Earth Surf. Dynam. Discuss.,
doi:10.5194/esurf-2016-37-RC3, 2016

Earth Surface Dynamics Discussions

Interactive comment on "Quantifying the roles of bed rock damage and microclimate on potential soil production rates, erosion rates, and topographic steepness: A case study of the San Gabriel Mountains, California" *by* Jon D. Pelletier

Anonymous Referee #2

Received and published: 7 October 2016

Anonymous review of Pelletier, eSurf submission 2016

I probably should start by disclosing the fact that also I reviewed a previous version of this manuscript that the author submitted for publication in Geology. This is relevant because It is very clear to me that the author took advantage of relaxed space constraints here at eSurf to address some of my and other reviewers' suggestions. This is commendable, and as a result, this version of the MS is significantly improved in my opinion. Nevertheless, there are still some major flaws – including an especially big one that I pointed out in my review of the Geology MS but that the author did not

address here.

Thus, despite the many changes that the author incorporated in this revision, and despite the fact that the manuscript is topically very well suited to the audience of eSurf, I do not think that the key conclusions are justified by the data and the analysis as presented. Hence, I do not think it is suitable for publication at this time. Big revisions are needed – and I frankly think a major overhaul of both the approach and discussion are needed to remove the impediments to publication in a revised submission to eSurf or a new submission to a different journal.

*********Here are my big general concerns in no particular order:

(1) First, I concur with reviewers Heimsath and Whipple on the fact that it is non-standard at best to assign a separate P0 value to each measured P value using exponential scaling relationship in Eq 1. It would be equivalent to predicting different values of a y intercept in a linear regression of y on x when you know the slope of the regression and the value of y and x for each data point. There is only one y intercept per regression through a cloud of data. This business of inferring one y-intercept per data point strikes me as – at best – a kooky way (i.e., that differs from established norms) of quantifying the uncertainty in the y intercept.

Based on what I can see in their reviewer comments, Heimsath and Whipple had the same reaction. And the author's response to their comment – i.e., "On a more practical level, I don't understand how we, as a community, could make significant progress on understanding the controls on P0 values if we accept the logic of Heimsath and Whipple that only two P0 values can be reliably determined from 57 CRN analyses" – is not compelling. The alternative logic of Pelletier seems to be that we should suspend conventions of statistics and stretch data farther than they can be stretched just to support some as yet non mechanistic formulation that he has presented here. I prefer the less radical option of recognizing the limits of data and working to overcome them in more traditionally acceptable ways – i.e., with new measurements and perhaps a more clever

analysis approach. For example, as an alternative to the methods presented here, the author might think of ways to model P rather than P0 using some sort of multiple regression analysis that includes h explicitly in a model of rock damage and microclimatic effects. This business of calculating a P0 effectively corrects for the exponential-with-depth variation in P before the modeling begins. In a true multiple nonlinear regression, one could account for h, D, A and everything else simultaneously, and as an outcome of the approach also quantify the relative importance (leverage) of each variable in the regression. If the outcome is that h dominates while D and A add little to the predictive power of the model, then the author would be forced to confess that D and A are not strong predictors of P and thereby P0. And as I point out below, there is good reason to suspect that that is precisely what he would learn.

(2) Second – and this is a bigger concern in my view – is the degree to which the predicted values of P0 ***diverge*** from the observed values in the dataset. It seems like a key goal in this paper is to use indices of rock damage and aspect to predict P0 and ultimately map the variations in P0 and some additional offshoots of it (E and E*L) onto the landscape. Starting at section 2 and continuing through to the end of this paper, this is actually ***the*** central focus of results and discussion. The trouble is, one must be willing to believe that the model in Eq 2 is a good predictor of P0 in order to confidently follow the author in this vital leap of faith.

Personally, after reading this paper, I am not willing to make that leap. Nor should any self-respecting data analyst, once he/she realizes that the predicted values are not actually a very good match to the observations. Sure – as the reviewer points out – there is a highly significant correlation between the measured and predicted values P0, but the existence of such a correlation is not sufficient in and of itself to demonstrate that the predictions are good enough to explain the variations in P0. Recognizing this challenge nearly fifty years ago, hydrologists Nash and Sutcliffe (1970) developed their own measure of model efficiency in their quest to objectively evaluate whether their models of river discharge were good predictors of observations. Though this Nash-

Sutcliffe (N-S) statistic – as it later became known – was developed for models of river flow, it has also been widely used to assess model efficiency for other natural variables, including erosion rates and nitrogen and phosphorus loading.

