Interactive comment on "Assessing the quality of landslide susceptibility maps – case study Lower Austria" *by* H. Petschko et al.

Authors response on Referee #1 comments

H. Petschko[1], A. Brenning[2], R. Bell[1], J. Goetz[1,2], T. Glade[1]

[1] Department of Geography and Regional Research, University of Vienna, Austria

[2] Department of Geography and Environmental Management, University of Waterloo, Ontario N2L 3G1, Canada

Correspondence to: H. Petschko (helene.petschko@univie.ac.at)

Reply to the specific comments

We numbered the comments given by the referee, as some comments refer to similar sections which were changed in the revised manuscript. This numbering should assist to make the cross-references within this reply easier to follow.

The comments by the referee are in presented in bold, whereas the reply is in normal font type. Please notice that some comments were split up into two comments if the original comment was referring to different lines or aspects. This is indicated with "..:" at the end and at the beginning of the comments concerned.

1. p1004 l4ff: In this paragraph the authors motivate their strategy of k-fold cross-validation. I feel that, in the introduction chapter, the generic strategy of 1-fold cross-validation (or "single hold-out") should be explained more generally, for the less experienced reader. In the methods section, the approach can (and must) be explained more in detail. Specifically, I suggest to briefly explain the term "single hold-out" as model estimation-validation-strategy in one short sentence instead of simply mentioning "single-holdout model performance measures".

The suggested brief explanation of the single-hold out performance assessment was included in section 2 of the revised manuscript to make the difference to the *k*-fold cross-validation clearer for the less experienced reader.

Added to page 1006, line 7

> In single hold-out methods the data set is split in one single training and test sample. The training sample is used to fit the model and the test sample is used to determine the model performance. This results in a single estimate of the performance measure (e.g. one single AUROC value) without providing a measure of precision of the estimator. The estimate depends on the (random) sample used for modelling the susceptibility and testing the model's performance, which may itself have "peculiar" random characteristics that would be different for a different test set. Repeated *k*-fold cross-validation solves this problem by using, one after another, different subsets or partitions of the data set as test and training sets, thus effectively using the entire data set for performance estimation (Brenning, 2012a, 2012b). In addition, repeating this procedure for different data partitioning reduces sampling variability and allows for the determination of the precision of the performance estimator (see section 5.3 for details).

2. p1004 l9: The abbreviation AUROC is used here without prior explanation what AUROC means and, even more importantly, what AUROC IS (the area under a receiver operating curve, and as such one possible concept to measure of model quality). As model validation is explained in the methods section, I suggest that the authors speak, less specifically, of "a range of possible validation outcomes instead of only one single, "random" outcome...". Alternatively, the validation concept of ROC and the quality measure of AUROC would have to be explained prior to using the abbreviation.

According to your suggestion the term AUROC (area under a receiver operating characteristic curve) is explained earlier in the manuscript and introduced as one possible performance measure. For the detailed explanation and interpretation of the AUROC value we refer in the introduction to the methods section (5.3). Furthermore, the sentence is altered according to your suggestion now being less specific. Generally, this paragraph was moved to section 2 of the revised manuscript. The first occurrence of the abbreviation AUROC is on page 1006 in line 4:

Added to page 1006, line 4

> Among the performance estimation techniques and measures, cross-validation using a single hold-out method and the area under receiver operating characteristic curve (AUROC) value based on ROC plots (e.g. Brenning, 2005; Beguería, 2006; Frattini et al., 2010) are usually applied (i.e. Chung and Fabbri, 1999, 2003; Fabbri et al., 2003; Remondo et al., 2003; Brenning, 2005; Beguería, 2006; Frattini et al., 2010; Rossi et al., 2010).

3. p1010 l12f: a comment: The question is also to what extent missing data (from blurred, reworked, removed landslides) influences the result of a susceptibility model. If the latter is good, the locations of previously existing but now invisible landslide should be "predicted" as susceptible. This cannot be checked, but: cross-validation (which reserves part of the inventory for evaluation purposes) compared to model goodness-of-fit (i.e. the model estimated from and applied to the same area/sample) should give an idea of how an incomplete inventory affects model quality.

We are thankful for the correct and helpful comment. This was included in the discussion of the results of the study:

Added to page 1026, line 14

The implications of an incomplete inventory on the model performance (shown by the AUROC value) were estimated by performing the repeated *k*-fold cross-validation using training and test sample. The results show rather high AUROC values for most modelling domains, which indicates that even with an incomplete inventory (training sample) the prediction of landslides of the test sample was successful for most cases. However, sample size is of importance for the model performance. For the discussion of this please see section 7.4.

4. p1010 l25ff: Please add some explanation on the possible physical meaning of the DTM-derived variables. For some, it is more obvious (slope) than for others (catchment height - climatic proxy for precipitation ? slope aspect - orientation relative to bedding?)

We understand the lack of description of the DTM derived variables. We included a short description of each variable in the revised manuscript.

Added to page 1011, line 6

Van Westen et al. (2008) had discussed the relevance of most of these terrain attributes to landslide susceptibility. The relationship of slope angle with landslide activity is well known from general slope stability literature (e.g. Crozier, 1986). Slope aspect can be used as a proxy for bedding orientation. It may also reflect differences in intensity of solar radiation, which controls the local temperature and evaporation and therefore soil moisture (van Westen et al., 2008). Curvature represents convex and concave surfaces related to local morphology (3x3 grid cells). The topographic wetness index was used as a proxy for the soil moisture and ground water level (Beven and Kirkby, 1979; Seibert et al., 2007). The position of the landslide on the slope and the distance from the ridge was represented by the variable catchment height. This was calculated for first order catchments. The catchment area was calculated for the sub-catchments and gives a local representation of the contributing area. Convergence indices were calculated to represent the slope morphology on two different scales by using two different window sizes for the calculation (10m and 50m). Positive values indicate ridges while negative values indicate local depressions.

