Nat. Hazards Earth Syst. Sci. Discuss., 1, C1799–C1824, 2013
www.nat-hazards-earth-syst-sci-discuss.net/1/C1799/2013/

Natural Hazards and Earth System Sciences
Discussions

Interactive comment on "Sample size matters: investigating the effect of sample size on a logistic regression debris flow susceptibility model" *by* T. Heckmann et al.

T. Heckmann et al.

tobias.heckmann@ku.de

Received and published: 18 November 2013

Reply to review by anonymous Reviewer 2

We thank the reviewer for his/her extensive and thoughtful review. Below, we address the issues raised in the review, and outline changes that we made in our manuscript. On the following pages, the comments are written in boldface, followed by our replies in standard text style, and manuscript revisions in "...".

1. The study aims at providing a framework for the assessment of the impact of sample size on landslide susceptibility maps (LSM) using quantitative methods and some new approaches. However, the hypotheses and methods are tested against a very limited target, represented by 2 very small study areas where a very high homogeneity in physical settings can be expected (a proof of such homogeneity could be the fact that using only 81 samples the authors attain an overall ROC AUC = 0.87, see e.g. Page 31 – rows 23-25). How can conclusions drawn from this very specific domain be translated to larger geomorphological contexts? Areas smaller than 10-20 km2 are not the usual target for LSM due to the fact that direct field survey and detail scale analysis (including field tests and deterministic modeling) are still possible with reasonable costs (see e.g. Berti and Simoni, 2010 JGR, 115). Most of the basic assumptions of the paper (such e.g. the sampling distance, the factor autocorrelation and so on) are influenced by this choice. A much larger area of study (e.g. > 100 km2 at least) would be needed to discuss the actual effects of sample size on LSM, an area where the relative importance of susceptibility factors could vary due to the presence of different types of movements, settings and so on.

We acknowledge that the small size of the study area, and its homogeneity lead to optimistic validation results.
In reply also to comments of the first reviewer, we have added some thoughts in the validation section:
"The different performance of the ZBT model in the LT area and vice versa is an interesting fact. This could be caused by different characteristics of the study areas (different range, spatial and statistical distribution of geofactor values); the two neighbouring areas, however, are regarded as very similar and homogeneous. Heckmann and Becht (2009) investigated the transferability of a debris flow susceptibility model among different study areas and reported that the predictive power of models is largely independent of the degree of similarity of training and test area; their model approach

(certainty factor), however, strongly differs from logistic regression. Besides computational and conceptual differences, continuous geofactors such as slope are classified using the same scheme in all study areas. Conversely, in our study, a different range of geofactors in training and test areas could lead to different coefficients and different model performance due to extrapolation. Another reason for the different performance could be the different debris flow density. In order to determine the controls of model performance, future research will have to use a larger number of different study areas with different debris flow densities. The methodological framework for the assessment of model variability and performance proposed here is considered useful for such investigations."

Furthermore, the small size of our study area and its characteristic settings naturally affect the transferability of our model and to some degree also the generality of our conclusions. Our aim, however, was to present a methodological framework (as also noticed by the reviewer: "...aims at providing a framework") rather than a transferable general model. On the one hand, we think it is plausible to expect that, irrespective of study area size, the sample-dependence of model composition (i.e. the number and type of geofactors remaining in the model after stepwise factor selection) will decrease, and there will be a sample size beyond which this sample-dependence will change only marginally. On the other hand, we cannot deny that things are possibly different in large, heterogeneous study areas. Therefore, we have re-formulated our conclusions and/or added remarks on the generality of the findings where applicable:

"While the typical scale of application of landslide susceptibility models is in the order of (many) tens to thousands of square kilometers, our study took place in a comparatively small study area. Considering the small size and the associated homogeneity of our study area with respect to the statistical and spatial distribution of geofactors, we add a note of caution to the interpretation of our findings. First, we

expect the necessary sample size to be larger in more heterogeneous areas, and we expect a larger variability of model selection and model coefficients. One possibility of dealing with large, heterogeneous study areas has recently been proposed by Petschko et al. (2012) who partition their study area in sub-areas based on lithological properties that are related to landslide activity. Second, the assessment of spatial autocorrelation from variograms of the geofactors is much less straightforward in larger, heterogenous areas. For example, different ranges of autocorrelation could exist for the same geofactor in different (sub-)regions of the study area, which calls into question the existence of a single sample size (and the associated average distance between sample points) below which the autocorrelation issue is mitigated. However, we are confident that our observation of a local minimum or plateau in model diversity will apply also at larger spatial scales (see, for example, Hjort and Marmion, 2008; Guns and Vanacker, 2012). Moreover, we uphold the general recommendation to investigate, through repeated sampling with different sample sizes, the behaviour of parameter selection in order to explore a suitable (small) sample size that both minimises sample dependence and facilitates a robust parameter selection."