My quick calculation of a N-S efficiency statistic for the model of predicted P0 yields a value of 0.18 based on data provided in the supplemental table. For context, realize that the maximum value of the coefficient, which is 1, would indicate that the model explains ***all*** of the observed variation in the data. By contrast a value of 0 would indicate that the model is ***just as good as*** the average value of the data at explaining observed variations across the data set. Values less than 0 imply that the model ***is worse than*** the average. In this case, my estimated value of 0.18 indicates that ***the model in Eq. 2 is a little bit better than the average P0*** at predicting the distribution of measured P0 across the dataset. For this reason, I think the machinations of the predictive modeling exercise (i.e., most of the paper) are not really warranted , irrespective of the significance of any inferred correlations between P0 and D and between P0 and A. Importantly the reader should only commit to believing those correlations to the extent that he/she can overlook the suspect exercise of calculating P0 for each measured value of P.

Ultimately, it is not clear to me that the author understands that there is a vital difference between documenting a statistically significant correlation between a measured and predicted value and demonstrating that a model is good at doing what it is supposed to do. If he does, he is hiding it at the top of page 7, where he seems to suggest that statistical significance in classical regression metrics is sufficient. I will not deny that the correlation coefficient and thus the coefficient of determination by themselves provide a very loose first approximation of model fitness. But even then, this is true only to the extent that high coefficients of determination (close to one) imply better correlations and low coefficients imply poor correlations (irrespective of whether they are statistically significant). To understand the problem with using Rˆ2 in the way the author seems to want to use it here, consider the toy example in which P0 observed

is exactly 0.2 times the value of P0 predicted for each inferred value of P0; in that case, the coefficient of determination of P0 predicted and P0 observed would be 1.0 with a very very low p value even though the predicted value of P0 is 5 times higher than the observed value at each site. P0 predicted is dead wrong but the coefficient of determination is fantastically good. This illustrates how simple correlation indices for predicted versus observed data sets can (and probably often do) fall short on gauging the predictive power of a model.

Disclosure: This comment is more elaborate but ultimately very similar to one that I made in my previous review of the Geology submission. Yet the author did nothing to address this concern in this revised manuscript. Although the predicted values are a bit different in this iteration (mostly reflecting changes in chosen values of L/k and Sc, I think), my comment still stands. Nothing about the revised paper has changed the fundamental problem that the model does little better than the average value of P0 at predicting P0 measured here. This seriously undermines the story. While it is probably true that the author is no longer quite as assertive as he was in round 1 about the role of microclimate and faulting in controlling landscape evolution, the words in the abstract and conclusion sections leave no doubt that he really believes – and wants his readers to buy the idea – that they play "subequal" roles along with tectonics in controlling P0. And he unambiguously (but unjustifiably) points to the statistical significance of Rˆ2 values of the relationships between predicted and observed values of P and P0 as his metrics of model fitness (page 7 lines 1 and 2) – which, as noted above, are not definitive, even when the Rˆ2 is much higher than the low values reported here.

—- On a side note, when I plotted the P0 measured and P0 predicted values in the supplemental table against each other, I get a pattern that looks slightly different than the one shown in the figure. The differences are not big enough to explain away the problem of low Nash-Sutcliffe statistics (Fig 3D and 4C), but it made me worry that the author has some version inconsistencies between his figures and the data he provided in the table. Not sure which version is "correct."

(3) Like Heimsath and Whipple, I am unimpressed with the theoretical basis of Eq 2, and moreover, I am not compelled by the author's response – i.e., "It is far beyond the scope of the paper to develop a comprehensive theory for how microclimate relates to vegetation cover, wildfire frequency and severity, and soil production rates, assuming such a theory is even possible." However, whereas Heimsath and Whipple rightly seem very worried about how D might connect to rock damage at landscape scales, and how those variations in D would actually connect to P0 in a mechanistic way, I am stuck on the fact that the authors never actually showed me that aspect should matter at SGM.

