Dear Associate Editor,

Thank you for the reviews and handling of our manuscript “What controls aftershock productivity?”. The main concerns of the reviewers and AE revolved around the definition of the relative productivity of earthquake sequences. We addressed this point by adding clarifications in the methods and demonstrating that an alternative definition reproduces the same major results. We also assessed confounding relationships in the data by analyzing the covariance among our predictors. Finally, we clarified technical aspects of the paper such as including additional explanatory text for the SVM and a fully reproducible trained model in the supplementary material.

We have now included Prof. Thomas Geobel as a co-author. His valuable contribution helped us solidify some of our statistical analyses and rapidly apply an alternative clustering and aftershock counting method.

You will find below detailed responses (italicized) to your remarks and those of the reviewers. Text in quotations indicates additions to the manuscript. In the supplementary, the additions were substantial enough that track changes was not appropriate.

We thank you and the reviewers for careful reading and thoughtful comments that have improved the paper.

Kelian Dascher-Cousineau

Associate Editor (Remarks to Author):

This is an interesting and well-written study that examines the factors affecting earthquake aftershock productivity on the global scale. The aftershock productivity is quantified with the parameter of Eq. (3), called relative productivity. A combination of techniques, including regression analysis and support vector machine approach, is used to establish statistical relation between the relative productivity and multiple attributes, including lithosphere age, plate boundary type, normalized rupture area, aspect ratio, dip, and volume of available brittle lithosphere. The topic of the study is timely and the results are of interest to the seismological community. At the same time, there are several methodological issues that need to be revisited before possible publication of the work.

Most importantly, the current definition of the relative productivity (Eq. 3) may suffer from statistical artifacts. Notice that the variance of the point scatter in Fig. 1 decreases with the main shock magnitude (heteroskedastic regression). This implies that the relative productivity is by definition more stable for large mainshocks and less stable for small mainshocks. Accordingly, the relative productivity might be affected by the spatial distribution of the maximal magnitude. I do not see how this effect is taken into consideration when interpreting the results.

This is an interesting point. We investigated the apparent heteroskedacity by measuring the interquartile range, which is a proxy for variance that is robust to censored data. We found that the interquartile range did not vary systematically with magnitude nor did the median. These findings are reported in a new figure, Fig. S1, and accompanying text. In the main text we added:

L. 298: *"...It is important that relative productivity be independent of magnitude, as we run the risk of confounding variations in the average size of earthquakes with variations in productivity. We show that the median and interquartile range of relative productivity are not magnitude-dependent (Figure 1b and Supplement Figure S1). Note that median and interquartile statistics which we use throughout ensure that we account for earthquakes with no aftershocks which have* $\Delta log(N) = -\inf$*."*

Next, the productivity regression of Fig. 1 does not take into account possible regional effects (regression line may differ in different settings) -- how will this affect the study results?

Variations in the regression could involve: (1) a change in productivity with setting, which is one of the targets of this study or alternatively, (2) a change in the variation of productivity with magnitude, commonly parameterized with the exponent alpha. If we have misfit alpha, we should observe a variation in results with magnitude. To directly address this point we reconstructed our synoptic figure for mainshocks greater than, and less than M_W7.5. We find that the results obtained in the more sparse but better constrained large earthquakes persists to lower magnitude. Naturally, given the larger sample size of smaller earthquakes, more subdivisions of the data by earthquake attribute yield a set of relative productivity measurements that are statistically distinct from the overall population of earthquakes.

We now discuss this point in the manuscript:

L. 491: "Similar results persist from large (M_W>7.5) mainshocks to smaller (6.5<M_W<7.5) mainshocks, indicating that there is likely no significant regional scaling differences that confound our results (Supplement Figure S3-4)"

The number of zero-aftershock events depends on the magnitude cut-off; may this dependence bias the study results?

