**Authors response to comments by referee #2**

Manuscript reference number: nhess-2018-85

Manuscript title: Evaluation of predictive models for post-fire debris flows occurrence in the western United States.

We would like to thank the reviewer for his/her constructive comments. We have revised the manuscript according to these comments and below we provide our detailed response to each of their comments. Comments from the reviewer are in black font and our response in blue font.

Note: Revised manuscript is attached as supplementary material.

**Reviewer #2**

**2.1** This paper is intended to present an improved model for forecasting debris-flow occurrence in recently burned areas. The presented material relies exclusively upon previously published data from other sources, and existing methodology. Given the relatively narrow target audience and lack of uniqueness of this topic, my criteria for recommending the publication of this type of study are as follows: 1) New insight into the mechanisms responsible for postwildfire debris-flow occurrence must be gained. 2) Unique methodology for improved predictions must be presented. 3) The methods used to compare model predictions must be defensible, and account for uncertainty in model predictions.

Response

We would like to thank the reviewer for providing his/her comments. Below we provide our detailed response to the points raised.

**2.2** While this paper is well-written, significant revision is required in order to meet the specified criteria for the following reasons:

1) No new insight into the prediction of postwildfire debris-flow occurrence or the pro- cesses responsible for debris-flow generation have been presented within.

Response

We respectfully disagree with the reviewer’s opinion that “no new insight into the prediction of post-fire debris flow occurrence” have been presented in this work. We have presented a systematic evaluation of state-of-the-art modeling procedures (two of the models developed in this work) and demonstrated the relationship of their predictive performance with data requirements (sample size and variables considered). Furthermore, in the revised version of the manuscript and per this reviewer’s suggestion, we present a comparison of the predicted DF probabilities from LR and RF-based models that allow us to demonstrate with more clarity the superiority of the proposed RF-based vs LR model (currently used by USGS in West US). Therefore we believe that the findings of this work provide collectively new information regarding the approaches for predicting post-fire debris flow occurrence.

It is true that no new insight is presented regarding the understanding of physical processes. We agree that better understanding of the processes responsible for debris flow generation is very important and that it is required to ultimately improve predictive modeling, however process understanding would require a vastly different research approach that is out of the scope of this work.

**2.3** The authors used an existing database of debris-flow occurrence, storm characteristics, and a select number of variables to test an existing modeling approach.

Response

Yes, we relied upon an existing USGS database, which was also used for the development of the LR model in Staley et al. (2017), for the following reasons:

1. It is a database with a large number of events and information associated to burned areas. To the best of our knowledge a similar database that is publicly available was not found for west US.
2. Given that one of the main objectives of the work was to compare against the current state-of-art model for post-fire DF prediction in west US (developed based on this database), it is appropriate to systematically evaluate the different models (proposed and existing) using the same dataset.

**2.4** The authors did not attempt to incorporate different metrics of topography, fire severity, or other physio- graphic properties of watersheds into this analysis.

Response

During the development of RF-based models we tested a number of combinations of existing variables. In the manuscript, we presented results for the “simple” RF-ED and best performing RF-all model. We did not attempt to generate new parameters (based on topography, fire severity etc.) for the watersheds. This is a significant effort that although could potentially prove beneficial, should be considered as a future step and most likely in collaboration with USGS to get access to the original dataset used to derive the parameters in database (e.g. GIS files of burned areas, satellite images before/after fire etc.).

We believe that the analysis based on the existing dataset still has merit and findings provide new information with important implication for future advances of post-fire DF prediction not only in the west US but elsewhere as well.

We revised conclusions to mention that investigation of new parameters should be considered as a future step:

*“Future work will be focused on examining the model performance using alternative sources of rainfall information (e.g. from weather radar, satellite-based sensors and numerical weather prediction models) and further investigating how extra physiographic parameters (not included in existing database) can potentially improve predictive ability of models.”*

**2.5** As such, the author’s solely relied upon someone else’s calculation of metrics and interpretations of the variables that are important for inclusion in this analysis. Hence, we did not learn anything about new factors or combinations of variables that may be relevant for the prediction of postwildfire debris-flow occurrence.

Response

We would like to note again that a direct and fair comparison of the proposed RF-based models to the currently used procedure for post-fire DF prediction in west US, required us to adopt fully the specificities of the currently used LR model. Otherwise, comparison would not be possible or consistent.

