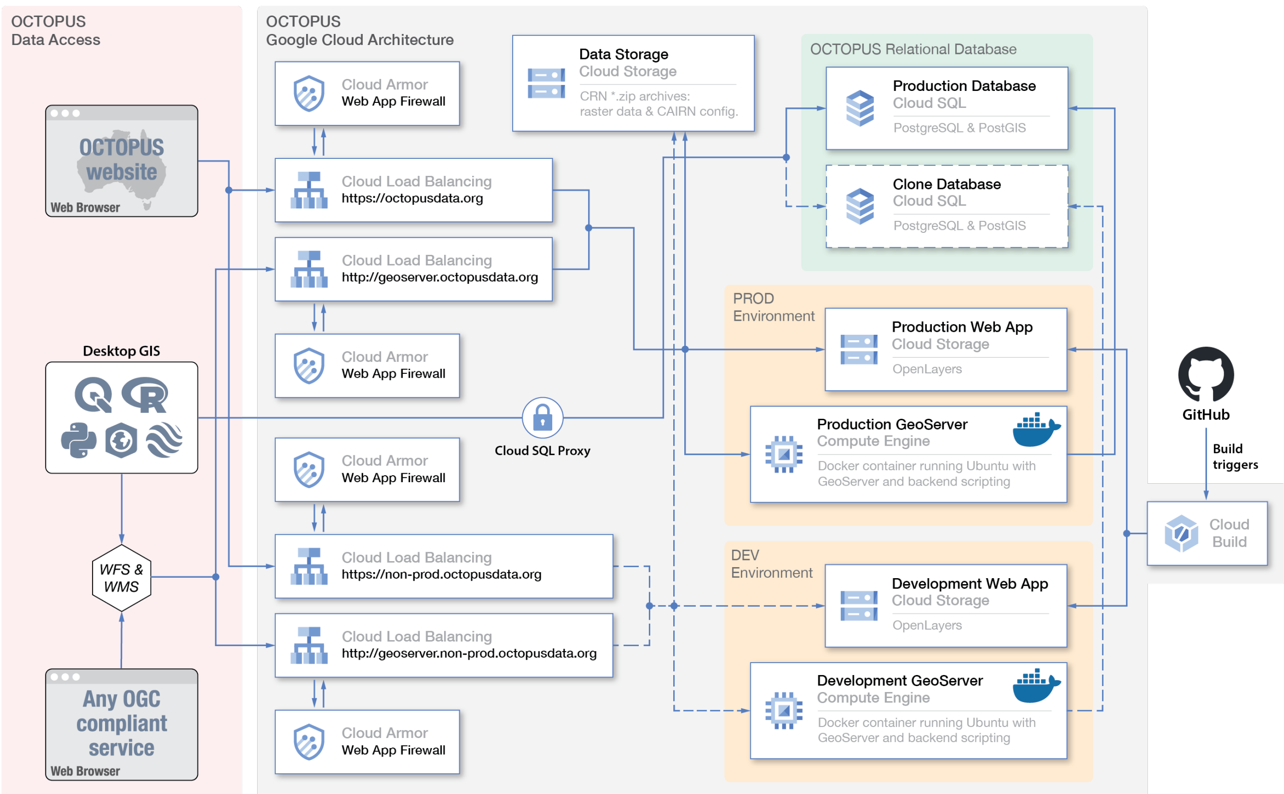
Preprint gchron-2024-28
Short communication: Updated CRN Denudation datasets in OCTOPUS v2.3

We wish to thank the reviewer for the detailed comments. We provide answers to each point below and, as with the first reviewer, we are happy to answer follow-up questions that may emerge from our responses.

RC – Reviewer comment
AR – Author response

RC13	*First, some general comments on reviewing papers about software and data sets. Basically, this should focus on whether the paper correctly describes what the software/data does/is. It's not really fair for reviewers to demand additional software features or tell the author to process the data in some totally different way. Unfortunately, this means that the most insanely maddening aspect of the OCTOPUS website -- the fact that I can't just click on a sample name displayed on the webpage and get a simple listing of the data associated with that sample -- is not in bounds for this review. Really, I think this is completely nuts -- it is bonkers to have to download a large data set, or hack the Geoserver XML responses, just to look at the data for an individual sample. And it is such a simple piece of functionality to add live links to data pages to the sample name popups that I am utterly mystified as to why this isn't done. However, none of this is allowable in a paper review, and I am obligated to stick to the basic function of evaluating whether the paper correctly describes the data set.*
AR13	Thank you for this acknowledgement. Indeed, it is frustrating when reviewers or commentators come with a shopping list of items that they would like to see added. The OCTOPUS system architecture is complex (see below) and adding new data or making changes publicly available is more involved than simply updating a spreadsheet.