This is indeed a potential issue with the study that we have addressed in the supplement by reproducing all of our figures with a more conservative estimate of the magnitude cutoff. We have added some clarification in the main manuscript pointing the reader to these supplemental figures:

L. 160: "... We test the sensitivity of our results to the magnitude of completeness, finding consistent results for M_c 4.5 to 5 (Supplement figures S5-15)."

We also added some discussion in the start of the interpretation summarizing the inferences from the supplemental figures:

L. 515: "...Our results are robust to aftershock selection method and catalog completeness, which are the primary two factors that could influence. A higher catalog completeness of M_W5 reproduces similar results to those we presented using a global completeness of M_W4.5, though far more mainshock have no observed aftershocks. When more than half of mainshocks considered in a group are censored, the estimation of a median productivity is obscured. Using an alternative clustering routine, the relationship between relative productivity and crust age is slightly less pronounced but still apparent."

Finally, I concur with comment #3 of Referee #2 that indicate possible problems with the current definition.

See the corresponding comment and new supplementary information documenting the effects of alternative definitions on the results.

In summary, the definition of relative productivity, which is the key technical elements of the study largely affecting its results, needs to be revisited. There exist multiple pitfalls with the current definition that need to be either removed or critically examined.

We have now reproduced all of the results with an alternative aftershock detection method and alternative parameter choices in the relative productivity definition. These results are in an expanded supplement as discussed below and concur with the original conclusions.

The work suggests multiple attributes statistically connected to relative productivity. Is it possible to examine which connections are causal, and which are spurious (confounding)? At least, a critical discussion of this issue is needed.

We now explicitly measure covariance for all of the parameters and present the results in a new figure which is now Figure 12, which shows the Pearson correlation coefficient between all pairs of parameters. We discuss these results with

L. 548: "Some source attributes strongly co-vary with each-other (see Figure 12). Width, length, rupture heterogeneity, aspect ratio, stress drop and dip are particularly correlated. Thus, separating their relative importance is complicated. However, we can distinguish the categories of parameters that appear to be related to relative productivity versus those that do not."

Establishing causality is not possible. All we can do is establish correlations within the data and assess the likelihood that these relationships are spurious (that is, a result of chance). We added some explanation to clarify exactly how we assess how likely our results are to occur by chance.

L. 389: "A potential problem with this analysis of 12 test parameters is that a spurious correlation could arise simply because of the numerous investigated regressions. We explicitly address this issue by investigating the probability of spurious correlations. To do so, we first remove any causal relationship between our 12 predictors and relative productivity by shuffling the aftershock measurements and randomly reassigning each relative productivity measurement to the parameters for a different mainshock. We then regress each parameter with the relative productivity and report the maximum variance reduction of any parameter in this shuffled set. This same routine is repeated 10000 times to generate a probability distribution function of the maximum variance reduction of 12 parameters should there be no causal relationship in the data. We refer to this evaluation of the extreme value for the full group, or family, of parameters as a family-wise test and it serves as a null hypothesis. We determine the percentile of our actual regression results within these random realizations, the family-wise p-value, for comparison (Figure 8a - top axis). The comparison yields the probability of obtaining equally good or better results by chance."

The other comments are of a technical nature. Please, define all statistical procedures used in the work (in particular, KS test and SVM) on the level sufficient for their independent reproduction.

Done - See comments below for specific additions to the text.

Also, the literature background can be strengthened. One paper (on the scaling exponent) is mentioned by referee #1 (Brengman, Clayton MJ, William D. Barnhart, Emma H. Mankin, and Cody N. Miller. "Earthquake-Scaling Relationships from Geodetically Derived Slip Distributions." Bulletin of the Seismological Society of America (2019).) A recent global earthquake cluster study (Zaliapin, I., & Ben-Zion, Y. (2016). A global classification and characterization of earthquake clusters. Geophysical Journal International, 207(1), 608-634.) is similar in spirit to the reviewed work.

