Thanks for Dr. Scaringi's valuable comments to the paper at first. His comments were very useful to increase the quality of the paper.

(1) line 145 - I understand that the inventories were made through visual interpretation. It would be good if the authors specify this here rather than at line 150 (which refers only to the most recent images). Furthermore, it would be good to specify if and how the authors evaluated the mapping uncertainties due to low imagery resolution and visual interpretation, for instance in terms of shape and size mismatch and amalgamation, and their propagation to landslide statistics (e.g.frequency-area distributions, classification by controlling factors).

Response: Indeed, we agree with your comment, and modified the text. The landslide inventory pre-2015 was based on three data sets. The pre-2015 inventory map was generated using topographic maps, multi-temporal Google Earth Pro images and Landsat ETM/TM images. We were able to digitize landslide polygons from the available 1:50,000 scale topographic maps, which cover only the Nepalese part of the Koshi River basin. These maps were generated from aerial photographs acquired in 1992, and active landslides with a minimum size of 450 m^2 visible on these images were marked as separate units. A set of pre-2015 Landsat ETM/TM images were available for the entire study area, from which the post 1992 and pre-2015 landslides were mapped. Pre-2015 landslides were also mapped from historical images using Google Earth Pro Historical Imagery Viewer which contains images from 1984 onwards. Although the oldest images are Landsat images, the more recent ones have much higher resolution, although not covering the whole study area in equal level of detail. By comparing the different images for the period between 1992 and 2015 we were able to recognize most of the landslides. We carried out field verification for a number of samples and could conclude that through the image interpretation we were able to map landslide with a minimum size of 50 m^2. Images from Google Earth were downloaded and geo-referenced and landslides were mapped using visual image interpretation and screen digitizing. A total of 5,858 rainfall induced landslides were identified in the Koshi River basin.

(2) line 168 - Also here, it would be good to specify how the rather low spatial resolution of the GlobeLand30 (30x30 m) affects the classification especially of landslides with small area (as low as 50 sq.m).

Response: We agree with your statement and we have also modified this in the text: Given the rather low resolution of the input data, the relation with landslides as small as 50m^2 may not be optimal, especially also considering the rather long time period over which land cover changes have occurred in many areas. But given the regional scale of this analysis, the use of higher resolution data was unfortunately not a viable option.

(3) line 176 - Here it would be nice to explain the 60%-40% choice (is it because of the sample size? is it arbitrary?) and to specify how the landslides are assigned to either set (e.g. randomly, but being sure that the size distribution and controlling factors classification are the same in both sets?).

Response: Thank you for your comment. It is a generally accepted method in literature to separate the landslide dataset into a training and validation set (e.g. Hussin et al. 2016; Reichenbach et al., 2018). We decided to select 60% of the landslide data as training data for the modeling, and 40% for the validation. Here is comment on this matter from an expert on ResearchGate: “A common practice is to split the data set into L and T as 2 : 1. There is no profound justification for this; neither there is it clear, whether different splits yield less precise results. The result of a split is an assessment of the quality of the prediction by the model. Such an assessment is subject to uncertainty because the split entails randomness. An ideal split is associated with very small variation of the results. By a split we balance the uncertainty associated with the model (large L is preferred for that) and with evaluation (large T is preferred)”. See also the below, from Hussin et al., 2016.

Citations	Size of study area	Pixel resolution	Nr. of landslide pixels	Model ratio landslide : non-landslide pixels	Performance or validation rates
Van Den Eeckhaut et al. (2006)	200 km^2	10 m	Training: 93 pixels Prediction: 23 pixels	1:5	AUC ROC 0.91 – 0.98
Hjort and Marmion (2008)	600 km^2	25 ha (500 m)	200 or more pixels	1:1	Mean AUC ROC 0.90
Blahut et al. (2010b)	450 km^2	10 m	Training: 21923 pixels Prediction: 21923 pixels	1:206	AUC SRC: 0.87 AUC PRC: 0.88
Regmi et al. (2010)	815 km^2	10 m	Training: 368 pixels Prediction: 369 pixels	1:22147	AUC SRC: 0.77 AUC PRC: 0.74
Van Den Eeckhaut et al. (2010)	1120 km^2	50 m	64198 pixels	1:1	AUC ROC 0.90-0.92
Piacentini et al. (2012)	7500 km^2	20m	Training: 617 pixels Prediction: 185 pixels	1:30389	AUC SRC: 0.80 AUC PRC: 0.76
Felicísimo et al. (2013)	140 km^2	10 m	340 pixels	1:2	Mean AUC ROC 0.76 – 0.78

