Note to editor/reviewers. Plain text represents the original comment, and bold text represents the response.

Reviewer 1:

The paper presents a valuable exploration of the sources of error and potential improvements of common predictive methods for debris flow parameters. I have a few suggestions below:

Figure 3 - isn't this backwards for (b)? Blue is deposition and red is erosion?

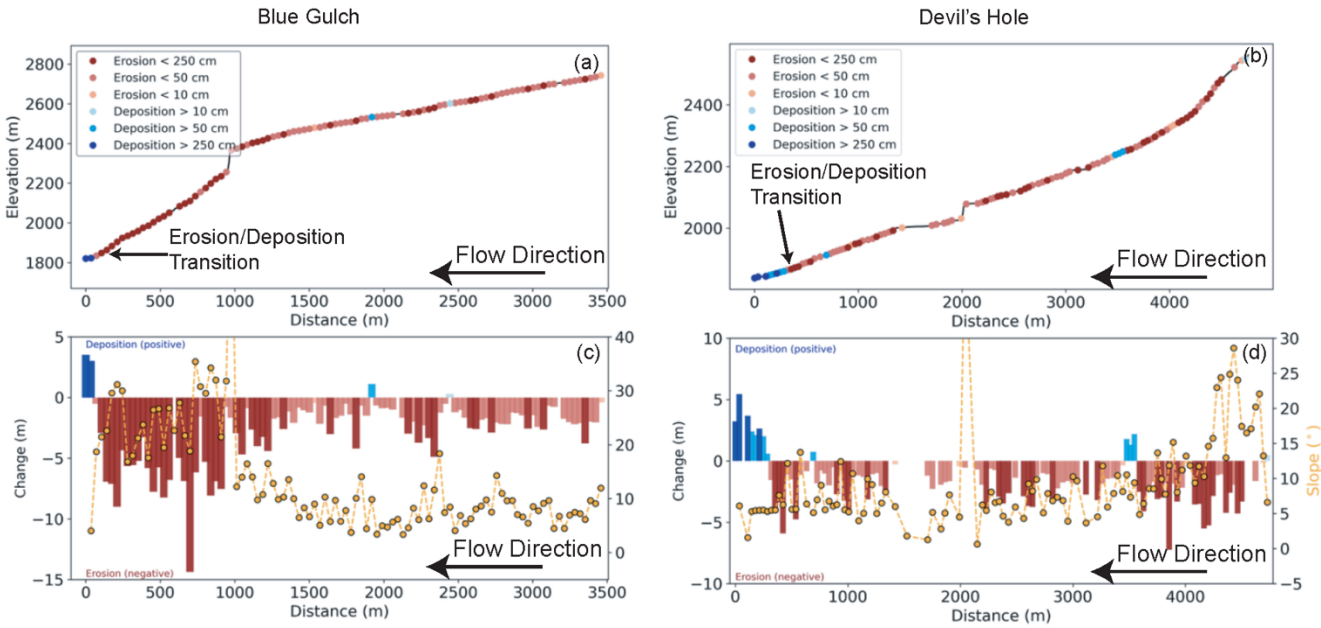


Figure 1. (a) Longitudinal Profile of the Blue Gulch watershed. (b) Longitudinal Profile of the Devil's Hole watershed In both (a) and (b) red dots and blue dots represent erosion/deposition, respectively. Areas without a dot did not experience change beyond the level of detection. (c) Measured deposition and erosion in the Blue Gulch watershed. (d) Measured deposition and erosion in the Devil's Hole watershed. In both (c) and (d) local slope is shown on the secondary y-axis.

One thing that may be making this confusing is that I didn't indicate the flow direction. I have now done this on the figure. The colors in Figure 3b are correct. In figure 3b flow goes from ~4800 m to 0m (this is the outlet), as flow moves from the upper reaches you see lots of erosion (red). There is a little bit of deposition (blue) at a low sloping spot around 3500 m, but the channel is mostly erosional (red) until the slope flattens out around 500 m. This can also be visualized in figure 3d. But in both cases the blue color is equal to deposition, and we are showing that as a positive value. We also added this text to make it more clear:

"Using the two lidar datasets (1 m pixel), we subtracted the elevation of the post event lidar from the pre event lidar ($Z_{2021} - Z_{2016}$) to create a DoD map of erosion and deposition throughout the Grizzly Creek burn area (Figure 2 and Figure 3)."

