Point-by-Point response to the reviews

Response to Review 1

We thank the Reviewer for her detailed review of our work. We take note of your suggestions to grammatically improve the body text. However, in your specific notes there are some questions and comments that demand a direct response:

Page 3, line 29: which information about landslides location did you collect? Did you digitalize the entire perimeter? Or location refers to the four GPS points? You can explain this better.

We used the four GPS points collected on the field as reference to draw landslide polygons using QGIS software and Google Earth satellite imagery.

To clarify this point, we modified the text as follows:

"We collected information about the location of each observed landslide, four GPS points (crown, toe, and two flanks), photographs, surrounding area features and information about the landslide type, according to the Varnes, (1996) classification. Each documented landslide was drawn and digitized using its four GPS waypoints recorded and photographs as a reference. QGIS and Google Earth satellite imagery were used for the purpose."

Page 3, line 5: why did you consider only shallow landslides?

We considered that different typologies of slope instabilities are triggered by different mechanisms, which means that the predisposing factors and the way they affect to the occurrence of a given type of landslide can be different. Before the field campaign, we reviewed bibliographical sources, finding that within our study area 753 landslides were inventoried in different studies (INGEMISA, 1995; Gipuzkoako Foru Aldundia, 2013; IDE de Euskadi, 2014). Among all of them, 75% were considered as shallow slide type of movement, 10% were rock falls or rock mass deposits and 3% were flows or complex movements, adding to the 12% of the subset that was labelled as landslide scarp but without specifying the type of landslide. Although those bibliographical sources were not considered appropriate for our study – because they were heterogeneous in type and quality – they showed the most frequent type of landslide, so we focused our research on shallow landslides.

We included the information in the text as follows (section 3.1, page 4):

"During several field trips, 793 individual landslides were collected, and 746 of them were classified as shallow movements. Our observations together with the revised bibliographical sources (INGEMISA, 1995; Gipuzkoako Foru Aldundia, 2013; IDE de Euskadi, 2014) confirm that shallow slides are the most frequent type of landslide in the study area. Consequently, in order to consider only landslides triggered by the same mechanisms, only shallow movements were used as landslide presence when defining the dependent variable in the susceptibility assessment."

Gipuzkoako Foru Aldundia (2013). Evaluación y gestión integada de riesgos geotécnicos en la red de carreteras de la Diputación Foral de Gipuzkoa. Technical report, Mugikortasun eta Bide Azpiegituren saila, Unpublished report.

IDE de Euskadi (2014). Infraestructura de datos espaciales de euskadi.

INGEMISA (1995). Inventario y Análisis de las Áreas sometidas a Riesgo de Inestabilidades del Terreno de la C.A.P.V. Technical report, Eusko Jaurlaritza.

Page 6, line 6: here it's not important that LR is implemented in LAND-SE software. Moreover, you already mentioned that you used this software to perform the analysis. What is important is that you applied the multivariate LR. Is it the most used statistical method for susceptibility in general or for landslide susceptibility? Please, specify.

It is actually relevant to mention that we used LR as implemented in the LAND-SE software, since the latter is actually a comprehensive package for data preparation, model training and validation, and visualization of the results. Concerning LR, what the review paper discuss is only its application to landslide susceptibility studies. In the text, we modified "susceptibility" to "landslide susceptibility", to avoid ambiguity.

Page 8, line 9: Why only three random set? What is the implication of using three or more?

The test the Reviewer refers to was carried out just to confirm that the random selection of the landslide inventory would not affect the model results in a relevant way. Indeed, before starting with the main analysis, three preliminary LR runs were performed only changing the training and validation data sets. In all the cases the model classification performances were very similar. So, in order to choose only one data set for further comparative analysis, we decided to select the one with the best classification result, although we believe that conclusions would not be affected if any other data set would have been used.

We included the information in the text as follows (section 4.3, page 8):

"This exercise allowed us to confirm that the random selection of the landslide inventory would not affect the model results in a relevant way, because in all the cases the model classification performances were very similar."

Page 8, line 21: Are you sure you mean >0.15% of unstable pixels?

The total inventoried landslides actually covered 0.15% of the surface of the whole study area. Nevertheless, the presence of one single landslide pixel within a slope unit was not considered enough to label this SU as unstable. Therefore, instead of arbitrarily defining a given threshold value in order to consider a SU as unstable, we decided to use the overall landslide density in the WA.

We included the information in the text as follows (section 4.3, page 9):

"The presence of one single landslide pixel within a slope unit was not considered enough to label this SU as unstable. Therefore, instead of arbitrarily defining a given threshold value in order to consider a SU as unstable, we decided to use the overall landslide density in the WA. For this reason, we considered as unstable those SUs containing equal or more 0.15 % of unstable pixels, and stable otherwise."

Page 8, line 25: do you mean that from your computation it results 304 unstable SUs? If so, please specify.

Yes, in 304 cases the SU contained 0.15% or more unstable pixels.

Page 9, line 32 and Page 10 line 4: why 152 validation SU? They are not 76?

As it was explained in section 4.3 (Page 8, lines 25-30), 76 SU labelled as unstable were used for validation. Then, the validation sample was obtained by selecting the same number of SU labelled as stable. Thus, the validation sample contained 152 SU (76 unstable + 76 stable).

We included the information in the text as follows (section 4.3, page 9):

“In 304 cases the SU contained 0.15% or more unstable pixels, so we selected at random 228 of them (75%) for training, and the remaining 76 (25%) were used for validation. Like in grid cell approaches, we created two different training samples where unstable SUs were exactly the same, and only the stable SUs were different in each case. The first training sample includes 228 stable SUs selected at random along the WA. The second training sample includes an equal number of stable SUs units selected at random among those that at least partially overlap the ESA.”

“Additionally, 76 SUs labelled as unstable were used for validation. Then, the validation sample was completed by adding a random selection of the same number of SUs labelled as stable and which at least partially overlap the ESA. Thus, the validation sample contained 152 SUs (76 unstable + 76 stable).”

Response to Review 2

We appreciate the Reviewer's comments about our research and we are pleased to see that our Manuscript was carefully reviewed. In the following we discuss the Reviewer's comments, including the suggestions on the bibliography and about adding new figures and tables. As a general remark, we would like to stress two points: (i) most of the references suggested by the Reviewer were known to us, and they were not explicitly included in the bibliography just because they appear in the revision papers cited, which represent a useful tool from this point of view; (ii) we believe that some comments contain misunderstandings about the very definition and meaning of the proposed idea of effective surveyed area (ESA), and occasionally they demand deeper explanations from our side. Thus, we will provide detailed answers to the Reviewer's comments below, and amend the Manuscript where needed.

Answers to comments

It is not clear the utility of using the r.survey code to define the ESA, instead of using a simple portable GPS to record the surveyed area.

A portable GPS device cannot be used to delineate areas, but only tracks or waypoints. During our data collection field campaigns we used GPS to record the route we followed. The point, here, is that even if one visits a given basin searching for landslide evidence, it is hardly possible to ensure that every single site of this basin was actually observed. This is due to the fact that some places can be not visible from the route followed during survey, typically existing roads. Thus, touring a basin does not ensure that the whole area was actually surveyed. This is the simple idea behind the ESA. The r.survey module is a tool that delineates the theoretically visible area from the points of view recorded during the field campaign by the GPS tracks. And, most importantly, the ESA, as delineated by r.survey, is an objective and reproducible portion of the study area directly observed by the geomorphologists, thus allowing to avoid arbitrary assumptions about which sites were actually surveyed and which ones were not.

Why do you used different stability thresholds for WA and ESA? How you defined the threshold values? You cannot compare these maps if you used different criteria to perform the analyses.

We did not use different stability thresholds. We carried out a few of the comparative analyses with a different stability threshold as an additional test, but we always performed pairwise comparisons between susceptibility maps obtained with the same threshold value.

The rationale behind testing two different thresholds is the following. The total inventoried landslides covered 0.15% of the surface of the whole study area (WA). Nevertheless, the presence of one single landslide pixel within a slope unit was not considered enough to label this SU as unstable. Therefore, instead of arbitrarily defining a given threshold value in order to consider a SU as unstable, we decided to use the overall landslide density in the WA for all the maps based on SUs, in order to consistently compare maps obtained training the susceptibility model within the WA and within the ESA. Additionally, as an additional control test, we prepared two additional susceptibility maps using as a threshold the landslide density within the ESA (0.33%). We observed that, even if the absolute value of the evaluation indexes (e.g., the $\mathrm{AUC}_{\mathrm{ROC}}$) slightly decrease with respect to the maps obtained using the threshold estimated in the WA (0.15%), the conclusions one can draw from the results of the two approaches are indistinguishable.

One of the main outcomes is that the use of slope units increases the quality of the results, but the performance differences between the different approaches are very low; I believe that such small difference do not justify the whole work.

In our opinion, pixel-based models and SU-based models are not simply comparable using absolute values of their validation performances, since they involve two different concepts of terrain subdivisions, with different meaning and application purposes.

The actual main question answered by our work is about the effect of training a statistical susceptibility model either in the portion of territory which was explicitly surveyed during a field campaign, the ESA, or in the extended area encompassing the ESA, the WA. As statistical methods are widely applied both using grid pixels and slope units, we addressed our question in both cases, since investigating only one of the two cases would not answer the question in the remaining one. Based on the results we obtained, we stated that pixel-based models (WA-PM an ESA-PM) showed more differences than SU-based models (WA-SUM and ESA-SUM). Our work is justified by the fact that we answered the question about relevance of the ESA, never addressed before in the literature, and not that one should use slope units instead of pixels as mapping units.

The small difference exhibited by the WA-SUM and ESA-SUM maps suggests that the use of SU as terrain partition mitigates the uncertainties introduced if the WA is used to calibrate the model. From this point of view, and from this point of view only, we conclude that SUs are best suited as a terrain partition approach. Of course the use of SUs has several additional advantages, but these are beyond the aim of the present work.

If a costs/benefits analysis would be performed, it could result that the WA-PM approach would be the best one.

To our knowledge, no similar analysis exists in the literature concerning landslide susceptibility studies. As a matter of fact, the very definition of costs and benefits associated with preparing landslide susceptibility maps is no easy task. Do we talk about the time spent on a single map, or we include the time during which a researcher builds his expertise, through his particular learning curve? Do we account for numerical models running time, and do we distinguish which software was adopted for the study? Do we talk about the economic cost of the equipment, and do we include the salary of the researchers? Do we distinguish between students and senior researchers? Not to talk about quantifying the benefits, that range from selfishly disseminating one’s own scientific production, to actually providing the scientific community with useful tools, and even to save lives, since we are dealing with hazardous and potentially life-threatening natural hazards. In our opinion, one should stick to the fact that scientific advance is a step-wise process, and every single bit is helpful in producing a more effective and reliable research methodology that, hopefully, can be adopted by practitioners and decision-makers for natural hazard mitigation.

Concerning the specific object of our investigation, the difference in training a statistical model within the ESA or within the WA amounts to collect a few hundred GPS points during the field campaign, which represent a negligible time with respect to the time necessary to conduct the whole field survey, and run the r.survey software for a few seconds to obtain the ESA. Training the statistical model within the ESA does not produce overhead with respect to doing that within the WA. In conclusion, there only are advantages in using the ESA, since the “costs” are negligible.

Moreover, working with field based landslide data, the use of the WA as a calibration area is conceptually incorrect, because there may be places in which the presence or absence of landslides must simply be inferred, with no actual evidence. By means of the ESA, we propose an alternative to revise this conceptual model, and our results suggest that this new approach produce performance advantages in validation.

Answers to comments in the supplementary material

Page 1, line 22: "Please consider also some other works, as: Ermini et al. 2009 (10.1016/j.geomorph.2004.09.025), Catani et al. 2013 (10.5194/nhess-13-2815-2013), Yalcin 2008 (10.1016/j.catena.2007.01.003), etc"

In page 1, line 22 we cited a review paper co-authored by one of us, in which "*an extensive database of 565 peer-review articles from 1983 to 2016*" was compiled. The large number of papers contained in the database is proof that the topic is widely discussed in the literature, and we believe that citing the review paper is the best thing to do, since citing a few papers about landslide susceptibility zonation would be a misrepresentation of the literature. Nevertheless, among the papers suggested by the Reviewer, we acknowledge that Ermini et al. 2009 (10.1016/j.geomorph.2004.09.025) is general enough to be included in the manuscript, also in consideration of its large citation records.

Page 2 line 12: "Please considere also some other works, as Ardizzone et al. 2012,(DOI:10.1080/17445647.2012.694271), Rosi et al. 2018 (doi: 0.1007/s10346-017-0861-4), etc."

As in the previous case, we believe that the references proposed by the Reviewer are not suitable in this specific part of the Manuscript, since they deal with two specific study areas and were aimed at producing specific landslide inventories, instead of investigating specific methodologies. Nevertheless, since the Reviewer points out that the list of references is not exhaustive of the literature, and consistently with what we stated in the previous comment, we prefer citing review papers in this specific point, namely:

Guzzetti et al., (2012) – already listed in the bibliography

Casagli N, Frodella W, Morelli S, Tofani V, Ciampalini A, Intrieri E, Raspini F, Rossi G, Tanteri L, Lu P. 2017. Spaceborne, UAV and ground-based remote sensing techniques for landslide mapping, monitoring and early warning. Geoenviron Disasters. 4(1):9. https://doi.org/10.1186/s40677-017-0073-1

And methodological papers, namely:

Fiorucci F, Cardinali M, Carlà R, Rossi M, Mondini A, Santurri L, Ardizzone F, Guzzetti F. 2011. Seasonal landslide mapping and estimation of landslide mobilization rates using aerial and satellite images. Geomorphology. 129(1–2):59–70. https://doi.org/10.1016/j.geomorph.2011.01.013

Fiorucci et al., (2018) – already listed in the biblography

F Catani, P Farina, S Moretti, G Nico, T Strozzi. On the application of SAR interferometry to geomorphological studies: estimation of landform attributes and mass movements. Geomorphology 66 (1-4), 119-131. https://doi.org/10.1016/j.geomorph.2004.08.012

Page 2, line 22: "Please consider also other papers. eg. Ba et al. 2018 (10.1007/s12145-018-0335-9), Zezere et al, 2017 (10.1016/j.scitotenv.2017.02.188)"

We have added the references to the bibliography.

Page 3, line 1: "It looks useless, since the surveyed area should be already know, once the survey is accomplished."

This comment is somehow difficult to understand, since it refers to the (simple) code developed in this work to define the ESA, which is the central part of the Manuscript. We believe we have addressed this point at length before, in this response to the Reviewer, and demonstrated the usefulness of both the software, to delineate the ESA, and the effectiveness of the ESA, in calibrating statistical susceptibility models.

Page 3, line 1: "This paragraph contains too many acronyms, it is hard to read. Please rephrase."

We modified "*instead of the WA, enhances*" to "*instead of the WA (the whole study area, encompassing the ESA), enhances*" and removed LR, writing explicitly "*logistic regression*", so that the WA is explained right away and at least one acronym is removed.