5. p1012 l25ff: Yes, tectonic faults are known to influence landslide (in a general sense) activity. In the present study, however, there is a focus on earth (and debris) slides

(p1008 l17), which means a granulometry of >80 (earth) or <80 (debris) per cent sand and finer. I wonder to what degree tectonics influence this type of process compared to, for example, lithology, degree of weathering, existence of cover beds, climate etc.

The occurrence of earth and debris slides is very much dependent on the availability of unconsolidated rock or subsequently formed soil of different thickness. The presence of tectonic fault lines is together with the parent material (lithology), climate, topography, vegetation, fauna and time one favouring factor for soil formation (Blume et al., 2010). The closer to a tectonic fault line, the stronger the material is mechanically stressed and fragmented. Therefore, the soil formation process might be faster compared to solid rock. This can also be observed in the field as stated by the Geological Survey of Lower Austria (pers. Comm. Schweigl, 2013). Although earthquakes occur in Lower Austria the influence of earthquakes on landslides was not analysed yet (Hammerl and Lenhardt, 1997). However, earthquakes along active tectonic lines are considered as a possible natural trigger of landslides (probably rather rock slides or rock falls than slides; Schwenk, 1992).

The variable “Euclidian distance to tectonic fault line” was selected in 9 out of 16 models using the entire landslide data for fitting the model. This shows that the data has some explanatory power for the occurrence of earth and debris slides. However, we found that the contribution of this variable to the overall model was marginal compared to other variables such as slope angle. No changes made.

6. p1014 l13ff: "landslide points" are mentioned here and in some other paragraphs, while throughout most of the papers, landslide cells are addressed. Please homogenise terminology. Generally speaking, "points" and "cells" are spatial units to which the sampling and the modelling are applied, and in "reality", you use raster cells, not point objects.

The authors acknowledge the main used terms and changed the wording to landslide cells in all cases throughout the revised manuscript.

7. p1014 l24: Pls define sampling rate (as the number of sampled cells per unit area) as opposed to sample size. …

The reviewer correctly points out a source of confusion in the manuscript. The authors added some sentences for clarification of the difference of sampling rate to sample size in the revised manuscript:

Added to page 1014, line 21

> The sampling of non-landslide cells for the entire study area was based on a density of 2% of all grid cells. An equal number of cases and controls (1:1) was used for each model fitted; the landslide samples were matched to an equal number of randomly selected non-landslide samples. This gives the sample size in the respective modelling domain.
>
> It was necessary to adjust each model's raw predictions based on the corresponding sampling rate to account for the general relative landslide susceptibility of each modelling domain. We adjusted the prediction by considering the sampling rate (τ_0/τ_1) of each lithology unit, using Eq. 3,
>
> $$odds^{*}(x) = t_0/t_1 \text{´ } odds(x) \quad (3)$$
>
> where,
>
> $$t_0 = \text{number of sampled non-landslide cells/total number of non-landslide cells} \quad (4)$$
>
> and
>
> $$t_1 = \text{number of sampled landslide cells/total number of non-landslide cells} \quad (5)$$
>
> and odds(x) is the unadjusted prediction, in our case, based on training a model with a 1:1 sampling ratio of landslides to non-landslides.

8. …Could you also comment on the justification of the 1:1 ratio of landslide to non-landslide cells? I suppose it has to do with the binary target variable and the cut-off (of 0.5) to distinguish predicted events from non-events.

The decision on a number of cases and controls is very important for the study design. Heckmann et al. (2013) gave a detailed review on which ratios have been chosen in landslide susceptibility modelling and on which grounds (including rare events logistic regression of King and Zeng (2001)). According to these, ratios between 1:2 and 1:5 are recommended to

better reflect the full range of values of the explanatory variables. Additionally, it was found that an increase in the ratio of controls to cases larger than four only marginally leads to an increase in precision or statistical power (Ury, 1975; Breslow and Day, 1980). It is rather recommended to increase the number of cases (by widening the study area geographically or temporally) to achieve a better precision (Wacholder et al., 1992). However, the cost of gathering more data on cases (landslides) is rather high in this study, as new landslides would have to be mapped from aerial photographs. As the number of cases and controls is large in most modelling domains the common selection ratio of controls to cases of 1:1 was selected (Breslow and Day, 1980). No changes made.

9. p1014 l26ff: Why is it necessary to adjust the predictions? The model predicts, in each domain, the probability [0,1] of landslide occurrence, no matter what the sample size or sample rates are. Could you explain this further ? …

The necessity to adjust the predictions according to the sampling rates of each modelling domain arises due to the need of providing a comparable landslide susceptibility map for the entire study area. Without adjusting the values every modelling domain would show probability values ranging from 0 to 1. However, the general differences in the susceptibility to landslides of each domain as expressed by the sampling rate or landslide density would not be taken into account. Therefore, the probabilities of each modelling domain have to be adjusted according to their sampling rates.

For the changes made in the text please refer to our reply on comment 7 and the inserted text in the revised manuscript.

