Referee #1 comments

RC1. Line 36: Remove "how" and Line 37: And "influence"

AC1. Thanks for these suggestions.

RC2. Line 316: I worry that ± 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^3 and the combined modeled debris flow volume from the Gartner et al. (2014) model was 320,000 m^3 (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 intend to add 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 intend to add 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 runout 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 runout 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.

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 intend to update the caption as follows: "(d) The similarity index, Ω, 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 intend to change 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.