Page 3, line 23: "this is a generic sentence and the paper is about a case of study (central Italy). This citation is wrong and has to be changed. More relevant paper can be found in literature. Rainfall is the primary triggering factor of shallow landslides, not of rock fall, DSGSD, etc. "

The citation is not wrong, since the whole cited paper deals with the effects of rainfall, and their projected changes, on landslide hazard for (rainfall-induced) shallow landslides. It is correct the cited study is about a case study in Central Italy, but it still is to the best of our knowledge the widest application of physically based slope stability models, in terms of study area extent. We also added one reference to a global analysis, Petley et al., (2005). Moreover, the Referee is more than right that rainfall is *not* the primary triggering factor for other types of landslides, such as rock falls, so we have modified "the primary triggering factor of landslides" to "the primary triggering factor of shallow landslides"

DN Petley, SA Dunning, NJ Rosser, O Hungr - Landslide risk management. The analysis of global landslide risk through the creation of a database of worldwide landslide fatalities Balkema, Amsterdam, 2005.

Page 4, line 20: "the list of variable should be provided before this sentence."

We have added the list of variables, in the revised version of the Manuscript.

Page 5, line 4: "These result have to be proved in some way. I suggest that you add some graphs and table to support these results."

We agree with the Reviewer that the description of the software and its usage should have been done in more detail. We added the following description and data in the Manuscript:

"In a 10 km^2 subset of the study area, we tested the software output using: i) maximum distance between sampled points of 50, 100, 200 and 500 m; ii) the original DEM at 5 m resolution and resampled versions of the DEM at 20, 50 and 100 m resolution; and maximum visible distance of 500 m (the later was dictated by the largest distance between the digitized field path and the farthest landslide pixel in the study area). Results of the test are summarized in Table 1.

Name	Resolution	D_{max}	Percentage of landslides within (%)
Survey_5	5	50	35
Survey_6	20	50	70
Survey_7	50	50	95
Survey_8	100	50	100
Survey_9	5	100	30
Survey_10	20	100	60
Survey_11	50	100	95
Survey_12	100	100	100
Survey_13	5	200	30
Survey_14	20	200	55
Survey_15	50	200	85
Survey_16	100	200	100
Survey_17	5	500	0
Survey_18	20	500	35
Survey_19	50	500	60
Survey_20	100	500	95

As target criteria, we considered that the best setting option was the one which allows covering the totality of the landslides but using the less possible points (bigger D_{max} value) and the lower possible resolution in order to optimize the calculation time.

In the case of the complete study area, the maximum visible distance was set in 1,100 m, in view that the largest distance between the digitized field path and the farthest landslide pixel was 1,092 m, and according to the results of Table 1, the rest of the settings were fixed: maximum sampling distance of 200 m, DEM resolution of 100 m."

Moreover, thanks to the revision process, we detected a mistake in the text of the manuscript. In page 5, line 5, it was written 100m as maximum sampling distance for the application of the r.survey code in the entire study area. The correct setting was of 2,000 meters, instead, as explained above. The mistake will be amended in the revised version of the Manuscript.

Page 5, line 10: "Even if the equation is correct, I hardly can believe that a 2.6 sq.km landslide can be notice from 1 km distance. Maybe in arid or desertic areas, not in this area. Please clarify which is the maximum distance between landslides and ESA borders."

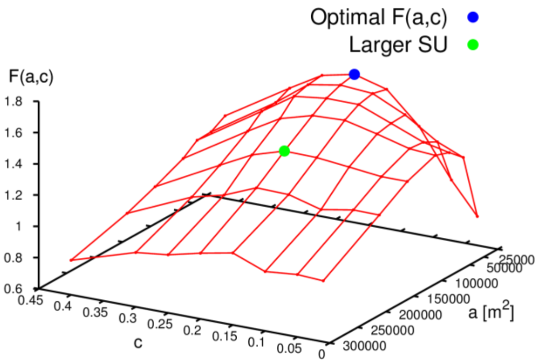
We based our justification in the equation proposed by Rodrigues et al. (2010), that as a published scientific article we consider it reliable. Nevertheless, the smallest shallow slide in our inventory covers 7.3 m^2, which is more than the double area of the theoretical minimum area of an object to be visible from 1,100 m of distance (always according to Rodrigues et al. 2010). So, we considered the maximum visible distance of 1,100 m in r.survey settings was appropriate for our purpose. Moreover, we stated in the Manuscript that the equation serves to "make sense" of the orders of magnitude of the involved quantities. Obviously the simple equation represents an estimate, and neglects vegetation, geometric effects, and so on.

Page 5, line 16: “I suggest that you remove from the code all the unnecessary comments, as unused strings, or comments in italian (if these comments are useful, please translate them).
Data preparation should be better descibed in the readme file (e.g. file format for points (shp, dxf, dwg, etc.) and DTM).”

We fully agree on all the suggestions of the Reviewer and provided a clean version of the python code, along with a detailed description of its usage in the readme file. We apologize for the poor content of these pieces of work.

Page 5, line 22: “These results (*of SU settings*) have to be supported by a more accurate description and, again, some graphs, tables, etc.”

We do not consider SU delineation as a “result” of the Manuscript, since the algorithm, software and optimization method were all described in already published pieces of work. To obtain an automatic delineation of the optimal SU partition in the study area, we used the the r.slopeunit software of Alvioli et al., (2016). We followed the optimization procedure illustrated at length in Alvioli et al., (2016), Schloegel et al., (2018) and Alvioli et al., (2018), so we did not include details in this Manuscript. For the sake of completeness, we can show in this Response to Reviewers some of the details. Optimization of SU involves finding optimal values for the most relevant input parameters of r.slopeunits, namely “circular variance”, *c*, and “minimum area”, *a*. The search for optimal parameters and as a function of (*a*, *c*) is shown in the figure on the left. The figure shows an “aspect segmentation” metric *F*(*a*, *c*), and the two bullets highlight the truly best (a, c) combination (blue) and the one selected for this study (green). The reason why we did not use the “optimal” (a, c) value is that we were working with a limited number of inventoried landslides, and the large number if (smaller, on average) SU contained in the “optimal” subdivision would have led to a critical unbalance between landslides SU and no-landslides SU. So, we selected an SU subdivision with analysis smaller number of polygons, labelled “Large SU” in the Figure.



We do not believe this information actually would add to the content of the paper; we actually believe that it would be off-topic and make the Manuscript’s main message difficult to follow. We consider SU delineation as a technical step and we would not like to include this discussion in the Manuscript.

Page 6, line 7: “If it is the most used, maybe more than 1 work can be cited, you cited only 1 review paper.”

We have here the same motivations as in the previous occurrence of this citations: see comments above. Moreover, in this particular point, we explicitly quote one of the results of the review paper, i.e. that logistic regression is the most used method in the literature, so we are confident that this is the correct citation. Moreover, it is not true that it is the only work we cited: in fact, one line below, we mentioned that logistic regression “proved to be useful” in several studies, among which the three further works we have cited (Nefesliogluet al., 2008; Van Den Eeckhaut et al., 2012; Trigila et al., 2015).

Page 6, line 16: “How do you accounted categorical variables in this equation?”

We addressed this issue in section 3.2 when we state that: “*For categorical variables, we computed frequency ratio (FR) values for each class, and used them as a relative value for their transformation into continuous variables (Lee and Min, 2001; Yilmaz, 2009; Trigila et al., 2015)”.*

So at the end, we accounted for categorical variables in the LR model's equation in the same way as continuous variables.

Page 6, line 23: "Did you perform a Student t-test? Please clarify how the p-value is calculated."

We apply statistical models contained in the LAND-SE software of Rossi and Reichenbach (2016). The software is an R script, and the logistic regression results are obtained using calls to glm (generalized linear model) function. The implementation of the glm function in R is such that it is possible to investigate the estimated standard error of a t-statistic for the null hypothesis of each of the coefficients of the linear model. The p-value represents the probability that the parameter is zero: for p-values much smaller than 0.05 the null hypothesis (vanishing coefficient) is rejected, thus the associated variable is significant for the final result.

Page 7, line 1: "Please add more references"

We have the same comment as above. Moreover, the mentioned metrics are very well-known in the community and need not be supported by additional references, in our opinion.

Page 7, line 15: "Why did you use 2 sigma, instead of 1? Please clarify"

We used 2σ following Guzzetti et al., (2006). The test represents an estimate of the variations obtained when input data is changed at random among many different, equally possible, data sets. Accounting for a 2 σ variation in a Gaussian model means accounting for 95% of the area under the probability density function. In this case, it represents the unit of variations in each pixel, or for each slope unit, which makes rather arbitrary the choice of 2 σ, σ or anything in the same order of magnitude.

Page 8, line 9: Why use three random sets of training/validation partition?

The test the Reviewer refers to was carried out just to confirm that the random selection of the landslide inventory would not affect the model results in a relevant way. Indeed, before performing the main analysis, three preliminary LR runs were performed only changing the training and validation data sets. In all the cases the model classification performances were very similar. So, in order to choose only one data set for further comparative analysis, we decided to select the one with the best classification result, although the preliminary test shows that conclusions would not be affected if any other data set would have been used.

Page 8, line 21: 0.15%? Are you sure? This value is very, very low. Please clarify how you defined this value.

See the second paragraph in Answer to comments section.

Page 8, line 25: Stable and unstable SUs have the same numbers (228 Unstable and 228 stable SUs for training, 76 for validation). Please check these values.

The choice of an equal number of stable and unstable SUs was done on purpose, and it is the standard procedure required by the LAND-SE software for landslide susceptibility assessment. The logistic regression model requires a balanced dataset, in which the number of stable and unstable cases are similar (Costanzo et al., 2014). There are other studies, like Felicísimo et al., (2013), that already tested the influence of introducing a larger number of negative data (no-landslide locations). They concluded that this strategy decreases the performance of the models, which is intuitively understandable. For this reason, in order to complete the training and validation samples, the same amount of SU defined as stable (and randomly selected) was added to each subset. So the training sample contained 456 SUs (228 unstable + 228 stable); and validation sample was composed by 152 SUs (76 unstable + 76 stable). We adopted the same procedure for the pixel based susceptibility maps. We have added the following reference to the bibliography:

Felicísimo, Á. M., Cuartero, A., Remondo, J., & Quirós, E. (2013). Mapping landslide susceptibility with logistic regression, multiple adaptive regression splines, classification and regression trees, and maximum entropy methods: a comparative study. Landslides, 10(2), 175-189.

Page 8, line 32: I believe you should use only those units that are totally within the ESA, you do not know what is going on outside the ESA.

We already considered this issue, because we agree that this way would be conceptually more consistent. However, there are some crucial aspects that have to be taken into account on this respect:

- The ESA is an automatic approximation of the real surveyed area that is likely to be contained in the actual ESA. Thus the fact that a given portion of a SU stay outside the ESA does not necessarily mean that it was not observed.
- Excluding all the SU that strictly were not within the ESA rejected most of the SUs in the training sample, because that prescription would include the SUs exceeding the ESA by a little portion. The opposite situation in which most of the SUs fall outside the ESA in our case occurred fewer times than the other way around.
- In those SUs where landslides were actually observed, they should be automatically considered for the model even if a portion stay outside the ESA, because the presence of instabilities was already confirmed, so the presence or not of instabilities in such theoretically not observed space would change nothing.

Thus, for these reasons, we concluded that the approximations of either excluding or not those SUs that falls partially within the ESA would produce an effect smaller than the difference between the results obtained by training the statistical model within the ESA or within the WA.

We tried to explain this idea in page 8, lines 33 - 35 when we state that "*if a portion of an SU falls within the ESA, that implies that at least one part of the SU was observed. Therefore, using this approach, we can remove at least those SUs that were not surveyed at all*".

Page 9, line 8: From table 1, it results that you did not use topographic parameters (slope, curvature, etc.) in WA-PM. For ESA-PM the only used continuous parameter is senoidal slope. Is it correct? How can you explain it? I suggest to perform some comparisons with susceptibility maps obtained with different sets of variable (all variables, same set of WA map, etc.) to verify if you selected the best set.

In table 1 we showed with an asterisk the final explanatory variables introduced in WA-PM, and two of these are the Surface Area Ratio and the Topographic Wetness Index, which are both topographic and continuous variables. In ESA-PM, the only continuous variable selected as final predictor was the Senoidal Slope. Results of the correlation tests were not included in the manuscript, but they indicated that Slope, Senoidal Slope and SAR are highly correlated between each other, so according to the procedure explained in section 3.2, only the one with the lowest p-value was selected as final predictor. Following the same rationale, TWI was rejected in ESA-PM because of its high p-value.

We applied this simplified and statistically oriented work-flow in order to objectively define the final set of predictors to be introduced in each pixel-based model. We maintain that our final aim is to show the effectiveness of training a statistical model within the ESA as compared to training within the WA, and not to obtain the “perfect” input data set or the “best” susceptibility map, as we stated in the Conclusions. We argue that the comparison is meaningful as long as it is performed between maps obtained consistently with the same input data sets and conditions, pairwise. As a matter of fact, the number of ways the two susceptibility maps in a pairwise comparison can be prepared is virtually infinite. Nevertheless, as we mentioned in the Abstract and in the Conclusions, we strongly believe that our results apply to any similar map pairs obtained with meaningful input data sets and statistical methods, as long as the significance of the statistical analysis is ascertained.

Page 9, line 22: This is not an objective approach. I suggest to perform some comparisons with susceptibility maps obtained with different sets of variable (all variables, same set of WA map, etc.) to verify if you selected the best set.

Se the next answer below.

Page 9, line 29: If Senoidal slope is derivative of slope, why do you not use slope in WA analysis? It should have the same significance.

According to the observations carried out by Santacana et al. (2003) and Amorim (2012), there are some types of landslides, like shallow slides, typically occurring in medium slope areas, decreasing their presence from 45º of slope on. Such a behaviour can be explained with the lack of surface formations in very steep areas, since rocks outcrop in such areas. Thus, depending on the type of landslides considered to the susceptibility analysis, the relation between them and the slope may not be completely linear and positive, because in some cases, from 45º of slope, the more is the slope the less is the probability of finding landslides. Considering that our landslide inventory is about shallow slides, we decided to consider the senoidal transformation of the slope as possible explanatory variable for landslide susceptibility modelling, according to Santacana et al. (2003).

Following the objective procedure to select explanatory variables explained in section 3.2, in ESA-PM senoidal slope was selected over slope because its p-value vas lower. In WA-PM instead, SAR was selected as most suitable explanatory variable over slope and senoidal slope for the same reason.

In the particular case of the SU-based models, explanatory variables are organized in a different way due to the irregular size of the terrain units. We used the mean and the standard deviation of the morphometric variables within each SU as explanatory variables, and the percentage of the area covered by each class of the categorical layers. This made impractical the application of the same variables selection approach, and we acknowledge that a specific variable selection approach for SU models would require further investigations. Nevertheless, in order to ensure the use of significant and non-redundant variables we decided to select at least those variables that were always selected in the pixel-based models. Then we added Slope as a morphometric variable because we considered it more suitable to describe the average morphology within a SU than the Senoidal slope or the SAR.

Santacana, N.; De Paz, A.; Baeza, C.; Corominas, J. y Marturià, J. (2003) A GIS-based multivariate statistical analysis for shallow landslide susceptibility mapping in La Pobla de Lillet area (Eastern Pyrenees, Spain). Natural Hazards 30(3):281–295.

Page 11, line 19: You should clarify how you defined the susceptibility classes reported in fig. 5.