10. … Furthermore, I could not comprehend the definitions of tau0 and tau1. tau0 is introduced as a "sampling rate for non-landslide points" and defined as the ratio of (land)slide to non-(land)slide points - but this ratio was described earlier as being unity. I understand "sampling rate" as the ratio of sampled cells and the total number of cells in the domain... Your "sampling rate tau1" is defined "for landslide points" and set to 1. This is confusing, because this seems to address the ratio of slide to non-slide cells that was explained earlier; your equation (4) simply becomes tau0*odds(x) because tau1 equals 1... As I understood it, the sampling rates of landslide and non-landslide "points" should be the same in each domain, because of the 1:1 ratio of sampled landslide and non-landslide points. Perhaps tau0 should be the sampling rate (ratio of slide and non-slide pixels to all

pixels) in the study area, and tau1 this ratio in the domain ? Or is tau0 the ratio of slide to non-slide pixels in the study area, and tau1 is the ratio of slide to non-slide points in the sample (so tau1 = 1:1 = 1)? To me, an "adjustment" only makes sense if a property of a domain (e.g. the ratio of sample size and total size, the ratio of landslide pixels to the total area, or the number of landslide pixels) is normalised with the same property of the complete study area. Please explain and clarify.

The referee points out an important source of misunderstanding. We mainly identify it as an misunderstanding of the terms sample size and sampling rate and some confusion we created in the equations for tau_0 and tau_1. The difference between sample size and sampling rate was clarified in our reply to comment 7. The equations of calculating tau_0 and tau_1 were updated and added in the revised manuscript (please refer to comment 7 and the added text in the revised manuscript). In the original manuscript we stated that tau_0 is calculated from the number of slide cells divided by the number of non-slide points. However, the number of sampled non-slide cells has to be divided by the total number of non-slide points. As the sampled number of non-slide cells is identical with the number of slide cells we did not differentiate here between the number of slide cells and sampled number of non-slide cells in the original manuscript. However, we acknowledge that this was rather confusing and therefore we changed this accordingly. Furthermore, we agree that the equation for the calculation of the adjusted odds could be cancelled to $odds^*(x) = \tau_0 \times odds(x)$. However, as future studies might decide to use a subsample of the slide cells we decided to present the entire equation.

For the added equations please refer to our reply to comment number 7 and the added text there.

11. p1016 l16: It is easily understood that the random (non-spatial) partition gives different results for every replication, and the strategy of taking the median and IQR of n partitions is feasible. Two questions arise for me: (1) did you check if the median and IQR of the performance measure are already reasonably stable with 20 replications, or has the number of 20 been chosen arbitrarily ? …

The number of 20 replications is a compromise between a desirable high precision of the cross-validation estimator and acceptable computational complexity given the large sample size and the computational complexity of generalized additive models. Some basic and

approximate back-of-envelope calculations based on Figure 4 (in the original manuscript) can give a general idea of the precision of cross-validation estimators. For simplicity we will have a look at the standard error of estimators of mean AUROC (as opposed to the median) across all 20 repetitions. We obtain the following statistical characteristics for domains 3786 ("worst case": high amount of variation between repetitions) and domain 35 (seems to be more "average") (IQR=observed sample (i.e. between-repetition) interquartile range; SD = sample standard deviation; SE = standard error, i.e. precision of the estimator of mean AUROC):

Domain 3786: IQR approx. 0.20; std.dev. approx. IQR/1.35 = 0.15; SE = SD/sqrt(100) = 0.033

Domain 35: IQR approx. 0.1, std.dev. approx. 0.074, SE approx. 0.017

Given the great amount of variation in AUROC among domains (median values from 0.52 to 0.98) this precision appears to be highly acceptable, and differences among domains cannot be explained by insufficient estimator precision. Figure 5 (in the original manuscript) also shows very little random variation, which supports this interpretation. We hope that the reviewer will be satisfied with this simplified consideration. No change made.

12. … (2) could you add a sentence explaining why/how the k-nearest-neighbour clustering of the "point" coordinates in the spatial partition approach leads to different results for every replication ? Are the k group centroids chosen randomly? Are the resulting spatial units similar in size or is it possible that models are estimated and validated in two partitions of very different size?

The algorithm used here is in fact the *k*-means algorithm, not the k-nearest-neighbour algorithm – apologies for the mistake, which has been corrected. The *k*-means algorithm essentially creates *k* Thiessen polygons. Since the algorithm starts with a random seed of initial cluster centres, different cluster locations and shapes result in each repetition. The resulting spatial units are of similar (comparable) size. Cross-validation ensures that the (approximate) proportion (*k*-1)/*k* of the data is used for training, and effectively (over all cross-validation folds and repetitions) all data are used for performance estimation, as indicated in the manuscript.

Added to page 1006, line 7

Repeated *k*-fold cross-validation solves this problem by using, one after another, different subsets or partitions of the data set as test and training sets, thus effectively using the entire data set for performance estimation (Brenning, 2012a, 2012b).

13. p1017 l1ff: Why is the IQR an estimated one? It is an empirical measure derived from 20 replications of a cross-validation procedure, i.e. of 100 empirical AUROC values (see p1016 l18f). …

Empirical "calculations" are formally referred to as "estimation" in statistics.

14. … Furthermore: Why does that measure have to be adjusted in order to be a measure of transferability? Perhaps you need to explain this more thoroughly, in a more step-by-step fashion. …

The IQR value is influenced by the sample size used for performing the *k*-fold cross-validation. As the sample size of each modelling domain is very different (104 to 12562 cells) the sampling variability of the AUROC varies. To provide a comparable transferability index for the entire study area this influence of the sampling variability has to be taken into account. Therefore, the IQR was adjusted for the contribution of the standard error *SE* of the AUROC estimator. Some more explanatory text was added in the revised version of the manuscript.