Based on that method readers should be able to see that wherever the post event data is positive, it should be equal to deposition, and those values are shown in blue in the legend.

Figure 5 - I'm still confused. It makes sense that blue, representing positive lidar differences., should be positive.

If you look in the legend of Figure 5, all shades of blue are positive indicating deposition, with the dash in-between just showing the range of values. All shades of red are negative indicating erosion.

Figure 6 - why would V3 be rectangular? Shouldn't it be trapezoidal?

This is a very good question. We simply don't have any way to estimate the slope angle reasonably. No bathymetry exists. Additionally, we tried propagating the angle of the slope bank into the river, but that resulted in unrealistic depths. In the end, we thought that the most honest approach would be to use the field estimates of depth that we have. Those estimates are shown in Table 2.

Figure 9 - Traditionally, shouldn't the observed (known) quantify be on the x-axis and the predicted (unknown) be on the y?

This is a good point. Typically, when fitting a model, you would definitely do it that way. In this case the predicted quantity of sediment volume is also known, so we are not looking at how an unknown varies as a function of the known. The reason we plotted the data in this way, is so that the coefficient of the line is what you multiply the predicted value by to get a corrected estimate. For example, taking the volume from the Gartner equation and multiplying by 0.21, as shown in our equation in the plot, is a little bit easier than if we plotted the opposite way because you would have to swap the sides of the equation to get a correction factor. It's not really hard, but we wanted to make the process as easy as possible.

I think part of this confusion is that we use the term "predicted", and create a variable called Vp. We have now changed the variable to Vg, representing Volume from the Gartner equation. In addition, we have changed to language throughout to say "estimated" volume rather than "predicted" volume. We think this will help clarify this issue for readers.

Figure 11 - "Damming can be observed in the discharge record of the Colorado River on 22 July 2021. (e)"

Good catch, thank you. Changed to have (e) at the beginning of the sentence now, which is now consistent with the rest of the caption.

"Figure 11. (a) View of debris-flow paths and deposits in the Colorado River looking southwest on 18 August 2021. Arrows indicate new debris-flow paths. (b) Sedimentation on the railroad in Glenwood Canyon on 1 August 2021. (c) Sedimentation on the lower deck of I-70 on 1 August 2021. (d) Photo of temporary damming of the Colorado River on 22 July 2021. (e) Damming can be observed in the discharge record of the Colorado River on 22 July 2021. The upstream discharge in the Colorado River (Red) rises steadily in response to the closest measured 15-minute rainfall intensity (Blue). When the debris flow at Devil's Hole temporarily dammed the Colorado River, a drop in discharge is observed at the downstream river gauge (Cyan)."

Line 61 - consider changing to "more frequent and higher overland flow"

Change to "increased likelihood of overland flow" because the frequency is dictated by the storms.

Line 92 - "are applicable" versus what? "No longer than the first two years"? Might help to clarify

Attempted to clarify this sentence by changing the text to say: "This work explored whether the current USGS operational models for debris-flow rainfall thresholds and volume successfully predicted debris flow occurrence and volume, respectively, at our study site during the first two years following wildfire."

Line 116 - capitalize "Quaternary"?

Changed.

Section 3.3 - you might comment on the fact that you used data fro rainfall gauges close to the measured debris flow events, whereas Gartner, et al, and M1 model relied on a more scattered network. In essence, they were forced to rely on widely spread input data while you have the advantage of comparing to local data. I would like to hear your thoughts on this difference. Clearly, it is better to you more local data, but what can you say about the need in other cases to rely on general data as in Gartner, et al.?

