We thank both reviewers for their constructive reviews. We have prepared a draft of the manuscript with minor revisions, which we can provide to the editors if requested.

Response to Reviewer 1

The authors define rainfall thresholds for landslides for an area within 2 km from the road network and sensitive infrastructure in Sitka (Alaska). This is a highly localized dataset, with very few historical landslide events exists (5 days with events). They use hourly rainfall and compare different accumulation periods, different antecedent rainfall windows, different methods for the definition of the threshold (frequentist and Bayesian).

They consider both probabilistic predictions (probability of occurrence, logistic regression model) and intensity (number of landslides, Poisson regression model). They find the frequentist approach with 3h rainfall to be the best statistical model. They find antecedent rainfall not to sufficiently improve the performances of a triggering rainfall only threshold. Finally, they carry out robustness tests by leave-one-out and splitting the available record into calibration and testing sets.

I believe the manuscript is very well written and organized, and the development of a methodology which seems robust even with extremely few historical events is interesting and surely worth publishing in NHESS.

I only have few minor comments on the manuscript and one concern which is probably worth elaborating on/discussing in the manuscript.

We thank the reviewer for their encouraging review and constructive suggestions. As the reviewer notes, validating the robustness of a data-scarce model was a primary goal of this paper.

I really appreciate all the work the authors put into verifying the robustness of the methods applied with such a limited number of historical landslide days. That said, I think some potentially critical aspects still could be further discussed. How representative are those events? I am thinking of two different things while posing this question which would both lead to overestimating triggering rainfall (by including only “extreme” events), the landslides used and the timing/triggering rainfall:

- While the landslide record seems to be reconstructed from areal imagery (which would capture events regardless of whether there were damages associated or not), could there still be biases towards stronger/more damaging events?

  We of course cannot entirely rule out the potential for bias in the existing landslide inventory. However, based on our familiarity of the region and the Tongass inventory, we are confident that potentially impactful landslides are well-represented in the study area. It is possible that very minor slope failures or localized erosional features (a few square meters in area) are not documented, but these types of events with minimal potential for impact to safety or infrastructure are not the focus of our study. We do acknowledge the potential for some bias in the landslide inventory in discussion section 4.3.

- How was the timing assigned? The authors mentioned eyewitnesses and precipitation record and based on Figure 2 it seems for all days except that of Harbor Mt. Slide the timing was well constrained (30min). But is precipitation used to constrain it to that time? Furthermore, the authors pick “the maximum cumulative precipitation in each sub-daily window”. This could result in considering rainfall occurring a lot earlier than the actual landslide or, more importantly, after its occurrence. Previous

studies already showed that it's typically not the strongest intensity that triggers landslides. E.g., Staley et al. (2013), looked at debris flow and showed that "there were statistically significant differences between peak storm and triggering intensities", confirming that it's not always the strongest rain to trigger them. If this is true, it could also apply to maximum cumulative rainfall. Figure 2 seem to suggest this shouldn't be a problem in general, but it's hard to really see the hourly intensities and still it could be the case e.g., for Harbor Mt. Slide where the maximum 3h cumulative is probably from 1h before the landslide time up to 1h after. This becomes potentially even more impactful if the rainfall record is used to narrow down the timing.

Thank you for this clarifying question. For events assigned a "precise" time in our manuscript (solid line in Figure 2), times were assigned based on eyewitness accounts or power outages (see details below). "Approximate" landslide times were estimated from the rainfall record, such that estimated landslide timing is ~1 hour after maximum hourly rainfall. These times were assigned manually for the sake of visualization in Figure 2, but are not relevant to the model set-up, which uses maximum 3-hour rainfall observed in each day. In a revised draft of the manuscript, we clarify in section 3.1 that precise timing is used for qualitative assessment, but that model set-up is less sensitive to precise timing information because we use maximum daily rainfall metrics.

The reviewer notes observations by Staley et al (2013) of post-fire debris flows triggered by non-peak rainfall. However, post-fire debris flows have unique characteristics and susceptibility, since they are triggered by infiltration-excess runoff. In our study, peak rainfall timing is well aligned with landslides for which precise timing information alone. Additionally, maximum daily rainfall reliably separates landslide and no-landslide events. This is consistent with other explorations of regional landslide data not presented in this manuscript, in which we found that landslide timing (when known) is typically within 1-3 hours following peak rainfall. Similarly, early studies by Roy Sidle found 2-6 hours.

