

## Referee #1 comments

- RC1. Line 36: Remove “how” and Line 37: And “influence”
- AC1. Thanks for these suggestions.
- RC2. Line 316: I worry that  $\pm 10\%$  may be too restrictive? Gartner et al. 2014 recognizes that their volume estimation is much less constrained than that, more on the order of +/- one order of magnitude.
- AC2. In general, we agree that plus or minus an order of magnitude is reasonable given the uncertainty in the Gartner et al. (2014) model. We think that plus or minus 10% is justified for our particular study site because we know that the Gartner et al. (2014) model did quite well at predicting debris flow volume. The combined observed debris flow volume for Oak and San Ysidro Creek was 307,000 m<sup>3</sup> and the combined modeled debris flow volume from the Gartner et al. (2014) model was 320,000 m<sup>3</sup> (Kean et al., 2019). This is a difference of less than 10% between the model and observation. We will add text to clarify this point.)
- RC3. Line 347: Could you explain more why this occurred? It is because the calculated volumes were too small? Or some uncontrolled artifact of the simulation and/or topography? I see this as a significant weakness that needs more explanation. Furthermore, although this may not be the ideal location in the paper to do it, I would like to see recommendations for compensating for this. Perhaps the solution is as simple as noting that one needs to run sufficient simulations to achieve a consensus in results while ignoring the “no inundation” results as non-converging analyses?
- AC3. There are a few competing phenomena that account for this behavior. 1) We intentionally chose a design of training runs that will result in a wide variety of behavior and we have added the following text in Sec 3.4: “Note that the training run design is intended to cover a broad range of possible scenarios that result in flows that vary from relatively small inundation footprint areas to those with relatively large inundation footprint areas.” The fan location in particular is inundated by some flows and not others. 2) The size of the current design is near its practical minimum. We have added the following text in Sec 4.2: “The rule-of-thumb for the size of a GP training set is at least 10× the number of input parameters varied (Berger and Smith, 2019).” 3) There is useful information in the flows that lead to no inundation. Yet extracting that information to be used in GP modeling requires a pre-processing step to distinguish between no-inundation flows that “just miss” versus those that are “not at all close” to inundating a particular location. Including this pre-processing step (or not) does not change the big picture of what we can investigate by using GPs of debris flows, but it should lead to more accurate GP models of debris flows. That said, this pre-processing step has been developed in the context of scalar GPs, but is an active area of on-going research for parallel partial emulators. We intend to elaborate on this in the discussion.
- RC4. Line 353: I would venture to say that this indicates that the parameters with the highest sensitivity are those they either 1) you know well, or 2) contribute most to volume and runoff potential. This agrees with my (probably biased) assumption that saturated hydraulic conductivity is too variable and hard to quantify accurately to be of use, and that soil thickness doesn’t matter because this is not a sediment-supply limited system. Just thoughts to keep in mind for your discussion later in the paper.
- AC4. Thank you for these suggestions for discussion. The parameters with the greatest sensitivity appear to be those that contribute most to volume, which plays an influential role in determining runoff and inundation extent. We hypothesize that soil thickness, which provides a limit on the maximum depth of incision, would be more influential if the model more accurately predicted erosion depths in the channel network. Currently, the modeled debris flows are sourced from more widespread shallow incision on hillslopes and low-order channels. Therefore, there are relatively few locations in the landscape where the maximum depth of incision is achieved. We added text to the discussion section: ‘Soil thickness, which provides a limit on the maximum depth of incision, could be more influential if the model

better reproduced observed erosion depths throughout the channel network. In simulations, debris flow sediment was primarily sourced from more widespread shallow incision on hillslopes and low-order channels. Therefore, there were relatively few locations in the landscape where the maximum depth of incision was achieved.’

- RC5. Figure 2: Indicate in the caption that black cross-hatching is the observed inundation and red is the modeled. Figure 2 caption: This sentence is confusing – I’m not sure what I’m supposed to be seeing. I suggest rewording for clarity.
- AC5. We have updated the caption as follows: “(d) The similarity index,  $\Omega$ , is greatest when the hydraulic roughness coefficient is low. This indicates a better fit between modeled and observed inundation extents when using lower roughness coefficients, especially values less than 0.1 (e) The similarity index generally increases with the volume of sediment eroded from the upper watersheds. Model performance, as quantified by the similarity index, is best when the modeled volume eroded is close to matching the observed volume eroded (dashed black line).”
- RC6. Line 381: I would be careful of suggesting that the peak flow is asymptotic to zero after a set time. The debris-flow hazard should drop asymptotically to the pre-fire hazard level, which is not zero. Perhaps include some wording to indicated that you are measuring the change due to fire and recovery after the fire and not measuring the overall long-term hazard.
- AC6. We have changed that phrasing to: “before dropping substantially after approximately nine months”
- RC7. Line 432: More rapid deposition but also easier detachment and erosion higher in the watershed, compared to more cohesive finer-grained sediments.
- AC7. We agree that increases in particle size that result in less cohesive soils would lead to easier detachment, but the model does not account for that type of effect.