Another comment from our side: The study by Brenning (2005, NHESS) that is cited multiple times, both in our study and in the reviewer comments, uses a study area of 11 km^2, so we feel that the small size of our study area has to be discussed, but is not prohibitive for our purposes.

The fact that this study only concerns debris flows does not exempt the authors to consider the practical fact that, actually, LSM application must be performed in most cases without any a-priori knowledge of active processes or at least without any possibility of discriminating different ensembles of environmental settings to perform limited-extent mapping. In this context, sampling for model calibration should also encompass the problem of statistical significance in different non-homogeneous areas with implications that are not negligible for the aim of the paper.

We strongly believe that knowledge of the (potentially active) processes, and/or focusing on specific types of landslides is essential for LSM tasks. Mass movements as different as rockfall, shallow and deep-seated landslides, slope- and torrent-bed type debris flows etc. are extremely unlikely to be predictable by the same model, as they occur under very different circumstances (for example, you can see that from the very different spatial distribution and appearance of their initiation zones). There are even differences within one type of mass movements: Wichmann et al., (2002), for example, show that a susceptibility model used for the prediction of slope-type debris flows is completely unable to do so for torrent-bed type debris flows. Finally, two sub-types of slope-type debris flows can be distinguished based on the mechanism of initiation (c.f. Zimmermann et al., 1997), and it cannot be taken for granted that one model can account for even these comparatively small differences. Our study area features almost exclusively slope-types debris flows of the second type (rockslope-talus-contact, progressive erosion etc.), as described in the study area section.
In large heterogeneous study areas, however, we contend that creating the inventory (the limited-extent-mapping addressed by the reviewer) is difficult, however it can be guided by having a look at aerial photos, landcover maps etc where applicable. It is clear that in large and heterogeneous settings larger sample sizes will be required in order to represent the diversity of the whole study area. One strategy for large areas has been recently proposed by Petschko et al. (2013, NHESSD) who separate sub-areas of their large study area based on groups of similar lithology. Due to the favourable homogeneity of our study area, we did not have to deal with such problems, however we clearly understand the need to critically appraise and to understate the generality of our conclusions. For modification of our discussion see above.

2 All the analysis and results are influenced by the impact of proportion between event and non-event pixels selected for the calibration process. That is well

discussed in the manuscript and this is also one of a very few papers that directly raise the issue and propose a solution. However, there is no accounting for "false negatives" in the study. We all know that landslide inventories are highly affected by missed landslides, which have not been recognized due to the absence of visible effects at the scale of mapping. This loss of information, due to the normal processes of geomorphic and environmental slope evolution after failure, is especially important for shallow landslides (such as the debris flows considered in the study) whose scars have a very low persistence in time. How the presence of such missed positives influences the results of analysis? Usually, one way of reducing the noise introduced by false positives would be to increase sampling dimension but, instead, this study seem to reveal that no gain is obtained by using larger samples. How is this conclusion influenced by the hidden presence of missed old debris scars within the non-event samples? How can this be translated (see also note 1) to larger areas where this problem could even be worse?

Morphological traces of debris flow initiation are visible very well, both on orthophotos (that have a ground resolution of 20 cm in our case) and on shaded relief representations of high-resolution DEMs (1m in our case). The slope-type debris flows we solely investigate here take place in an area where essentially no human impact takes place (such as ploughing in case of shallow slides in an agricultural area, c.f. Bell et al. 2012, Geografiska Annaler A 94.1, pp. 135-156) that could effectively remove the traces of debris flow initiation. Moreover, we argue that debris flow activity in environments like our study area tends to persist once it has started (because a newly formed incision forces convergence of overland flow, and because debris flows appear to be transport-limited) for a long time (until either sediment storage is depleted or slope gradient has become too low for debris flow initiation). The problem of overlooking debris flow events is, from our experience, more related to the deposits than to the initiation zones (because the former in many cases progressively and quickly change their colour due to weathering, or are covered by more recent deposits). Last but not least, a good LSM is expected to indicate susceptibility at the site of former landslides even if the corresponding event was not part of the training dataset (provided that the topography has not changed towards more stable conditions). We have included some text on false negatives in the inventory section of the methods chapter:

"Guzzetti et al. (2012) discuss the importance of landslide inventory maps and report on advantages, limitations and new methodological developments. With respect to susceptibility mapping, the quality of the underlying inventory is a limiting factor for the reliability of predictive models (Ardizzone et al., 2002). While fresh landslides are readily detected, post-event modifications such as human impact (e.g. ploughing), landcover change, erosion and landslide reactivation etc. can hamper the identification of landslides and thus jeopardise the completeness of the inventory (Bell et al., 2012, e.g., analyse persistence and change of landslide morphology depending on age). For debris flows in our study area, however, we argue that the risk of false negatives (i.e. the risk of an incomplete inventory due to overlooked debris flow scars) is small: The activity of debris flows tends to persist once it has started, because an incision enhances and sustains the convergence of surfcace runoff. Due to the transport-limited conditions of debris flow initiation in our study area, this is supposed to hold for a long time, until either sediment storage is depleted or slope gradient has become too low. Conversely, debris flow deposits are frequently modified by renewed activity, and less pronounced depositional lobes can loose contrast on aerial photos due to progressive weathering. Additionally, human activities that could potentially modify the appearance of debris flow scars are completely absent in the relevant regions of our study area."

