Solid Earth Discuss.,
https://doi.org/10.5194/se-2020-104-RC1, 2020

Solid Earth
Discussions

Interactive comment on "A systems-based approach to parameterise seismic hazard in regions with little historical or instrumental seismicity: The South Malawi Active Fault Database" *by* Jack N. Williams et al.

Richard Styron (Referee)

richard.h.styron@gmail.com

Received and published: 19 August 2020

The paper by Williams et al. provides a high-quality map of active faults in southern Malawi, and presents a clever method for partitioning regional deformation rates onto the rift structures, with a thorough exploration of the uncertainties.

The work (the mapping, rate estimation, and manuscript) are executed competently, and there is nothing that is strictly incorrect, although some topics would benefit from a bit more explanation if not revision. The mapping is quite high quality, and is the most

solid contribution made here.

Although the authors may not want to do this, I think that the work could benefit from being split into two different, shorter papers: one that presents the fault mapping and discusses it in a bit more detail and context (although not too much more), and another that presents the parameter estimates, ideally as a part of a PSHA. (Note that I am not asking that this be done for major revisions–it's just something to consider doing.)

Major issues:

Separation of data and estimates

The first issue is that there is no apparent separation in the fault data and attributes between observation (and associated interpretation firmly based in observations) and rough estimation based on little to no data. Although the authors are helpfully conforming to the schema laid out by the GEM Faulted Earth (GFE) project (e.g. Christophersen et al. 2015), I believe that the GFE is intended to hold observations or measurements, rather than the estimates made in this project. For example, recurrence intervals would be derived from paleoseismology rather than calculated from the slip rate and assumed magnitudes. Only considering measured recurrence intervals makes the recurrence intervals independent of assumptions of earthquake magnitude, scaling relations, or other factors. Christophersen et al (2015) state that the database is often sparse where observations don't exist.

The coupling of the fault traces (which are observations or data, as far as I am concerned, even if there is an interpretive component) with the parameter estimations crosses a traditional boundary. Typically, the fault data are considered to be somewhat objective and immutable (though subject to revision) while the derivation of earthquake rates is considered to be part of the modeling process and often done while making fault sources for a PSHA. Model results are a bit more subjective and mutable, as they are dependent on assumptions that are explored during the modeling process and/or are project-dependent. A good modeler will have a process for testing and re-

fining some of these assumptions (i.e. magnitude scaling relations or slip partitioning) through comparisons with instrumental seismicity or other observations. However the observed data are usually not revised to improve a data-model fit. A user who wants to incorporate this dataset into a seismic hazard model, but who may not agree with some of the model assumptions used here (i.e. scaling relations, magnitude ranges, or the style of partitioning between internal and border faults) may have a hard time knowing what to keep and what to discard, without reading a long paper. This can be a big challenge for the many seismic hazard modelers who do not have a great facility with the English language. Similarly, if this data were incorporated into other fault databases, the end user may not be able to cleanly separate data from model results.

This does not mean that what the authors have done is wrong or necessarily needs to be changed. It is just to raise their awareness of a potential concern (that these estimates may be confused for observations) and that many hazard modelers would prefer to redo the estimation rather than rely on these results.

I am not sure of the best course of action. If it were me doing the work, I would separate these processes and release both 'only data' and 'data plus estimates' datasets. I would also consider publishing them independently, and perhaps incorporating the rate estimation work into a PSHA rather than going part way as is done here. But there is no 'right' or scientifically optimal decision here, and a lot depends on the particular circumstances of the authors.

Another possibility is to keep the parameter estimation through the slip rate estimation but stop there, which would avoid the problems of choosing a magnitude-frequency distribution, estimating the seismogenic thickness of the crust, etc. In a typical project, these tasks are often done by the hazard modeler rather than the geologist who prepares the fault data up through slip rate estimations.

I can state that as a fault data compiler, I am a bit hesitant to bring any of the estimated parameters into the GEM Global Active Faults Database, as they are too poorly con-

strained and data-limited, and I don't want users to confuse them for measurements.

Estimation of uncertainty:

The second issue is that the logic tree framework used to propagate uncertainties and explore the parameter space is perhaps more complex than it needs to be based on the lack of input data. It is a clever method and there seems to be nothing incorrect in the implementation, but I question the wisdom of using it southern Malawi where there is essentially no data to feed in. The old saying in modeling is "garbage in, garbage out"; in this case it's more like "nothing in, nothing out" (I am not suggesting the work is garbage!). The exercise seems to simply quantify the obvious, that each fault slips somewhere between 0-5 mm/yr. I am not sure that it provides much value. The further work, estimating recurrence intervals, has larger theoretical issues (discussed below) in addition to adding several more layers of uncertainty into the results. It could easily be removed from the database (though perhaps kept in the paper for discussion).

As a subordinate issue, I don't think a logic tree framework is really the most appropriate method of propagating uncertainty as used; it is more appropriate when the parameters that make up the branches in the tree are discrete variables with a few choices, rather than continuous random variables. For example, an appropriate use of logic trees is to consider different scaling relationships. With continuous random variables (i.e., extension rate or dip), the use of unweighted logic trees considers the lower, mid, and high values to all have equal probability. Do the authors consider this to be the case? Do the authors believe that the resulting low, mid and high values are equally probable? Even if the inputs are all equal, if there are no correlations between the different parameters, the middle values should be more probable (see for example the Central Limit Theorem).