Heckmann et al. (2014)	19 km^2	5 m	81 pixels	1:3.7 - 1:4.3	Mean AUC ROC 0.83
Petschko et al. (2014)	15850 km^2	5 m	50 to 12562 pixels	1:1	AUC ROC 0.76 – 0.84

(4) line 216 - Here you classify the landslides into small and large depending on "field experience" and on the basis of the frequency-area distributions. You choose 6000 m2 as your threshold which is more or less the cut-off value in the frequency-area distribution of the earthquake-triggered landslides but is much smaller than that of the rainfall-triggered landslides. However, the cut-off (or rollover point) may be affected by under sampling of small landslides, which you should be able to rule out explicitly. Also, what field experience means in this context remains unclear. So, this threshold area seems quite arbitrary. I would encourage the authors to introduce a physically-based justification for this choice, which you did in part already in the introduction. On the other hand, I would also suggest that you run your model multiple times with different thresholds, to show if there is an optimal (data-driven) threshold that can best differentiate the statistics of RTL and ETL in your study area. This threshold will certainly have a hidden physical meaning, which could be then discussed

Response: The landslide inventories in the Koshi River basin show similar cut-off values, around 30,000 m^2 for different triggers (rainfall and earthquake). Here we should take in mind, however, that the two rainfall-triggered landslide inventories are not event-based inventories (Guzzetti et al., 2012). The two inventories differ in the sense that the 1992 inventory is based on landslides that were large enough to be mapped on the topographic map, where as the inventory between 1992 and 2015 represents the landslides that could be mapped from multi-temporal images over a number of years. Although the two inventories differ substantially with respect to the number of small landslides, it is striking to see that the cut-off values, and β values are similar. It is very difficult to obtain a complete event-based landslide inventory for rainfall inducedlandslides in Nepal, as landslides are generally generated by a number of extreme rainfall events during the monsoon, which can not be separated, as the area is cloud covered through most of the period. The size-frequency distributions for both ETL and RTL show very similar behaviour for landslides above the cut-off value of 30,000 m^2. Landslides are generally classified in terms of area and volume. But landslide volume is very difficult to measure, as it requires high quality multi-temporal Digital Elevation Models, and knowledge on slip surfaces (Jongmans and Garambois, 2007). In practice , landslide classification is mostly based on area, and in China the Tong et al. (2013) proposed a classification with landslides with an area smaller than 10,000 m^2 as small, those with an area between 10,000 m^2 and 100,000 m^2 as medium, and those with larger sizes than 100,000 m^2 as large size landslides. Based on the results of the FAD analysis, that resulted in similar cut-off values for the RTL and ETL and similar β values, we subdivided them into two size-groups, with 30,000 m^2 as threshold value (Table 1). The results will therefore be more reliable for the class above the threshold of 30,000 m^2 , where under sampling is not an issue, then for the small landslide class, which have different rollover points, and completeness levels.

References:

Guzzetti, F., Mondini, A.C., Cardinali, M., Fiorucci, F., Santangelo, M. , Chang, K.T. (2012) Landslide inventory maps: New tools for an old problem. Earth-Science Reviews 112:42–66

Hussin HY, Zumpano V, Reichenbach P, Sterlacchini S, Micu M, van Westen CJ and Balteanu D (2016) Different landslide sampling strategies in a grid - based bi - variate statistical susceptibility model. Geomorphology, 253: 508-523

Jongmans D, Garambois S (2007) Geophysical investigation of landslides: a review. Bulletin Société Géologique de France 178(2):101–112

Reichenbach P, Rossi M, Malamud BD, Mihir M and Guzzetti F (2018) A review of statistically-based landslide susceptibility models. Earth-Science Reviews, 180: 60-91

Tong L.Q., Qi L.S., An G.Y. et al. (2013) Large Scale Geological Hazards investigation by Remote Sensing Technology[M]. Science Press.