This explanation actually is in page 20, lines 20-23. "*The probability of landslide occurrence resulting from each model estimate (trained either within the ESA or within the WA) and for each considered mapping unit (either grid cells or slope units), can be reclassified in five landslide susceptibility classes which were labelled as Very low (for susceptibility values in the range 0-0.2), Low (0.2 0.45), Medium (0.45-0.55), High (0.55-0.8) and Very high (0.8-1)*".

Page 11, line 31: How can you state that the landslide inventory you used in not complete? How is it possible? This point is crucial, please clarify

We state in page 12 line 24 that the moderate prediction capacity of the resulting susceptibility maps could be, among other reasons due to **"***the lack of more complete landslide inventory*", but it doesn't mean that our landslide inventory was considered incomplete. Even though we acknowledge that more landslides than those included in our inventory probably exist, we maintain that the data collected by direct field observation offers very reliable information, and considering that the landslides size distribution is in agreement with what is expected from a complete inventory (Figs. 2b and c; see Malamud et al. (2004)). We believe that our landslide inventory is complete for the purpose of this research that, we stress, is about showing the effectiveness of training a statistical model within the ESA as compared to training within the WA, and is not about obtaining the "perfect" input data set or the "best" susceptibility map. We added the following text and references the revised Manuscript:

Completeness refers to the proportion of landslides shown in the inventory compared to the real (and most of the times unknown) number of landslides in the study area (Guzzetti et al. 2012; Malamud et al., 2004).

BD Malamud, DL Turcotte, F Guzzetti, P Reichenbach (2004). Landslide inventories and their statistical properties. Earth Surface Processes and Landforms 29 (6), 687-711 https://doi.org/10.1002/esp.1064

Page 12, line 8: This is not trure: in fig.4, 3 out of 4 parameters of WA-SUM are better than those of ESA-SUM.
-Does training ares correspond to ESA? If they are different you should provide a map of the training area.

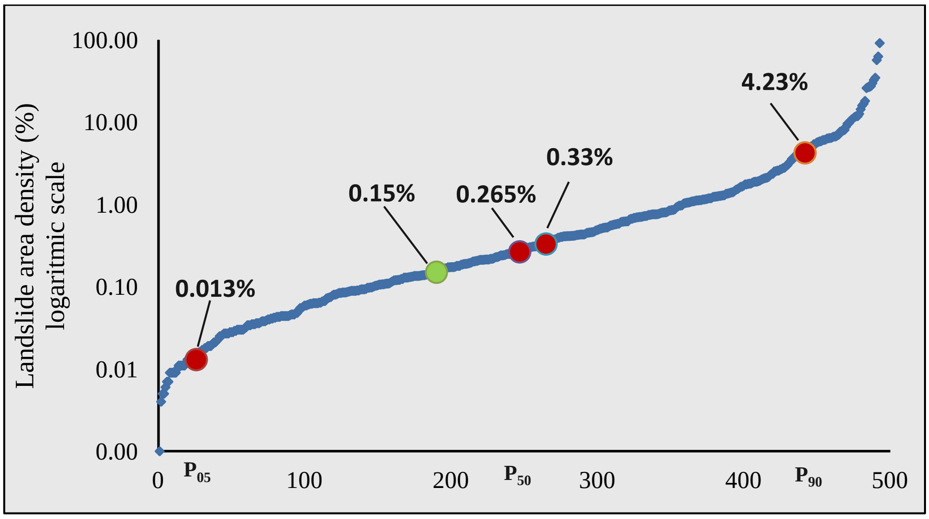
We rephrased this paragraph as follows:

"A straightforward comparative analysis using standard prediction performance metrics revealed that the ESA-based approach is better than the WA-based, at least in grid-cell mapping units based approaches, for what concerned the training area (i.e. within the ESA or WA). Introducing different mapping units in the comparison, we further found that using slope units slightly reduces the gap between results obtained training a statistical model within the ESA versus WA. Thus, the capacity of the slope unit mapping subdivision to mitigate this error was demonstrated, as suitable alternative to the conventional pixel-based approaches."

Response to the Editor

We agree with the Editor about the importance of the threshold limit to consider a SU as unstable, which in our work was defined in 0.15%, the landslide areal density in the whole study area. However, we believe that what the Editor suggested about shifting from constant to spatially variable threshold could result in important and, most importantly, poorly known biases in the results. The use of the average density as a threshold is justified by the fact that measuring landslide presence relative to the average provides an implicit relation between the behaviour each mapping unit and the remaining of the study area. Using a variable threshold would mimic a normalization for landslide presence in the different regions of the study area, while the final objective is precisely the detection of spatial differences about the landslide presence or absence. Additionally, considering that the presented research shows a pairwise comparative study where the effect of other variables is investigated (mainly the effect of using the ESA instead of the WA for training the models), we maintain that a constant threshold limit should be the most appropriate approach.

The other main point of the Editor's comment was the selection of that particular threshold value, which we partially answered above. As we stated in the paper, "instead of arbitrarily defining a given threshold value in order to consider an SU as unstable, we decided to use the overall landslide density". The number of SUs containing at least one landslide pixel is 497. The distribution of the landslide density area among all of them can be shown in the following graph. In order to confirm that the definition of the threshold does not affect the conclusions of the paper, we performed additional calculations considering the Percentile 05, Percentile 50 and Percentile 90 of the landslide density distribution as presence/absence threshold, in addition to the already cited test in the paper about the average landslide density within the ESA, i.e. 0.33%. The resulting values for the area under the ROC curve between ESA and WA maps for these tests are summarized in the table below.



	P_{05} (0.013%)	$WA_{density}$(0.15%)	P_{50}(0.265%)	$ESA_{density}$(0.33%)	P_{90}(4.496%)
ESA-SUM	0.659	0.71	0.679	0.619	0.556
WA-SUM	0.657	0.69	0.672	0.586	0.692
Percentage of unstable SUs	96.4%	61.2%	50.1%	46.5%	10.06%

The results in the Table suggest that for all the cases, except in P_{90}, the model test show better performance for ESA-SUM than for WA-SUM, which is in concordance with the main conclusions of the paper. The case of P_{90} demonstrates the importance of having an adequate amount of data for training and validating a statistically driven model. Because of the high threshold defined in P_{90}, the model was trained with only 74 SUs and validated with 26 SUs, which gives to the result a very poor reliability. On the other hand, the smaller the threshold, the higher is the probability of considering as landslide SU some with really scarce landslide presence, which causing an overestimations of the actual landslide presence in the study area, and that's why the overall AUC values decrease.

In summary, it is true that the definition of the threshold plays a key role in the model performance, but our test shows that our conclusion about the advantage of training the statistical model within the ESA always hold, as long as the mapping units in the training sample are well balanced between stable/unstable. This is intuitively correct, since the statistical model needs enough information about both landslide presence and absence and their relation with the predictors, in order to formulate s correct classification. In addition, a threshold of 0.15% is considered appropriate because it is an objective value obtained from the average landslide density in the study area, while it maintains a balance between an appropriate number of cases (456 for training and 152 for validation) without compromising the introduction of false positive SU classification with a too low threshold.

To clarify this point, we modified the text as follows (we would not like to add new figures or tables):

6 Discussion

(…)

Moreover, since the threshold value for distinguishing stable and unstable SUs could affect the LR model performances, we performed a sensitivity test evaluating the LR models, for both the WA and ESA, using different presence/absence thresholds. We carried out calculations using as a threshold the 5th percentile (P_5, threshold 0.013%), the 50th percentile (P_{50}, threshold 0.265%) and the 90th percentile (P_{90}, threshold 4.5%) of areal landslide distribution, along with the average landslide density calculated within the ESA, i.e. 0.33%. We observed that for all the cases, except in P_{90}, the model tests showed better performance for ESA-SUM than for WA-SUM, which is proof that the conclusions obtained following any approach were indistinguishable. We note that because of the high threshold defined in P_{90}, the model was trained with a very small sample of unstable SUs, which gives to the result a very poor reliability On the other hand, for in the P_5 case, the unbalance does not take place, since each SU where at least one landslide pixel exists belongs to the unstable class, resulting in minimum yet relevant number of unstable SUs. Therefore, we maintain that results of the test confirm that SUs mitigate the relevance of the calibration area (ESA versus WA) when building an SU-based susceptibility model with a field-based landslide inventory, independently of the landslide presence threshold value. However, we acknowledge that the search of an optimal threshold value that ensures a balanced sample is a relevant point, though it is beyond the scope of this work.

List of all relevant changes made in the manuscript

- Some relevant references were added.
- Figure 1. was modified.
- More detailed explanation about the definition of the Effective Surveyed Area and the setting options used in this research were included. Results of exploratory tests for different setting options in a smaller portion of the study area were added to the manuscript.
- A clearer explanation about the data selection for landslides susceptibility was included.
- A more in depth discussion about the relevance of the threshold limit defined to classify SUs as unstable was added.

Effective surveyed area and its role in statistical landslide susceptibility assessments

Txomin Bornaetxea[1], Mauro Rossi[2], Ivan Marchesini[2], and Massimiliano Alvioli[2]

[1]Department of Geography, Prehistory and Archaeology, Faculty of Arts of the University of the Basque Country UPV/EHU. c/ Tomás y Valiente, s/n, 01006, Vitoria-Gasteiz.

[2]Consiglio Nazionale delle Ricerche, Istituto di Ricerca per la Protezione Idrogeologica, via Madonna Alta 126, 06128 Perugia, Italy.

Correspondence: Txomin Bornaetxea (txomin.bornaetxea@ehu.eus)

Abstract. Geomorphological field mapping is a conventional method to prepare landslide inventories. The approach is typically hampered by the accessibility and visibility, during field campaigns for landslide mapping, of the different portions of the study area. Statistical significance of landslide susceptibility maps can be significantly reduced if the classification algorithm is trained in unsurveyed regions of the study area, for which landslide absence is typically assumed, while ignorance about landslide presence should actually be acknowledged. We compare different landslide susceptibility zonations obtained by training the classification model either in the entire study area or in the only portion of the area that was actually surveyed, which we name effective surveyed area. The latter was delineated by an automatic procedure specifically devised for the purpose, which uses information gathered during surveys, along with landslide locations. The method was tested in Gipuzkoa Province (Basque Country), North of the Iberian Peninsula, where digital thematic maps were available and a landslide survey was performed. We prepared the landslide susceptibility maps and the associated uncertainty within a logistic regression model, using both slope units and regular grid cells as reference mapping unit. Results indicate that the use of effective surveyed area for landslide susceptibility zonation is a valid approach to minimize the limitations stemming from unsurveyed regions at landslide mapping time. Use of slope units as mapping units, instead of grid cells, mitigates the uncertainties introduced by training the automatic classifier within the entire study area. Our method pertains to data preparation and, as such, the relevance of our conclusions is not limited to the logistic regression but are valid for virtually all the existing multivariate landslide susceptibility models.

1 Introduction

Landslide susceptibility is defined as the likelihood of a landslide occurring in an area on the basis of the local terrain and environmental conditions (Brabb, 1984; Guzzetti et al., 2005). Landslide Susceptibility Zonation (LSZ) is important for landslide mitigation plans, since it supplies planners and decision makers with essential information (Van Den Eeckhaut et al., 2012). A large number of LSZ studies based on statistical methodologies (Reichenbach et al., 2018) and comparative studies (Cascini, 2008; Das et al., 2010; Schicker, 2010; Amorim, 2012; Blais-Stevens et al., 2012; Trigila et al., 2015; Wang et al., 2015) were published in the last decades. Many statistical methods, aimed at estimating the propensity of a territory to experience slope failures, rely on landslide inventory maps and spatial thematic layers as predisposing factors (Ermini et al., 2005; Van Den Eeckhaut et al., 2006; Camilo et al., 2017).

In statistical landslide susceptibility models, as the logistic regression (LR) model adopted in this work, the preparation of the training data set is a fundamental and critical step. Commonly, this requires the selection of a sample of stable (without landslides) and unstable (with landslides) mapping units. While assuring the presence of a landslide is straightforward, and it can be supported by the geomorphological signatures on the slope or by direct observation of the events, the selection of landslide-free areas is more critical. Assuming as landslide-free the locations of a study area where no landslides were reported in a field survey is correct only in the unlikely circumstance that the landslide inventory has been prepared surveying every single site of the study area, and following homogeneous criteria. In other words, any landslide-free location in an inventory map should have been explicitly checked to be free from landslides.

Nowadays, there are methods based on the visual interpretation of aerial photographs or digital processing of remotely acquired optical and radar imagery (Catani et al., 2005; Herrera et al., 2009; Fiorucci et al., 2011; Casagli et al., 2017; Mondini, 2017; Fiorucci et al., 2018; Alvioli et al., 2018b), that allow to prepare historical and event landslide inventories. However, the adoption of such methods can be hampered by the lack of imagery or image interpretation ~~/classification~~ expertise, low performance of automatic classification, and other factors. Alternatively, bibliographic sources like newspapers and news feeds, administrative reports or scientific literature can be used to obtain landslide information. Nevertheless, the downside of these type of data is that they hardly are as accurate as required by LSZ studies. As a consequence, sometimes the best option to obtain a reliable landslide inventory is a straightforward geomorphological field mapping. A detailed discussion about the characteristics, advantages and limitations of different approaches for landslide mapping can be found in Guzzetti et al. (2012); Santangelo et al. (2015); Fiorucci et al. (2018).

An operational disadvantage of field-based landslide mapping is the difficulty in surveying the whole area where the LSZ must be carried out since some places can be inaccessible or not visible from the accessible places. Difficulties in surveying the landscape affect the completeness and the spatial representativeness of the landslide inventory and, as a result, inclusion of non-visible areas within a landslide inventory introduces a bias since presence or absence of landslides cannot be ascertained in such portions of landscape. This uncertainty has hardly been considered in existing studies that use field-based landslide inventories (Yesilnacar and Topal, 2005; Murillo-García et al., 2015; Wang et al., 2017).

On the other hand, selection of an appropriate terrain subdivision is also a critical ~~phase~~ step in LSZ analysis. The land surface can be divided in portions following geomorphologic features using terrain units, topographic units, geo-hydrological units or slope units, but also considering thematic layers resulting in unique condition units or administrative units, as well as regular grid cells partitions (Van Den Eeckhaut et al., 2006; Reichenbach et al., 2018). Selection of different mapping units can result in considerable differences in the susceptibility assessment (Carrara et al., 2008). In this work, we considered grid cells and slope units (Carrara et al., 1991, 1995; Guzzetti et al., 2006; Alvioli et al., 2016; Zêzere et al., 2017; Rosi et al., 2018; Ba et al., 2018), and investigated the effect of the different ways of training LSZ models within both types of mapping units.

We propose an automatic and reproducible procedure to delineate the actual area which was explicitly surveyed in preparing a landslide inventory by geomorphological field mapping, *i.e.* the effective surveyed area (ESA), and to use such relevant information in statistical analyses. The procedure allows to carry out the calibration of ~~the~~ a statistical model within the ESA and then to apply the resulting susceptibility model to the whole area (WA) under investigation. Moreover, we implemented an automatic approach for the delineation of the ESA in a newly developed ~~software~~ GRASS GIS module named *r.survey.py* ~~(see section 3.3)~~. The software ~~is able to delineate the ESA selecting the visible area from a given set of points of view in an objective and reproducible way.~~ delineates the theoretical visible areas from the points of view recorded during a field campaign by the GPS tracks. Most importantly, the ESA delineated by r.survey.py is an objective and reproducible portion of the study area directly observed by the geomorphologists, thus allowing to avoid arbitrary assumptions about which sites were actually surveyed and which ones were not.