Added to page 1017, line 1

The non-spatial and spatial transferability were expressed by the interpretation of the estimated interquartile range (IQR) of the AUROC values resulting from the non-spatial and spatial cross-validation of each modelling domain. The lower the estimated IQR the better we considered the non-spatial and spatial transferability of the model within the modelling domain. Sample size differences among modelling domains result in differences in sampling variability of AUROC estimators, which then has an influence on the IQR of AUROC among cross-validation repetitions. In order to account for this contribution to sampling variability and be able to provide a transferability measure that was comparable among modelling domains, the IQR has to be adjusted according to the sample size.

15. … Moreover, "Eq (1)" refers to the corresponding equation in the paper of Hanley and McNeil, not to Eq(1) in your study... I feel that this equation should be given here, or that "Eq(1)" should be removed. …

According to your suggestions the equation (1) of Hanley and McNeil (1982) was included in the revised manuscript.

Added to page 1017, line 5

> For this purpose, we calculated the approximate standard error SE of the AUROC (AUC) estimator on a test set of *N* landslide and *N* non-landslide samples using the equation presented by Hanley and McNeil (1982):
>
> $$SE = (AUC \times (1 - AUC) + (N - 1) \times (Q_1 - AUC^2) + (N - 1) \times (Q_2 - AUC^2)) / N^2 \quad (5)$$
>
> The quantities Q1 and Q2 were calculated from the AUROC (AUC) value as shown by the following equations:
>
> $$Q_1 = AUC/(2 - AUC) \quad (6)$$
>
> $$Q_2 = 2 \times AUC^2/(1 + AUC) \quad (7)$$

16. … Concerning equation (5) for the calculation of the T index: Not knowing "Eq 1 of Hanley and McNeil", I suspect that the SE of the AUROC is estimated from the standard deviation of AUROCs, and will decrease with larger n. I do not see why T should be a better indicator of transferability than the empirical IQR, for example. A larger IQR means that a model could be very good, but also really bad, while a smaller IQR indicates that the models predict similarly well (or poorly). I feel that T should be better justified and explained. I understand from the paragraph that you "correct" a non-parametric empirical measure of AUROC variability (IQR) by subtracting a parametric, estimated measure of variability (that is multiplied by 1.35 to supposedly have the same value as IQR under the assumption of normality...). But why ? It may be correct, but it needs more explanation.

We agree on the need for a clarification about the motivation of calculating a transferability index instead of using the estimated IQR values directly as a transferability measure. Please refer to our reply to the comments 14 and 15 as we clarified the questions of the current comment in our reply to the previous comments and in the corresponding changes to the manuscript.

17. p1017 l17ff: assessing the thematic consistency with an index is a good idea.

The authors are thankful for the positive feedback on the thematic consistency index.

18. p1019 l24: The classified susceptibility map is mentioned here, but the classification rules are introduced in the results chapter (p1020 l7ff). This should be done before/at the first instance when the classified map is introduced (here, or somewhere else in the methods section).

We agree with the referee that the methodology of classifying the map has to be introduced in the methods section. However, we would like to refer to page 1015 of the original manuscript where the classification rules of the final susceptibility map were presented in the paragraph starting from line 12. There we point out that we selected the classes according to the percentage of slides contained in each class. The final thresholds of 5% of slides in the lowest susceptibility class and 70% of slides in the highest susceptibility class was a result of testing different thresholds which were checked in the field according to the best geomorphic and planning plausibility.

Please refer to page 1015, line 12ff of the original manuscript for more details on the classification rules applied within this study. No changes made.

19. p1021 l10f: Considering that AUROC for the spatial (more meaningful) partition is 0.53 (very close to useless), the contrast to AUROC=0.79 (acceptable) for random partition is very good evidence for the consequences of using over-optimistic validation strategies!

The referee correctly identifies one principal finding of our study. We extended our statement on the performance of the spatial cross-validation in the discussion section of the revised manuscript.

Added to page 1028, line 6

> The median AUROC values estimated by spatial and non-spatial cross-validation were similarly high in this study. However, the median AUROC values and the transferability index clearly showed that non-spatial cross-validation provided a more optimistic or maybe also over-optimistic assessment of the model performance and transferability in contrast to spatial cross-validation.

20. p1021 l13: You mention a (1st quantile) AUROC value of 0.35 - but the AUROC only takes values between 0.5 and 1 (see also p1016 l23)!

The authors are thankful for pointing out the source of confusion we created in the original manuscript. We rephrased the sentence on page 1016 in order to be more specific about which values the AUROC can take and which are describing the model's ability to discriminate landslide and non-landslide cells. Values below 0.35 are worse than what would (on average) be expected to be achieved by chance alone, but of course actual estimated values can be worse than that, adding to our argument that resampling-based performance estimation by cross-validation is needed to achieve AUROC estimators of high precision (see comment 11).

Added to page 1016, line 22

> The AUROC takes values between 0 and 1 where a value of 0.5 would be achieved by pure chance agreement between predictions and observations and a value of 1 represent perfect discrimination (Brenning, 2005; Guzzetti et al., 2006); however, this may also indicate overfitting. Thus, the AUROC measures the model's ability to discriminate landslide and non-landslide cells (Hosmer and Lemeshow, 2000).

