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 2

General comments

This study deals with the relevant problem of establishing an early-warning system at a small regional scale with few observations. This problem is approached by testing different models with different inputs and tested for robustness. I think the study will be interesting to early-warning system developers as this system has been implemented in practice and therefore had to tackle many practical problems from how to establish thresholds to how to issue warning levels.

Generally, the manuscript is well written and organized. My main criticism is that the comparison/evaluation/validation is at times confusing and not as streamlined as other parts of the manuscript. Although I very much appreciate that much effort is put in the validation, I think section 2.3 (and 2.4 maybe) need some justification why so many different approaches are being taken. These should maybe also be presented more clearly later. You may have good reasons for choosing many different validation methods (leave-one-out for events, train/test split, different criterion metrics, etc.) but it's not clear to me from the text and I think it will be confusing to readers. For example, there is leave-one-out for landslide events and train/test split. Wouldn't it be simpler and similar added value to do leave-one-out with the recorded years (e.g. train with 2002-2018 and test with 2019)? Or why do you need the Brier skill score? Can't you use the same skill score as for the other models and compare these? Anyway, I don't think you go into details with the results of this part so it could be cut. To systematically compare the predictive power, you could also compute the area under the curve and assess if a model is better than random, as it is commonly done to assess predictive model skill.

I think if these issues could be solved, the manuscript will be more accessible to readers and that the authors are in a good position to solve this. Please find more specific comments below.

We thank the reviewer for their positive assessments that our study will be interesting to early-warning system developers and that our manuscript is well written and organized.

We appreciate the constructive feedback that the motivation for some steps of our approach are not sufficiently clearly described. Our model selection and evaluation strategy can be grouped into two parts. The first part concerns only the statistical models and their predictions. The second part treats the decision boundaries and their performance. We argue that by keeping the statistical and decision making parts of the analysis separate, we are able to provide more transparency and information for decision makers (see, for example, https://hbiostat.org/blog/post/classification/index.html). We elaborate on the motivation for each step below, and will update the text accordingly.

Statistical models

1. **Model selection. First, we used information criteria (AIC, BIC, LOOIC) to select the most appropriate rainfall timescale from a range of possible options. This step was necessary to decide which rainfall metric to use for the LEWS. AIC and BIC both assess goodness of fit while penalizing complexity for the frequentist models, but have different**

drawbacks. Using both overcomes some of these limitations. LOOIC assesses out of sample predictive accuracy for the Bayesian models. The agreement between all three criteria that the 3-hourly timescale is most appropriate gave us additional confidence in our choice to use this timescale for the LEWS.

2. **Model evaluation. After selecting the timescale, we evaluated the 3-hourly models (FL-3H, BL-3H, FP-3H, BP-3H) further. This consisted of two steps:**
    a. **We performed leave-one-out cross validation to assess how sensitive the parameter estimates were to individual landslide events. Since AIC, BIC, and LOOIC evaluate fit across the entire dataset (including the many non-landslide days), we wanted to specifically check the sensitivity to a potentially missed landslide event, and to assess how well the model might be able to predict a not yet observed landslide event. Because we estimate daily probability, we argue that leaving out a single day is a more appropriate check than leaving out an entire year, as the reviewer suggests.**
    b. **We used the BSS to compare the FL-3H model's skill to a simpler alternative model based on historical frequency. This is the first step of our workflow that evaluates skill. This approach is akin to a common strategy for evaluating weather forecasting models, in which the model's skill is compared to climatology: average weather conditions over long time periods (e.g. Wilks, 2011). In our case, climatology is replaced with historical landslide frequency. We note the important distinction between evaluating a model's predictive skill vs. evaluating its skill as a classifier. The Brier Score checks the predicted probability against the outcome (e.g. model predicts 0.8, and landslide occurs). This provides us with different information than the AUROC, which evaluates the model's skill as a classifier, and is thus a complementary metric. Classification requires choosing a decision boundary, which we will discuss below, but by providing and evaluating predicted probabilities, we offer more information for decision making than by only providing a classification. Therefore, we disagree that this section should be cut, but we will clarify the motivation for it in the updated manuscript, and present the results more clearly.**

Decision boundaries

3. **Threshold selection. Based on the statistical modeling results (FL-3H) and input from the Sitka community, we selected decision boundaries for implementation in the warning system by evaluating precision and recall. The reviewer suggests computing the AUROC to assess whether the model is better than random. The AUROC for FL-3H is 0.9993, and we include the ROC curve below (Figure AC2). This metric indicates that FL-3H far outperforms a classifier with no skill, and indeed is a near perfect classifier. However, in a strongly imbalanced dataset with near perfect separation between landslide days and non-landslide days like ours, we argue that the AUROC is a somewhat misleading metric, as its high value partially results from the large number of potential thresholds with a true positive rate of 1. While we could report this, it would be yet another step, which we do not feel adds much information. While the ROC curve can help to choose an optimal threshold, we argue that the Precision-Recall curve is more useful for this application in our case, because it better illustrates the tradeoffs between missed alarms and false**

alarms (Figure 11). The optimal threshold found with the ROC curve (upper left corner) is identical to the threshold that maximizes recall.