In line section 2.2 of the revised manuscript, we clarify our assumption that triggering rainfall is well-represented by peak rainfall.

The timing of each of the landslides was assigned as follows. This information can be added to the supplemental document in a revised submission.

- **The South Kramer landslide event on 8/18/2015 was assigned "precise" timing based on eyewitness accounts that stated the fatal landslide occurred at 9:30 am. We assume that other landslides during this storm occurred near that same time.**
- **The Halibut Pt landslide on 9/4/17 was assigned "approximate" timing at 12:00 pm based on a news report of a landslide that occurred "around noon" (https://www.kcaw.org/2017/09/04/landslide-closes-halibut-point-road-sitka/#:~:text=Officials%20in%20Sitka%20have%20closed,been%20evacuated%20as%20a%20precaution)**
- **The Medvejie slide on 9/20/19 was based on a news report of a power outage caused by the slide at "shortly before 1 pm." We therefore assigned the time to 12:50 pm. (https://www.kcaw.org/2019/09/20/slide-cuts-off-green-lake-road-hatchery-access/). The timing of the S. Kramer Landslide was assigned based on an eyewitness account which stated the time as 9:30 am.**
- **The Harbor Mountain landslide event on 10/26/2020 was labeled "approximate" because eyewitness accounts could only constrain the event to the night of occurrence. We assigned the time of "early morning" based on peak rainfall totals. Two landslides**

occurred this night. The timestamp plotted on Figure 2 is estimated as occurring shortly after the timing of peak rainfall.

- **The Sand Dollar Drive landslide event included at least two periods of landsliding. Eyewitness accounts constrained timing to one "precise" event at 6:00 pm and one "approximate" event between 9:30 pm and 5:00 am (https://www.kcaw.org/2020/11/02/back-to-back-landslides-block-sitkas-sand-dollar-drive/).**

These aspects are important because while the authors really show the robustness of the method with respects of the available landslide events, having missed one event triggered by a small(er) amount of rainfall, could have a strong impact on the threshold (but also possibly increasing the added value of considering antecedent rainfall). That would be the case if either some events have been missed/not reported or if the timing of any of the used events would be off by some hours. If the Harbor Mt. Slide happened 5-10 h earlier than the estimated time (the uncertainty is 12h) and only rainfall prior to that time was considered, how would it impact the results? This case study appears to be the best I have seen in terms of separation between rainfall on days with landslides or without, even for small domains, which could either be due to exceptional local properties (e.g., very homogeneous region) or indicative of the maybe non-representativeness of the landslide occurrence. I really don't think any of these aspects invalidates the work presented or the methodology used, but it is probably worth discussing and adding some more information, especially about how the timing is determined and about whether rainfall after the estimated time is ever considered.

This is an important question that we have addressed in several ways. First, we train the model on maximum daily rainfall totals, not rainfall preceding the landslide, so potential error in landslide timing does not impact model results. Additionally, evaluating model robustness in the case of missing landslide events is the primary goal of our leave-one-out analysis presented in Figure 10 (first column) and Supplemental Figure S1. We find that excluding the lowest landslide event from the model training results in surprisingly similar parameter estimates. As Bayesian analysis demonstrates, (Figure S1) excluding the low-rainfall landslides does not substantially change the median posterior parameter estimates, but the uncertainty for those values does increase. However, we recognize that sparse datasets may not always capture events in the tails of a distribution (for example, a landslide occurrence at low rainfall rates). This is part of what motivated us to select a conservative lower warning threshold at probability = 0.01, rather than the value which maximizes precision and recall (Fig. 11, first paragraph of section 3.4).

Based on the Reviewer's comment, we have now performed a counterfactual scenario analysis in which we *added* a hypothetical landslide at a lower rainfall value than landslides have been observed in the past (3-hourly rainfall = 18.0 mm), and re-fit the frequentist logistic regression (FL-3H-CF). We seek to test the sensitivity of our regression results to a potentially "missed" lower intensity landslide observation. We compared these results with FL-3H, the logistic regression model on which we based the thresholds. We find that parameter estimates change little (Figure AC1), although the uncertainties are reduced by virtue of having 6 landslide events instead of 5. From a practical perspective, the 3-hourly precipitation value associated with a probability of 0.01 (lower threshold) is 19.5 mm from this model, compared to 21.3 mm from FL-3H. 35.5 mm gives a probability of 0.7 (upper threshold), compared to the 34.0 mm used in the warning system. We surmise that in the unlikely case that a landslide that occurred at a lower rainfall value went unreported, it would not have impacted the parameter estimates in a way that is meaningful for the warning system. This robustness likely results from the relatively large number of days with lower rainfall values on which no landslides were reported. We can add details for this case in the supplemental file of a revised manuscript.