	Our long-term goal is to democratise the OCTOPUS project and so we are more than happy to invite the reviewer (and any other member of the scientific community) to join the OCTOPUS GitHub repository (https://github.com/octopus-db) and contribute with code. For security reasons the OCTOPUS code base is currently sitting in a private repository.
RC14	*1. The relationship of 'OCTOPUS' and the various OSL, palynology, and C-14-related data sets that are briefly touched on in the introduction is unclear. It is clear that the interests of the author lie primarily in the area of erosion-rate applications of cosmogenic-nuclide data, which means that in the context of this paper the OSL and other data sets appear basically an afterthought. In addition, the title of the paper is 'updated cosmogenic-nuclide data' and not 'updated OSL data', but the discussion around line 30 indicates that there are some updates to the OSL data as well. As cosmogenic-nuclides-in-detrital-sediment data, OSL, and paleoecology data are really not very similar, and in many ways the challenges of storing OSL data in an organized way are much greater, my suggestion is for the author to just write papers about these things separately -- cover only the cosmogenic-nuclide data in this paper and then write a different paper in which the other data sets can be discussed in useful detail.*
AR14	Our aim in the introduction section is to provide a brief history of OCTOPUS releases. We mention the updated OSL data and the other collections for completeness and clarity and point the reader to papers or manuscripts that describe these collections in detail: • SahulArch: Saktura et al. 2023 https://doi.org/10.1080/03122417.2022.2159751 • SahulChar: Rehn et al. 2024 https://doi.org/10.5194/essd-2024-328 • IPPD: Herbert et al. 2024 https://doi.org/10.5194/cp-2024-44 A manuscript focusing on the OSL and TL data in SahulSed is in preparation. This *Geochronology* manuscript focuses on the updated CRN data and so we are already doing what the reviewer is recommending.
RC15	*2. What are 'partner' datasets (around line 23), and why were these two specific data sets selected from the much larger universe of geochronology databases? Why not just include anything with a geodata type feed (ICE-D, Earthchem, USGS Geochron)? Is this, like, an endorsement deal? Did money change hands?*
AR15	Partner collections are data that we have agreed to host on the OCTOPUS platform but have not committed to maintaining. We are more than happy to include other collections, and for example ICE-D would be a nice addition and perhaps we can have a conversation on how best to do this. USGS Geochron is an awesome resource – we were not aware of this and wish to thank the reviewer for pointing it out to us. There is a similar effort led by AuScope (https://www.auscope.org.au/ausgeochem) and we had some preliminary conversation about making AuScope Geochem and OCTOPUS work together. At the end of the day, however, we have finite temporal and financial resources and cannot do everything without help from the community.
RC16	*3. The relationship of 'collections' to data sets is very ambiguous. In one context (line 30-ish), a data type (e.g., OSL or paleoecology) is a 'collection,' whereas later discussion (line 45, 55 areas) then reveals that a single data type is composed of multiple 'collections.' Why is this?*

	This confusing nomenclature is inherited from the earlier versions of OCTOPUS, where it was also confusing, and it would be helpful if it was explained more clearly here.
AR16	We will make changes to the text in order to improve clarity.
RC17	*The discussion around line 68 of why there are some studies that are not included in the database is mystifying, and somewhat concerning. The whole point of developing a centralized online database is that the 'return on investment' (line 69) is very high no matter how few data points there are in a study, because you only have to read the paper and ingest the data once, and then you are done. Also, of course, the computational effort scales with the number of samples, so studies with fewer data are actually less of an investment. Thus, this argument is not at all compelling; in fact, at face value it seems kind of bizarre. Furthermore, this discussion gives the idea that data sets with fewer data are deemed to be of lesser quality, which of course is a terrible approach from the science perspective. From the perspective of using the data for actual Earth science (rather than, e.g., bibliometrics) applications, how the data are grouped into 'studies' is totally arbitrary and irrelevant, so data selection should not be based on this grouping. In my view this section of the paper reveals a significant weakness of the database implementation. As this is written, it's also kind of a weird threat: include more than ten data points in your paper or else it will not be included in the database. As inclusion of data in databases of this sort is a significant element in subsequent data discoverability and reuse, this is a terrible message to send not only from the scientific perspective, but also from the perspective of outreach and student/early career development. Honestly, if I were the funding agency I would squelch this approach immediately.*
AR17	There are a number of misconceptions in this comment that we unpack below: • To clarify, the funding that we have received from the Centre of Excellence for Australian Biodiversity and Heritage, and that we acknowledge in the manuscript, was to cover migration to Google Cloud, database running costs (i.e., monthly fees for Google Cloud hosting) and the development of SahulSed, SahulArch, SahulChar and IPPD. In 2016 we obtained funding to develop the CRN collections that were released as part of OCTOPUS v.1. Since then maintaining the CRN collections has been a voluntary effort by the two authors of this manuscript with no funding available for this purpose. In this context '*return on investment*' is referring to the time we volunteer to maintaining the database. • The computational effort does not always scale with the number of samples. In fact, for the reasons described next, papers with fewer data often need more effort to ingest. Recalculating basin-wide denudation rates involves a lot of detective work to do with identifying and delineating the basins from where samples were collected. This works best when a publication (1) includes sample coordinates with sufficient precision, a detailed map identifying each sampled basin, and information such as basin areas and/or sample site elevations, and (2) the different pieces of information match. In our experience, most studies (although there are exceptions) with small number of CRN data points (n ~ 3 or 4) do not allocate a lot of space to documenting this data in sufficient detail for reproducibility. We do our best to ingest all available data, and contact the corresponding authors for help, but this is often futile. As we mention in our manuscript, we have identified about 47 studies that we may never be able to incorporate due to lack of information.