In my opinion, a more appropriate method for representing the uncertainty in the results (i.e., slip rates or recurrence intervals) would be to define distributions for each continuous random variable (i.e., dip or total geodetic extension at that latitude) and

then randomly sample from these distributions, and then characterize the resulting distributions for the results parameters. This is a simple Monte Carlo method. The major strength of this method is that the sampling will cover far more of the parameter space than a coarse 'low/med/high' sampling method. It is also quite trivial to introduce distributions for each parameter that may reflect prior knowledge (i.e., a PDF of regional dips based on focal mechansisms or structural measurements).

One way to think of this is that the representation of uncertainty in the model should reflect the real uncertainty of the parameter. Continuous variables should be represented through continuous distributions, while discrete variables (i.e. the choice of scaling relationships) should be represented through discrete distributions (i.e. lists or arrays, perhaps weighted).

The strategy employed here does a good job of defining the absolute range of the results based on the inputs, but a worse job of defining the central values (broadly like the mean and one standard deviation rather than three standard deviations). If the authors believe this is the better choice, that is fine, but I would like to hear their arguments.

The calculations of recurrence rate:

The authors choose to calculate recurrence rates under the assumption that all of the seismic moment that accumulates on each fault is released during earthquakes of identical magnitude. This is essentially the "characteristic earthquake hypothesis" which featured quite prominently in mid-late 20th century paleoseismology and PSHA but was always quite contentious (for example see "Characteristic Earthquake Model, 1884-2011, RIP" by Kagan et al. 2012, Seismological Research Letters). This hypothesis is believed by fewer and fewer scientists with each passing year, as our observations of variable rupture segmentation and per-event displacement accrue. The few remaining national-level PSHA models that still use a 'pure' characteristic earthquake model (not a distribution that includes aleatoric variability) do so primarily because it

simplifies time-dependent hazard analysis.

The modern state of practice is to consider a range of earthquake sizes on each fault, and to distribute moment throughout the range of earthquake sizes by specifying the relative frequencies of different magnitudes of earthquakes, and then calculating the absolute frequencies through moment rate balancing. The canonical reference for this is Youngs and Coppersmith (1985 BSSA), which provides equations for multiple magnitude-frequency distributions. GEM's OpenQuake Engine and OQ Model Building Toolkit also has some Python code for this purpose, if the authors are interested in using or studying a functional implementation (https://github.com/gem/oq-engine/tree/master/openquake/hazardlib/mfd; https://github.com/GEMScienceTools/oq-mbtk/blob/master/openquake/mbt/tools/fault_model

If the authors favor the pure characteristic earthquake hypothesis, then they should provide some supporting arguments. Otherwise they may either calibrate the magnitude-frequency distributions, or simply drop this part of the estimation procedure (even if they retain the estimates up through the slip rate calculations).

Minor issues:

Data license, distribution and updates:

One of the promises of 21st century science is that new technologies enable rapid and low-friction sharing, integrating, and updating of data. However, it raises some new topics that have been heretofore ignored by most. The first is the license of the data. As the creators of a nice dataset, the authors are entitled to specify the terms and conditions under which others may use it. A good "open-data" choice is the Creative Commons Attribution license, which is what the articles published by the EGU/Copernicus journals use.

However, the authors may wish to specify a different license, such as a non-commercial license (meaning that it can't be sold or used for other commercial purposes), a share-

alike license (meaning that any modifications to the data, which are allowed by the Creative Commons licenses, must be redistributed under the same conditions), or various others. There are also more and less restrictive licenses, but these may start to conflict a bit with the release of the data in this journal.

It may sound like a bit of boring lawyer stuff, but it's very important to many of us that deal with others' data regularly. If the authors want the data to be most useful, please explicitly state what the license is, so the potential users can have some clarity about what they can or can't do with it. It's an easy process: just put a 'license.txt' file in the zip with the GIS data.

Similarly, the data will probably see a lot more use if it is easy to get to, and in a place where it's easy for the authors to update. The easiest here is using GitHub (github.com) which has turned into the default small data distribution channel for many, including the GEM Global Active Faults Database. GitHub, or other similar services such as GitLab, provide a great platform for licensing, distributing and updating data, in a way that makes the history of the data transparent to the users by being integrated with a version control system.

Something else to consider is whether the authors would welcome updates or extensions to the mapping (and perhaps parameter estimation). It may be that other users who are interested or have some need for a fault database over a wider area than just that covered in this dataset, and may want to expand along strike. This is the kind of collaborative science that is quite easy to do now, especially with services such as GitHub, but I don't think the academic publication process, and allotment of credit (citations etc.) has caught up. Nevertheless, if the authors support this (in principle, no need to blindly accept changes) they could write a sentence or two in the manuscript or in a text file with the data describing this.