4. **Threshold evaluation. Finally, we evaluate the threshold performance by splitting the dataset into training and test sections. This is necessary to understand how well the warning system might perform in the future. In this case, we wanted to evaluate how many true alarms / false alarms / missed alarms / true no alarms would result from using the thresholds during the test period, rather than test the sensitivity of the statistical models to individual landslide days.**

We have clarified these three separate goals for statistical evaluation in the abstract (lines 18-20). We have updated the methods introduction, as well as sections 2.3 and 2.4 to better explain and motivate each step. We have also added a more thorough treatment of the BSS results to the results in section 3.3.

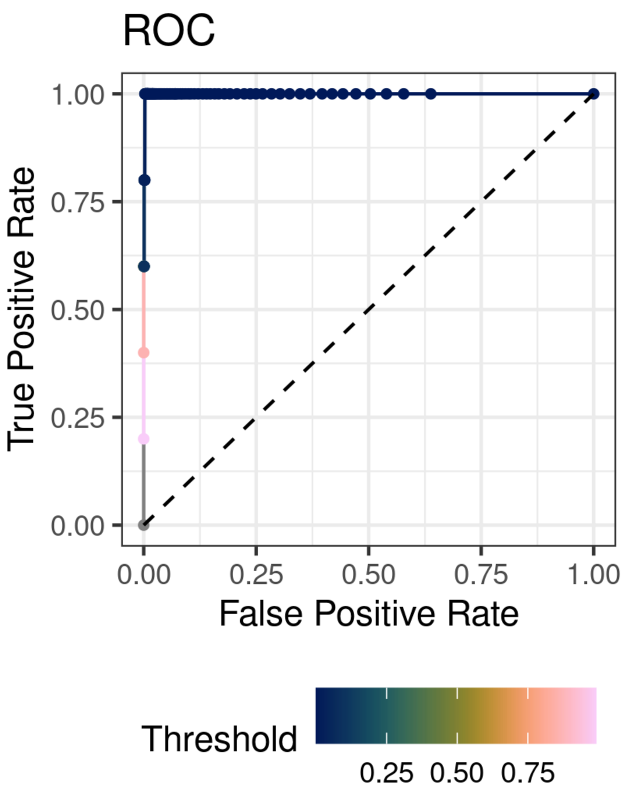

Figure AC2. ROC curve for FL-3H. The dashed line represents a model with no skill for comparison.

Specific comments

L17: Please specify «136 statistical models». When only reading the abstract I cannot imagine what that means. How do models differ?

We mention in the abstract that the 136 models have different timescales of precipitation variables. In our manuscript revision, we can clarify the wording of this statement. Further detail is provided to the readers in the methods section.

L18: does that mean that your data is in daily resolution? If not, I would count the number of non-triggering rainfall events instead of the days.

Our models are based on cumulative precipitation over a range of timescales and estimate daily landslide probability and intensity, therefore, we report the number of days. Our method avoids the need to define rainfall "events."

L23-25: seems more like a conclusion to me. I would state this later in the abstract.

In our revised manuscript we have moved this conclusion later in the abstract.

L34-72: the intro nicely shows that there is a need for a LEWS in this region but that it's difficult to establish with current methods

Thank you!

L53-54: I think this sentence should end with a citation

Several studies exist which can support this statement ("Accurately predicting rare events like landslides remains challenging because the complex and spatially heterogeneous processes that drive landslide initiation are difficult to characterize at sufficiently high resolution across broad regions") We add a few citations in our revised manuscript.

L153: general question: what is the added value of the number of landslides if you don't know where it's going to happen?

We estimate the number of landslides to provide information on whether widespread landsliding with multiple failures could be expected, or whether isolated landslides are more likely. This is valuable information for planning, even if the exact locations of the landslides aren't modeled.

L217: So 1-day antecedent precip is the total daily rainfall at the day of the landslide?

1-day antecedent precip is the total daily rainfall on the day *before* the landslide.

L230: Please give a citation for these equations

We have added a citation to the beginning of section 2.2 (McCullagh and Nelder, 1989, Generalized Linear Models) which describes both models.

L235: Please integrate this sentence in another paragraph where you discuss this problem

We have integrated this sentence into section 2.4.

L247: more intuitive than what?

Added that we mean more intuitive than frequentist confidence intervals.

L271: What is a chain?

Hamiltonian Monte Carlo, the sampler used to estimate the Bayesian posterior distributions, relies on Markov chains. The sampler progresses through the parameter space from point to point: a Markov chain is the record of these points (e.g. Betancourt, 2018). Essentially, what we mean when we say that we ran four chains for 2000 iterations is that four independent instances of the sampler progressed through 2000 points. We have specified that we mean Markov chains in the text.