This work aims at demonstrating that the calibration of a landslide susceptibility model within the ESA, instead of the WA (the whole study area, encompassing the ESA), enhances the performance of model itself. In a test study area, we calibrated the multivariate logistic regression model for landslide susceptibility~~, implemented in the LAND-SE module of Rossi and Reichenbach, 2016~~ in four different ways, combining two different calibration areas (ESA and WA) with two different mapping unit types: (i) a regular grid cell partition with a ground resolution of 5 m x 5 m and (ii) a~~n~~ slope unit (SU) partition (consisting in irregular terrain subdivisions bounded by drainage and divided lines)~~obtained using the GRASS GIS module *r.slopeunits* developed by Alvioli et al, 2016 (see section 3.4)~~.

The paper is organized as follows. Section 2 provides an overview of the study area. Section 3 shows the details about ~~the~~ data acquisition; in particular, the r.survey is described in 3.3 and SU delineation in 3.4. Section 4 contains a general description about the multivariate method applied to model the landslide susceptibility and the approach followed ~~for~~ to validate ~~it~~ model results, as well as a detailed description about the set-up of the different model assessments. Results are described in Section 5, and are further discussed in Section 6. Eventually, our conclusions are drawn in Section 7.

2 Study Area

The Gipuzkoa Province was selected as test study area. It is located in the north of the Iberian Peninsula along the western end of the Pyrenees, and covers an area of 1980 km^2, with altitude ranging from the sea level to 1528 m a.s.l. Six watersheds of different size drain the study area and reach the Cantabrian Sea (Fig. 1a). The Province is characterized by a steep morphology with 55% of its surface having a slope larger than 15°.

The investigated area is lithologically heterogeneous (Fig. 1b), with materials ranging from Paleozoic rocks to Quaternary deposits (EVE, 2010), and it corresponds to a hilly and mountainous Atlantic landscape (Mücher et al., 2010). The average annual precipitation is 1,597 mm (González-Hidalgo et al., 2011) with two maximum periods: 34% in November-January and 10% in April. Even though rainfall is the primary triggering factor of shallow landslides (Petley et al., 2005; Alvioli et al., 2018a), anthropogenic slope modifications such as slope clearings and forest extraction activities also strongly affect landslide occurrence (Corominas et al., 2017) in the area.

3 Data preparation

3.1 Landslide inventory

We prepared a landslide inventory by a direct geomorphological field survey, during the period from June to August, ~~of~~ 2016. We collected information about the location of each observed landslide, four GPS points (crown, toe, and two flanks), photographs, features of the surrounding area ~~features~~ and information about the landslide type, according to the Cruden and Varnes (1960) classification. Each documented landslide was drawn and digitized using its four GPS waypoints recorded and photographs as a reference. The QGIS software and Google Earth satellite imagery were used for the purpose. Moreover, and most importantly ~~for the aim of this work~~ to define the ESA, we digitized the route followed during the field survey. This information ~~will be later~~ was then elaborated ~~for the definition of the ESA,~~ using a GRASS GIS module developed for the purpose and included in this work as supplementary material.

~~Each documented landslide was drawn and digitized using a combination of the QGIS module and Google Earth satellite imagery.~~ As a result of several field trips, 793 individual landslides were collected; 746 of them were classified as shallow movements (Fig. 2a). Our observations together with the existing literature (INGEMISA, 1995; Gipuzkoako Foru Aldundia, 2013; IDE de Euskadi, 2014) confirmed that shallow slides are the most frequent type of landslide in the study area. Consequently, in order to consider only landslides triggered by the same mechanisms, only shallow movements ~~The latter~~ were used to determine landslide presence when defining the dependent variable in the susceptibility assessment. Figures 2b and 2c show the distribution of landslide sizes, highlighting that a difference of five orders of magnitude exists between the smallest and the largest inventoried shallow slide.

3.2 Explanatory Variables

The selection of the appropriate explanatory variables to build a landslide susceptibility model is an important step (Ayalew and Yamagishi, 2005; Schlögel et al., 2018), and no universal criteria nor guidelines exist for the purpose.

We obtained relevant environmental digital layers from the Spatial Data Service of the Basque Country[1], and created 13 maps describing the different explanatory variables (see Table 2). To produce derived morphometric continuous variables, such as *Slope*, *Sinusoidal slope*, *Surface area ratio (SAR)*, *Terrain wetness index (TWI)*, *Curvature*, *Plan curvature* and *Profile curvature*, we used a DEM raster layer with 5 m x 5 m spatial resolution. *Sinusoidal slope* is a derived morphometric variable proposed by Santacana Quintas (2001) and Amorim (2012) to emphasize the fact that shallow slides typically occur in medium slope areas, while they seldom occur on slopes steeper 45°. For categorical variables, such as *Lithology*, *Permeability*, *Regolith thickness*, *Land Use*, *Vegetation and Aspect*, we computed frequency ratio (FR) values for each class, and used them as a relative value for their transformation into continuous variables (Lee and Min, 2001; Yilmaz, 2009; Trigila et al., 2015). We acknowledge that the FR values can vary depending on the portion of the territory considered as the total area (ESA or WA).

[1] http://www.geo.euskadi.eus

In order to perform a direct comparison, we decided to maintain the same FR values (calculated considering the WA) in all regular grid cell-based susceptibility analysis.

In this work, we first adopted grid cells as mapping units, and we applied a simplified and statistically oriented work-flow that ensured that only significant variables were taken into account as well as the non-redundancy of the contributed information by each covariate (Ayalew and Yamagishi, 2005). To do so, the whole set of 13 variables were considered within the LR analysis, and correlation coefficients were computed. We considered as collinear two variables when their correlation coefficient is greater than 0.5 with a significance level of 0.01. In such a case, as an objective criterion for variable selection, the variable with highest p-value between the two (see section 4.1), was not taken into consideration in the final run of the susceptibility LR model. Additionally, variables with p-value higher than the threshold of 0.05 were rejected.

Then, considering the variables actually used for the application of the statistical models with grid cells, we have further restricted the set of variables to be used with slope units (see section 5.2).

3.3 Definition of the effective surveyed area

In this work we suggest the concept of ESA, and training of statistical models therein, as an approach to be used to train a landslide susceptibility model avoiding assumptions about the presence or absence of landslides in areas not explicitly observed. We ~~defined~~ delineated the ESA by means of the newly developed GRASS GIS python module *r.survey.py* (see supplementary material). ~~The module makes use of the map of~~ Input data to define the visible area (*i.e.* ESA in our case) are: i) a sample of points to be considered as points of view; ii) a DEM of the area; iii) the maximum visible distance. ~~the routes followed by the geomorphologists during field surveys and few additional inputs; ii) the maximum distance between points sampled from the path; iii) a DEM of the area; iv) the maximum visible distance.~~ The sample of points of view, in our case, was defined re-sampling a given number of points along the recorded path during the field campaigns. This number of points depends on the maximum distance set between them, and together with the DEM resolution selected the results can be directly affected. In a 10 km^2 subset of the study area, we tested the software output using: i) ~~GPS points recorded along the path followed during th field campaign ; i~~ i) maximum distance between sampled points of 50, 100, 200 and 500 m; ~~i~~ii) the original DEM at 5 m resolution and resampled versions of the DEM at 20, 50 and 100 m resolution; ~~iv~~iii) maximum visible distance of 500 m (the later was dictated by the largest distance between the digitized field path and the farthest landslide pixel in the subset of the study area). Results of the test are summarized in Table 1 .

We considered that the best setting option was the one which allows covering the totality of the landslides using the smallest number of points (~~bigger~~ larger D_{max} value) and the lower DEM resolution in order to optimize the calculation time. In our case, considering the whole study area, the maximum visible distance was set to 1,100 m, in view that the largest distance between the digitized field path and the farthest landslide pixel was 1,092 m. Then, and according to the results of Table 1, we set the maximum sampling distance to 200 m and adopted a DEM resolution of 100 m. ~~In a 10 km^2 subset of the study area, we have tested the module output using: i) maximum distance between sampled points of 50, 100, 200 and 500 m; ii) maximum visible distance of 500 and 1100 m (the later was dictated by the largest distance between the digitized field path and the farthest landslide pixel in the study area, 1092 m); ii) the original DEM at 5 m resolution and resampled versions of~~

~~the DEM at 20, 50 and 100 m resolution. From the experiment, we concluded that the best settings in terms of accuracy and processing time were: maximum sampling distance of 100 m, maximum visible distance of 1100 m, DEM resolution of 100 m. We used these values for the analysis of the entire study area to define the overall ESA.~~

We can make sense of the numerical values of the parameters used in the *r.survey.py* ~~software~~ module considering that the minimum size A of an object visible from a distance Δ is given by Rodrigues et al. (2010) and Minelli et al. (2014):

$$A = \frac{25\,\Delta^2}{c}, \tag{1}$$

where c is a steradiant to square minutes conversion factor, $c \simeq 1.18 \cdot 10^7$. Using Δ = 1,100 m in Eq. (1), we get A = 2.6 m^2, meaning that the smallest landslide in our inventory, with size 7.3 m^2, would actually be identifiable from at least one point along the route, if the landslide sits within the ESA. The resulting ESA covers 44.24% of the entire study area and it is shown in Fig. 2a.

3.4 Slope units delineation

For SU delineation we have adopted the *r.slopeunits* software described in Alvioli et al. (2016). The software is a GRASS GIS module, as the *r.survey.py* code presented in this work, and it was designed for the automatic and adaptive delineation of SUs, given a DEM and a set of user-defined input parameters. The code can be used to produce several SU partitions, using different combinations of the input parameters, which can thus be tuned according to user-defined criteria. We partially followed Alvioli et al. (2016), in that we selected the best SU partition ~~maximizing~~considering the quality of terrain aspect segmentation. In addition, we have performed preliminary tests using the LR susceptibility model, showing that the use of very small SUs provides unrealistic results, which can be understood considering the limited variability of variables within such small SU polygons. We concluded that, in the case of the Gipuzkoa Province the most suitable SU partition for landslide susceptibility zonation should be obtained with the following *r.slopeunits* input parameters: flow accumulation area threshold t = 1 km^2; minimum SU planimetric area a = 0.15 km^2; minimum circular variance of terrain aspect within each SU c = 0.2; reduction factor r = 5; threshold value for the cleaning procedure *cleansize* = 0.025 km^2. As a result, we obtained a set of SUs which range in size from 0.026 km^2 to 3.6 km^2 with average 0.28 km^2. A discussion of SU delineation and optimization of input parameters can be found in Alvioli et al. (2016) and Schlögel et al. (2018), and it is out of the scope of this work.

4 Modelling framework

We prepared f~~F~~our landslide susceptibility maps (LS maps), ~~were prepared~~ by means of a ~~statistical approach~~ multivariate LR model. ~~All the maps were obtained using LR and their c~~Classification performances were measured by means of a set of validation tests explained in the following sections. We prepared the first two maps using 5 m x 5 m regular grid cells as mapping units. The two maps differ because in one case the LR model was calibrated within the ~~ESA~~ WA, and within the ~~WA~~ ESA (described in Section 3.3) in the other case. The third and fourth LS maps, instead, were prepared with different mapping units, namely with SUs (described in Section 3.4) instead of grid cells, where calibration data were also changed considering data within WA in one case and within ESA in the other. We end up with four maps, which we name as follows: WA-PM (whole area, pixel map), ESA-PM (effective surveyed area, pixel map), WA-SUM (whole area, slope units map) and ESA-SUM (effective surveyed area, slope units map).

4.1 Logistic regression

We used logistic regression (Hosmer Jr et al., 2013), one of the multivariate statistical approaches available in the LAND-SE software (Rossi and Reichenbach, 2016), to build the landslide susceptibility model in the test study area. The method is the most used in the scientific literature (Reichenbach et al., 2018) and proved to be useful and reliable in several studies (Nefeslioglu et al., 2008; Van Den Eeckhaut et al., 2012; Trigila et al., 2015). The LR model works with either continuous or categorical independent variables, or a combination of the two types, regardless whether they ~~present a normal distribution~~ are normally distributed or not (Costanzo et al., 2014).

The ~~underlying~~ mathematical relationship between the dependent dichotomous variable (presence/absence of a landslide in the mapping unit; Y in the following) and the n independent variables (e.g., slope, lithology, etc.; X_1, ..., X_n), within th eLR model, reads as follows:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n \,, \tag{2}$$

where β_0 is the intercept of the model and $\beta_1, ..., \beta_n$ the linear regression estimate coefficients. The independent (explanatory) variables, X_1, ..., X_n, included in our case both continuous and categorical layers (the latter were previously transformed into continuous variables, as described in section 3.2); see Table 2 for ~~a~~ the full list of variables used in this work. Calibrating an LR model amounts to selecting numerical values for the $\{\beta_i\}_{i=1}^{i=n}$ coefficients in Eq. (2) that maximize the agreement between model output, i.e. landslide probability:

$$P = \frac{1}{1+e^{-Y}} \,, \tag{3}$$

and empirical landslide data, in a training area. The same values of the coefficients can then be used to validate the model prediction skills in a different area, where landslide conditions are unknown to the model but the same explanatory variables layers exist.

In addition to the β coefficients, the LR method also offers a significance p-value for each explanatory variable. The implementation of the `glm` function of the R programming language library[2], used in the LAND-SE software, is such that it is possible to investigate the estimated standard error of a t-statistic for the null hypothesis of each of the coefficients of the linear model. The p-value represents the probability for the parameter to be zero: for p-values smaller than 0.05 the null hypothesis (vanishing coefficient) is rejected, thus the associated variable is significant for the final result. ~~A p-value higher than 0.05 indicate a weak relation between explanatory variables and the dependent variable. In such a case the variable is not statistically significant and it can be removed from the analysis. Conversely, in statistical terms, those predictors with p-value under the~~

[2] https://www.r-project.org

~~threshold value of 0.05 are all equally significant.~~ So, the p-value can be considered as an objective indicator for the selection of the most relevant variables to be used in the statistical model (Schlögel et al., 2018).

4.2 Evaluation of model performance

The performance of statistical susceptibility models, *i.e.* of multivariate binary classifiers, can be evaluated comparing ~~its~~ their predictions with the landslide data used in the model calibration/training step ~~phase,~~ (*i.e.* model fitting performance), or with independent landslide data (*i.e.* model prediction performance). The definition of training and validation input samples is crucial to detect how well each model fits input data, but also how good is the model at predicting new data.

The statistical metrics commonly used in the literature (Corominas and Mavrouli, 2011; Van Den Eeckhaut et al., 2006; Lombardo et al., 2015; Reichenbach et al., 2018) for that purpose are (i) confusion matrices (contingency tables) and their graphical representation (four-fold or contingency plots), (ii) Receiver Operating Characteristic (ROC) curves and their associated Area Under Curve (AUC) value, (iii) classification error plots and (iv) Cohen's kappa index.

Four-fold (or contingency) plots are visual representations of the confusion matrices reporting the percentages of the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). ROC curves are more complex representation of the classification performance based on different probabilistic threshold values. The area under the ROC curve (AUC) is an indicator of the model performance in predicting landslide susceptibility. AUC values vary between 0 and 1, with higher values indicating better prediction skills (Fawcett, 2006).