3 There is, throughout the paper, a diffuse contradiction: in the methodology part (and somewhere else as well) you report that step-wise model selection

(or, I would say, factor selection since you always use one GLM model) is not appropriate for testing theory and to draw conclusions on the physical significance of single factors with respect to debris flows. In several other parts (mainly results and discussion) you nevertheless use this step-wise selection results to make hypotheses on the role of susceptibility factors.

On the one hand, it was not our aim to rank geofactors with respect to their relative physical importance and discuss the implications of the specific ranking. We only state that certain factors form part of most models, and compare this selection to other studies.

On the other, we understand that the assertion that step-wise model selection (which is, by the way, common terminology in statistics) was inappropriate for theory testing may cause undue confusion and is not too relevant for the aims of our paper. The complete quote by Menard reads “...there appears to be general agreement that the use of computer-controlled stepwise procedures to select variables is inappropriate for theory-testing because it capitalizes on random variations in the data, and produces results that tend to be idiosyncratic and difficult to replicate in any sample other than the sample in which they originally were obtained”. In our paper, we are dealing with such sample dependence (which we aim to minimise). Therefore, studies applying stepwise variable selection on the basis of only one single or too few samples run two risks with respect to sample dependence: i) variable selection and ii) model spatial structure and performance. Menard later reserves that stepwise selection may be appropriate for purely predictive research (where “there is no concern with causality, only with identifying a model, including a set of predictors, that provides accurate predictions of some phenomenon”) and explorative research (where “there may be a concern with theory construction, when the phenomenon is so new or so little studied that existing ‘theory’ amounts to little more than empirically unsupported hunches about explanations for the phenomenon."). In different LSMs published, there is some agreement on some important factors, however the relative role of the factors is not the same everywhere, and may vary spatiotemporally.

Facing the fact that the relative role of factors is indeed part of many studies, and stressing once more that the main focus of our paper is not the investigation of predictors, we decided to remove this possibly confusing statement. The corresponding paragraph now reads:
"(...) The results of stepwise logistic regression have often been used to rank the controlling factors by importance (e.g. vandenEeckhaut et al.,2006). While we assume that the methodological framework of our study would also be suitable for the assessment of sample size effects in such investigations (Guns and Vanacker, 2012, e.g., suggest a "robust detection of controlling factors" based on repeated sampling and stepwise model selection), the latter are not the aim of our present study. (...)"

4 Furthermore, basically, I wonder why would you choose to perform all these preliminary analyses on sampling dimension and factor selection to reach an overall performance (AUC=0.8-0.9) similar to those found by many other authors without recurring to it if you are also unable to use the results of factor selection to draw some interesting conclusions on the physical processes actually producing debris flows? Please clarify this point.

What we intended to tackle were the issues of i) model "stability", i.e. the dependence of model results on the single sample that many studies still rely on and ii) sample (in-)dependence. Independence of observations is an important prerequisite for the statistical method of log. regression (the parameters are fit using a maximum likelihood approach; for the likelihood function, probabilities for the pixels that belong to the sample are multiplied, which is only admitted under the assumption of stochastic independence, p. 2739 in the original manuscript). Our results indicate that the violation of

statistical assumptions does not appear to affect model quality in terms of predictive capacity evaluated using the AUC (with the restriction: "in our study area"). However, the violation is known to produce false significance estimates (see Brenning, 2005, NHESS, p.857) that might negatively affect significance-based parameter selection (used e.g. in van den Eeckhaut, 2006, NHESS). Then, similarly, the selection-based interpretation of the relative importance of geofactors (which is not so much the aim of our paper, but common practise and also advocated by the reviewer) would be compromised.

We decided to clarify our intentions in more detail, especially with reference to autocorrelation in the non-event sample (section "why the sample must not be too large":

"Atkinson and Massari (2011) explain that (spatial) autocorrelation of the geofactors causes the model residuals to be spatially autocorrelated (which is not acceptable as model residuals have to be uncorrelated), and that this may lead to "incoherent significance estimates for the parameters" (see also Brenning, 2005). Consequently, such incoherent estimates compromise both significance-based model selection and the assessment of parameter importance that is based on the latter."
"(...) In some instances, the risk of autocorrelation is dealt with for "events" only, as geofactors tend to be homogeneous (and consequently strongly autocorrelated) on landslide terrain (Atkinson and Massari, 2011). However, the independence assumption refers to all observations of the dependent variable (Hosmer and Lemeshow, 2000), in our case to the occurrence and non-occurrence of debris flow initiation. As the geofactors used as independent variables are supposed to be associated with the dependent variable, we argue that the degree of autocorrelation of these geofactors should be accounted for in the sampling procedure. In order to mitigate the issue of spatial autocorrelation, (...)"