Regarding the distance from a rain gage, the Gartner et al., 2014 paper uses gages that are within 2km of a watershed, and the Staley et al., 2017 paper uses gages that are within 4 km of a watershed. In this study, our gages are on average 4.2 km from the watersheds of interest, but we have a maximum distance of 6.8 km. I have amended the text to indicate this now:

"The maximum distance between an observation and its associated gauge was 6.8 km, the minimum distance was 0.2 km, and the average distance was 4.2 km (Rengers et al., 2023b). These distances are similar to other studies. Staley et al. (2016) used rain gauges within 4 km of an observation to generate the debris-flow inventory used to develop the M1 model, and Gartner et al. (2014) used distances of 2 km to assign storms to debris-flow activity."

This leads to a bigger issue. The Gartner et al. and M1 models rely on a dataset with limited accuracy, but smoothed over an area and temporally. You are using more accurate data, which one would expect to produce better results, but in some ways compares apples to oranges. For instance, earlier volume prediction models relied on StatsGo or other generalized soil data for things like Liquid Limit. More accurate liquid limit measurements would not improve that model because they represent micro-data that is very different from the area-averaged data that the model used. The area-averaged data is not very good, but simply adding focused accurate data does not improve the roughness of the model. I want to make sure that your analysis does not fall in the same trap - providing more precision on a dataset that lacks in accuracy. I think it would help to comment on this disconnect.

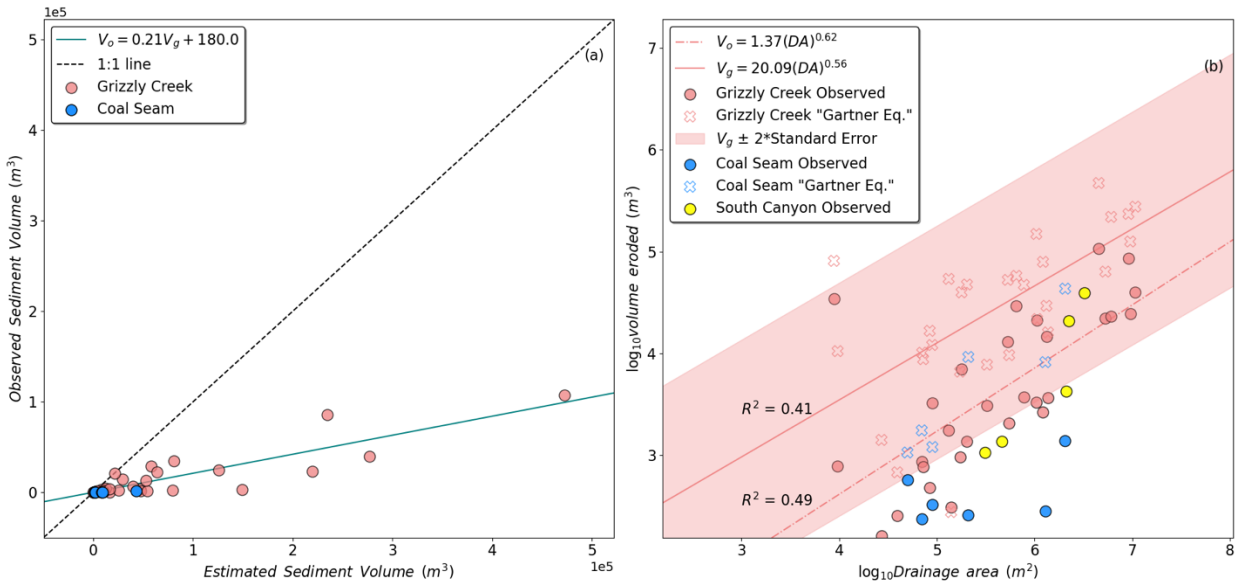
The equation developed in Gartner et al., 2014 to predict postfire debris-flow volumes in the first year after fire was based on data that had considerable uncertainty. The methods that were used in that paper to obtain volume estimates were the following:

"Volumes of sediment removed from debris- retention basins were measured by counting the number of trucks filled during the cleanout of a debris-retention basin or by comparing field or aerial photographic surveys of full and empty debris-retention basins."