21. p1022 l12ff: "thus the transferability" - does "thus" also apply for n between 200 and 400 as in line 10f ?...

We agree on the need for more specific wording. The "thus" refers to the increases of the interquartile range at a sample size of about 400 and 200. As the IQR increases the transferability of the model decreases substantially, a trend which becomes stronger the smaller the sample size is.

Added to page 1022, line 12

> ; thus the transferability of the model decreased substantially for sample sizes smaller than 400 (spatial cross-validation) and 200 (non-spatial cross-validation).

22. … Moreover, can you recommend from your findings a minimum sample size?If so, is it related to i) absolute sample size, or ii) to the corresponding sampling rate ? This question can possibly be answered with 16 large domains with different landslide densities...

As stated in the discussion section (pages 1028 & 1029) both the sample size and the sampling rate have an influence on the transferability and on the thematic consistency of the model as analysed within the *k*-fold cross-validation. We found that a small sample size and a small sampling rate referring to a large modelling domain with only few landslide occurrences leads to a low transferability and consistency index. Furthermore, the minimum prediction' standard error is lower with a large sample size.

However, there are exceptions from this general trend (Tab. 3). Therefore we have to state that the minimum sample size is also always related to the topographic and geotechnical characteristics of the study area. For a domain of similar size and homogeneity in local terrain conditions, a minimum sample size around 400 sample cells is recommended. Therewith the transferability and thematic consistency showed a distinct change using spatial cross-validation at this value (Fig. 5 in the original manuscript). However, the influence of a smaller sample size on the geomorphic plausibility (Bell, 2007) of the susceptibility map is unclear. It might be possible that with a smaller sample size the geomorphic plausibility of the map is lower. Nevertheless, the analysis of this was beyond the scope of this study.

We inserted a new small section (7.4) on this in the discussion:

Added to page 1030, line 29

7.4 Considerations on sample size

Summarizing the previously discussed findings some considerations on a minimum sample size might be possible. While the transferability index is less strongly related to sample size or sampling rate the thematic consistency index shows a stronger relationship to them. Generally, larger sample sizes and sampling rates result in better thematic consistency and transferability of the model. Furthermore, the minimum standard error of the prediction was lower with larger sample size (Table 3).

The effect of a reduced sample size on the median and interquartile range AUROC values was assessed in the Flysch domain. We found that the median AUROC remained satisfactory high but decreased as sample size decreased, while the interquartile range of

the AUROC increased. Even with the smallest sample size the model still achieved a good discrimination between landslide and non-landslide cells according to the median AUROC value. Summarizing the results, a minimum sample size with a sum of around 400 slide and non-slide cells might be recommended for the methods applied in this study. This size leads to an acceptable transferability and thematic consistency of the model in spatial cross-validation. However, examples from successfully fitting a susceptibility model with smaller sample sizes (10 landslides with 15 cells each in an area of 177km²; Demoulin and Chung, 2007) give a very contrasting result. Furthermore, the sample needs to be substantially complete which might be difficult to estimate for small samples (Malamud et al., 2004). However, increasing the sample size can only be done by enlarging the landslide inventory (e.g. by selecting a larger study area). This is challenging, as in some regions no additional data on landslides (or resources for mapping landslides) might be available.

This study showed that the general trends found for sample size and sampling rate do not apply for all modelling domains. Therefore, we highlight that the resulting quality estimates (transferability index, consistency index and prediction uncertainty) might additionally be dependent on a combination of the domain size and the landslide density (landslides per km²). Also, dependencies on local terrain conditions and their homogeneity in the modelling domain might exist.

Moreover, the geomorphic plausibility of the susceptibility map has to be analysed. Previous studies found that high performance measures do not always guarantee high geomorphic plausibility of the map (Bell, 2007; Trigila et al., 2013). It might be possible that with a smaller sample size the geomorphic plausibility of the map is lower. However, the influence of a small sample size on the geomorphic plausibility of the susceptibility map is unclear. Nevertheless, analysing this was beyond the scope of this study.

23. p1022 l20f: "one specific random sample and variable selection repetition" - you should delete "repetition" (because the results are based on one sample and subsequent variable selection).

“Repetition” was deleted in the revised manuscript according to your suggestion.

24. p1022 l27: the possible physical meaning of the variable "catchment height" is not explained, neither here nor in the section introducing the variables. See also my comment on p1010 l25ff)

Please refer to our authors reply on the referee's comment number 4.

25. p1023 l8: why "on average" ? You have x model runs, and p percent of them included the variable.

Changed to "on average over all modelling domains."

26. p1024 l20ff: the propagation of uncertainty to susceptibility classes is a very good idea in order not to over-interpret uncertainty while at the same time giving end-users the chance of having a closer look where uncertainty crosses the boundary of one or even two classes.

The authors are thankful for the positive feedback on our proposed methodology. No changes made.

27. p1028 l23ff: what does "adverse effects" mean? Does that mean that the performance measure is biased? or wrong? or that the performance could be better than estimated?

This comment refers to criticism of Guzzetti et al. (2006) on the spatial partitioning of the landslide data into training and test sample. Their discussion states "splitting the study area into two adjacent sub-areas can be problematic" as this approach assumes similar characteristics of the explanatory variables (Guzzetti et al., 2006). However, the statement of Guzzetti et al. (2006) is more understood as a word of caution, pointing out the possible pitfalls of spatial partitioning of the study area. Splitting the study area into modelling domains is a major step in the modelling towards reducing possible differences in the characteristics of the explanatory variables. However, as we stated in the discussion section (7.2), it might still be possible that in some samples one terrain condition (e.g. flat areas) is overrepresented relative to others.