Publication of code to perform parameter estimation:

I think that by default, any code used in a scientific work should be published with the

paper. This would definitely include any code used to perform the parameter estimation (one assumes it wasn't done on a hand calculator). There may be some extenuating circumstances where publication of code isn't a good idea, but this would involve prior intellectual property restrictions or something. I wouldn't consider messy scripts to be exempted here. Detailed inspection of methods and reproducibility is central to the scientific process, and code is perhaps the most perfect form of scientific inquiry that allows for this. Please publish the code, even if it's a messy script of zip file of them. (I also think that EGU/Copernicus asks for this but I could be wrong.)

Line edits:

Line 5 (and throughout manuscript): Superscripts are formatted as subscripts. This is particularly annoying with exponents.

Line 18: All seismically active areas on earth have instrumental records much shorter than the 'repeat times' of the larger earthquake produced in these regions (hundreds to tens of thousands of years).

Line 56: Actually, active fault databases have been developed for close to all seismically active regions; the GEM Global Active Faults database is referenced elsewhere in the paper, which has global coverage. Some areas (like the EARS) need better mapping and slip rate measurements, but active fault data does exist.

Line 79 (and elsewhere): I would be more careful with the suggestions that PSHA based on instrumental seismicity is likely to underestimate seismicity in moderately low strain rate regions. The cited references don't do a good job of backing this assertion up, which is not surprising as many earthquake scientists who are not actively involved in PSHA overestimate their knowledge of it (Stein being a prime offender). The justification that this study will provide better constraints on earthquake rates than PSHA models that incorporate instrumental seismicity (which, when done correctly, is quite capable of dealing with incomplete catalogs) is cringe-inducing in light of the extremely poor constraints on earthquake rates produced in this work.

Line 160: The GEM Global Active Faults Database has now been through peer review, and the citation should be changed to Styron, Richard, and Marco Pagani. "The GEM Global Active Faults Database." Earthquake Spectra, Aug. 2020, doi:10.1177/8755293020944182.

Line 324: Note that the GEM neotectonics database is part of the GEM Faulted Earth project, which ended around 2015, and is quite distinct from the GEM Global Active Faults Database (Styron and Pagani, 2020). Please more explicitly refer to the earlier neotectonics database as the Faulted Earth database for clarity.

Line 325: It is worth noting (but not necessarily changing the fault data or the manuscript) that the hierarchy developed by Christophersen et al (2015) as part of the GFE is a bit contentious and has been abandoned at GEM. The newer Global Active Faults database does not incorporate it, as I decided it was too cumbersome and instead chose a 'flat' system where the 'trace' units in the GFE system would be mapped as a single, continuous trace (in most cases it's somewhat obvious that the traces connect in the bedrock regardless of surface expression, as most faults in these databases have a kilometer or more displacement which can't geologically drop to zero where the traces don't quite join). This simplifies the mapping, drastically reduces the file size of the fault database, and makes for easier hazard modeling as the maximum earthquake can be calculated from the area of a single feature rather than manual joining of multiple features. Many other institutions, such as the USGS, are considering following suit if they have not done so already–the simplicity of the system allows for easier updates and more automated pipelines for incorporating faults into PSHA.

Line 383: The calculations here are another instance of what many would consider to be modeling decisions rather than something incorporated directly into fault databases.

Line 639: This is not in any way a test of the results. The comparison of very broadly estimated rates with data-based estimates for faults hundreds of kilometers away does not meaningfully indicate the validity of the rate estimates here.

Line 651: This is also not a very meaningful comparison. The reasons that the projected date of initiation of the rifting derived from geodetic data (an extrapolation of 1,000,000x) don't match geologic data are manifold to the point where it may not be worth discussing; consider removing this paragraph.

Line 669: Why exactly are only half of the 128 parameter combinations considered in this? How were these 'carefully selected' in a way that is not cherry picking? Computers are pretty fast these days and if this analysis is worth doing (it is interesting) it is worth doing with all of the combinations. Surely it wouldn't take more than a few seconds.

Lines 691-730: I don't think these bits of discussion add anything to the paper, and removing them would improve the focus of the paper.

The digression about fault growth is interesting but not very relevant. The second paragraph has some sloppy scholarship; the 30-60 km long normal faults here are not at all on the long side of normal faults worldwide, as is clearly evident in the GEM Global Active Faults Database which is cited a few times. The Jackson and White reference is very out of date.

The paragraph on seismic risk is important but could be tightened up and placed in the introduction, where it is more appropriate. The next paragraph, comparing the lengths of faults in this database to earthquakes also suffers a bit because it compares a small number of global earthquakes to a local fault database, which isn't a good point of comparison (longer normal faults exist in several orogens and generally have similarly slow slip rates, i.e. the Basin and Range in the US).

Line 758: This paragraph is troubling. It seems to discourage others from attempting to collect real data to use in PSHA, though there is no reason to think that the rough estimates provided in this work are superior to field measurements.

Line 798: The probability distributions listed here describe aleatory variability in recur-

rence, but the topic under discussion is epistemic uncertainty. In this case these are not comparable.

Interactive comment on Solid Earth Discuss., https://doi.org/10.5194/se-2020-104, 2020.