The final multiple linear regression model, referred to as the "Emergency Assessment Model" in the Gartner et al., 2014 paper has a residual standard error of 1.04 (the equation predicts ln(Volume) and

thus this value has no unit) and the Gartner et al. 2014 training data generally fit within bounds of one order of magnitude of the predictions. I think our data are likely more accurate and precise than this because I imagine that the uncertainty associated with counting trucks could be quite high. However, we don't have a good measure of the uncertainty from the approach used in Gartner et al., 2014. I don't think this impacts the results of this manuscript though. Functionally, the "Emergency Assessment Model" was developed to estimate volumes, and users take the volumes from that equation to make predictions. Although the data we are comparing may be higher precision and accuracy, I'm not sure that has any implications on the model. We are not trying to re-calibrate the model, and that is where I think we would fall into the trap that you point out. I think we would need more observations of volume across the Colorado region to attempt a model recalibration.

One addition that I have now made, is that I added a band of plus or minus 2*standard error, to Figure 9 for reference (see below) because 95% of the data should fit within this. Also, this is close to +/- one order of magnitude and the Emergency Assessment model observations fit within plus or minus one order of magnitude. So we now give readers a sense of the range that they should consider when looking at the volume estimates from Gartner et al., 2014.



Line 195ff - even more important than the side of the canyon (north or south) and the proximity, is the elevation of the rain gauge, to account for orographic effects. Can you include a comment on this?

I agree that both elevation and the side of the canyon are important. Our mapping of debris-flow initiation showed that most of the debris flows started toward the canyon rim. In terms of elevation, this puts the initiation points closer to the elevation of the gauges, even if they are not as close distance-wise. Furthermore, due to the gage locations, we had to use the data available and didn't often have the luxury of choosing between both canyon side and gauge elevation. I made a map of the initiation points and the rain gauges that I have now added to the supplement.