## Referee #2 comments

Spiller et al. present a simulation study designed to evaluate the ability of paired complex model-surrogate model to generate probabilistic postfire debris-flow runout hazard assessments both immediately after a fire and as recovery progresses. The manuscript is technically sound and the result are not over interpreted. The authors present an innovation that has the potential to substantially improve the process of postfire hazard assessment and I look forward to seeing the manuscript published in NHESS. My primary concerns are mostly suggestions for clarification or areas that I think more discussion is warranted.

### Primary concerns

- RC1. I suggest refining how the volume emulator introduced at L304 is described. As best as I can tell, the volume emulator takes either a subset or the full parameter vector (clarify) and produces only one value, the total eroded sediment (define  $s$  in L312). The volume emulator is then used as a filtering/screening step for the probabilistic analysis (described starting L313) to ensure consistency between the parameter sets and the Gartner et al. (2014) emergency assessment model.
- AC1. Thank you for this suggestion. We have added the following text at the bottom of the paragraph you are referencing here, “Rather than a map, the output from this GP, which we refer to as the volume emulator, is a single scalar value (e.g.,  $s = 1$ ) – the total volume of eroded sediment for a given debris flow simulation. Note that this volume emulator is trained on the same  $N = 64$  design simulations as the PPE inundation depth emulator and can likewise be evaluated at any untested combinations of parameters,  $\mathbf{q} = [k_s, r_{max}, \eta_0, F, \delta, RI_{fac}]$ . We then use this volume GP emulator in a screening step to identify parameter sets that are consistent with the Gartner et al. (2014) emergency assessment model.”

- RC2. The process for generating probabilistic hazard assessments (starting L313) assumes that some subset of parameters is conditioned on the Gartner et al. (2014) emergency assessment model and some subset of parameters is sampled randomly (state whether uniform distribution or something else). This seems like a reasonable first pass at generating a probability distribution at each map pixel; however, I also think there are a number of limitations of interpreting the results as a probability. For example, it assumes you have a decent prior distribution of the randomly sampled inputs. I suggest discussing the implications of these limitations, and potential ways forward, in the discussion section. What would it take to not condition on matching the Gartner model and what implications would meeting that bar have for portability (see point 5).
- AC2. Thank you for this suggestion. We have briefly elaborated on our process in the paragraph toward the end of Sec 3.5 with the following text. “Emulators offer a significant advantage to probabilistic hazard analysis as compared to approaches that rely directly on debris flow simulations. If a GP is trained on a sufficiently broad design covering the support of any reasonable prior distribution, then the probabilistic modeling of aleatory (scenario) uncertainty can be developed *independently of the simulation set*. Here we looked at two approaches, one where we screened a uniformly sampled input parameter space for parameter combinations that matched the Gartner model (using the volume GP) and a second that relied on parameters samples from a validation study of postfire recovery from the Liu et al (2021) work (using the inundation depth PPE).” Additionally, we have added the following text toward the end of the Discussion, “Further, as emulators do not depend on a prior distribution characterizing scenario randomness, modeling scenario randomness can be done in parallel or after emulator construction.”
- RC3. The results of Section 4.2 make me wonder how few simulations you can get away with. It would likely depend on where (in map view) you are located. Consider discussion how you might determine the minimum number consider the spatial variability of the runout area.
- AC3. This suggestion is closely related to referee #1’s RC3 and our response to it. Additionally, we have added the sentence “The rule-of-thumb for the size of a GP training set is at least 10× the number of input parameters varied (Berger and Smith, 2019).” Further, we agree that a discussion of future work to address design issues for PPE-based hazard analysis will be useful to the reader and we have added text to expand on that at the end of the last paragraph in Sec 5.
- RC4. It is hard to interpret the results of the recovery experiment without seeing the curves of Liu et al. (2021). I think it would be worth reproducing those curves in this contribution, as well as stating the values for hydraulic roughness and saturated conductivity used for each panel in Figure 6. In the discussion you might also describe the potential to stretch curves based on satellite-based recovery metrics (Graber et al., 2023) or even use spatially distributed measures of vegetation recovery directly within the simulations.
- AC4. We have added a figure to the SI showing changes in hydraulic roughness and  $k_s$  with time since fire and updated the caption to Figure 6. We have also added the following to the discussion: ‘In addition to being parameterized based solely on time since fire, changes in hydrologic model parameters, such as hydraulic roughness or saturated hydraulic conductivity, could be tied to changes in remotely sensed vegetation indices.’
- RC5. The text at L403 describes the spatial pattern of erosion. This is very cool, and an important point to make to motivate future improvement and to highlight that the correct runout results can be obtained even if the upland recruitment is not quite right. However, I suspect it would benefit from a figure illustrating the finding.
- AC5. A figure of spatial erosion patterns is in the supplemental materials, apologies that you did not have those at the time of your initial review.