We believe that results from the analysis provide new information that can be useful for the development of predictive models for post-fire DF. Specifically, results from this work provide insight on the relative importance of data requirements (both in terms of size and variables) and model complexity for predicting DF occurrence. Additionally, the revised version of the manuscript provides a more detailed comparison between LR and RF-based models highlighting a) the issues to be considered for further development of LR-based procedures (if possible) and b) an alternative way forward in DF prediction with RF-based models quantifying also their corresponding significance in performance improvement.

**2.6** 2) The authors present a method that utilizes an existing methodology (random forest analysis). It’s unclear to me how or why this is significantly different, or a significant advancement, over the machine learning method presented by Kent et al., 2017. It seems that the comparison of model predictions focuses almost exclusively on those between the models presented by the author and the USGS logistic regression model.

Response

Random forest algorithm is an existing algorithm that was used to develop the models presented in this work RF-ED and RF-all. These specific models have not been previously presented. As we argue in the conclusions section, there may very well be other machine learning algorithms that can potentially generate even further improved results. To be able to investigate this, one needs to carry out a systematic evaluation for all the possible models, which is not within the scope of this work. The models evaluated in this work were selected based on a very specific context. First, the ED model is the most widely used threshold for rainfall-induced mass movement phenomena (landslides, debris flows etc) and at the same time the simplest in terms of data requirements and development. The LR model was chosen because it is the current state-of-art and the RF-based models were developed to demonstrate the potential advancements that can be achieved by considering machine learning techniques for post-fire debris flow prediction.

In addition to this, the work presents aspects of the performance evaluation relative to data requirements (sample size and number of variables considered) and in the revised version of the manuscript, it further explores in more detail the differences in DF probability prediction between LR and RF models. Such an evaluation, contrasting in particular the currently operational model in west US does not exist in previous studies.

**2.7** 3) The manuscript provides methodology for a binary classifier model used to predict the occurrence of a postwildfire debris flow. The authors compare their predictions to the USGS logistic regression model, but significantly simplify the USGS model predictions. Logistic regression analysis produces a continuous estimate of the likelihood of debris flow occurrence with values that range from 0 - 1, with 1 representing the highest likelihood. In this study, the authors consider the LR model as a binary classifier model, failing to consider that the model is not inherently binary. The analysis then proceeds to use ROC metrics for binary classifier models to compare the results. The Staley et al 2017 manuscript that the authors draw their comparative statistics from does use ROC metrics to compare findings for a portion of the results and discussion. In my opinion, this is a significant flaw in the Staley et al. 2017 manuscript.

By only viewing the USGS model as a binary classifier, the authors oversimplify the predictions made by the USGS logistic regression model by not accounting for the implicit uncertainty in the logistic regression model predictions. For example, the USGS model might estimate a 0.51 likelihood for a debris flow (i.e. just slightly better than a coin toss), but no debris flow might have occurred. This manuscript considers this to be a failed prediction, but is it really? I would argue that a model that estimated a 0.49 likelihood of not having a debris flow is not incorrect. Hence, the conclusion that the presented model significantly outperforms the USGS model predictions based upon the analysis of the model as a binary classifier is misleading.

Response

We would like to thank the reviewer for providing a thorough discussion on this issue.

Indeed, the example mentioned by the reviewer was one of the reasons that we allowed the probability threshold (used for binary classification) to be set through optimizing TS instead of assigning it a fixed value (e.g. 0.5). This means that if we were dealing with a systematic difference between the LR and RF predicted probabilities, the use of a dynamic probability threshold would take care of this effect and we would not notice a difference in the binary classification performance.

However, we agree with the reviewer that a more thorough evaluation of the DF probabilities predicted by LR and RF-based models would benefit the analysis and strengthen the finding of this work. For this reason, we carried out extra analysis that investigates this aspect. In the revised version of the manuscript we have added a new section (5.3) that presents and discusses results from the comparison between the distribution of the predicted DF probabilities for the different models (LR, RF-ED and RF-all).

As it is shown in the results of the new Figure 7, the superiority of RF-based models in better separating probabilities associated to DF and noDF events is evident. The LR model suffers from considerable underestimation of the probabilities associated with DF events and overestimation of those associated to noDF events, which creates the issue of “overlap” and the difficulty in accurately identifying between the DF and noDF events.