The references cited on page 3 have nothing to do with the effect of aspect on vegetation or the effect of either vegetation or aspect on fire intensity or severity in the SGM. Where is the proof that vegetation, fire frequency, and slope steepness vary with aspect in the SGM? It seems it would be crucial to demonstrate this is the case before motivating the paper and the formulation of equation 2 more specifically. The aspect story fits with some of the author's work in other landscapes but not here – at least not according to the references cited here. If anything, the Keeley and Zedler study seems to suggest that the current regime – in which the landscape is prone to large fires that sweep through the landscape with indifference to aspect – has been the norm for a long time

Additionally, this study seems to hang a lot of its motivation on the idea that fire promotes weathering. But - despite the good investigative work cited on page 3 - I am not sure I concur that the connection has been well documented at SGM. All of the studies cited here are fascinating but ultimately just anecdotal investigations of weathering of boulders – not weathering of rock under soil, which presumably is important here since much of the SGM area is covered by soil. Moreover, they do not report faster weathering rates on fire-prone versus not-fire-prone slopes. In fact, none of the studies actually report rates (focusing instead on processes) and none compare fire-prone versus not-fire-prone slopes. Shtober-Zisu et al. comes close to reporting a rate but ultimately says it is hard to say how the boulder spalls in carbonate outcrops influence denuda-

tion rates across the landscape. And again, there is no comparison to a landscape that is not fire prone, so there is no control in the experiment – and importantly no support for the author's claim here that weathering is faster in fire-prone versus not-fire-prone landscapes.

However strong the correlation between P and A may be, I think it is very important for the author to step back from this generic claim that aspect-driven differences in wildfire are driving the show and more precisely drill in on how anecdotal studies from the SGM in particular support the slope aspect idea. Bottom line is there needs to be some stronger motivation here – hopefully shored up some sound mechanistic explanations for why both D (measured in the S&B 2011 approach) and A should matter. I do NOT think it is "beyond the scope" of this paper to justify the formulations that it presumes to impose broadly on the landscape.

(4) The statistical analyses are nonstandard. My discomfort with them is very high. My discomfort started with the first indication – I think on page 5 – that the author thinks of statistical significance as the logical and quantitative complement to a calculated p value. This is not the case, of course. Rather "significance" is commonly reserved referring to the threshold false positive rate that is allowed in a statistical hypothesis test. So the idea that the author thinks that a calculated p = 0.001 corresponds to a "statistical significance" of 99.9% set me on edge. This misappropriation of terminology was repeated many times throughout the text. But that was just the start. The author also evidently thinks it is ok to calculate a y-intercept for each measured value in a dataset using an overall regression slope that was calculated from the entire data set – and which also yields an overall regression intercept. To be honest, this seems akin to data fabrication to me, but I can settle on the gentler view of Heimsath and Whipple that it is really just of a crude way to estimate the uncertainty in the intercept. Next, the author follows a rather stilted approach to quantifying the relationship between P0 and D and A. I personally think it should be P versus D, A and h, thus recognizing h as a factor regulating P and avoiding the problem of getting just two P0 values from

57 values of P. In addition, I think the author missed an opportunity to perform a very standard multiple regression analysis on log-transformed variables and instead opted for a multi stage approach that undoubtedly underestimates errors and fails to produce vital outputs like leverage plots and partial regression coefficients which would help the audience gauge the relative importance of the different factors in the regression. In addition, there is no attempt to propagate uncertainties through any of this. This is a major oversight that needs to be fixed. Last and not least, the author also thinks it is ok to use the significance of Rˆ2 for the relationship between predicted and observed values to judge the performance of his model. In the hydrology community that idea has been rejected for nearly half a century. I am very concerned about the strength of the analyses for these reasons.

*********Specific comments keyed to page numbers and line numbers. For example 5.2 is line 2 on page 5 of the PDF. Note: Although some of these comments pertain to minor issues, others are just as important to address as the ones outlined above.

2.10. I see that Heimsath and Whipple have provided a review of the manuscript and will defer to them as experts on evaluating this paragraph as a motivating theme for the paper. They did not call attention to any problems here. However, as I read line 21 on this page, I guess I have to say that this was not the take home message I got from Heimsath et al., 2012. Higher frequency of disturbance?

3.6-3.10. This study seems to hang a lot of its motivation on the unsupported idea that aspect promotes differences in vegetation which in turn promote differences in fire that promote differences in weathering in the SGM area. See general comment above.

3.23-3.25. This is non-standard to say the least. See general comment above.

4.11. I think I understand what the author is trying to do here (correct the measured P0 for the hump in the SPF), but on reading this, I am confused. You used 1.78P for P0? Not 1.78P0? The way I want to read it is the author is correcting the "measured" P0 – which is inferred from the exponential function to the data – by some correction factor.

But again, I am confused by this statement.

4.12. "This modification of equation (1) affects 4 of the 57 data points." This would only be comforting if there was actually a very strong trend across all the data. Instead, it seems that the data form really loose clouds of correlations that are hinged entirely on a few points. So the fact that this affects 4 of the points is actually troubling – not comforting – to me.