Another, somewhat minor, conceptual issue: in the introduction it is not clear what are the novel aspects proposed by the paper with respect to recent similar works such as e.g. Brenning (2005) who already attempted sample size sensitivity analysis. Please add some more explanation here.

Perhaps the most important novel aspect is the strategy to minimise (not prevent) autocorrelation in both the “event” and “non-event” pixels. While several studies recognise the autocorrelation issue for the “event” sample (e.g. by taking one pixel for each landslide within the inventory), this has not been applied to the “non-event” fraction. Brenning (2005, NHESS) and Atkinson (2011, Geomorphology 130, pp 55-64) follow the strategy of including spatial autocorrelation in the regression (autologistic models).

Another novelty is the number of replications. Brenning (2005), for example, uses 50 replications, while we use 1000 on the grounds of a much higher stability of the results. With only 100 replications, visible differences appeared when comparing two diversity diagrams (Fig. 4) – this suggests that only 25-50 replications are not enough to achieve stable results. However, the required number of replications depends on data characteristics and can probably not be generalised.
Modified/added:
"We analyse model diversity by repeating the stepwise model selection with 1000 independent samples of a given sample size. Such a high number of replications is novel compared to existing studies that employ multiple samples; we chose the number of 1000 because we noticed in first experiments that the model diversity assessment was too unstable with a lower number of replications (e.g. between 25 and 50 in the studies of Brenning, 2005; Begueria, 2006, Guns and Vanacker, 2012)."

Another novel aspect is the quantification of model variability (as one expression of sample dependence) using indices from information theory and ecology (biodiversity

assessment). This is an alternative for the recently proposed "thematic consistency" index (Petschko et al., 2013); the latter uses variable-selection frequencies in model replications and is based on the Gini impurity index.
Modified/added:
"Therefore, we propose the "model diversity" as a measure of model quality in terms of reproducibility; similarly, Petschko et al. (2013) recently proposed a "thematic consistency" index that uses variable-selection frequencies in model replications and is based on the Gini impurity index."

A minor novelty is that we use only DEM-related parameters (prerequisite: availability of high-res and high-quality DEM), which is admittedly facilitated by the absence of significant vegetation cover in the relevant parts of the study areas.
A note concerning this was already contained in the original manuscript: "Although geological and landcover maps were available, we tried to use only geofactors that can be derived from (high-quality) digital elevation models (DEM) in order to test the feasibility of DEM-based modeling (such high-quality DEMs are increasingly available for large parts of the world)."

5. The authors test model stability against random sampling after selecting the optimal sample size. This comes a little bit too late in the paper, in my opinion. Why not try to understand which is the impact of local mapping heterogeneities at different sample sizes? I would think that, the smaller the sample, the larger the impact of outliers on the model calibration should be. Your results seem to contradict this quite common belief. But you only perform random tests for 1 sample size. For example, how errors in the estimation of local DEM elevation influence the results of your analysis? The origin of the data (LiDAR acquisitions) usually implies high precision but high frequency of local errors, especially on mountain topography.

We are not sure what the reviewer meant with "testing model stability against random sampling AFTER selecting the optimal sample size". In order to find the optimal sample size (that sample size i. above which model diversity does not decrease considerably any more, and ii. for which the expected distance between neighbouring pixels in the sample is above the autocorrelation range of as many geofactors as possible) we conducted random sampling (1000 random samples for each of more than 15 different sample sizes) and related the measured model diversity to sample size. In observing that model diversity (and hence, sample dependence) decrease for larger n (fig. 4), we find both our working hypothesis (page 2750 lines 11ff) and the "common belief" addressed by the reviewer (which we do share) confirmed, not contradicted (given that a low model diversity/sample dependence implies a low sensitivity to "outliers").

Then, having identified the optimal sample size, we take 100 random samples of that size to generate 100 models (and the corresponding susceptibility maps, figs. 5,6,7), and for model validation (fig.8). We did that not only for the optimal sample size, but also for n=81, according to a "standard" recommendation of a 1:1 ratio of event and non-event units (fig. 8).