To estimate the uncertainty associated with the landslide susceptibility value assigned to each mapping unit, it is possible to run multiple instances of the LR model varying, randomly, the input data. In each run, the input is prepared sampling the original training data set with a bootstrap technique, consisting in a random sampling with replacement (Efron, 1992; Davison and Hinkley, 1997) (Rossi et al., 2010; Rossi and Reichenbach, 2016). Classification error plots summarize the distribution of multiple results and show the mean probability estimate of landslide spatial occurrence for each mapping unit (x-axis), ranked from low (left) to high (right) values, related to the variation of the model estimate (y-axis), measured by 2 standard deviations (2σ) of the probability estimates obtained by the different model runs (Guzzetti et al., 2006). The parabolic model fitting equation resulting from the point cloud (*i.e.* using a non-linear least square method), describes analytically the overall model prediction performance variability. Cohen's kappa index (k) is an additional measure of the reliability of a classification model (Cohen, 1960; Rossi et al., 2010), whose higher values also indicate a more accurate prediction skill.

In this study the probability of landslide occurrence resulting from each model estimate (trained either within the ESA or within the WA) and for each considered mapping unit (either grid cells or slope units), ~~can be~~ was reclassified in five landslide susceptibility classes which were labelled as Very low (for susceptibility values in the range 0-0.2), Low (0.2-0.45), Medium (0.45-0.55), High (0.55-0.8) and Very high (0.8-1).

~~In this work~~ Moreover, in order to spatially identify the pairwise matching degree between different model estimates, we additionally adopted a simplified classification of the landslide susceptibility. Each mapping unit was reclassified as stable or unstable considering a threshold value of 0.5. The different maps, all prepared with the same mapping unit partition, were overlapped. Then, the mismatch degree between grid cell and SU susceptibility maps was quantified in terms of number of mismatched mapping units and overall mismatched area.

4.3 Data selection for landslide susceptibility

The DEM available for the study area consists of $7.91 \cdot 10^7$ cells with 5 m resolution. For landslide susceptibility assessment, both using ~~pixels~~ grid cells (*i.e.* pixels) and SUs, we prepared raster layers corresponding to each available explanatory variable, aligned to the DEM grid cells.

We devised a rigorous sampling procedure to minimize possible statistical biases during training/validation partition. The procedure is slightly different for the ~~pixel~~ grid cell and SU mapping units cases.

In the first case, a grid cell was considered unstable if it is located within any landslide ~~area~~ polygon, and stable if it is outside the landslide boundaries. In the second case, an SU was considered unstable depending on the percentage of landslide area present within it. In any case, the 75% of the unstable mapping units together with a similar amount of stable mapping units were used to train the LR model, and the remaining 25%, also together with a similar amount of stable mapping units, for validation. The choice of an equal number of stable and unstable mapping units was done on purpose, and it is the standard procedure required by the LAND-SE software for landslide susceptibility assessment, because the LR model requires a balanced data set, in which the number of stable and unstable cases are similar (Felicísimo et al., 2013; Costanzo et al., 2014).

For regular grid cell-based models, we selected at random 558 landslides (75%) for model training, and converted them into raster layers (84,623 unstable pixels). The remaining 188 landslides (25%), used for validation, were also rasterized (29,247 unstable pixels). This is at variance with the usual random selection of unstable pixels, in which a given percentage of grid cells are sampled within landslide. Here we select whole landslides, and consider all the pixels encompassed by the landslide bodies as training/validation samples. We ran the experiment with three different training/validation random sets, containing the above percentages~~, and selected the one with the best classification results~~. This exercise allowed us to confirm that the random selection of the landslide inventory does not affect the model results in a relevant way, because in all the cases the model classification performances were very similar. In order to choose one single data set for further comparative analyses, the data set with the best classification result was selected. Then, training sets were selected as follow: 84,623 unstable pixels and an equal number of stable pixels were selected as training set. Two different sets were selected at random first within WA and then within ESA. We made sure that unstable pixels were exactly the same in the two cases, because we wanted the only difference to be that the stable pixels were sampled within the WA, in the first case, and within the ESA, in the second case. Finally, in order to guarantee the comparability of the prediction performances, one unique validation sample was created as follow: the remaining 29,247 unstable pixels together with an equal number of stable pixels selected at random within the remaining stable pixels within the ESA.

~~Then, two different training samples were created. The entire training pixel sample includes 84,623 stable pixels (selected at random within the WA) and an equal number of unstable pixels. The effective training pixel sample includes again 84,623 stable pixels (selected at random within the ESA) together with the same unstable pixels as in the other case. Additionally, we selected at random 29,247 grid cells among the remaining stable pixels and an equal number of unstable pixels; we named~~

~~this set as validation pixel sample and we used it to evaluate the model prediction performances. This is to guarantee the comparability of the prediction performances for both the models trained in WA and in ESA, respectively.~~

Concerning the SU-based models, we first partitioned the study area in 6,907 SUs with the technique outlined in Section 3.4. SU boundaries do not match those of the dependent or explanatory variables layers, allowing the presence of different classes, or values, inside each SU. Moreover, the presence of one single landslide pixel within a slope unit was not considered enough to label this SU as unstable. Therefore, instead of arbitrarily defining a given threshold value in order to consider an SU as unstable, we decided to use the overall landslide density in the WA. For this reason, we considered as unstable those SUs containing 0.15% or more unstable pixels, and stable otherwise. We used as explanatory variables the mean and the standard deviation of the morphometric variables for each SU and the percentage of the area covered by each class of the categorical layers. In 304 cases the SU contained 0.15% or more unstable pixels, so we selected at random 228~~, out of the 304 unstable SUs,~~of them (75%) for training, and the remaining 76 (25%) were used for validation. Like in grid cell approaches, we created two different training samples where unstable SUs were exactly the same, and only the stable SUs vary in each case. The first training sample includes 228 stable SUs selected at random along the WA. The second training sample includes an equal number of stable SUs units selected at random among those that at least partially overlap the ESA. Additionally, 76 SUs labelled as unstable were selected from the hwole set, for validation. The validation sample was completed by adding a random selection of the same number of SUs labelled as stable and which at least partially overlap the ESA. Thus, the validation sample contained 152 SUs (76 unstable + 76 stable). ~~The entire training SU sample includes 228 stable SUs selected at random along the entire study area together with the training unstable SUs. On the other hand, effective training SU sample includes 228 stable mapping units selected at random among those SUs that at least partially overlap the ESA together with the training unstable SUs. In order to complete the validation SU sample, we also selected 76 stable SUs at random among those that partially overlap the ESA.~~

~~Finally, considering that~~ Eventually, since the ESA is an approximation of the real surveyed area, we stress that we always selected stable mapping units for validation only if they are fully or partially within the ESA, because no evidence exists that a mapping unit falling entirely outside the ESA is actually free from landslide. Moreover, if a portion of an SU falls within the ESA, that implies that at least one part of the SU was observed. Therefore, using this approach, we can remove at least those SUs that were not surveyed at all.

5 Results

5.1 Susceptibility maps using grid cells

We ran the LR model using the pixel-based data sets twice: once using the entire training pixel sample and once using the effective training pixel sample as dependent variables. We defined the obtained results as whole area pixel map (WA-PM) and effective surveyed area pixel map (ESA-PM), respectively.

Both in WA-PM and ESA-PM, we first used the same 13 explanatory variables, listed in Table 2, and then we selected for each model assessment the most relevant explanatory variables considering the collinearity between each pair of variables and the significance (p-value) of the regression estimates (see section 3.2). As a result, for each case, only the variables marked with an asterisk in Table 2 were introduced in the final LR analysis.

Using the validation pixel sample, we evaluated the prediction skills of the pixel susceptibility maps. Inspection of the four fold, or contingency, plots (Figs. 3a, d) reveals that WA-PM predicted correctly the 63.58% (TP+TN) of the observed unstable and stable mapping units, whereas ESA-PM was capable to correctly predict a higher amount of mapping units (65.45%). The ROC curves (Figs. 3b, e) also indicate better prediction skills in ESA-PM (AUC = 0.7) than in WA-PM (AUC = 0.68) and the same happens for the Cohen's Kappa index (Fig. 3; k = 0.309 versus k = 0.272). Moreover, ~~the almost flat classification error plots in both cases (Figs. 3c, f) show high stability of model results, which confirms the reliability of the validation tests obtained in WA-PM and ESA-PM.~~ the classification error plots (Figs. 3c, f) provide an estimate of the error associated with the predicted susceptibility values, which does not exceed 0.1 standard deviations in any case, highlighting the reliability of the results. And finally, the mutual mismatch map (Fig. 5e) shows that 14.8% (corresponding to an extension of 293 km^2) of the mapping units flipped their landslide susceptibility class in WA-PM and ESA-PM.

~~Thus, considering these metrics, we can infer that the pixel susceptibility map which considers only stable mapping units inside the ESA exhibits the best prediction capacity.~~

5.2 Susceptibility maps using slope units

Due to the subdivision of categorical variables in ~~different~~ classes, and to the use of both mean and standard deviation of morphometric variables, the introduction of the original 13 explanatory variables would result in 56 new variables in which many of them (all those classes belonging to the same categorical variable) would be highly correlated. For this reason, ~~for model assessment with SUs,~~ the variable selection approach used in the pixel-based case is not viable when working with SUs and a specific variable selection approach for SU models would require further investigation. Thus, for this work, the most appropriate set of explanatory variables, among those considered as the most relevant in pixel-based model assessment, was selected by expert criteria. Considering such set of variables as a starting point, we selected new sets of explanatory variables to evaluate landslide susceptibility using SUs, *i.e.* to calculate the whole area slope unit map (WA-SUM) and the effective area slope unit map (ESA-SUM). Taking into account that the automatic procedure for the SUs definition already included the flow accumulation calculation, used for *TWI* estimation, and the aspect component, we rejected *Aspect* and *TWI* to avoid spurious correlations. We selected the following set of variables used to produce both pixel-based maps such as *Lithology*, *Permeability*, *Regolith thickness* and *Vegetation*, and we added *Slope*. The reason for choosing *Slope* over *Sinusoidal slope* or *SAR* is due to the fact that these two are derivative variables of the former. Moreover, we consider *Slope* more suitable feature to describe the average morphology within SU than *Sinusoidal* slope or *SAR*, so we decided to select it in order to simplify interpretation of the results.

Using the validation SU sample, we assessed the prediction skills of the SU maps. For the WA-SUM the 65.13% of the 152 validation mapping units were correctly classified (TP+TN) (Fig. 4a). The ROC curve provides ~~an~~ AUC = ~~value of~~ 0.69, and the corresponding Cohen's Kappa ~~is~~ k = 0.302 (Fig. 4b). Concerning the classification error plot (Fig. 4c), it can be observed that in the SUs with ~~h~~High and ~~l~~Low landslide susceptibility probability (probability $>$ 0.8 and $<$ 0.2) the 2σ value stays below 0.2, but variability in the estimates becomes larger for intermediate susceptibilities. This reveals a considerable variation in the stable/unstable classification of the territory, which implies a low reliability, at least for the intermediate probabilities (Guzzetti et al., 2006). For the ESA-SUM, ~~the~~ 63.82% of the 152 validation mapping units were correctly classified (TP+TN) (Fig. 4d) with $AUC = 0.71$, slightly larger with respect to the other SU model assessment, whereas, the Cohen's Kappa index performed slightly worse, being $k = 0.276$ (Fig. 4). The classification error plot shows a considerable variation in intermediate probabilities (Fig. 4f) while the uncertainty is lower for ~~h~~High and ~~L~~low probabilities. Nevertheless, the quadratic fit curves indicate a lower overall variability for ESA-SUM than for WA-SUM.

Visual inspection of the SU susceptibility maps (Figs. 5b, d) shows similarities between WA-SUM and ESA-SUM. The difference is graphically presented through the mismatch map (Fig. 5f), where 12.6% of the mapping units (corresponding to an extension of 247 km^2) change their landslide susceptibility class, ~~from~~ between WA-SUM ~~to~~ and ESA-SUM.

~~Taking into account the AUC and the variability of the model results, ESA-SUM could be singled out as the best method, but if only the Four fold plot and the Cohen's kappa are not used as validation metrics.~~

6 Discussion

The number of scientific publications focusing on landslide susceptibility zonation has notably increased during the last decades (Gutiérrez et al., 2010; Rossi and Reichenbach, 2016; Liberatoscioli et al., 2017; Valagussa et al., 2017; Zhou et al., 2018; Reichenbach et al., 2018) and, nowadays, there is a huge variety of applications and comparisons which provides an enormous range of approaches to prepare a landslide susceptibility map. Differences between these approaches can be summarized in (i) the type of landslide inventory, (ii) the environmental variables used, (iii) the mapping unit partition, (iv) the method used to prepare susceptibility maps and (v) the scale of application. The existence of such a big production of papers investigating these aspects is proof that no fully consolidated standard exists for all the steps involved in landslide susceptibility analysis.

In this work, we showed that the information contained in a field-based landslide inventory for landslide susceptibility analysis should be critically examined, also in combination with the mapping unit of choice.

A field work-based landslide inventory is by definition a source of uncertainty in statistical analysis, owing to various reasons, including mapping errors, accuracy, subjectivity, and others. The focus of this work is the analysis of an additional uncertainty due to use of field mapping, namely the fact that it is impossible to ensure that the study area was surveyed in a homogeneous way. An objective delimitation of the surveyed area by means of the ESA, proposed in this paper along with a module to objectively delineate the ESA (see supplementary material), is one way to reduce this uncertainty.

The hypothesis tested in this work is that any statistical landslide susceptibility model trained inside the ESA is by definition more correct than considering the entire study area for training the model. The statement was borne out by the results of multivariate LR model calculations. We acknowledge that the ESA is only an approximation of the real surveyed area, though a much more realistic one than using the whole study area. Our definition of the ESA depends on the maximum distance between points along the field, trips paths and the selected resolution of the DEM. Preliminary tests in a reduced portion of the territory provided the most suitable settings for a satisfactory definition of the ESA in the particular case of Gipuzkoa Province (section 3.3).

In the case of the pixel-based susceptibility maps, the metrics of model prediction performances are in agreement with our main statement about the relevance of ESA. As a matter of fact, all the validation performance tests (confusion matrix metrics, the area under the ROC curve and Cohen's Kappa index) present an improvement if the stable pixels used for training the LR model are selected within the ESA (like in ESA-PM, Fig. 3a) than if they are taken from the WA (like in WA-PM, Fig. 3b). In addition, ~~the classification error plots provide an estimate of the error associated with the predicted susceptibility values, which does not exceed 0.1 standard deviations in any case, highlighting the reliability of the results~~ the almost flat classification error plots in both cases (Figs. 3c, f) show high stability of model results. The spatial distribution of the susceptibility classes are different as well between ESA-PM and WA-PM (see Figs. 5a, c), and such differences are highlighted in the mutual mismatch map (Fig. 5e). ~~Moreover, the mutual mismatch map (Fig. 5e) shows that 14.8% (corresponding to an extension of 293 km^2) of the mapping units flipped their landslide susceptibility class in WA-PM and ESA-PM.~~ Another difference between the two pixel maps is the set of explanatory variables selected as predictors. The variable selection approach presented in this paper, and previously adopted in a similar way in Schlögel et al. (2018), demonstrated to be effective and capable to detect presence of redundant information, as well as offering an objective way to choose between collinear explanatory variables.