	• Grouping data into studies offers a practical way of organising the large amounts of raster data and also the CAIRN input/output files that we include in the database. The actual tabular data stored in the relational database is, of course, seamless and the link to studies is achieved using a unique ID (STUDYID). The raster layers come in various spatial resolutions and are projected to different UTM zones depending on the size and location of each study area. Therefore grouping data into studies allows for these raster layers to be containerised and served for download. We discuss this point in previous papers describing OCTOPUS.
RC18	*5. The discussion about how data sets from cratonic areas are sparse should probably make note of the fact that river systems in these areas are largely depositional rather than erosional systems, in which case cosmogenic-nuclide erosion rate measurements actually don't work. Really what is wanted here is an assessment of how representative these measurements are of the area in which the measurements could be made, not the area in which the measurements could not be made.*
AR18	We see the point of the reviewer here. However, all we are trying to do in the offending sentence (i.e., "*However, data from low-gradient, tectonically passive regions remain sparse, particularly in Africa*") is to point out the lack of data from the African continent.
RC19	*6. The discussion in line 105-ish should be more clear about the fact that the mean atmospheric pressure is not derived from an atmosphere model (as one might reasonably expect) but by inverting the scaling model. That is, one should not assume that the mean atmospheric pressure in this field has any meteorological significance. It sort of says this, but this point should be made clear.*
AR19	This is a good point. We will change the last sentence of the paragraph starting at line 105 to make this point more clear.
RC20	*7. Near line 118, later on this page, and in Figure 3, I don't understand why calculations are being made at the location of the basin outlet. If you are able to calculate the mean elevation, then you obviously know where the basin is, and can also calculate the centroid latitude. In what circumstance would you ever care about the location of the basin outlet, or want to use that in a calculation?*
AR20	This is also a good point. Our aim here was to see how bad things can get – but we acknowledge that this is probably an unlikely scenario and most (if not all) users will calculate basin centroid coordinates and use those.
RC21	*8. Line 120-123 is slightly misleading. These differences are not because the calculation methods (Balco 2008 vs. CAIRN) are different, they're because the mean basin elevation is different from the effective elevation. This should be clarified.*
AR21	Good point, again. We will clarify this in the text.
RC22	*9. The section at the bottom of p. 6 (lines 135-140 area) is a little bit incoherent, because it is not clear what each comparison is supposed to test. It would be helpful to rewrite this to make clear what assumption is being tested with each comparison, e.g., something like, 'To quantify the effect of using the mean elevation vs. the effective elevation, we did X and here*