We have added these references and discussed both:

For Zaliapin & Ben-Zion (2016)

L. 91 "...Zaliapin & Ben-Zion (2016) find similar geographic patterns in clustering statistics and related them to global heat flow."

and

L. 616: "Previous work has found that clustering statistics correlate well with heat flow with high heat flow reducing aftershock productivity and interpret their findings in the context of a temperature dependent rheology (Ben-zion et al., 2003, Zaliapin et al, 2016). The observations are consistent with fault availability playing an important role. Regions with thin lithosphere have high heat flow and so are predicted to be aftershock poor in both frameworks. However, the correlation of focal mechanism (Figure 9a) with aftershock productivity is more challenging to explain by rheological changes alone. The direct correlation between aftershock productivity and heat flow may be a special case of the influence of fault availability. High heat flow can reduce the number of available faults for aftershocks."

For Brengman et al., 2019:
We have added a new Figure S2 that compares the results using Brengman's scaling for aftershock detection with the results using Wells and Coppersmith (1994) scaling, as

originally implemented. The two scalings give nearly identical results. We point the reader to this discussion in the main text with the statement:

L. 245: We also check the results by using the more recent geodetic-derived scaling relationship of Brengmann et al. (2019). Supplemental Figure S2 demonstrates that the results are nearly the same with 79% of the mainshocks having identical aftershock counts.

Reviewer #1 Evaluations:
Significant: Yes, the paper is a significant contribution and worthy of prompt publication.
Supported: Mostly yes, but some further information and/or data are needed.
Referencing: Mostly yes, but some additions are necessary.
Quality: Yes, it is well-written, logically organized, and the figures and tables are appropriate.
Data: Yes

Reviewer #1 (Formal Review for Authors (shown to authors)):

This paper deals with the timely question of aftershock productivity and provides a systematic analysis of global seismicity. In particular, it provides insights into what determines aftershock productivity other than main shock magnitude. Overall, I find the results sufficiently interesting and novel for publication in JGR. However, the authors need to address a number of points before final acceptance:

1. The first key point goes beyond what is stated in the abstract in terms of the key attributes. I would suggest listing all key attributes in the abstract explicitly.

Done (L. 26)

2. Eq.(2): There is no general consensus on the exact value of this scaling exponent. The most recent study I am aware of is the 2019 paper entitled "Earthquake-Scaling Relationships from Geodetically Derived Slip Distributions" by Miller and co-workers in BSSA. This paper 10.1002/2014JB010940 also contains a review of recent findings. This should

be discussed and the possible implications for the study at hand should be highlighted. Also, you do need to add units to Eq.(2).

As discussed in response to the AE, we now include in the supplement a comparison of aftershock counts using rupture length scaled from mainshock magnitude follow both Wells and Coppersmith, 1994 and Brengman et al., 2019. The number of aftershocks inferred in both cases is nearly the same. We point the reader to the result in the main text with the statement:

L. 245: "We also check the results by using the more recent geodetic-derived scaling relationship of Brengman et al. (2019). Supplemental Figure S2 demonstrates that the results are nearly the same with 79% of the mainshocks having identical aftershock counts."

3. l184: What about the other 1%? I am not sure why such events are even selected given the described procedure.

The reviewer here is referring to the 1% of mainshocks that appear to have overlapping sequences. These rare events can happen in the aftershock detection algorithm for mainshock magnitudes that are different and aftershock sequences that partially overlap. Since these cases are unusual (1%), it seemed unwise to modify the algorithm further to eliminate them. We explain this with new text that reads:

L. 260: "...The non-isolated cases occur when two mainshocks of different magnitudes have overlapping, but not coincident, aftershock sequences. These occurrences are rare (1%) and therefore we do not complicate the algorithm further to eliminate them."

Section 2.2: I strongly believe that the order of section 2.1 and 2.2 should be reversed since, currently, 2.1 uses the data that are only explained in 2.2.

We follow your recommendation and reversed the order of these sections.