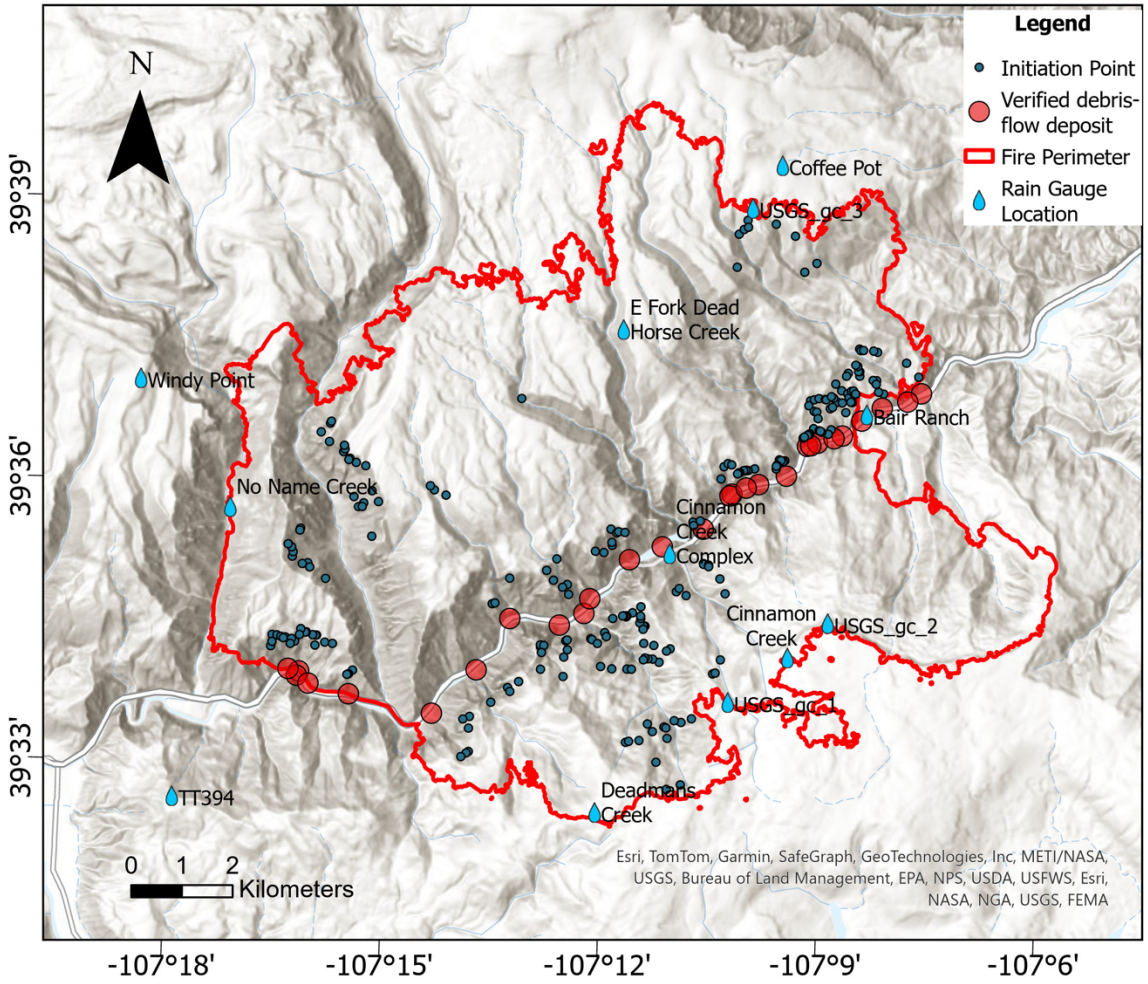


Figure S1. Map showing initiation locations of debris flows with respect to rain gauges and verified debris-flow deposits.

Line 237 - was there notable deposition in the form of levees along the flow channel?

We did see levees during some field investigations. Here is a photo, that I have now added to the supplement. I have also added the following text: "Some deposition was observed in the field in the form of levees, however, levees were not systematically mapped in the lidar because they were often smaller than the pixel cells and therefore difficult to verify without field verification."


Flow Direction
Debris-Flow Channel Cross Section
Coarse-grained Levee

Figure S4. Debris-flow levee with coarse grains shown adjacent to a debris-flow channel. The largest grain sizes in the levee were approximately 30 cm.

Line 255 - not clear to me how V1 differs from V2. They both sound like subaerial deposition. Perhaps distinguish between them better? Also, I would like to see error or confidence estimates on your method of calculating the total volume.

Yes we have now revised the text to make it more clear how we differentiate V1 and V2.

The original text said: "Next, we divided the volume deposited in the Colorado River into two pieces. V2 is the subaerial volume above the river water surface elevation (WSE) at the time of the post-event lidar collection."

The new text has been changed to better explain that V2 is simply the volume of sediment that sits in the area that used to be fully occupied by the river before the fan was deposited. The new text says:

"We divided the volume deposited in the Colorado River into two pieces. These volumes are stacked (one on top of the other) in an area that was fully occupied by the Colorado River prior to fan deposition. V_2 is the subaerial volume above the river water surface elevation (WSE) at the time of the post-event lidar collection."

As far as uncertainty, this is shown in Table 2 for all of the fans.

Line 333 - I would argue that it is vegetation recovery and that sediment supply is not a limiting factor. Perhaps include a reference or two, and maybe shift the focus more on vegetation? Not to push my own papers, but you might check this one for an example of sediment supply independence:

Santi, P. and MacAulay, B., 2021, Water and Sediment Supply Requirements for Post-Wildfire Debris Flows in the Western United States, Environmental and Engineering Geology, vol. 27, pp. 73-85, doi.org/10.2113/EEG-D-20-00022.

Good suggestion. Changed the text to say: "This observation likely results from vegetation recovery in susceptible basins as has been observed elsewhere (Santi and Macaulay, 2021)."

Equation 6 - line 353 - wouldn't you normally want an equation that uses the observed value on the right side ((x-axis) to predict the values on the left side?

Good point, we partially answered this in our response to Figure 9 above, but I want to just emphasize that this is a different situation and we have sought to make it more clear to readers by changing the language from "predicted" to "estimated" and we have changed the variable from Vp to Vg (see reasoning above). The advantage of our approach is that it allows us to use the slope (0.21) as the correction factor. That is, we multiply the volume from the Gartner equation (Vg) by 0.21 to obtain an estimate that better fits our observations. If we rearranged the equation, we'd have to divide the observed volume by the slope of the equation, and that is a little less straightforward.

Lines 406-413 - It is great to see this vegetation-specific analysis. I've never seen this before.

Thanks!

Paul Santi