5.3. This equation does not include the fault specific constant of Savage and Brodsky. So I think this assumes that the constant is the same across the study area. Is this justified? Also, to make D dimensionless wouldn't delta x need to be raised to the 0.8 power too?

5.10. ***This is very important.*** The line plotted in Fig. 3A is a log-log regression that ignores the cluster of five data points circled in the figure. There is NO justification for ignoring these points!!! He says in line 5.20 that they occur in an area of unusually dense landslides. I do not see this in figure 1!!! Even if I did, it would not justify excluding them from the analysis. Heimsath and Whipple seem to agree. I think it is complete nonsense. Makes the line look steeper than is should be. Sweeping these points under the rug does not make them go away. Including them in the regression would undermines his story that D plays a "subequal" role with tectonics.It not only looks suspicious. It is suspicious. Author needs to HONOR the data in this study and in his other work and not try to sweep data points away like this.

6.10. I do not understand why the correlation would shut off on north-facing slopes. Is there a mechanistic/theoretical basis for this? If not than the relationship is purely empirical.

6.20. Some more non-standard statistical machinations. The author does a regression that suggests that the power law exponents of A and D are 1.1 +/- some error. Then he reanalyzes things assuming that they are 1 to determine the value of c – the constant in front of A and D in Eq. 2. I am at a loss here. I know the author to be very bright and

competent quantitatively. Yet here he invoking using some unnecessary, non-standard, and potentially misleading steps to avoid what would be a fairly straightforward multiple regression analysis of all of the parameters (slopes and intercepts) implied by a power law formulation of Equation 2. Doing this in a more standard way would yield some very useful metrics like partial correlation coefficients and leverage plots. Perhaps his approach seemed easier to explain at the time he wrote it. But I would argue that the community deserves and expects more.

7.1-7.2. This is actually not a very good correlation for predicted versus observed – especially since it is strangely for a log-log plot. To understand this, look at the plot. There is almost an order of magnitude of variation in predicted P0 at any given value of P measured. To evaluate this model, rather than see an Rˆ2 for a log-log observed versus predicted plot, I think we need to see something like a Nash-Sutcliffe statistic, which would tell us how good the model is compared to simply assuming that we could use the average P measured to estimate P everywhere.

7.5 What are the assumptions inherent in simplifying the equations in this way? Simply citing off to previous work here is not sufficient. What are the assumptions inherent in doing this? For equation 6 you assume slopes are planar, right? Is that reasonable here? What are the limitations of removing the higher order terms of Roering et al.?

7.19. Why 0.03? Just because this is the minimum finite thickness measured? But the whole point is they have no thickness!!! The mathematical inconvenience of having a value of 0 on what you want to plot on a log scale does not justify making up a value that ***drives*** a regression that you then plot through the data. Importantly it is very true that these points have a lot of leverage on the regression. Since calculating understanding the relationship between h and S is vital to calculating E from topography, this ends up being key to the paper. And I really do not think it is well justified.

8.8-8.17. I don't buy the predicted values of P0 so I guess none of this mapping of E and E*L onto the landscape really resonated with me.

9.8. If this is the key result, then you need to demonstrate it using more conventional statistical approaches. A multiple linear regression of the log of P versus log D and h and log A would be a good place to start. This would avoid the strange – and thus hard-to-justify – correction of P to P0 that you have employed here. It would also avoid the strange practice of finding a 1.1 +/- error power slope and then redoing the regression assuming the slopes are 1 to find the best fit intercept term. This whole analysis seemed like a contorted and potentially error-prone way of doing what could have been a textbook application of multiple linear regression analysis on transformed variables.

10.12. This is misleading at best. I see a factor of 2 to 3 in either direction, so a factor of 4 to 6 overall. For example, in Fig 3D, at a value of P0 observed of $\sim$150 m/My I see a range of predicted values running from 85 to 450 m/My. That's a factor of nearly 6 range in predictions for a single value of P observed. That is NOT a good prediction in my book and my assertion is asserted by the very low N-S statistic for this modeling exercise.

11.6. I disagree. This has ***not*** been documented. The fundamental data are suspect (calculations of P0 from P and slope of exponential regression) and the analysis of relative effects of D and A is non-standard.

11.8. Definitely not shown. The business of mapping E*L onto the landscape and comparing it to topography is fundamentally undermined by the lack of predictive power of equation 2, which is demonstrated by the very low N-S stat for the comparison between predicteds and observeds.

Interactive comment on Earth Surf. Dynam. Discuss., doi:10.5194/esurf-2016-37, 2016.