	are the results. To quantify the differences between CAIRN and Balco 2008 with the same input parameters, we did Y and here are the results.' You get the idea.
AR22	Indeed, the last paragraph on page 6 is somewhat incoherent. We will change the text to better explain what we are trying to achieve here.
RC23	*10. I agree with the other review that the paper should make clear that the glacier cover fraction and quartz-occurrence-inferred-from-lithology fields are suitable for general guidance, but probably not very quantitative.*
AR23	We feel that we are already doing an adequate job here. We mention on lines 153-155 and 160-162 that our GLIMS corrected rates are end-member maximum values and are meant for providing an insight "*into the potential influence of glaciers on the overestimation of ^{10}Be-derived denudation rates*". Regarding quartz occurrence we provide ample caution including this sentence at lines 181-183: "*To reiterate, for most basins, these data are too coarse to enable precise corrections to the recalculated denudation rates; however, they may still offer valuable insights and help identify basins where such corrections could be warranted.*"
RC24	*11. The y-axis in Fig 5 is labeled such that it looks like GLIMS is being subtracted from CAIRN, which doesn't make any sense. Also, if we assume that it is really two erosion rates that are being subtracted, then the units should be m/Ma, not percent. This needs to be labeled in a way that conforms with the units (something like 100 * (E_glaciers - E_noglaciers)/E_noglaciers ?).*
AR24	The y-axes in Figs. 3 and 5 and both x- and y-axes in Fig. 4 represent percent difference between two denudation rate values, calculated as: [(D1-D2)/mean(D1,D2)] x 100. In the case of Fig. 5, spelling this out would make the label: [(E_glaciers – E_nonglaciers) / mean(E_glaciers, E_nonglaciers)] * 100 The above would be too long to fit. As a compromise we went with CAIRN – GLIMS [%] and we explain in the figure caption what this actually means. We could substitute with 'Percent difference [%]' in all figures but then we lose information on what the sign (negative or positive) means.
RC25	*12. Again following discussion of quartz-fraction estimate in other review, the discussion in line 180-ish is not helpful without some idea of which lithologies are and are not quartz-bearing. Perhaps the easiest way to handle this would be just to have a supplemental table indicating which GLiM classifications are and are not considered to have quartz.*
AR25	Good suggestion. We will include a table indicating the GLiM classification.
RC26	*13. In line 190-ish, the discussion here basically says that the topographic shielding calculations are wrong, but we did them anyway, which is rather odd. Again, the computational-resources argument is not super compelling, but on the other hand this*

	paragraph does appear to correctly describe what has been done, so even if what was done is questionable, this section passes the test of whether it is accurately described.
AR26	Strictly speaking (1) calculating topographic shielding by not accounting for the change in shielding with depth – what is done in OCTOPUS, and (2) avoiding topographic shielding corrections altogether, including in steep basins where quartz distribution and / or denudation rates are not uniform, are both incorrect. What we are saying in AR11 is that (1) the effect of calculating or ignoring shielding is trivial (~99% of the data have differences between shielding and no-shielding that are below ~6%, and below the median external uncertainty on the calculated denudation rates), and (2) this does not warrant spending two months on recalculating everything with CAIRN. Furthermore, users are provided with the means of doing the recalculations themselves if they wish.
RC27	*14. Around line 205, there needs to be more explanation of whether basins with internal drainage issues are (i) not recorded at all in the database (which in my view is bad practice) or (ii) recorded in the database with Be-10 concentrations, etc., but with no associated calculated erosion rate (preferable). Also, frankly, I don't really understand what the problem is here, because Figure 6 makes it clear that you do know what the actual correct drainage basin boundary is in lat/long coordinates - why is it possible to project the black lines into UTM but somehow impossible to project the blue lines? Also, of course, for pixel-based production rate calculations, you don't actually have to project out of lat/long - you can just weight by the actual area of each cell in real units. But that would probably require a redesign of the whole thing.*
AR27	We only exclude basins (and data) where we are unable to reproduce with confidence the drainage basin as described in the source publication, and we have convinced ourselves (for example based on satellite imagery, etc) that our inability to reproduce the basin is not due to mistakes in the source publication. The number of such basins is small (n < 20). The issue described in the final paragraph of Section 5 (lines 197 – 205) describes a limitation of CAIRN: it delineates the basins from a DEM that needs to be projects in one of the WGS84 UTM zones. One cannot provide basin outlines as a vector file to aid in basin identification (although one could clip a DEM using basin outlines and this provides a solution in some but not all cases). All that CAIRN takes as input is the DEM and the sample locations. It then delineates basins etc by snapping the sample locations to the nearest channels – as defined using stream order and upstream contributing area thresholds. Fig. 6 shows drainage basins **as derived** from two different DEMS (one hydrologically enforced and the other after the former was projected to UTM). The basin boundaries are not the ones being projected as the reviewer seems to believe in RC27. Following reprojection of the DEM, CAIRN is using its own sink filling and flow routing algorithm to derive a hydrologically connected drainage net that results in internally drained areas being connected to the main drainage. Using the blue lines in Fig.6 to clip the DEM and remove internally drained cells prior to reprojecting may result in CAIRN not being able to produce a drainage net that connects all upstream areas to the basin outlet.