Response to reviewer #1

Thank you very much for the constructive comments on our manuscript. We are very thankful for your huge efforts. Converning your comments, we reply below.

General Comments

This manuscript uses a Quantile Regression Forests statistical tool to model sediment concentrations and yields from logged watersheds in Chile. The authors find that the QRF model outperforms the more traditional sediment rating curve approach to modeling sediment yields, in that sedigraphs predicted by the QRF model more closely agree with measured values than sedigraphs predicted by the rating curve approach. Given the promise shown by the QRF method, this work will likely garner the interest of scientists and land managers that are engaged in the prediction of sediment yields from a wide range of landscapes. Overall, I think this is interesting work that appears to be methodologically sound, but I do have several general comments/questions:

The rating curve method uses only one variable (discharge) to predict sediment concentrations. In contrast, there are 21 variables in the QRF model and it would be useful to know if pared-down versions of the model perform as well as the full model, with respect to the predictions from the rating curve method. In other words, with only discharge data as an input, does the model still outperform the rating curve method?

Thank you for this comment. We agree that the seemingly better performance of Quantile Regression Forests (QRF) over sediment rating curves (SRC) will partly be due to the higher number of predictor variables used, although this (perhaps not surprising) finding is far from the main point that we emphasize in our study. Both SRC and QRF are regression techniques. However, the latter is using randomized subsampling of the data in order to avoid overfitting. In addition, QRF is a decision-tree based method and thus not suitable for using a single predictor. Thus a pared-

down comparison of SRC and QRF is not straightforward, as the SRC approach would need to be based on multivariate regression instead.

Besides the seemingly better performance, our QRF approach has two major advantages: First, the majority of predictors, i.e. 18 out of 21 are derived from only two time series, i.e. rainfall and discharge. Consequently, our model is based on five independent predictors and 16 sub-predictors. The sub-predictors were derived such that the time windows do not overlap in order to avoid potential correlation effects. The predictors were selected in order to describe and quantify the antecedent conditions and the intra-event dynamics. Our results indicate that both antecedent preconditions and intra-event dynamics are essential for understanding sediment flux under the given disturbed conditions of a recent clear cut. These conditions cannot be captured by sediment rating curves though they provide a straightforward method for a first-order estimate. Second, the comparatively low temporal, e.g. daily, resolution of SSC sampling in many previous studies has hardly captured any intra-event dynamics. Thus, we also argue that a presumably good SRC fits may be biased because of hysteresis effects, i.e. differences in suspended sediment concentration between raising and falling discharge limbs. Our QRF approach, in contrast, is suitable to simulate such discharge-concentration hysteresis.

Comparison of modeled and measured annual sediment fluxes are presented in Supplementary Table 3, but little of the manuscript text is devoted to discussing the differences. It would be beneficial for the authors to more fully explore/explain these results and to explain (if possible) the underlying differences between the modeled and measured values.

We agree that the differences between SRC- and QRF-predictions are important to address. However, we are convinced that we already addressed this issue in the manuscript. One page 321, lines 11-21 we write: "*For one, we treat our bulk sediment flux measurements as minimum estimates given their low temporal resolution compared to the fast hydrological response. Under such restrictions, we assume that they do not*

fully capture potentially high but short-lived SSC during intense rainfall events. Thus, the total sediment yields based exclusively on bulk samples are a lower bound estimate (Table 2). Furthermore, we find that conventional sediment rating curves (SRCs) are sensitive to outliers, resulting in implausible high SSC (e.g. 10-15 g s^{-1}*; Figure 7e), but remain below our QRF predictions on average (Table 2, Supplementary Table 5). Under the recorded low-flow regime (Huber et al., 2010), SRCs underestimate the hydro-geomorphic work of more frequent though lower sediment fluxes, while they overestimate the less frequent higher-magnitude events. This finding supports earlier work arguing that SRCs significantly underestimate sediment fluxes (e.g., Asselman, 2000).***"***

In order to increase the visibility of our data, we further moved Supplementary Table 3 into the results section of the main text, where it is now referred to as Table 2.