Added to page 1028, line 25

> Here, similar characteristics of the explanatory variables in both training and test sample are assumed and necessary (Guzzetti et al., 2006). If this assumption is not met

by the data (e.g. a rock type or land use class is missing in the test sample) the transfer of the fitted model to the test sample and the estimation of the model performance are difficult (or impossible) (Guzzetti et al., 2006). In our study some model domains might have high contrast between stable (e.g., large flat areas) and unstable (e.g., steep areas) terrain which gives potential for greater variation of sampled terrain conditions; it may be possible that in some samples one terrain condition is overrepresented relative to others.

28. p1028 l28: I slightly doubt that serious over- or underrepresentation is possible with large samples in the order of hundreds to tens of thousands of pixels. Perhaps in very inhomogeneous study areas - but that is being dealt with in your approach by establishing domains (at least with respect to lithology).

The total sample size (double the number of landslides) is rather small compared to the total number of cells available in each modelling domain (represented by the area in km² of the respective modelling domain in Table 1). Therefore, we did not want to exclude the possibility of over- or underrepresentation of terrain conditions in one sample. However, the detailed analysis of this was beyond the scope of this study. No changes made.

29. p1029 l27f: Does an underestimation not occur, for example, in the medium class as well ?

The authors are thankful for pointing out the possible source of misunderstanding. With this sentence we wanted to refer to the main source of underestimation which is represented by the percentages indicated in Figure 6 (in the original manuscript). The overlaps mainly occur between the low and medium susceptibility class (PP→ULCI 6%; LLCI→PP 5%) compared to the percentage of overlapping cells of the medium and high susceptibility class (PP→ULCI 2%; LLCI→PP 2%).

Added to page 1029, line 27

Special attention should be given to the low susceptibility class. Here, the highest percentage of overlapping classes and underestimation of the susceptibility were detected.

30. p1030 l3ff, especially l9ff: I feel that the comparison of your susceptibility map with the hazard zonation plans is not fully feasible. Risks are induced by mass movements not exclusively where they initiate, but also where they stop (in case of mass movements with a considerable runout). For some types of movement, the hazard zonation map needs to assess the runout zone as well. This might or might not be the case for earth and debris slides that represent the main focus of this paper.]

The authors agree with the expressed reservations of the referee. The general limitation of susceptibility maps is the input data used for the modelling. In this study the landslide inventory consisted of a point inventory representing the main scarps. Therefore, the modelled susceptibility is also only showing the probability of the presence of main scarps. The runout zone is not considered explicitly as no runout modelling was performed. However, it might happen more or less accidently that the possible runout zone might be located in the same susceptibility class as the main scarp, due to the adjustment of the probability by the sampling ratio and the later on classification in three classes only. We did an analysis of the coverage of in the classified landslide susceptibility map by landslide polygons available in some parts of the Flysch Zone. We found the majority of the landslide polygons (65%) were located in the high susceptibility class. Furthermore, 29% were located in the medium and only 6% in the low class. Naturally, this is dependent on the selected classification method or thresholds.

Accordingly, a sentence was included in the discussion section of the revised manuscript. The discussion section 7.3 was changed significantly as we put the focus more on the need of visualizing and communicating the results than on the comparison with the hazard zonation plans. This comparison was removed from the revised manuscript. The changes of the discussion in section 7.3 is presented here:

Added to page 1029, line 25

> Some model form uncertainties within this method arise from using the lookup table for transferring the prediction standard error to all grid cells as shown by the range of resulting R^2. This method might be improved or substituted by a function assigning the standard errors to all grid cells.
>
> It was found that in the classified map the majority of grid cells did not change. However, there are differences between the modelling domains where some domains had larger overlaps of different susceptibility classes than others. Special attention

should be given to the low susceptibility class. Here, the highest percentage of overlapping classes and underestimation of the susceptibility were detected.

The visualization of these spatially varying uncertainties is of special interest for future land-use and development planning usually performed by non-landslide experts. In the aftermath of this study each landslide susceptibility class will be related to, not legally binding, recommendations for the designation of new building areas. Therefore, a misclassification (e.g. low instead of medium susceptibility) might lead to an interpretation by the municipality or landowner that underestimated landslide susceptibility. Knowledge about the susceptibility class overlaps might outline where more caution and detailed investigations are necessary. Additionally, it also shows where no uncertainties are expected, which might help to avoid costs for slope investigations.

There is also a need to communicate the research results and their quality with appropriate explanations for the local officials, environmental managers and the public to raise awareness and knowledge on it which leads to an easier understanding and incorporation of the results into the decision-making process (Knuepfer and Petersen, 2002; Rogers, 2006; Brierley, 2009; Hill et al., 2013). This analysis might aid to a good acceptance of the landslide susceptibility maps in the local governments, as instead of a fuzzy statement on involved uncertainties these are clearly shown in a map on grid cell level (Guzzetti et al., 2006; Luoto et al., 2010). Furthermore, the preparation of the susceptibility maps showing the class overlaps contributes to an easier understanding of the possible effects of the prediction uncertainties.