Autocorrelation. The authors rise here a very important issue that concerns almost all LSMs published so far. Their approach of using average variogram range as the limiting inter-distance for sampling (as similarly proposed by Brenning, 2005) is very understandable and should become common practice in environmental analysis.
However, I believe that the whole geostatistical issue is in general treated with too much superficiality in the manuscript. In particular, I have two doubts regarding the specific case which I hope the authors will be able to clarify:

a. Many of the commonly used variables (factors) influencing landslides are not continuous in space and cannot be treated as random space functions (or fields). Very clear examples are geology and land cover, whose spatial representation (also for mapping constrains) is that of areal objects with properties which are constant within polygons and suddenly vary when a polygon boundary is crossed. Due to the very nature of DEMs, this also applies (even though only partially) to elevation data and their derivatives (slope and so on) who are always measured (or computed) over finite spatial domains (pixels of pixel clusters). This implies the existence of very specific and complex spatial autocorrelation patterns that cannot be captured by simple pixel-based variogram analysis. A corollary of this is that strong spatial asymmetries are present in geological and geomorphological data, which in most of the cases would render a simple representation of average range completely useless. I believe the solution of taking the lowest range for staying on the safe side is not the right solution here.

We agree with the reviewer that some of the variables cannot be treated as fields. Geology, however, is not included in our model (though we contend that it must be included in a model for a larger, more heterogeneous area). Landcover is not included as well, except the binary variable "rough class" which is computed from slope and roughness (Fig. 2 shows the variogram of roughness, not of the categorical variable used in the model). While we would like to stick to the "simple" variogram-based estimation of autocorrelation range in our paper (in a small, homogeneous area), we agree that different analyses have to be used in larger, more heterogeneous settings. For example, a large study area should be subdivided, for example based on lithology (see Petschko et al., 2013 NHESSD), landuse, climate, or combinations of such "large-scale" factors, and autocorrelation analyses should be conducted for each of the subareas.

A note of caution has been added to the discussion (see also above):
"Second, the assessment of spatial autocorrelation from variograms of the geofactors is much less straightforward in larger, heterogenous areas. For example, different ranges of autocorrelation could exist for the same geofactor in different (sub-)regions of the study area, which calls into question the existence of a single sample size (and the associated average distance between sample points) below which the autocorrelation issue is mitigated."
The sentence before the one cited here contains a reference to Petschko et al. and their approach to subdividing a large heterogeneous study area in more homogeneous subareas.

Furthermore, we are not suggesting to choose that non-event sample size which leads to an average distance just within the lowest autocorrelation range in order to stay on the safe side. In terms of sample sizes, this is a “bottom up” approach in that we start with small sample sizes (with associated low reproducibility, high sample dependence, high sensitivity to outliers and inability to represent the variability geofactors with in the study area) and aim at reaching either a plateau or a local minimum of model diversity, ideally before we come into the autocorrelation range of any geofactor. We assume that the smaller the sample size (and the larger the average distance between observations), the more geofactors will be uncorrelated, and the smaller the risk of violating the assumption of independent observations. However, if the autocorrelation ranges of only few parameters are crossed, we have to accept that as inevitable with respect to sample dependence.

b. The authors discuss autocorrelation and its physical meaning in the landscape as a possible negative influence on factor selection due to lack of sample independence. The problem is, that the authors modify the (already artifact) autocorrelation properties of the factors by arbitrarily averaging some of them to

different scales. It is clear, and I agree on that, that slope curvature calculated at the 1 m pixel scale has no relationship with debris flow initiation. However, why use 5 m aggregation and not e.g. 10 m or 2 m? This choice clearly influences the variogram analysis. I would be grateful if the authors could clarify this point and how they think this influences their analysis and conclusions.

We agree that the degree of smoothing will influence the issue of autocorrelation. In our study, we have not rigorously checked several possible radii of smoothing filters (which would be an interesting study on its own). The reason why we used a radius of 10 m (not 5) is the (coarse) estimation of the “typical” scale of channels within the rock faces and talus cones that are both prone to and indicative of debris flow activity. With a larger window, topography is smoothed too much as to indicate locations of concentrated flow; moreover, radii of profile curvature become too large, for example, to represent the boundary of rockface and talus). While we wanted to “smooth out” detail for the curvature calculation, we kept the window smaller for the roughness calculation, as we were more interested in material properties within smaller surroundings (for example on the sidewall of a gully rather than capturing the medium or large-scale roughness of the gully itself). See also reply to similar comments by the first reviewer.

In all the manuscript the authors use the diversity and the diversity indexes as measures of model performances. While I agree on the fact that a model should be parsimonious and that over-parametrisation is to be avoided, it is still not clear to me why they automatically assume that low model diversity is perforce better.

The diversity measures indicate neither model performance (this is measured by AIC and above all by AUC) nor parsimony vs. possible over-parametrisation (the indices do not measure the number of parameters in the models). If a model with many parameters appears as the “best” (in terms of AIC) in the vast majority of the 1000 simulations, the diversity/selection stability will be low, and we would conclude that in this case the dependence on the specific sample is low. That also has to do with reproducibility (if we agree that a sample size leading to very different results for two different samples of the same area jeopardizes the reproducibility and applicability of the model). Looking at the fact that those studies that involve sampling at all mostly take only one sample, a low diversity indicating a low sample dependence is very important in our opinion, also with respect to “robust detection of controlling factors” as concluded by Guns and Vanacker (2012). Hence, we address sample dependence as a different aspect of model quality (besides predictive capacity).