There should be more information about the timing of logging and post-logging treatments, as the introduction seems to make the case for a need to assess the impact of clear-cutting, but the discussion indicates that decreased suspended sediment yields following dry season logging may be due to replanting.

We agree that more information on the timing of the logging is required. We expanded the text (study site section) as follows: "*Two catchments previously planted with Pinus radiata were logged by the same clear-cutting technique during different seasons: catchment #3 was clear cut during the winter rainy season (Jul-Aug 2009), and remained bare for ~1 year, whereas catchment #4 was harvested during the end of dry summer season (Mar-Apr 2010), and replanted in early spring 2010 (Sep-Oct 2010) (Figure 2a). Both catchments were reforested by Eucalyptus globulus (Schuller et al., 2013). Although clear-cutting is permitted under the Chilean standards, the forest companies are requested to adopt best management practices in accordance with Forest Stewardship Council certification agreements. Among others, these consider cable harvesting on slopes >30%, the use of ground skidders in areas of lower slopes, the maintenance of riparian buffer strips (which in the study sites are ~7.5 m wide both sides of the channel network) and piling up forestry residues along contour lines at the end of*

*the harvesting operations. The logging of catchment #4 severely damaged the riparian buffer strip whereas the buffer strip in catchment #3 remained unaffected by the timber harvest. Overall, ~88% of the area of catchment #3 were logged, and more in catchment #4. The clear cut was done using heavy rubber-tired skidders to drag logs uphill to landings whereas cable logging was limited to steep slopes (Mohr et al., 2013) (Figure 2b). The loggings covered the entire catchment area including their ridges. Catchment #1 remained unlogged and covered with P. radiata, and served as a control catchment.***"***

In addition, we added a recent paper by Schuller et al. (2013) who worked in the same study sites.

The implications of this study for geomorphic work do not seem to be as clear as those presented in the manuscript. The highest discharges measured during the study may be extreme with respect to the other values in the dataset, but a longer-term record is not presented (for either discharge or precipitation) that demonstrates that the discharges are extreme with respect to annual exceedence probabilities. More context is needed to demonstrate that these findings differ in a substantial way from, for example, the view put forth by Wolman and Miller (1960).

We acknowledge that there might be an ambiguity regarding the definition of "extreme events". In this study, we define extreme events as the values exceeding the 95th percentile of our data set. It is correct that we cannot provide evidence that such events are extreme with respect to annual exceedance probabilities. However, our intention was to show that the relevance of extreme events in terms of geomorphic work can be scaled down to high-frequency time series as afforded by our dataset. Clearly, we are not capable of judging with our data the role of extreme events on longer (annual to millennial) timescales.

Specific comments Page 313

Lines 1-2: Avoid leading off the manuscript by bringing up a discussion about man-made forests, especially because this topic is not addressed again in the paper.

Thank you for this advice. We agree and have deleted this sentence.

Line 8: I don't believe the road-related landslides documented by Montgomery et al., were triggered by the failure of road cuts, but the change in upstream drainage area caused by the construction of the roads.

Thank you very much for this constructive comment. We have changed the wording accordingly.

Lines 20-25: It is unclear why replanting specifically, requires a technique capable of dealing with few samples collected under varying conditions. Put another way, the tools introduced in this paper can likely be put to a much broader use than for assessing sediment yields from replanted clear-cuts in Chile, as a small number of samples, high variance, and changing environmental conditions are inherent to a broader range of scenarios where sediment yields need to be quantified. It would be worthwhile to present a broader utility of the techniques developed as part of this study.

Thank you very much for this very reasonable comment. We agree that our technique is promising for different types of disturbances with limited available sample size. We therefore modified the manuscript as follows: "*Reliable knowledge of pre- and post-disturbance sediment fluxes is vital in this regard, and may be acquired by physics-based modelling or statistical treatment of field data. This holds particularly for Chile where law mandates immediate reforestation after clear-cuttings. However, in many situations sample size for a robust assessment remains limited, because both time and resources for sampling hydro-geomorphic impacts are often tightly constrained; hence the acquired field data may not represent the full range of water and sediment fluxes. This limited data availability requires an analysis technique capable of dealing with few samples of high variance under changing environmental conditions (Figure 1a).***"**

Line 7: SSC measurement were not made every three minutes, but every 30-60 minutes, which was not the impression one gets from reading lines 10-11 of the Abstract, which states sediment concentrations were measured every three minutes. I'd encourage you to report the data that were collected in an unambiguous manner.