4. Fig. 2: I did not appreciate that the figure was rotated. In terms of content, it shows that there is a systematic variation in alpha with the selection parameters. This is, however, not mentioned or discussed in the text. Please add this and discuss possible implications. It is also not clear whether the SAME larger space-time window (see l180-182) is used in all cases to eliminate events. Please clarify.

Orientation:

Agreed, we changed the figure orientation to facilitate viewing.

Variations in alpha:

We now explain with additional text in the caption of Figure 2:

Figure 2 caption: *"...However, with increasing time and space windows background events inflating aftershock counts and reducing α-values by overestimating the productivity of smaller events becomes increasingly prevalent."*

Most importantly, the new supplement including Figure S1 shows that there is no magnitude dependence to the scatter, as would be expected if alpha was incorrectly assessed.

Larger space-time window:

We apologize for not explicitly specifying this parameter before. We now do this in the caption of Figure 2:

Figure 2 caption: *"...Corresponding larger space and time windows are 4/3 and 5/3 of the selection windows."*

5. Fig. 4: Why are some events clustered exactly at 10km depth? This seems to be an artifact.

Yes, it is a default location depth for 10% of the events. We have now taken care to make that clear in the text with the added text:

Figure 4 caption: *"...Note: Discretization of depth is apparent in this plot as some events have default values. Depths of 33km, 5km, 10km and 15km are reported for 6%, 1%, 10% and 0.7%, respectively, of the catalog."*

6. l267-270: Please give the actual numbers if you mention a specific example. How productive were these events?

Good point, done
L. 330: *"With 170 aftershocks within 792 km, ..."*

7. Fig. 6: Please add a figure to the sup mat that shows the same but for the more conservative choice of the magnitude of completeness of magnitude 5.

This figure was already in the supplement. It is now highlighted in the text with the new second paragraph of the interpretation section that includes the sentence:

L. 513: "...A higher catalog completeness of M_W5 reproduces similar results to those we presented using a global completeness of M_W4.5, though far more mainshock have no observed aftershocks. When more than half of mainshocks considered in a group are censored, the estimation of a median productivity is obscured. Using an alternative clustering routine, the relationship between relative productivity and crust age is slightly less pronounced but still apparent."

We have also added a table of contents to the supplement to make it easier to find material.

8. Fig. 8a: The grey shaded area is not well-defined since it crucially depends on the number of permutations but it also depends on the set of realization - the maximum obeys extreme value statistics and it is highly variable. Estimating the 99% quantile, for example, is much more robust. Please revise accordingly.

The description of how the grey area was defined was not clear, we clarify:

L. 389: "A potential problem with this analysis of 12 test parameters is that a spurious correlation could arise simply because of the numerous investigated regressions. We explicitly address this issue by investigating the probability of spurious correlations. To do so, we first remove any causal relationship between our 12 predictors and relative productivity by shuffling the aftershock measurements and randomly reassigning each relative productivity measurement to the parameters for a different mainshock. We then regress each parameter with the relative productivity and report the maximum variance reduction of any parameter in this shuffled set. This same routine is repeated 10000 times to generate a probability distribution function of the maximum variance reduction of 12 parameters should there be no causal relationship in the data. We refer to this evaluation of the extreme value for the full group, or family, of parameters as a family-wise test and it serves as a null hypothesis. We determine the percentile of our actual regression results within these random realizations, the family-wise p-value, for *comparison (Figure 8a - top axis). The comparison yields the probability of obtaining equally good or better results by chance.”*

9. l366: Please mention here how you quantify "best prediction". Similarly, in the captions of Fig. 10 states precisely in which measure the SVM model "outperforms" the other model.