In the case of SU-based susceptibility maps, validation metrics do not present us with clear-cut results as in the pixel-based maps. As a matter of facts, AUC performs better in ESA-SUM while Confusion Matrix and Cohen's Kappa index present higher prediction performance in WA-SUM (Fig. 4). The classification error plots show considerable variations in intermediate susceptibility probability values, but the quadratic fit curves suggest a slightly lower variability in ESA-SUM (Figs. 4c, f). We interpret these results as an indication of a smaller effect that proper usage of the ESA can have in SU-based susceptibility maps, with respect to pixel-based maps. ~~Visual inspection of the SU susceptibility maps (Figs. 5b, d) also shows similarities between WA-SUM and ESA-SUM. The difference is graphically presented through the mismatch map (Fig. 5f), where 12.6% of the mapping units (corresponding to an extension of 247 km^2) change their landslide susceptibility class, from WA-SUM to ESA-SUM.~~ Despite the small difference in model prediction performance between WA-SUM and ESA-SUM, the reduction of the mismatch degree (Fig. 5f) suggests that the usage of the ESA is equally recommendable for SU susceptibility maps carried out by field work landslide inventories.

The pixel- and SU-based maps obtained within the method presented in this work are inherently different from a conceptual point of view. We maintain that an SU-based map probably represents a better option, for SUs bear a clear relation with topography, they reduce mapping errors and are more useful for practical (planning) purposes. Nevertheless, for the sake of completeness and to show differences between the two approaches, we discussed pixel-based and SU-based maps independently. The uncertainty introduced by a field work-based landslide inventory can be mitigated by using SUs, resulting in more similar susceptibility maps and validation performances in WA-SUM and ESA-SUM than in pixel models.

~~Moreover, knowing that the threshold value for the distinction of stable and unstable SU could affect the LR model performances, we decided to build the SU based maps using a threshold value equal to 0.33% (*i.e.*, the landslide ratio in ESA) and compare the results with the ones obtained with 0.15% (*i.e.*, the landslide ratio in the WA). We observed that,~~

~~even if the absolute value of the evaluation indexes (*e.g.*, the ROC area) slightly decrease with respect to the maps obtained using the threshold estimated in the WA (0.15%), the performance of the two approaches is indistinguishable. Therefore, we maintain that results confirm that SU mitigate the relevance of the calibration area (ESA versus WA) when building a SU based susceptibility model with a field-based landslide inventory, independent of the landslide presence threshold value.~~

Moreover, since the threshold value for distinguishing stable and unstable SUs could affect the LR model performances, we performed a sensitivity test evaluating the LR models, for both the WA and ESA, using different presence/absence thresholds. We carried out calculations using as a threshold the 5th percentile (P_5, threshold 0.013%), the 50th percentile (P_{50}, threshold 0.265%) and the 90th percentile (P_{90}, threshold 4.5%) of areal landslide distribution, along with the average landslide density calculated within the ESA, i.e. 0.33%. We observed that for all the cases, except in P_{90}, the model tests showed better performance for ESA-SUM than for WA-SUM, which is proof that the conclusions obtained following any approach were indistinguishable. We note that because of the high threshold defined in P_{90}, the model was trained with a very small sample of unstable SUs, which gives to the result a very poor reliability. On the other hand, for in the P_5 case, the unbalance does not take place, since each SU where at least one landslide pixel exists belongs to the unstable class, resulting in minimum yet relevant number of unstable SUs. Therefore, we maintain that results of the test confirm that SUs mitigate the relevance of the calibration area (ESA versus WA) when building an SU-based susceptibility model with a field-based landslide inventory, independently of the landslide presence threshold value. However, we acknowledge that the search of an optimal threshold value that ensures a balanced sample is a relevant point, though it is beyond the scope of this work.

7 Conclusions

We explored the effects of training an LR classifier (Rossi and Reichenbach, 2016) for landslide susceptibility zonation within the area that was actually surveyed at landslide mapping time, the ESA, and the extended study area, WA, encompassing the ESA. We prepared four susceptibility maps combining variables (*cf.* Eq. (2) and Table 2) sampled strictly within the ESA or from the WA with two different mapping unit partitions, *i.e.*, 5 m x 5 m grid cells and slope units (Alvioli et al., 2016), delineated for the purpose.

A straightforward comparative analysis using standard prediction performance metrics revealed that the ESA-based approach is better than the WA-based, at least in grid-cell mapping units based approaches, for what concerned the training area (*i.e.* within the ESA or WA). Introducing different mapping units in the comparison, we further found that using slope units slightly reduces the gap between results obtained training a statistical model within the ESA versus WA. Thus, the capacity of the slope unit mapping subdivision to mitigate this error was demonstrated, as ~~powerful~~ suitable alternative to the conventional pixel-based approaches.

The results illustrated above support the following statements:

(i) ~~when~~ working with pixel mapping units, training ~~the LS model~~ a statistical classifier for LSZ within the ESA is the correct approach to reduce the uncertainty inherent to the landslide inventory;

(ii) ~~when~~ working with slope unit terrain partition this uncertainty can be mitigated, even though it is still advantageous to train the LS model within the ESA;

(iii) use of ESA should be considered, if sufficient information is available, in preparing landslide susceptibility maps with any multivariate statistical model;

(iv) collecting information about the path followed during field campaigns for landslide mapping is a meaningful procedure for estimating the ESA, at model assessment time, using the GRASS GIS module *r.survey.py* presented in this work.

We acknowledge that the overall performances of the landslide susceptibility maps presented in this paper are of moderate to low prediction capacity, with AUC values ranging between 0.68 to 0.71 and an overall accuracy which hardly overcomes ~~the~~ 65% in the best case (Figs. 3 and 4). This could be due to (i) the lack of a more complete landslide inventory (Guzzetti et al., 2012; Malamud et al., 2004) or (ii) the use of not up-to-date thematic layers. Nevertheless, the preparation of a definitive landslide susceptibility map for the study area was out of the scope of our investigation. Instead, we performed pairwise comparative analyses in which we only changed, across the compared model assessments, the region of logistic regression training.

Code availability.

- The software developed in this work to delineate the effective surveyed area, *r.survey.py*, is contained in the supplementary material
- The software developed in Alvioli et al. (2016) to parametrically delineate slope units, *r.slopeunits*, is available at: `http://geomorphology.irpi.cnr.it/tools/slope-units`
- The software developed in Rossi and Reichenbach (2016) for the statistical assessment of landslide susceptibility zonation, LAND-SE, is available at: `https://github.com/maurorossi/LAND-SE`

Competing interests. No competing interests are present.

Acknowledgements. This work has been funded by the Cultural Landscape and Heritage UNESCO Chair of the University of the Basque Country (UPV/EHU), and the article could not be possible without the support of the IT1029-16 research group funded by the Basque Government, whose head Iñaki Antigüedad has provided his direct collaboration. Authors are also grateful to professor Orbange Ormaetxea for her collaboration during the manuscript corrections. Part of this work was carried out by TB during a scientific visit at CNR IRPI.

References

Alvioli, M., Marchesini, I., Reichenbach, P., Rossi, M., Ardizzone, F., Fiorucci, F., and Guzzetti, F.: Automatic delineation of geomorphological slope units with *r.slopeunits v1.0* and their optimization for landslide susceptibility modeling, Geoscientific Model Development, 9, 3975–3991, https://doi.org/10.5194/gmd-9-3975-2016, 2016.

Alvioli, M., Melillo, M., Guzzetti, F., Rossi, M., Palazzi, E., von Hardenberg, J., Brunetti, M. T., and Peruccacci, S.: Implications of climate change on landslide hazard in Central Italy, Science of The Total Environment, 630, 1528 – 1543, https://doi.org/10.1016/j.scitotenv.2018.02.315, 2018a.

Alvioli, M., Mondini, A. C., Fiorucci, F., Cardinali, M., and Marchesini, I.: Topography-driven satellite imagery analysis for landslide mapping, Geomatics, Natural Hazards and Risk, *In Press*, https://doi.org/10.1080/19475705.2018.1458050, 2018b.

Amorim, S. F.: Estudio comparativo de métodos para la evaluación de la susceptibilidad del terreno a la formacion de deslizamientos superficiales: Aplicación al Pirineo Oriental, Ph.D. thesis, Universidad Politécnica de Catalunya, http://futur.upc.edu/10953986, 2012.

Ayalew, L. and Yamagishi, H.: The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan, Geomorphology, 65, 15–31, https://doi.org/10.1016/j.geomorph.2004.06.010, 2005.

Ba, Q., Chen, Y., Deng, S., Yang, J., and Li, H.: A comparison of slope units and grid cells as mapping units for landslide susceptibility assessment, Earth Science Informatics, pp. 1–16, https://doi.org/10.1007/s12145-018-0335-9, 2018.

Blais-Stevens, A., Behnia, P., Kremer, M., Page, A., Kung, R., and Bonham-Carter, G.: Landslide susceptibility mapping of the Sea to Sky transportation corridor, British Columbia, Canada: comparison of two methods, Bulletin of Engineering Geology and the Environment, 71, 447–466, https://doi.org/10.1007/s10064-012-0421-z, 2012.

Brabb, E. E.: Innovative approaches to landslide hazard and risk mapping, 4th International Symposium on Landslides, Toronto, pp. 307–324, 1984.

Carrara, A., Cardinali, M., Detti, R., Guzzetti, F., Pasqui, V., and Reichenbach, P.: GIS techniques and statistical models in evaluating landslide hazard, Earth surface processes and landforms, 16, 427–445, https://doi.org/10.1002/esp.3290160505, 1991.

Carrara, A., Cardinali, M., Guzzetti, F., and Reichenbach, P.: GIS technology in mapping landslide hazard, in: Carrara, A., Guzzetti, F. (Eds.), Geographical Information Systems in Assessing Natural Hazards, pp. 135–176, Kluwer, Dordrecht, https://doi.org/10.1007/978-94-015-8404-3_8, 1995.

Carrara, A., Crosta, G., and Frattini, P.: Comparing models of debris-flow susceptibility in the alpine environment, Geomorphology, 94, 353–378, https://doi.org/10.1016/j.geomorph.2006.10.033, 2008.

Casagli, N., Frodella, W., Morelli, S., Tofani, V., Ciampalini, A., Intrieri, E., Raspini, F., Rossi, G., Tanteri, L., and Lu, P.: Spaceborne, UAV and ground-based remote sensing techniques for landslide mapping, monitoring and early warning, Geoenvironmental Disasters, 4, 9, https://doi.org/10.1186/s40677-017-0073-1, 2017.

Cascini, L.: Applicability of landslide susceptibility and hazard zoning at different scales, Engineering Geology, 102, 164–177, https://doi.org/10.1016/j.enggeo.2008.03.016, 2008.

Catani, F., Farina, P., Moretti, S., Nico, G., and Strozzi, T.: On the application of SAR interferometry to geomorphological studies: estimation of landform attributes and mass movements, Geomorphology, 66, 119–131, https://doi.org/10.1016/j.geomorph.2004.08.012, 2005.

Cohen, J.: A coefficient of agreement for nominal scales, Educational and Psychological Measurement, 20, 37–46, https://doi.org/10.1177/001316446002000104, 1960.

Corominas, J. and Mavrouli, O. C.: Living with landslide risk in Europe: Assessment, effects of global change, and risk management strategies, Tech. rep., SafeLand. 7th Framework Programme Cooperation Theme 6 Environment (including climate change) Sub-Activity 6.1.3 Natural Hazards, 2011.

Corominas, J., Mateos, R. M., and Remondo, J.: Review of landslide occurrence in Spain and its relation to climate, Slope Safety Preparedness for Impact of Climate Change, p. 351, 2017.

Costanzo, D., Chacón, J., Conoscenti, C., Irigaray, C., and Rotigliano, E.: Forward logistic regression for earth-flow landslide susceptibility assessment in the Platani river basin (southern Sicily, Italy), Landslides, 11, 639–653, https://doi.org/10.1007/s10346-013-0415-3, 2014.

Cruden, D. M. and Varnes, D. J.: Landslide types and processes, in: Turner, A.K., Schuster R. L. (Eds.) Landslides: Investigation and Mitigation. National Research Council, Transportation and Research Board Special Report 247, pp. 36–75, National Academy Press, Washington, DC, 1960.

Das, I., Sahoo, S., van Westen, C., Stein, A., and Hack, R.: Landslide susceptibility assessment using logistic regression and its comparison with a rock mass classification system, along a road section in the northern Himalayas (India), Geomorphology, 114, 627–637, https://doi.org/10.1016/j.geomorph.2009.09.023, 2010.

Davison, A. C. and Hinkley, D. V.: Bootstrap methods and their application, vol. 1, Cambridge university press, 1997.

Efron, B.: Bootstrap methods: another look at the jackknife, in: Breakthroughs in statistics, pp. 569–593, Springer, 1992.

EVE: Mapa Geológico del País Vasco Escala 1:100.000, Basque Energy Agency-Basque Gobernment, 2010.

Fawcett, T.: An introduction to ROC analysis, Pattern Recognition Letters, 27, 861–874, https://doi.org/10.1016/j.patrec.2005.10.010, 2006.

Felicísimo, A. M., Cuartero, A., Remondo, J., and Quirós, E.: Mapping landslide susceptibility with logistic regression, multiple adaptive regression splines, classification and regression trees, and maximum entropy methods: a comparative study, Landslides, 10, 175–189, https://doi.org/10.1007/s10346-012-0320-1, 2013.

Fiorucci, F., Cardinali, M., Carlà, R., Rossi, M., Mondini, A., Santurri, L., Ardizzone, F., and Guzzetti, F.: Seasonal landslide mapping and estimation of landslide mobilization rates using aerial and satellite images, Geomorphology, 129, 59–70, https://doi.org/10.1016/j.geomorph.2011.01.013, 2011.

Fiorucci, F., Giordan, D., Santangelo, M., Dutto, F., Rossi, M., and Guzzetti, F.: Criteria for the optimal selection of remote sensing optical images to map event landslides, Natural Hazards and Earth System Sciences, 18, 405–417, https://doi.org/10.5194/nhess-18-405-2018, 2018.

Gipuzkoako Foru Aldundia: Evaluación y gestión integada de riesgos geotécnicos en la red de carreteras de la Diputación Foral de Gipuzkoa, Tech. rep., Mugikortasun eta Bide Azpiegituren saila, Unpublished report, 2013.

González-Hidalgo, J. C., Brunetti, M., and de Luis, M.: A new tool for monthly precipitation analysis in Spain: MOPREDAS database (monthly precipitation trends December 1945–November 2005), International Journal of Climatology, 31, 715–731, https://doi.org/10.1002/joc.2115, 2011.

Gutiérrez, F., Soldati, M., Audemard, F., and Bălteanu, D.: Recent advances in landslide investigation: issues and perspectives, Geomorphology, 124, 95–101, https://doi.org/10.1016/j.geomorph.2010.10.020, 2010.

Guzzetti, F., Reichenbach, P., Cardinali, M., Galli, M., and Ardizzone, F.: Probabilistic landslide hazard assessment at the basin scale, Geomorphology, 72, 272–299, https://doi.org/10.1016/j.geomorph.2005.06.002, 2005.