Following a comment along the same lines (why the lowest diversity value corresponds to a minimal dependence of model selection on the sample) by the first reviewer, we added some explanation on why we take the diversity index as a measure of model quality (in terms of reproducibility, not in terms of predictive ability):
"(...)Shannon’s Entropy has been interpreted in terms of the "average surprise a probability distribution will evoke" (see e.g. Thomas, 1981, p.7). The result of a stepwise selection with a sample size for which low diversity (low H) has been measured is not expected to be surprising, because one or few species have a very high probability of occurrence. We hypothesize that the diversity of model species, and the degree of surprise with which we see one particular outcome of the selection given the results of 1000 models, will reflect the sample-dependence of the stepwise selection. Therefore, we propose the "model diversity" as a measure of model quality in terms of reproducibility; similarly, Petschko et al. (2013) recently proposed a "thematic consistency" index that assesses variable-selection frequencies in model replications and is based on the Gini impurity index."

In some conditions low diversity could also be a warning of possible overfitting problems, especially at large sample sizes.

We assume that overfitting is of much less importance in logistic regression (see Brenning, 2005) compared to machine-learning approaches.
We have added a reference for this statement to the text:
"Brenning(2005), however, states that overfitting is ’not a serious problem for logistic regression’, contrary to machine-learning methods (c.f. Petschko, 2013, and references therein)."

A very big concern: at page 29-30 row 28 to 37 you say that GLM models (and in particular the logistic regression used in the study) generates some errors in the susceptibility maps because “linear modeling approach is not capable of modeling complex non-linear relationships such as the one of slope and debris flow...”. Now, slope angle, if I am not mistaken, is one of the most important factors resulting from your study. Why then did you choose to adopt logistic regression in the first place? A methodology more flexible towards nonlinearities such as machine learning non-parametric models could have been a better choice.

It was not our aim to find out the best model approach in terms of predictive ability. If this had been the purpose of the study, we should have conducted a multi-approach comparative work. This has been done in a number of studies, and recent work indeed seems to favour approaches like GAM (e.g. Hjort and Luoto, 2011, ESPL 36, pp. 363-371) or machine learning because of their flexibility in modelling nonlinear response. In comparative studies, however, logistic regression ranks among the best performing approaches (e.g. Brenning, 2005, NHESS, Carrara et al. 2008, Geomorphology), and it is definitely among the most frequent approaches used (original manuscript p2735 l2ff). Conversely, sample independence, which is one of the comparatively few assumptions of GLM (but very important with reference to the maximum likelihood parameter estimation), has rarely been addressed, and not in context with sample size. We think that it must be possible to highlight a weak point of the method without having to resort to another method, especially if the best-performing LSM is not the main focus of our study.

However, we realised from this comment that we have not made it sufficiently clear in the original manuscript that our study focuses deliberately on logistic regression because of the widespread application combined with the disregard of the independence assumption in many if not most studies. We have made an effort to clarify this in the introduction chapter:

"The present study has two main foci that will be developed in detail in the following subsections. It is not the aim of our study to find out the best performing method for a debris flow susceptibility model (comparative studies of predictive models were carried out, for example, by Brenning(2005), Marmion et al. (2008), Carrara et al. (2008) , Vorpahl et al. (2012); we deliberately chose logistic regression for its widespread use, and for the relevant assumption of sample independence which we found to be frequently neglected in previous studies. First, we explore the sensitivity of stepwise model selection to sample size. Sections 1.1 and 1.2 will explain why the sample size must neither be too small nor too large. Here, the main aim of the study is to investigate if an "optimal" sampling size can be found as a compromise between samples too small and too large. Second, we quantify the uncertainty inherent in a stepwise modelling approach, with respect to i) the selection of geofactors, ii) model parameters, and iii) the spatial pattern of uncertainty in the resulting susceptibility map. This study aim will be developed in section 1.3."

The entire analysis set could have gained in clarity had the authors introduced, among the geofactors, also a dummy (completely random) parameter to act as a benchmark. The latter should have resulted as discarded in all model species so as to ensure that the metrics used to measure diversity and performances were correctly working in the model tests

We acknowledge that such a strategy may be of much value in machine learning approaches (see e.g. Catani et al., 2013, NHESSD), but is in our opinion not transferable to logistic regression. Specifically, listing a “non-sense” candidate variable (i.e. a variable which is known NOT to have any relationship with the observed process) would constitute an intentional mis-specification of the model (one of the assumptions of GLM is that “no irrelevant predictors of the dependent variable are included in the analysis” ; Menard 2002, p.5).