- RC6. I would be interested in the authors thoughts on the implications of this result for the portability of the emulators and the need for site-specific calibration. I could imagine that on the basis of this work, the authors might articulate a set of behaviors any emulator-based approach would need to demonstrate in order to be portable. This list of candidate criteria could be highly useful for guiding future research.
- AC6. Thanks for this input. For good or ill, emulators inherit properties of the simulator. If a simulator needs site-specific calibration, an emulator will need something akin to site-specific calibration. That said, generally we want the emulator to be calibrated on a relatively wide range of plausible (even if unlikely) and physical realistic scenarios (in that their input rainfall timeseries + input parameters map to reasonably physical inundation patterns). Given this wide set of plausible inputs, we rely on probabilistic analysis to quantify the chances of a particular event and its likely impact. Nevertheless, there is likely some measures that can be transferred from one site to another, even if imperfectly. Although the precise features of one watershed are different from those of a different watershed, approximate characterizations of those regions – for example, soil type or fraction of vegetation cover – will provide an approximation of response to an intense rain event. The challenge is to better understand the measures that can be transferred and the approximate accuracy when transferring those concepts. Articulating strategies for easy transfer of emulators will be a suitable topic for future investigation.
- RC7. I am also interested in the authors sense of how sensitive the result are to the shape of the design storm. I do not think additional work is needed to elaborate on this point. However, different storm shapes are commonly discussed in operational work (e.g., USDA SCS 24-hour storms, alternating block method). Consider elaborating on this point in the discussion.
- AC7. We think this is a good avenue for future work and have elaborated on it in the discussion as an avenue for future work. This could be explored using an emulator if storms could be characterized by several criteria. Consider, for example, storms with a time series of rainfall that is shaped like a normal distribution and defined by a mean intensity and standard deviation. We added the following to the discussion: ‘Since the spatial and temporal distributions of rainfall can influence flood and debris-flow processes future work could consider sensitivity of debris-flow runout to storm characteristics beyond peak intensity.’

### Figure comments

- RC1. Consider showing the location of KTYD on Figure 1.
- AC1. Thanks for this suggestion. The location of the KTYD gauge is approximately 5 km west of the watershed centroid, which makes it awkward to include in panel c and still focus the image on the watershed and downstream deposits. We state in the revised caption that “The KTYD gauge is located approximately 5 km west of the watershed centroid.”
- RC2. I think the manuscript would benefit from a plot of the rainfall timeseries as well as the XX time series used to generate...
- AC2. We apologize that the SI was not visible for review. The rainfall time seres is shown in figure S1.
- RC3. Figures S1, S2, and S3 are referenced but not present.
- AC3. We apologize that the SI was not visible for review. The SI includes these three figures.
- RC4. In Figure 3, I suggest adding a legend that labels the red circles as the left out points, and the blue/black as predicted. It took me a while to understand this figure because I was confused by the large range of the red values. It might be helpful to remind the reader in Section 4.2 that the training simulations are designed to generate a large range of flow depth values and a good test of emulator performance is the ability to confidently predict the left out value using the other values.