Guzzetti, F., Reichenbach, P., Ardizzone, F., Cardinali, M., and Galli, M.: Estimating the quality of landslide susceptibility models, Geomorphology, 81, 166–184, https://doi.org/10.1016/j.geomorph.2006.04.007, 2006.

Guzzetti, F., Mondini, A. C., Cardinali, M., Fiorucci, F., Santangelo, M., and Chang, K. T.: Landslide inventory maps: New tools for an old problem, Earth-Science Reviews, 112, 42–66, https://doi.org/10.1016/j.earscirev.2012.02.001, 2012.

Herrera, G., Fernández-Merodo, J., Mulas, J., Pastor, M., Luzi, G., and Monserrat, O.: A landslide forecasting model using ground based SAR data: The Portalet case study, Engineering Geology, 105, 220–230, https://doi.org/10.1016/j.enggeo.2009.02.009, 2009.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X.: Applied logistic regression, vol. 398, John Wiley & Sons, 2013.

IDE de Euskadi: Mapa geomorfológico de Euskadi, www.geo.euskadi.eus, 2014.

INGEMISA: Inventario y Análisis de las Áreas sometidas a Riesgo de Inestabilidades del Terreno de la C.A.P.V., Tech. rep., Eusko Jaurlaritza, 1995.

Lee, S. and Min, K.: Statistical analysis of landslide susceptibility at Yongin, Korea, Environmental geology, 40, 1095–1113, https://doi.org/10.1007/s002540100310, 2001.

Liberatoscioli, E., van Westen, C. J., and Soldati, M.: Assessment of landslide susceptibility for civil protection purposes by means of GIS and statistical analysis: lessons from the Province of Modena, Italy, Revista de Geomorfologie, 19, 29–43, 2017.

Lombardo, L., Cama, M., Conoscenti, C., Märker, M., and Rotigliano, E.: Binary logistic regression versus stochastic gradient boosted decision trees in assessing landslide susceptibility for multiple-occurring landslide events: application to the 2009 storm event in Messina (Sicily, southern Italy), Natural Hazards, 79, 1621–1648, 2015.

Malamud, B. D., Turcotte, D. L., Guzzetti, F., and Reichenbach, P.: Landslide inventories and their statistical properties, Earth Surface Processes and Landforms, 29, 687–711, https://doi.org/10.1002/esp.1064, 2004.

Minelli, A., Marchesini, I., Taylor, F. E., De Rosa, P., Casagrande, L., and Cenci, M.: An open source GIS tool to quantify the visual impact of wind turbines and photovoltaic panels, Environmental Impact Assessment Review, 49, 70–78, https://doi.org/10.1016/j.eiar.2014.07.002, 2014.

Mondini, A. C.: Measures of Spatial Autocorrelation Changes in Multitemporal SAR Images for Event Landslides Detection, Remote Sensing, 9, 554, https://doi.org/10.3390/rs9060554, 2017.

Mücher, C. A., Klijn, J. A., Wascher, D. M., and Schaminée, J. H.: A new European Landscape Classification (LANMAP): A transparent, flexible and user-oriented methodology to distinguish landscapes, Ecological Indicators, 10, 87–103, https://doi.org/10.1016/j.ecolind.2009.03.018, 2010.

Murillo-García, F. G., Alcántara-Ayala, I., Ardizzone, F., Cardinali, M., Fiourucci, F., and Guzzetti, F.: Satellite stereoscopic pair images of very high resolution: a step forward for the development of landslide inventories, Landslides, 12, 277–291, https://doi.org/10.1007/s10346-014-0473-1, 2015.

Nefeslioglu, H., Gokceoglu, C., and Sonmez, H.: An assessment on the use of logistic regression and artificial neural networks with different sampling strategies for the preparation of landslide susceptibility maps, Engineering Geology, 97, 171–191, https://doi.org/10.1016/j.enggeo.2008.01.004, 2008.

Petley, D., Dunning, S., Rosser, N., and Hungr, O.: The analysis of global landslide risk through the creation of a database of worldwide landslide fatalities, Landslide risk management. Balkema, Amsterdam, pp. 367–374, 2005.

Reichenbach, P., Rossi, M., Malamud, B., Mihir, M., and Guzzetti, F.: A review of statistically-based landslide susceptibility models, Earth-Science Reviews, https://doi.org/10.1016/j.earscirev.2018.03.001, 2018.

Rodrigues, M., Montañés, C., and Fueyo, N.: A method for the assessment of the visual impact caused by the large-scale deployment of renewable-energy facilities, Environmental Impact Assessment Review, 30, 240–246, https://doi.org/10.1016/j.eiar.2009.10.004, 2010.

Rosi, A., Tofani, V., Tanteri, L., Stefanelli, C. T., Agostini, A., Catani, F., and Casagli, N.: The new landslide inventory of Tuscany (Italy) updated with PS-InSAR: geomorphological features and landslide distribution, Landslides, 15, 5–19, 2018.

Rossi, M. and Reichenbach, P.: LAND-SE: a software for statistically based landslide susceptibility zonation, version 1.0, Geoscientific Model Development, 9, 3533–3543, https://doi.org/10.5194/gmd-9-3533-2016, 2016.

Rossi, M., Guzzetti, F., Reichenbach, P., Mondini, A. C., and Peruccacci, S.: Optimal landslide susceptibility zonation based on multiple forecasts, Geomorphology, 114, 129–142, https://doi.org/10.1016/j.geomorph.2009.06.020, 2010.

Santacana Quintas, N.: Análisis de la susceptibilidad del terreno a la formación de deslizamientos superficiales y grandes deslizamientos mediante el uso de sistemas de información geográfica. Aplicación a la cuenca alta del río Llobregat., Ph.D. thesis, Universitat Politècnica de Catalunya, Barcelona, https://www.tdx.cat/handle/10803/6213, 2001.

Santangelo, M., Marchesini, I., Bucci, F., Cardinali, M., Fiorucci, F., and Guzzetti, F.: An approach to reduce mapping errors in the production of landslide inventory maps, Natural Hazards & Earth System Sciences, 15, 2111–2126, https://doi.org/10.5194/nhess-15-2111-2015, 2015.

Schicker, R. D.: Quantitative landslide susceptibility assessment of the Waikato region using GIS, Ph.D. thesis, The University of Waikato, 2010.

Schlögel, R., Marchesini, I., Alvioli, M., Reichenbach, P., Rossi, M., and Malet, J. P.: Optimizing landslide susceptibility zonation: Effects of DEM spatial resolution and slope unit delineation on logistic regression models, Geomorphology, 301, 10–20, https://doi.org/10.1016/j.geomorph.2017.10.018, 2018.

Trigila, A., Iadanza, C., Esposito, C., and Scarascia-Mugnozza, G.: Comparison of Logistic Regression and Random Forests techniques for shallow landslide susceptibility assessment in Giampilieri (NE Sicily, Italy), Geomorphology, 249, 119–136, https://doi.org/10.1016/j.geomorph.2015.06.001, 2015.

Valagussa, A., Frattini, P., Crosta, G. B., Valbuzzi, E., and Gambini, S.: Regional landslide susceptibility analysis following the 2015 Nepal Earthquake, in: Workshop on World Landslide Forum, pp. 1035–1042, Springer, 2017.

Van Den Eeckhaut, M., Vanwalleghem, T., Poesen, J., Govers, G., Verstraeten, G., and Vandekerckhove, L.: Prediction of landslide susceptibility using rare events logistic regression: a case-study in the Flemish Ardennes (Belgium), Geomorphology, 76, 392–410, https://doi.org/10.1016/j.geomorph.2005.12.003, 2006.

Van Den Eeckhaut, M., Hervás, J., Jaedicke, C., Malet, J. P., Montanarella, L., and Nadim, F.: Statistical modelling of Europe-wide landslide susceptibility using limited landslide inventory data, Landslides, 9, 357–369, https://doi.org/10.1007/s10346-011-0299-z, 2012.

Wang, F., Xu, P., Wang, C., Wang, N., and Jiang, N.: Application of a GIS-Based Slope Unit Method for Landslide Susceptibility Mapping along the Longzi River, Southeastern Tibetan Plateau, China, ISPRS International Journal of Geo-Information, 6, 172, https://doi.org/10.3390/ijgi6060172, 2017.

Wang, Y. T., Seijmonsbergen, A. C., Bouten, W., and Chen, Q. T.: Using statistical learning algorithms in regional landslide susceptibility zonation with limited landslide field data, Journal of Mountain Science, 12, 268–288, https://doi.org/10.1007/s11629-014-3134-x, 2015.

Yesilnacar, E. and Topal, T.: Landslide susceptibility mapping: a comparison of logistic regression and neural networks methods in a medium scale study, Hendek region (Turkey), Engineering Geology, 79, 251–266, https://doi.org/10.1016/j.enggeo.2005.02.002, 2005.

Yilmaz, I.: Landslide susceptibility mapping using frequency ratio, logistic regression, artificial neural networks and their comparison: a case study from Kat landslides (Tokat–Turkey), Computers & Geosciences, 35, 1125–1138, https://doi.org/10.1016/j.cageo.2008.08.007, 2009.

Zêzere, J., Pereira, S., Melo, R., Oliveira, S., and Garcia, R.: Mapping landslide susceptibility using data-driven methods, Science of the Total Environment, 589, 250–267, https://doi.org/10.1016/j.scitotenv.2017.02.188, 2017.

Zhou, C., Yin, K., Cao, Y., Ahmed, B., Li, Y., Catani, F., and Pourghasemi, H. R.: Landslide susceptibility modeling applying machine learning methods: A case study from Longju in the Three Gorges Reservoir area, China, Computers & Geosciences, 112, 23–37, https://doi.org/10.1016/j.cageo.2017.11.019, 2018.

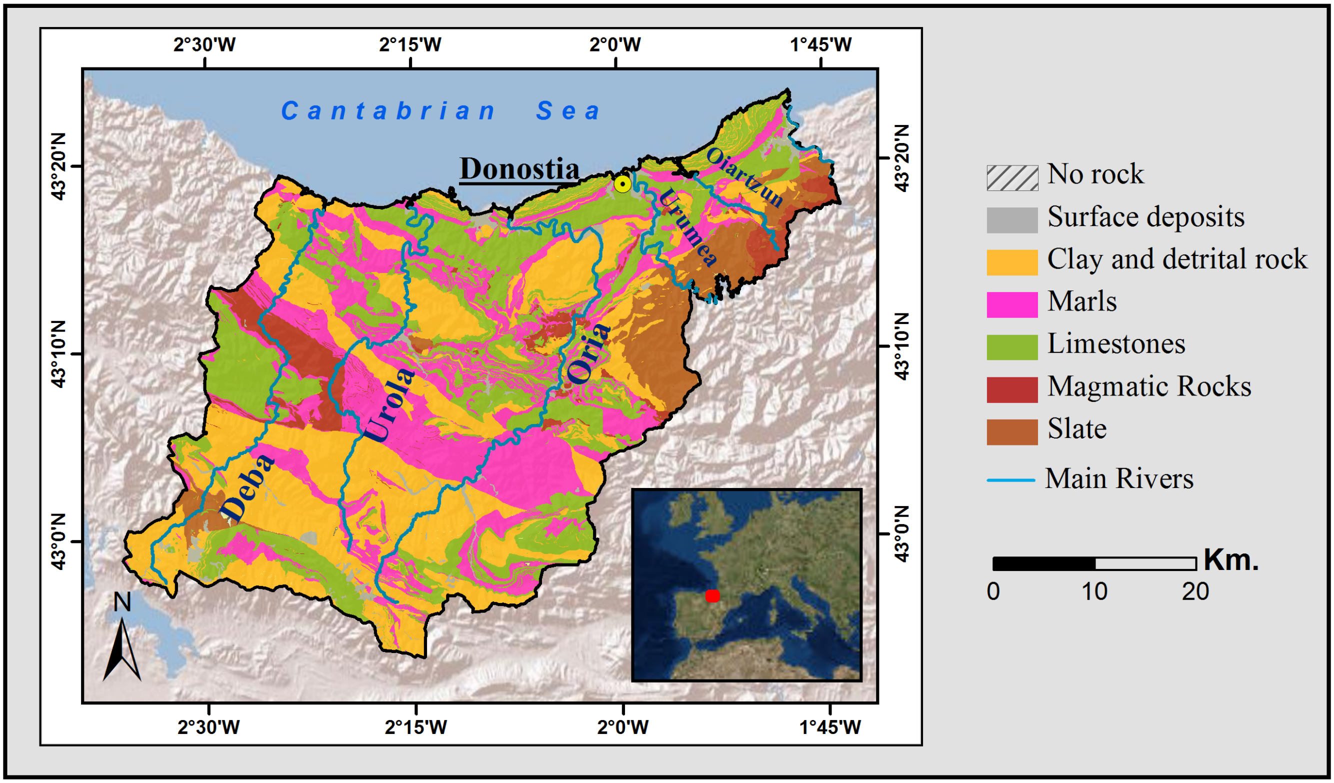


Figure 1. Location of the Gipuzkoa Province study area and simplified lithological map developed according to the original map of the Spatial Data Service of the Basque Country. Coordinates in degrees, Universal Transversal Mercator (UTM) Zone 30N, European Datum ETRS 1989.