L276-279: how are these criterions defined?

We have added citations to the text in section 2.3 which present these criteria in detail.

L275-289: why was not the same validation performed on each model?

After selecting the three-hourly time scale from among all considered models, we focus our efforts on evaluating these models for the sake of simplicity and timely production of scientific products.

L332.335: I agree about the inadequacy of accuracy in imbalanced datasets but ROC shows exactly true alarms and false alarms in relative terms.

We will remove the reference to ROC from this sentence.

Fig. 3&4: please make sure these figures have the same layout (axis limits, labels, font size, order of plotting lines, etc.) (same for fig. 5&6). As these figures look very similar, you could consider e.g. only showing figs 3&5 here and move the others to the supplement.

We find it informative to be able to compare the results of frequentist and Bayesian analysis, so we prefer to keep figures 3 & 6 in the main text. We have prepared modified versions of figures 3-6 to ensure that they have the same color scheme and font.

Fig. 7/8: blue stands for better and red for worse. Better and worse compared to what?

"Better" and "worse" refer to the AIC and BIC values used for model selection, where lower values (lower predicted error) are preferred (Kuha, 2004). The red and blue coloring is simply intended to help visualize the values in the tables for readability. Following your previous suggestion, we have added citations for AIC and BIC to the main text. We also adjusted the captions to indicate that model fit relates to estimated prediction error.

Fig 11: Generally a very nice figure. Some comments:

Thanks!

- Maybe the color scheme for the landslide probability could be optimized, e.g. change in color where you set your threshold (at values mentioned in L508)
  - **The color scheme is "batlow," a scientific color map that is optimized to be perceptually uniform and universally readable (Crameri, 2018). We prefer to mark**

the threshold transitions using symbols instead.

- Why is the lower threshold not slightly higher where recall is still 1?
  - **We chose to lower the threshold below the optimal threshold as a conservative approach to account for uncertainty in the probability estimates, as described in section 3.4 and discussed in 4.3.**

- When following this line from left to right, are you sure you can increase recall and precision at the same time? This is a univariate model, right? I can't think of how this would happen. Since you have 5 events, should't the steps be in intervals of 0.2 for precision?
  - **Yes, it is possible to increase recall and precision at the same time in this univariate model. For example, at the step where the number of true alarms increases from 3 to 4, the number of false alarms stays at 7, and the number of missed alarms decreases from 2 to 1. This leads to a simultaneous increase in recall from 0.6 to 0.8 and precision from 0.30 to 0.36. With 5 events, the steps for recall (TA/(TA+MA)) are 0.2, but the steps in precision (TA/(TA+FA)) are influenced by the number of false alarms as well.**

- Caption: I would simply refer to the equations for definitions of recall and precision, but there you could mention the alternative names (e.g. precision=true positive rate)
  - **Thank you for the suggestion, we will modify the caption accordingly.**

L579: yes, compared to some shallow landslide thresholds this is rather low. Could it also be because some of the events are debris flows? The most predictive thresholds for runoff-triggered debris-flows can be at the 10-min timescale.

Interesting question! Field observation suggests that the landslides in our study are shallow landslide failures which transitioned into debris flows. We have not observed runoff-generated debris flows in this study area. Similarly, we did not observe any signs of overland flow in the study area, but there are other small, shallow landslide scars in the region.

L610: please specify "hydrologic monitoring". In this context, I assume soil moisture measurements.

We can specify in section 4.1 that hydrologic monitoring includes soil moisture, groundwater level, and soil water potential.

L619: I would say with "few landslide events" instead of "without". I don't think you investigated threshold determination without triggering events.

We have modified this language in section 4.2 to refer to "few landslide events."

L620-L625: you are of course right that negative events should be considered and in practice it may still be done only occasionally. However, this has been well-known for a while. The first ones I can think of are Staley et al. (2017, https://doi.org/10.1016/j.geomorph.2016.10.019) and Gariano et al. (2015, https://doi.org/10.1007/s11069-019-03830-x) and since then many others have adopted this procedure, some of them you cite earlier.

This section intends to emphasize the potential for precipitation thresholds even when very few landslide events can be used to train the models, which is best reflected in the recent Peres and Cancelliere (2021) paper we cite here. The works by Staley et al. and Gariano et al. you note are good examples of landslide prediction; we added reference to Staley et al in section 1.3 (publication year for the link you included is 2017, *Geomorphology*). We cite Gariano et al. in section 4.1 (publication year for the link you included is 2020, *Natural Hazards*)

L639-644: isn't this contradicting the earlier statement in L623-625 about the value of low precipitation totals? By using precision and recall you get rid of exactly these.

As described in section 3.4, we selected thresholds using a heuristic approach which incorporated several sources of information, including Precision-Recall as well as a confusion matrix and qualitative assessment of risk tolerance.

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.