The question if the policy makers or stakeholders are really interested in knowing more about the uncertainty is discussed conversely. The study of Brugnach et al. (2006) pointed out that the confidence in modelling results is dependent on the way the uncertainties are addressed. Policy makers were missing more information on the uncertainty of any model result. Therefore, the modelling results should be presented with a measure of uncertainty or confidence indicator (Brugnach et al., 2006). In habitat suitability modelling the visualisation of uncertainty was identified as relevant to inform decision-makers about areas with extreme error, but also about areas which are particularly well modelled (Elith et al., 2002). This openly addresses the uncertainties involved in the maps instead of giving an impression of certainty (Elith

et al, 2002). However, interviews of Klimeš and Blahůt (2012) showed that local governments do not want any information on uncertainties.

Nevertheless, these uncertainties might have severe consequences on buildings and their inhabitants if an event occurred within the uncertainties of the method used to delineate the hazard zones. The converse discussion shows, that more or better communication with the stakeholders or policy makers (also during the modelling process) is necessary to learn about uncertainties and enlarge confidence into the modelling (Brugnach et al., 2006). However, the way how the uncertainties are presented to the stakeholder has to be adapted by the scientist to ensure the success of the communication. The visualization of some aspects of the quality of landslide susceptibility maps, such as the spatially varying prediction uncertainty, can enhance the communication among experts and decision-makers to facilitate informed decisions (Kunz et al., 2011).

Additionally, further aspects of considering and communicating the effects of epistemic uncertainty are still open research fields in susceptibility modelling. A clear assessment of these is necessary to evaluate on their consequences on the susceptibility (or hazard or risk) map.

31. p1047 Fig. 4: Why does the AUROC for domain 230 scale on a 0-to-1 axis, while the AUROC has a range of [0.5,1]?? This is one of two inconsistent uses of AUROC range (see comment p1021 l13). …

The authors are thankful for pointing out the inconsistencies in the original manuscript. We apologize for the inaccuracy and inserted some details on the AUROC. Please refer to our reply on comment 20 for the changes made in the revised version of the manuscript.

32. … Moreover, the legend for each boxplot should be changed: either use "spCV and nspCV" or (shorter and probably better) "sp and nsp".

The legend of each boxplot was changed according to the suggestion of the referee.

33. p1048 Fig. 5b: Something is wrong with the y axis labels. Either it should be the numbers from 0.00 to 0.10, or from 0.00 to 1.00.

We are grateful for the careful check of our figures. The y axis labels were adjusted to range from 0.00 to 0.10.

Technical Corrections

34. p1013 l13ff: I suggest to split this sentence: "Among the currently available methods for landslide susceptibility modelling a GAM shows a compromise between the flexibility of machine learning algorithms and the smooth representation which results from GLMs such as logistic regression; meanwhile, it still gives the opportunity of a transparent and easy interpretable model (Brenning, 2008; Goetz et al., 2011). Alternatively: ...such as logistic regression while still giving the opportunity..."

The sentence was changes according to the suggestion:

Added to page 1013, line 13

> Among the currently available methods for landslide susceptibility modelling a generalized additive model (GAM) is a compromise between the nonlinear predictive flexibility of machine learning algorithms and the smooth, yet linear, predictions of GLMs such as logistic regression. The model fit of the GAM can be easily interpretable unlike most machine learning algorithms (Brenning, 2008; Goetz et al., 2011).

References

Beguería, S.: Validation and Evaluation of Predictive Models in Hazard Assessment and Risk Management, Nat Hazards, 37(3), 315–329, 2006.

Bell, R.: Lokale und regionale Gefahren- und Risikoanalyse gravitativer Massenbewegungen an der Schwäbischen Alb, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn., 2007.

Beven, K. J. and Kirkby, M. J.: A physically based, variable contributing area model of basin hydrology, Hydrological Sciences Bulletin, 24(1), 43–69, 1979.

Blume, H. P., Horn, R., Brümmer, G. W., Kandeler, E., Kögel-Knabner, I., Kretzschmar, R., Stahr, K., Schachtschabel, P. and Wilke, B. M.: Scheffer/Schachtschabel: Lehrbuch der Bodenkunde, Spektrum Akademischer Verlag.

Brenning, A.: Spatial prediction models for landslide hazards: review, comparison and evaluation, Natural Hazards and Earth System Sciences, 5(6), 853–862, 2005.

Brenning, A.: Statistical Geocomputing combining R and SAGA: The Example of Landslide susceptibility Analysis with generalized additive Models, SAGA—Seconds Out, 19, 23–32, 2008.

Brenning, A.: Improved spatial analysis and prediction of landslide susceptibility: Practical recommendations, in Landslides and Engineered Slopes, Protecting Society through Improved Understanding, edited by: Eberhardt, E., Froese, C., Turner, A. K., and Leroueil, S., Taylor & Francis, Banff, Alberta, Canada., 789-795, 2012a.

Brenning, A.: Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: the R package "sperrorest", IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 23-27 July 2012, Munich, 5372–5375, 2012b.

Breslow, N. E. and Day, N. E.: Statistical methods in cancer research, Iarc Lyon, France. [online] Available from: http://com.iarc.fr/en/publications/pdfs-online/stat/sp32/SP32_vol1-0.pdf (Accessed 16 July 2012), 1980.

Brierley, G.: Communicating Geomorphology, J. of Geography in Higher Educ., 33(1), 3–17, 2009.

Brugnach, M., Tagg, A., Keil, F. and Lange, W. J.: Uncertainty Matters: Computer Models at the Science–Policy Interface, Water Resources Management, 21(7), 1075–1090, 2006.

Chung, C. J. . and Fabbri, A. G.: Validation of spatial prediction models for landslide hazard mapping, Natural Hazards, 30(3), 451–472, 2003.