We provide clarification to this point throughout the section:

L. 446: *“...We measure the performance of these tools by computing the root mean squared value,* $RMSE = \sqrt{\sum(\hat{f}_i - f_i)^2/N}$, *where* $\hat{f}_i$ *is a prediction of relative productivity and* f_i *are one of N observed values. To avoid over-fitting the data, we perform leave-one-out cross-validation---individual predictions are calibrated on the remainder of the data (Witten et al., 2011).”*

L. 455: *“...The chosen number of nearest neighbors produces the lower root mean squared error between predictions and observed relative productivity measurements.”*

L. 463: *“Support Vector Machines (SVM) yield the lowest root mean squared error (Figure 10b)”*

Figure 10 caption: *“The SVM model provides a 20% improvement in the root mean squared error when compared to k-nearest neighbor model”*

Reviewer #2 Evaluations:
Significant: The paper has some unclear or incomplete reasoning but will likely be a significant contribution with revision and clarification.
Supported: No
Referencing: Yes
Quality: Yes, it is well-written, logically organized, and the figures and tables are appropriate.
Data: Yes

The manuscript presents an analysis of aftershock productivity of global seismicity with magnitudes M W ≥ 4.5 from 1990 to 2019. Mainshocks are defined by M w ≥

6.5. The main focus is on the relation of the dependence of the productivity on features of the regional tectonic regime, e.g. lithosphere age, stress drop etc. It is concluded that some factors are more influential than others and that this result is helpful for the goal of aftershock prediction.

I find the paper overall interesting and well presented. I have, however, concerns regarding the aftershock model and the statistical robustness of the results.

1. My main concern is related to the aftershock model and the relative aftershock productivity, which is defined in Eq. (3). There is a broad agreement that seismicity consists of two components: first background activity that follows more or less a homogeneous Poisson process, and second aftershock clusters, which have their own statistical laws. In the present work, an aftershock is defined solely by the spatial and temporal vicinity to a mainshock. The whole catalog is decomposed hierarchically into mainshocks and aftershocks. Background seismicity is not taken into account. In my opinion, this is an inappropriate model of seismicity. It would be better to use a thinning or declustering algorithm to define aftershocks, although this will result in an overall smaller number of aftershocks.

Sensitivity to the aftershock model is an important issue and therefore we repeated our entire analysis using the declustering routine presented by Zaliapin et al., 2008. These results are presented in the supplementary section S1.3 which includes every figure from the main text side-by-side with the same calculation derived from Zaliapin's declustering. Supplementary Figure S16 shows that the aftershocks productivity measured by both methods correlates well and therefore, the subsequent results are nearly the same. It is important to recognize that the catalog completeness threshold (M_c=4.5) is a rather high value with a correspondingly low background rate. It is therefore unlikely to have a significant effect on the counting here. If one were considering much smaller events, background rates could be significant, not in this global analysis.

We discuss the alternative declustering in a new paragraph in the main text:
L. 264: "For our alternative method, we use a clustering routine following Zaliapin et al. 2008, Geobel et al., 2019. This approach seeks to build earthquake families by linking earthquakes to parent events based on a distance metric that combines magnitude, space and time. Pairs of parent and daughter events exhibit a statistical distribution with two modes: one that corresponds to clustered events and another that arises from a

Poissonian background of seismicity. Separation of earthquake clusters is achieved by defining a decision boundary between these two modes and cutting all links that exceed this threshold. The largest event in each cluster is identified as a mainshock and aftershocks are counted as the number of events that follow it. See Zaliapin et al., 2008, for a detailed overview of the method, distance metrics, and theoretical connections to other schemes (ETAS). The specific parameters selected are consistent with previous implementations (Zaliapin et al., 2008) and are documented fully in supplementary material (see Supplement Section S3.3). Supplement figure S16 shows that the aftershock productivity measured from both methods correlate well (R^2=0.96)."

Background is inherent in the mainshock definition and we also added some clarification to this point in the text:

L. 223: "...The hierarchical approach captures both mainshocks that arise naturally from background seismicity; and mainshocks in more complex chains of seismicity in which later earthquakes in a sequence become the largest earthquake and thus the mainshock."

