final results in terms of F_T Would a 5µm resolution scan provide similar results, with the added advantage of a significantly higher sample throughput? A short paragraph discussing this would be of great value to the manuscript.

  We are very interested in pursuing further experimentation to explore how the utility of this application can be maximized by increasing throughput of samples. This does include lowering resolution, but doing so comes along with other complications and uncertainties, as discussed in the text (e.g. FT uncertainties, issues at grain edge, imaging small inclusions and highly complex zonation, etc.). We discuss this in section 3.6 and have added some text about tuning tomographic parameters to maximize utility, but at this point we have not completed enough experiments to know exactly how increasing to 5um resolution would impact results overall.