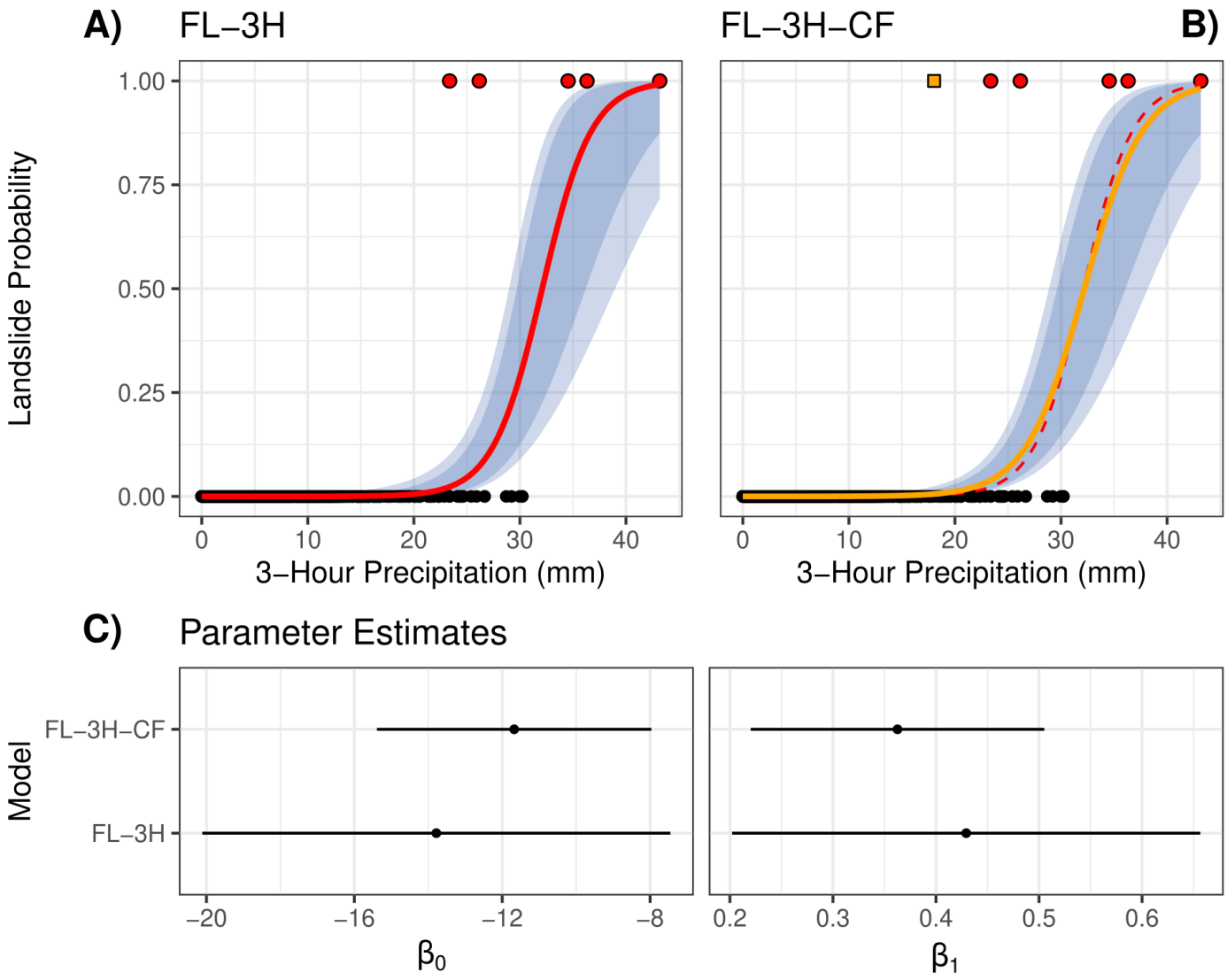

Figure AC1. "Missed" landslide counterfactual scenario. (A) Results of FL-3H, including the five reported landslide events in Sitka (red points). (B) Counterfactual scenario with an additional landslide event (orange square) at 18 mm in three hours (FL-3H-CF). The orange line shows the results of FL-3H-CF, the red dashed line FL-3H. The confidence intervals refer to FL-3H-CF. (C) Comparison of parameter estimates for FL-3H-CF and FL-3H. Error bars show 95% confidence intervals based on standard errors.