- AC4. Thanks for this suggestion. We have added to add a legend to the figure and some more text in section 4.2 to explain the the leave-one-out experiments. Also note that upon revision of this figure, we found a mistake in the plotting script that impacted part of figure 3a. None of the information that we were intending to plot nor the related summary statistics changed, but the corrected figure is slightly different than the original.
- RC5. I cannot easily see the flow edges in the figures that use satellite base maps (Figures 5, 6, 7, 8). Suggest revising.
- AC5. We appreciated this suggestion, but have decided to leave them as they are since the change from colormap (or black) to the background effectively creates a visual boundary.
- RC6. Suggest stating the other parameter values used for each panel in Figure 2.
- AC6. Thank you for the suggestion, we have added those values.
- RC7. Suggest stating that no recovery is considered in the Figure 5 caption.
- AC7 We plan on revising to “Probability of inundation for three rainfall scenarios assuming parameters associated with immediate postfire conditions...”

### Suggested references

In the time since this was submitted, Dunne et al. (2025) was published. As it complements Alessio et al. (2021) and Morell et al. (2021) work, consider incorporating it.

We agree that this is a relevant reference and have added it to the revised text.

### Line level comments

L9: Elements before ‘grain size’ are missing.

We have changed it to ”Simulation results are most sensitive to hydraulic roughness and grain size.”

L10: The contents of the sentence starting “Sensitivity analysis” seems to duplicate the content of the sentence starting on L9. Consider stating only once in the abstract.

Thank you for the suggestion, we have adjusted that part of the abstract.

L16: Parentheses are missing around nearly all of the citations in the manuscript.

Thanks. We have made most citations parenthetical in the revision.

L31: Clarify what is meant by ‘sources’. Do you mean debris-flow volume? As you show later, that is itself parameterized by other inputs (e.g., rainfall).

We have deleted the word “sources” as our later analysis includes a scaling of the rainfall.

L39: Consider saying that this is infiltration excess overland flow.

Thanks. Have changed it to “...spatially-distributed, infiltration-excess overland flow...”

L67: In this paragraph you might also point out that not relying on specification of a volume permits exploration of recovery.

This is a good point. We do mention this in the following paragraph so have opted not to include it in this paragraph (“Since rainfall, runoff, and erosion processes are related to model parameters known to change following fire, such as saturated hydraulic conductivity, hydraulic roughness, and vegetation cover, this framework can be used to explore how postfire recovery affects debris-flow initiation, growth, and runoff.”).

- L75: Although not strictly necessary, consider consistently quantifying what “computationally intensive” means throughout. E.g., core-hour per fine grid cell?  
We have added a parenthetical note after “computationally intensive”
- L105: I suggest stating why San Ysidro and Oak creeks were selected over other runoff paths in the region (Montecito Creek, Buena Vista Creek, Romero Creek).  
These two watersheds were chosen over adjacent ones because they have been included in other modeling studies (e.g., Bessette-Kirton et al., 2019; Gibson et al., 2022), which allows for comparisons, and because observed debris flow volumes closely match those predicted by a commonly used empirical model that we employ as part of this study (Kean et al., 2019). We added this explanation to the end of the paragraph referenced here.
- L125: Here and throughout, I suggest giving a name to McGuire et al. (2017) as implemented by Titan2D. E.g., Titan2D is not one model, but many and so in section 5.1 on model performance, it is a bit confusing discuss Titan2D performance. You could call it T2D-M2017 (or something better). Whatever you do, this is the section to introduce the terminology.  
Thanks for this insight. We used the suggested name in the revised manuscript.
- L134: Define what is meant by ‘new material models’. I understand you are discussing different rheologies, or similar. However, I think most readers will benefit from a bit more explanation.  
We changed the wording from “new material” just called it “different representations of the physical processes: in the revision.
- L135: Because this manuscript is not about the computational speedup, I suggest the text describing details of implementation are extraneous.  
We agree that the details of TITAN features can be taken out. We modified the text to two summary lines near L135 that highlight the added benefits of using the TITAN framework.
- L141: Define  $U$ ,  $F$ , and  $G$ .  
We have added text to describe that  $U$ ,  $F$ , and  $G$  denote the vector of conserved variables and their corresponding flux functions in the  $x$ - and  $y$ -direction, respectively.
- L159: Define  $c$  as the sum of  $c_1, c_2, \dots, c_k$ .  
Thanks for pointing out the need to clarify the sediment concentrations here. Volumetric sediment concentration is denoted by  $c$  whereas  $c_1, c_2, \dots, c_k$  denote sediment concentration for different particle size classes in units of  $\text{kg m}^{-3}$ .
- L189: Cite the source of this DEM.  
We have added the reference.
- L205: Given that debris flows in general, and this debris flow, in particular, had a wide range of grain sizes, I suggest adding some explanation as to the interpretation of  $d$ .  
Thanks for the suggestion to provide some additional detail about interpretation of the representative particle size,  $\delta$ . The representative particle size plays an important role in determining the balance of entrainment and deposition in runoff through its influence on the particle settling velocity. We interpret  $\delta$  as a parameter that effectively modulates the rate and style of sediment transport by influencing settling velocity. The model, for example, does not treat bed load and suspended load fluxes separately but the balance of entrainment and deposition fluxes, the latter of which is strongly influenced by  $\delta$ , influence how frequently particles are in contact with the bed as they are transported. For these reasons, we do not attempt to link the value of  $\delta$  to any particular metric (e.g.,  $d_{50}$ ) related to the grain size distribution. We have added text to the methods section to clarify.