2. The metric to measure aftershock productivity provided in Eq. (3) is based on a comparison of a observed and expected aftershock numbers. The latter is taken from Eq. (1) with values α and k fitted in Fig. 1. If I understand correctly, the values are α = 1 and k = 10 −6,2 . However, Fig. 1 shows enormous scatter of the data points. It would be therefore fair to take somehow uncertainties of these values into account.

The scatter is in fact what we aim to better understand and therefore it would be circular to propagate uncertainties the measurements of relative productivity. We now specify this reasoning:

L. 295: "The use of Δlog(N) as a measure of relative productivity provides a means to assess whether scatter in aftershock production is related to specific mainshock parameters on an earthquake by earthquake basis."

3. The space window for the aftershock identification in Eq. (2) is based on Wells & Coppersmith (1994) and increases exponentially with the mainshock magnitude. If I consider an earthquake model without aftershocks, in which earthquakes are distributed uniformly in space and in time, the earthquake number in a circular space window with radius R will increase according to N (R) ~ R 2 . With Eq. (2) I get N (M) ~ 10 1,18M. Inserting this into Eq. (3) with α

= 1 results eventually in Δ log(N) ~ 0, 18M . In particular, if I consider artificial data without any aftershocks and without any physics, the measure Δ log(N) returns “more aftershocks than predicted”. This seems hardly reasonable. Maybe it will be more appropriate to use number of aftershocks per area.

Although one might suspect the aftershock measure to behave as the reviewer outlines if background activity is high, it is not the case in this dataset because of the rather high magnitude of completeness. Figure 1b shows that there is no observable trend with magnitude. This is a natural outcome as we define relative productivity as the difference relative to the magnitude trend. The low background is also demonstrated by the random shuffles in Figure 2 that are more than an order of magnitude lower than the inferred productivity. The magnitude independence is also now demonstrated by a new Figure S1 that show the interquartile ranges.

4. At various stages, the authors use statistical tests like the Kolmogorov-Smirnov (KS) test, without saying, what they do exactly. For example, on page 16 I read that a 1KS test “suggests that the relative productivity of events on oceanic transform faults and continental convergence boundaries have small probabilities (...) of being sampled from the overall distribution by chance”. It seems that the authors just look at a specific value of the relative productivity in an empirical distribution and report the corresponding probability. This has nothing to do with a KS test. If indeed a KS test has been used, please provide the details of the test.

Yes, this is a good point. We did not clearly explain the use of the KS test in various instances and had inaccuracies in some cases where we used a two-sample KS test instead of a 1KS test. See elaborations and corrections below:

Estimating the magnitude of completeness:

L. 154 “...We select earthquakes with moment magnitude exceeding global catalog completeness. We determine a global completeness utilizing the Kolmogorov-Smirnov test to evaluate the goodness of fit between a theoretical Gutenberg-Richter distribution and the data at a range of possible completeness magnitude. The lowest magnitude that produces a local minimum in the Kolmogorov-Smirnov metric is selected as the magnitude of completeness (following Clauset et al., 2009, Goebel et al., 2017).”

Comparing data:

L. 362: "...A two-sample Kolmogorov-Smirnov test comparing the cumulative distribution of each subset and that of the entire set suggests that the relative productivity of events on oceanic transform faults and continental convergent boundaries have small probabilities, $1/10^9$ and $3/10^4$ respectively, of being sampled from the overall distribution by chance; the remainder of the subsets are not significantly different from the overall distribution ($p > 0.05$ or, equivalently, a 1/20 chance)."

5. Figure 8: In panel a it is claimed that normalized rupture width W ∗ , aspect ratio are correlated with the relative productivity, while for the log stress drop Δσ this is almost the case. If I look, however, at the scatter plots in panels b-d, I see (at least in b and c) completely uncorrelated data. Please clarify. And again: the statistical testing should be explained in more detail.