There are several sentences in the manuscript that are not clear and difficult to understand (and also many small grammar errors and typos). The readability of the manuscript would gain a lot from a careful English spelling and grammar revision.

We have checked the manuscript and applied some changes that we think improve readability (e.g. splitting long sentences etc). Moreover, we hope that the copy-editing process will further improve the readability if need be.

MINOR CORRECTIONS

1. (all manuscript) It would be useful for improving readability and comprehension, to present sample size not only in frequency terms (n) but also in relative areal extent. For example, how large is the proportion of the area sampled (over the total) for n=81? And for n=350?

Although commonly absolute sample sizes are considered more important than

relative ones, we now report additionally the relative sample sizes (% of study area) where applicable, e.g.
"In this study, when we speak of sample size, we always address a sample of "non-events", i.e. a sample of raster cells without debris flow initiation. (...) Besides the non-event sample size, the relative sample size (i.e. the areal extent of the total sample divided by the size of the study area) will be reported."
"We analyse model diversity by repeating the stepwise model selection with 1000 independent samples of a given sample size. Sample size varies between 50 and 5000 non-event raster cells; together with the sample of n=81 initiation areas in the ZBT area, the samples cover between 0.02 and 0.68 percent of the study area (ZBT)."
The percentage appears to be very small, but look at the following example: With a relative sample size of 0.01, one out of 100 raster cells would be selected, which would be on average one out of a 10x10 cells area, which equals 50x50 m (2500 m^2), which is already smaller or close to the autocorrelation range of several geofactors.

2. Page 3 – row 11-14: the sentence is not clear, please rephrase
The (spatial) probability of occurrence of an event forms an important factor of the hazard term in quantitative risk assessment, although for a complete formulation one also needs to consider the temporal probability and the magnitude–frequency relationship of events (Guzzetti et al., 2006).

3. Page 3 – row 22: please add some more recent literature here
We added 5 references here (Glade and Crozier 2004, Brenning 2005, Huabin et al. 2005, Luoto and Hjort 2005, Carrara et al. 2008).

4. Page 5 – row2: as far as I know, the first important applications of ANN to LSM studies are those by Lee et al (2004, Env.Geol.) and Erminni et al. (2005, Geomorphology)

We thank the reviewer for his/her suggestions, and we have included them in the text. We used Lee et al., 2003, ESP&L as an earlier reference than Lee et al. 2004.

5. Page 8 – row 11-14: the sentence is not clear, please rephrase
We were not 100% sure which sentence was meant here; p. 8 has one sentence from row 12 to 15 which we slightly rephrased: “In order to limit the sample size and to mitigate the rare-events issue (see below), the literature suggests different ratios of event:non-event sample sizes, mostly without justifying the particular choice of this ratio. Instead of merely adopting one of these suggestions (which generally range from 1:1 to 1:10), our paper aims at an empirical analysis of sample dependence and performance of the susceptibility model as a function of sample size.”

6. Page 11 – row 23-26: a recent example that you may refer to and cite here is probably Catani et al., (2013,NHESS Discussions Online) which has recently been out. The authors perform some tests which are quite similar to those presented in your study.

While we took the opportunity of explaining the issue in more detail and give some logistic regression-related references, we found that the approaches used in our study and in Catani et al. are too different to relate to here. The new, slightly extended paragraph reads:
"This is important because in the majority of studies employing sampling for model calculation, only one sample is taken, and no account is given of uncertainty beyond the standard errors of the parameters. On the other hand, most studies involving repeat sampling (e.g. Brenning 2005, Begueria 2006, van den Eeckhaut et al. 2010, Guns and Vanacker, 2012) concentrate on the set of geofactors, the parameters and the predictive ability of the models, and do not investigate how this affects the spatial distribution of susceptibility. Only rarely has the spatial distribution of model

uncertainty been addressed using multiple replication approaches (e.g. Guzzetti et al. 2006, Luoto et al. 2010, Petschko et al. 2013 NHESSD)."

7. Page 15: you did not discuss the problem of noise introduced in the DEM by the presence of low standing vegetation. You say that the areas are not much vegetated but on the other hand (Tab. 1) you say that almost 20-25% has some patchy, shrub and woodland cover that could potentially introduce errors of an order of magnitude compatible to the DEM local differences. Such errors are known to occur even after LAS data filtering through last-impulse selection.

We agree that the DEM may contain noise in presence of even low-standing vegetation. This would affect presumably the calculation of roughness (and the classification in bedload and sediment that is based on roughness and slope). The classification, however, appears to be reasonably good. Moreover, the 20-25% of the study area that are covered with sparse vegetation are located mostly in the lower parts of the study area, while the debris flow initiation takes place mostly in the upper parts (at the contact of talus slopes/cones and steep rock alls), therefore DEM uncertainty is not supposed to seriously affect the LSM in our case study.

Page 15 – row 26: DHM5 should probably read DEM5
Yes; has been corrected.