Finally, I have some very minor suggestions the authors could consider:

- Line 435: while 0.7 is a value commonly used to recognize a model that cannot be trusted, it could be interesting to report the Pareto-k values for the landslide days (since it's only 5 days)

  We will report the Pareto-k values for the landslide days in section 3.4 of the updated version of the text.

- Comparison to "weighted coin toss": while this is presented as a baseline very simple approach and only used to report the BSS (and not a focus of the work presented), it would probably be more meaningful to compare to something more realistic (e.g., accounting for seasonality of the landslides, which all occurred in the AugustNovember timeframe).

We consider our simple model a useful baseline for comparison, as it offers the simplest alternative to a precipitation-based model. This is similar to a common evaluation strategy in weather forecasting, in which forecasts are compared to climatology (e.g. Wilks, 2011; see more detailed response to Reviewer 2). This step also allowed us to evaluate if our model performed better than "random" guessing, but for rare events a comparison against 50:50 isn't particularly meaningful. However, instead of our historical frequency (5 / 6,606 = 0.0007), we could compare against landslide frequency on days in the rainy-*est* season (most ARs occur in August, September, October and November, but heavy rainfall can occur any time of year). In this case we obtain a daily frequency of 5 / 2215 rainy-season days in our record = 0.002. The BSS between FL-3H and wet season daily frequency is 0.54, indicating that our model still offers notable improvement.

- While all components of the figures are explained in the captions, legend are usually helpful for the reader (e.g., landslide red lines in Figure 2, Figure 10).

  Thank you for the suggestion, we added legend items to describe the red lines in Figure 2 and added a legend to Figure 10.

- In Figure 10 the comparison among the plots is very difficult. While it clearly conveys the message that removing each landslide day does not have a strong impact, it might still be worth either replacing the 5 graphs (5 for probability, 5 for number) with just one where the all the estimated probabilities (and another for number of events) can be easily compared. Either removing the CIs or plotting only the edges with a different color (consistent with the probability) for each landslide removed.

  Thank you for this suggestion. We experimented with this visualization and found that overlaying 5 model fits, particularly with the confidence intervals, becomes too busy to easily interpret. We prefer our layout which allows readers to compare how the model changes when specific landslide events are omitted.

- Figure 12 is a bit complicated to read. The authors could consider either splitting it into two different figures, because it looks like B would be a "zoom in" of A, or a calibration/test split, whereas they refer to different models. Furthermore, I would suggest removing the light blue area (instead just showing the edges, in an empty box around the timeframe in A and around the plot in B) and being more consistent in what is shown (e.g., in A and B the y axis show two different things). They might also remove the black vertical lines for events above the threshold(s). Finally, I am not sure what "the gray field shows the 95% standard error" refers to, but that probably will become more visible removing the light blue background.

  Thank you for this feedback. We have prepared a revised version of figure 12 in which the panels are more clearly differentiated to clarify that they show different values, the blue shading has been removed, and the gray shading is more visible.

Staley, D.M., Kean, J.W., Cannon, S.H. et al. Objective definition of rainfall intensity–duration thresholds for the initiation of post-fire debris flows in southern California. Landslides 10, 547–562 (2013). https://doi.org/10.1007/s10346-012-0341-9.

References (for both author replies)

Betancourt, M.: A Conceptual Introduction to Hamiltonian Monte Carlo, https://doi.org/10.48550/arXiv.1701.02434, 15 July 2018.

Crameri, F. (2018). Scientific colour maps. Zenodo. http://doi.org/10.5281/zenodo.1243862

Kuha, J., 2004, AIC and BIC: Comparisons of assumptions and performance: Sociological Methods and Research, v. 33, p. 188–229, doi:10.1177/0049124103262065.

Wilks, D. S.: Forecast Verification, in: Statistical Methods in the Atmospheric Sciences, vol. 100, Elsevier, 301–394, https://doi.org/10.1016/B978-0-12-385022-5.00008-7, 2011.