Relationships between stress drop, normalized width, and our target variable, relative productivity, are indeed low correlations, but significant with the p-values as indicated in the text. We hope that these subplots can provide some transparency and used them to emphasize the nature of the correlations. We now emphasize this fact in the text:

L. 386: "At face value, this analysis shows that aspect ratio, width, and stress drop are best correlated with the relative productivity. Figures 8b-d show that these correlations are seen in the raw data, albeit with significant scatter."

Note, however, that repeating our analysis with the alternative clustering routine yielded much cleaner relationships for these exact parameters. We now note this in new text:

L. 519: "Interestingly, we find betters relationship between aftershock productivity and normalized rupture area (p=0.004), stress drop (p=0.0008) and normalized width (p=0.002). Note that these changes occurred among variables that are strongly co-varied (as shown in Figure 12). The correlation between source properties and aftershock productivity are otherwise unchanged using this alternative definition of aftershock clusters. Finally, only subtle changes are incurred when examining our major results (Figures 11 and S25). Consistency with conservative catalog completeness, mainshock size, and aftershock selection method suggest that our major results are robust."

To improve clarity, we have added more detail regarding how we critically assess the statistical significance and likelihood of relationships within our data in the text. Also, your comment motivated the restructuring of results to clearly separate marginal conclusions and from more definite ones.

L. 389: "A potential problem with this analysis of 12 test parameters is that a spurious correlation could arise simply because of the numerous investigated regressions. We explicitly address this issue by investigating the probability of spurious correlations. To do so, we first remove any causal relationship between our 12 predictors and relative productivity by shuffling the aftershock measurements and randomly reassigning each relative productivity measurement to the parameters for a different mainshock. We then regress each parameter with the relative productivity and report the maximum variance reduction of any parameter in this shuffled set. This same routine is repeated 10000 times to generate a probability distribution function of the maximum variance reduction of 12 parameters should there be no causal relationship in the data. We refer to this evaluation of the extreme value for the full group, or family, of parameters as a family-wise test and it serves as a null hypothesis. We determine the percentile of our actual regression results within these random realizations, the family-wise p-value, for comparison (Figure 8a - top axis). The comparison yields the probability of obtaining equally good or better results by chance."

L. 412 "Our analysis suggests that correlations with log-stress drop (p = 0.07) and normalized width (p = 0.05) are marginally significant. Aftershock productivity negatively correlates with the logarithm of stress drop and positively correlates with normalized rupture width (see Figure 8b-c)

L. 416: "...Slip-zone aspect ratio (p = 0.007) is related to relative productivity in a statistically significant sense. Aftershock productivity negatively correlates with aspect ratio (see Figure 8d)."

6. I have serious problems with Section 3.5. The authors present an aftershock prediction scheme based on Support Vector Machines (SVMs), which is a label for a class of models from machine learning. Here, SVMs are introduced as a black box followed by results, which are qualitatively presented in the text and in Fig. 1b. However, the reader has no chance to reproduce or even to understand these results. Therefore, I suggest one of the following options: Either remove the section on SVMs or provide detailed information on the method including parameters, assumptions etc. in order to enable the reader to reproduce the results.

We add a more intuitive explanation for how the models function and include our trained model in the supplement for reference.

L. 466: *“…Key differences between SVM regression and a typical linear regression are 1) a tolerance for a margin of error, 2) an explicit minimization of model complexity using a penalty for each additional input parameter and 3) an embedded non-linearity enabled by the transformation of the coordinate system via prescribed kernel functions. Required model hyperparameters are the order of the kernel polynomial transformation, the margin width, and the complexity trade-off. SVMs are robust to over-fitting data and are therefore well suited for small datasets (Witten et al., 2011). We present results from an SVM model trained using a quadratic kernel. The trained model is included in the supplemental files with full documentation of hyperparameters.”*

In summary, the manuscript is interesting and well suited to JGR, but the statistical evaluations are not entirely convincing, and the definition of aftershocks should be reconsidered.

We hope that we have been able to bolster our analysis to demonstrate that our results are sound. We would like to thank once more the associate editor and reviewers for their comments.