We are sorry for the ambiguity which was not intended. We changed the manuscript text accordingly.

Page 321

Lines 11-12: As presented, it is unclear why the bulk sediment flux measurements are minimum values.

In order to improve the clarity of the text, we added one sentence to the end of this paragraph: "*For one, we treat our bulk sediment flux measurements as minimum estimates given their low temporal resolution compared to the fast hydrological response. Under such restrictions, we assume that they do not fully capture potentially high but short-lived SSC during intense rainfall events. Thus, the total sediment yields based exclusively on bulk samples are a lower bound estimate (Table 2).*"

Page 322

I'm not convinced that the sediment transport events that were measured are extreme. These data are not put into the context of a longer record, but the recurrence intervals seem to be < 1 yr. More context is needed to demonstrate that these results differ in a substantial way from the view put forth by Wolman and Miller (1960).

Please see our previous reply. In this study, we treat extreme events as the values exceeding the 95th percentile of our data. This is what we had stated earlier by noting that our results "significantly expand down to the process time scale": earlier work has largely shown the dominance of extreme events on much longer observations intervals.

Supplementary Table 3

The results presented in Supplementary Table 3 contain information that most readers will want to know: how do total sediment yields predicted by the QRF model compare to those predicted by the sediment rating curve method, and how do both model predictions compare with measured data. These data need to be more fully integrated with the main text of the manuscript. Currently, the only table in the manuscript presents the number of samples, whereas the comparison of the model predictions are much more important.

In order to better integrate the data presented in Supplementary Table 3, we moved the table into the results section as Table 2. For model and observation comparison, please also refer also to our previous reply.

Editorial comments Page 313

Line 10: I suggest revising this sentence, as it is not clear what is meant by "the long-term decay of soil conservation functions".

We changed the sentence to "*As a result, boosted erosion and re-deposition of soil promote the long-term degradation of soil and water resources not only on harvest patches, but also often in downstream areas (Sidle et al., 2006).*"

Page 315

Line 18: Suggest changing "gauges" to "weirs".

Done and changed accordingly.

Line 22: Suggest inserting "rain gauge" following "bucket".

Done and added.

Line 22-24: Suggest revising to: "*A Wilcox rank sum test was used to assess whether hourly rainfall intensities differed significantly ($p \leq 0.05$) between each year.*"

Done and changed accordingly.

Line 24: It is unclear what "bulk monitoring data" refers to, total sediment yield?

In order to improve clearness, we changed the sentence to: "*Total sediment yields estimated by bulk samples from June 2008 to September 2009 in these and adjacent catchments indicate that pine plantations were more prone to soil erosion than eucalyptus plantations (Huber et al., 2010).*"

Page 316

Line 20: replace "larger" with "longer"

Done and corrected.

Lines 22-24: It is unclear what information is trying to be conveyed by this sentence.

We agree and have simply deleted this sentence.

Line 25: It would be useful to explicitly define what an "integrated sample" is, as this may clear up the somewhat confusing text in this paragraph.

We changed "integrated" to "bulk" as follows: "*We considered these data as first-order benchmarks for the modelled sediment fluxes. Any given bulk sample merged four samples each day over a period of one week (Huber et al., 2010).*"

Page 323

End of line 12: it is unclear what is meant by “they”.

We changed this to: “*When logged, pine roots rapidly decay in root-strength (Sidle, 1991).”*

Figure 2 caption: Change “base” to “basis”.

Done and changed accordingly.

Figure 7: It is difficult to see red crosses in many of the panels.

Thank you very much for this comment. In fact, only one panel shows the red crosses indicating SSC measurements. We agree that the red crosses are not easy to identify, and have slightly enlarged the crosses for better legibility. We also improved readability of Figure 6.