Chung, C. J. F. and Fabbri, A. G.: Probabilistic prediction models for landslide hazard mapping, Photogrammetric Engineering and Remote Sensing, 65(12), 1389–1399, 1999.

Demoulin, A. and Chung, C.-J. F.: Mapping landslide susceptibility from small datasets: A case study in the Pays de Herve (E Belgium), Geomorphology, 89(3-4), 391–404, 2007.

Elith, J., Burgman, M. A. and Regan, H. M.: Mapping epistemic uncertainties and vague concepts in predictions of species distribution, Ecological modelling, 157(2), 313–329, 2002.

Fabbri, A. G., Chung, C. J. F., Cendrero, A. and Remondo, J.: Is prediction of future landslides possible with a GIS?, Natural Hazards, 30(3), 487–503, 2003.

Frattini, P., Crosta, G. and Carrara, A.: Techniques for evaluating the performance of landslide susceptibility models, Engineering geology, 111(1-4), 62–72, 2010.

Goetz, J. N., Guthrie, R. H. and Brenning, A.: Integrating physical and empirical landslide susceptibility models using generalized additive models, Geomorphology, 129, 376–386, 2011.

Guzzetti, F., Reichenbach, P., Ardizzone, F., Cardinali, M. and Galli, M.: Estimating the quality of landslide susceptibility models, Geomorphology, 81(1-2), 166–184, 2006.

Hammerl, C. and Lenhardt, W.: Erdbeben in Österreich, Leykam, Graz, 1997.

Hanley, J.A. and McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology, 143, 29-36, 1982.

Heckmann, T., Gegg, K., Gegg, A. and Becht, M.: Sample size matters: investigating the effect of sample size on a logistic regression debris flow susceptibility model, Natural Hazards and Earth System Sciences Discussions, 1(3), 2731–2779, 2013.

Hill, L. J., Sparks, R. S. J. and Rougier, J. C.: Risk assessment and uncertainty in natural hazards, in Risk and uncertainty assessment for natural hazards, edited by J. C. Rougier, R. S. J. Sparks, and L. J. Hill, pp. 1–18, Cambridge University Press, Cambridge, 2013.

Hosmer, D. W. and Lemeshow, S.: Applied logistic regression, Wiley, New York, NY, 2000.

King, G. and Zeng, L.: Logistic regression in rare events data, Political analysis, 9(2), 137–163, 2001.

Klimeš, J. and Blahůt, J.: Landslide risk analysis and its application in regional planning: an example from the highlands of the Outer Western Carpathians, Czech Republic, Natural Hazards, 2012.

Knuepfer, P. L. . and Petersen, J. F.: Geomorphology in the public eye: policy issues, education, and the public, Geomorphology, 47(2–4), 95–105, 2002.

Kunz, M., Grêt-Regamey, A. and Hurni, L.: Visualization of uncertainty in natural hazards assessments using an interactive cartographic information system, Natural Hazards, 59(3), 1735–1751, 2011.

Luoto, M., Marmion, M. and Hjort, J.: Assessing spatial uncertainty in predictive geomorphological mapping: A multi-modelling approach, Computers & Geosciences, 36(3), 355–361, 2010.

Malamud, B. D., Turcotte, D. L., Guzzetti, F. and Reichenbach, P.: Landslide inventories and their statistical properties, Earth Surf. Process. Landforms, 29(6), 687–711, 2004.

Remondo, J., González, A., De Terán, J. R. D., Cendrero, A., Fabbri, A. and Chung, C. J. F.: Validation of landslide susceptibility maps; examples and applications from a case study in Northern Spain, Natural Hazards, 30(3), 437–449, 2003.

Rogers, K. H.: The real river management challenge: integrating scientists, stakeholders and service agencies, River Research and Applications, 22(2), 269–280, 2006.

Rossi, M., Guzzetti, F., Reichenbach, P., Mondini, A. C. and Peruccacci, S.: Optimal landslide susceptibility zonation based on multiple forecasts, Geomorphology, 114(3), 129–142, 2010.

Schwenk, H.: Massenbewegungen in Niederösterreich 1953 - 1990, in Jahrbuch der Geologischen Bundesanstalt, vol. 135, pp. 597–660, Geologische Bundesanstalt, Wien., 1992.

Seibert, J., Stendahl, J. and Sørensen, R.: Topographical influences on soil properties in boreal forests, Geoderma, 141(1-2), 139–148, 2007.

Trigila, A., Frattini, P., Casagli, N., Catani, F., Crosta, G., Esposito, C., Iadanza, C., Lagomarsino, D., Mugnozza, G., Segoni, S., Spizzichino, D., Tofani, V. and Lari, S.: Landslide Susceptibility Mapping at National Scale: The Italian Case Study, in Landslide Science and Practice, edited by C. Margottini, P. Canuti, and K. Sassa, pp. 287–295, Springer Berlin Heidelberg. 2013.

Ury, H. K.: Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data, Biometrics, (31), 643–649, 1975.

Wacholder, S., Silverman, D. T., McLaughlin, J. K. and Mandel, J. S.: Selection of controls in case-control studies: III. Design options, American Journal of Epidemiology, 135(9), 1042–1050, 1992.

Van Westen, C. J., Castellanos, E. and Kuriakose, S. L.: Spatial data for landslide susceptibility, hazard, and vulnerability assessment: An overview, Engineering Geology, 102(3-4), 112–131, 2008.