L201: State the AMR refinement criteria, levels, and any other specifications used.

Given your L135 comment, we removed the section of text that mentioned AMRs. That said, the number of AMR levels is defined in the code and in most computations we have restricted it to 2 levels. A simple absolute magnitudes of fluxes have been used as the refinement criteria. More formal error estimators are expensive and in this class of problems do not provide much advantage in our experience.

L237:  $W$  has another usage (introduced at L179). I suggest using a different symbol.  $g$  is also used (with subscripts) in equations (5), (6), and (7). Disambiguate the symbology. This equation also differs from that introduced by Heiser et al. (2017) in that beta and gamma should be subtracted from alpha. I find it useful to point out that this misfit metric is commonly called other things after undergoing linear transformation (trimline ratio, threat score, critical success index).

Thanks for pointing this out. We have updated the equation for the similarity index and changed the notation for clarity. This also included an update to the notation used in Figure 2.

L258: Here and elsewhere: I found inconsistent use of tense. The run inputs were chosen. I suggest revising to consistently describe work done in the past in the past tense.

Thank you for pointing this out. We have revised the manuscript to consistently use past tense for work done. Note in the attached track changes manuscript, we highlighted additions of text in red, but not changes in tense.

L271: Earlier, the subscript  $k$  was for size classes. Consider using unique symbology for size classes and input parameters.

Thanks for this suggestion. We have updated the symbols here to be indexed by  $m$  instead of  $k$  in the revision.

L270: ‘Range parameter’ may be the standard name for Theta. However, because all the other parameters are inputs and the range parameter is a measure of sensitivity, I got confused. If it is not a standard name, consider using a more descriptive name. If it is a standard name, consider introducing it in a bit more detail (e.g., it appears like a variance such that low variance parameters are more important).

Thanks for pointing this out. ‘Range parameter’ is a standard terms. We have added text in Sec 3.4 to provide more detail.

L274: Here is a place where a consistent definition of computational efficiency would benefit. It would permit direct comparison between the complex model and the surrogate at a prediction pixel level.

Thanks for this suggestion. We think core hours of simulation time between minutes on laptop should be sufficient. We have added details of the processors in both cases.

L289: The text in this paragraph does not seem like a numerical experiment to me. Rather, it is the analysis of the emulator fit.

Thanks. We will update the title of Section 3.5 to ‘Emulator analyses’. Section 3.5 is meant to be a summary of various emulator studies whose results are presented in Section 4.

L293: State whether the emulator here is fit with  $N=64$  or is something else.

Thanks. We have done so.

L316: This 10% range is larger than the uncertainty in the Gartner et al. (2014) emergency assessment model. Consider explaining in more detail why 10% was chosen and what it means in light of the larger uncertainty on that model.

In general, there is a roughly an order-of-magnitude of uncertainty associated with volume estimates predicted by then Gartner et al. (2014) model. We used a range of 10% around the Gartner et al.

(2014) model prediction since that model performed well at predicting debris flow volume in the two watersheds that we modeled (Kean et al., 2019). The observed volume was  $307000 m^3$  compared with a modeled volume of  $320000 m^3$ . We have added a sentence to clarify our thinking for the range of 10% given the greater uncertainty with the model in more general circumstances.

Fig 4 caption:  $I_{15}$  subscript formatting issue.

Thanks, this is corrected in the revision.

L433: ‘with increasing delta’.

Thanks, has been changed to “decreases with increasing  $\delta$ .”

L477: What would be required for a pre-fire emulator?

A pre-fire emulator would also require selecting a range of potential values (or a distribution) for model parameters. Nearby calibration sites, if they exist, would be beneficial for refining these parameter distributions.

L505: Consider providing an example driver file.

We appreciate the suggestion. Part of this team is in the process of writing a surrogate for mass flow tutorial/review which will provide much more detail/background along with hands-on examples.