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.

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?
- 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.

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.

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)
- 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 August-November timeframe).
- 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).
- 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.
- 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.

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.