9. Page 16 – rows 10-23: the parameter SCA is quite important in your study, but you say it has been computed by using the multiple-direction algorithm proposed by Freeman (1991) which is notoriously known to work only for convex shaped areas. The same algorithm, in concave areas (such as channels and hollows for debris flow initiation) always underestimate flow accumulation because it does not account for

convergence prevalence. A different algorithm should thus be used for concave areas.

We accept that the reviewer challenges the use of the MFD algorithm. The assertion that it only works for convex shaped areas, however, is too strong and too restrictive in our opinion. From our experience, the described disadvantages of the MFD algorithm appear above all in larger valleys with channel systems that can be much wider than one raster cell. Here, the MFD algorithm models flow divergence that is not expected to exist in reality. The debris flow initiation areas in our study area, however, are located on steep slopes, partially in deeply incised, narrow channels where unrealistic divergence is either widely absent or only of very limited extent. We are confident that the algorithm is not grossly inappropriate, forcing us to switch algorithms and conduct the whole analysis again. Therefore, we chose to stick to the procedure as described in the paper (and used, for example, by Begueria, 2006, Geomorphology). Please note also that the calculation of flow accumulation in many other papers is either not specified with respect to the algorithm used, or the D8 algorithm is taken (which, conversely, is not appropriate on steep slopes where D8 causes unrealistic parallel flow paths).
By the way: SAGA GIS would offer the possibility to switch to the D8 algorithm where specified thresholds (of flow accumulation, or of convergence) are exceeded

10. Page 18 – row 28 and Page 19 - rows 1-3: I personally do not agree with this sentence which is in my opinion very strong and contradicts most of the published literature on LSM. I would advise the authors to use a different argument or to sustain the present one with strong proofs.
The sentence has been deleted in order to avoid confusion; it was considered not too relevant for the aims of our paper (see reply to major comment nr. 3)

11 Page 22 – rows 8-9: this is not necessarily so. Points separated by distances

shorter than the range can still have very different values and be not autocorrelated

We agree. We have contended that our approach does not really guarantee independence because, by minimising the sample size, the AVERAGE distance between sample pixels is maximised (so close neighbours can always be part of a sample; page 2752 line 16ff). Then, the statement of the reviewer is in favour of our approach; the variogram analysis reports an average of all point pairs, so close neighbours can indeed be very different, and not all close neighbours within the sample cause a sample dependence issue.

12. Page 24 – rows 7-8: after devoting a lot of space to the problems of factor collinearity it is not contradictory that you propose the combined SCA*Slope factor?

With multicollinearity, we have taken up an issue that is dealt with in a number of LSM publications. Looking at the literature, collinearity and correlation are sometimes used synonymously, although a correlated pair of variables is not necessarily collinear. The "sensu stricto" definition of collinearity refers to a strong **linear** relationship between factors in a regression, where one factor could be determined from another (or others), and where we cannot differentiate the factors with respect to their influence on the target variable: Consider three geofactors x_1, x_2, x_3 so that x_1 can be expressed with the help of two real parameters β_2 and β_3 as $x_1 = \beta_2 x_2 + \beta_3 x_3$. In case of the stream power index (see next comment), $x_1 = x_2 * x_3$, i.e. x_1 is not a linear combination of x_2 and x_3. Hence, the interaction term SCA*Slope is not supposed to be collinear sensu stricto with SCA or Slope: SCA*Slope cannot be determined from only one of the two interacting factors (what would be the case if the SCA*Slope was collinear with either SCA or Slope).

13. Page 25 – row 2 and also Page 24 – rows 10-23: the combined parameter

SCA*Slope is better known as distributed (or discrete) stream power index and has been early used and proposed by Bagnold (1966) for rivers and then by Moore at al. (1991) and Seidl and Dietrich, (1992) for slopes. Please add some refs.

We thank the reviewer for this information. We now refer to SCA*Slope as the stream power index rather than referring to the CIT index, and we've added the suggested references:

"These findings are consistent with previous work on (slope type) debris flow susceptibility: Heckmann and Becht (2009) and Wichmann et al. (2009), for example, use slope, landcover, and a variable called the CIT index (Montgomery and Foufoula-Georgiou, 1993). The latter is calculated as the specific catchment area times the squared tangent of slope. The interaction term slope*SCA used in our study can be interpreted physically (mathematically, it is the product of the two geofactors) as the compound topographic index indicating stream power (catchment area and slope as proxies for the abundance and energy of surface runoff, Moore et al., 1991)(...)"

14 Page 32 –row 21: not clear, spelling error?

There was indeed a spelling error ("if" instead of "it") that we have now corrected.

15. Figure 3 : could this figure be compressed into one single plot to improve readability in comparing different geofactors?

We have implemented this suggestion.

Interactive comment on Nat. Hazards Earth Syst. Sci. Discuss., 1, 2731, 2013.