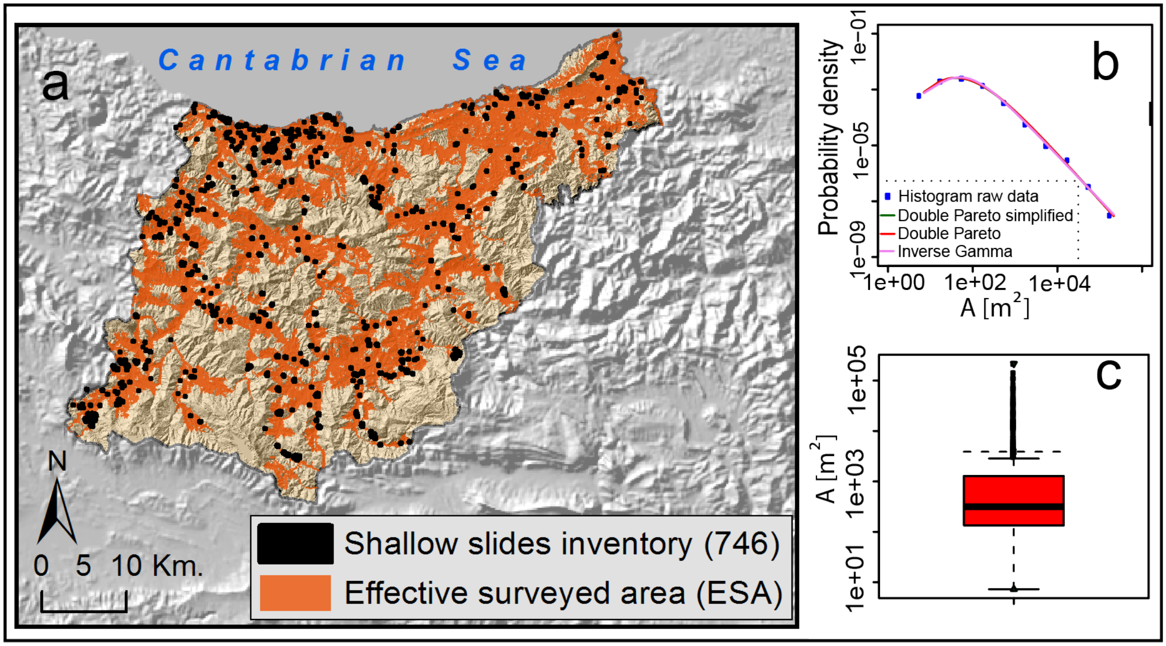
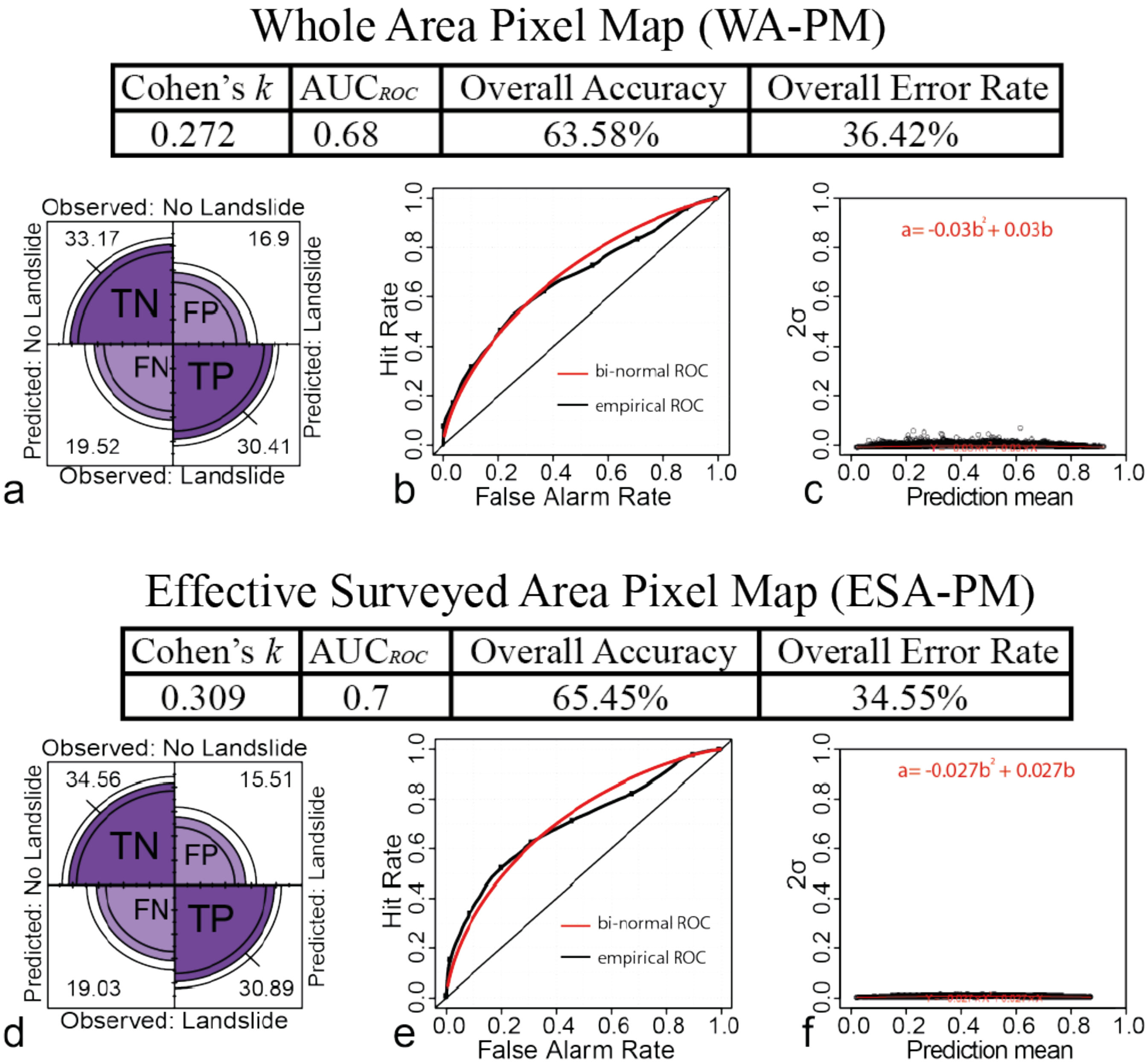


Figure 2. (a) Distribution of the shallow slides inventory along the study area and extension of the effective surveyed area (ESA). (b) Probability density plot of the shallow landslide size (Area in m^2) distribution. (c) Box plot of the same distribution.



Whole Area Pixel Map (WA-PM)

Cohen's *k*	AUC$_{ROC}$	Overall Accuracy	Overall Error Rate
0.272	0.68	63.58%	36.42%

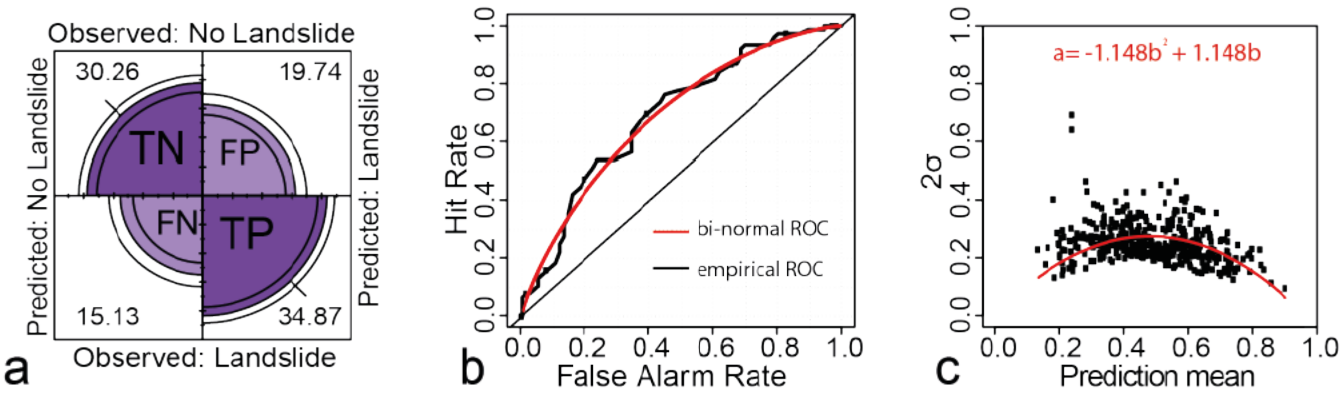
Effective Surveyed Area Pixel Map (ESA-PM)

Cohen's *k*	AUC$_{ROC}$	Overall Accuracy	Overall Error Rate
0.309	0.7	65.45%	34.55%

Figure 3. Pixel-based LR models prediction performance results: summary tables of the Cohen's kappa index, area under the ROC curve (AUC), overall accuracy ((TP+TN)/(TP+TN+FP+FN)) and overall error rate ((FP+FN)/(TP+TN+FP+FN)); (a,d) four fold or contingency plots; (b,e) ROC curves; (c,f) classification error plots and the quadratic regression fit curves (red line).

Whole Area Slope Unit Map (WA-SUM)

Cohen's *k*	AUC$_{ROC}$	Overall Accuracy	Overall Error Rate
0.302	0.69	65.13%	34.87%



Effective Surveyed Area Slope Unit Map (ESA-SUM)

Cohen's *k*	AUC$_{ROC}$	Overall Accuracy	Overall Error Rate
0.276	0.71	63.82%	36.18%

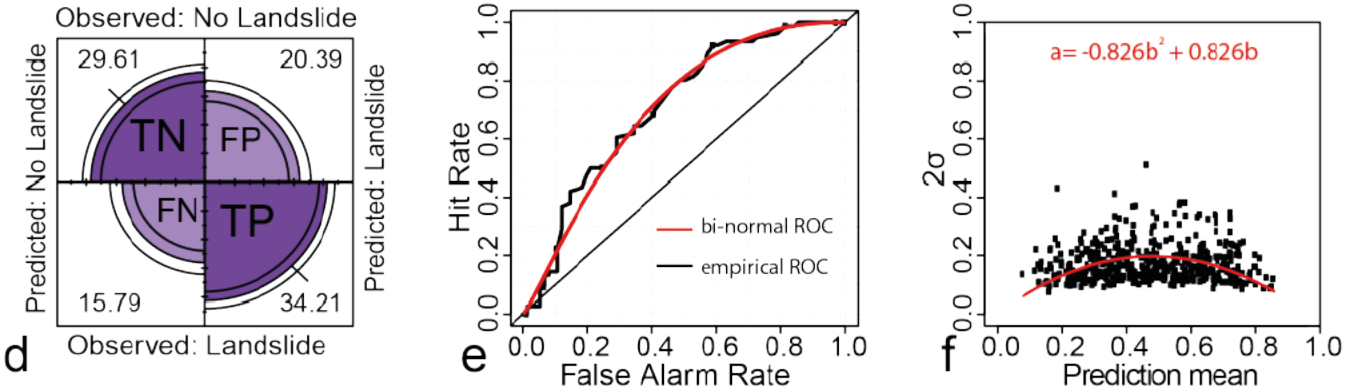


Figure 4. SU-based LR models prediction performance results: summary tables of the Cohen's kappa index, area under the ROC curve (AUC), overall accuracy ((TP+TN)/(TP+TN+FP+FN)) and overall error rate ((FP+FN)/(TP+TN+FP+FN)); (a,d) four fold or contingency plots; (b,e) ROC curves; (c,f) classification error plots and the quadratic regression fit curves (red line).

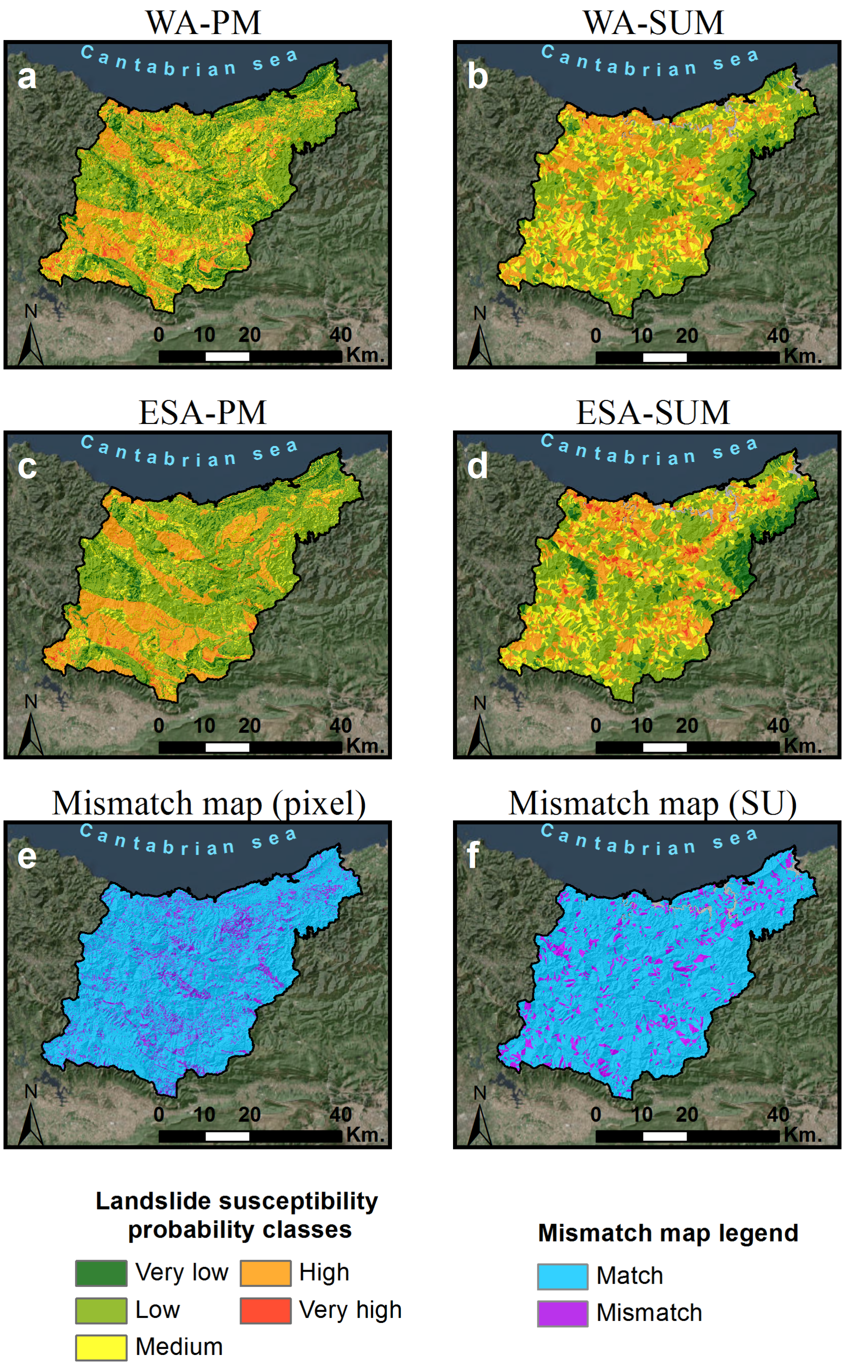


Figure 5. (a-d) Landslide susceptibility maps represented in five classes for WA-PM, WA-SUM, ESA-PM and ESA-SUM. (e,f) Mismatch maps representing the spatial distribution of the mapping units differently classified using ESA between pixel models and slope unit models.

Table 1. Results of the setting test of r.surbey in a 10 km^2 subset of the study area.

Name	**Resolution (m)**	**D$_{max}$**	**Percentage of landslides within (%)**
Survey 5	5	50	35
Survey 6	20	50	70
Survey 7	50	50	95
Survey 8	100	50	100
Survey 9	5	100	30
Survey 10	20	100	60
Survey 11	50	100	95
Survey 12	100	100	100
Survey 13	5	200	30
Survey 14	20	200	55
Survey 15	50	200	85
Survey 16	100	200	100
Survey 17	5	500	0
Survey 18	20	500	35
Survey 19	50	500	60
Survey 20	100	500	95

Table 2. Set of environmental variables introduced for the whole area pixel-based (WA-PM) and effective surveyed area pixel-based (ESA-PM) models calculation, together with the significance p-value estimate corresponding to each explanatory variable (*cf.* Section 4.1). The best predictors were labelled with an asterisk.

Name	**Description**	**Significance p-value**	
Continuous		WA-PM	ESA-PM
Slope	The slope gradient in degrees.	$1.17\cdot10^{-189}$	$1.06\cdot10^{-111}$
Sinusoidal Slope	The sinusoidal mathematical transformation applied to the slope variable (Amorim, 2012)	$1.00\cdot10^{-155}$	$7.57\cdot10^{-134}$ *
Surface area ratio	The relation between the theoretical volume and the surface of each pixel.	$3.743\cdot10^{-203}$ *	$1.89\cdot10^{-99}$
Terrain wetness index	The spatial distribution of soil moisture or saturation (Yilmaz, 2009)	$9.864\cdot10^{-10}$ *	0.126807342
Curvature	The spatial variation of the slope gradient.	0.909592654	0.525989188
Plan curvature	The curvature of the surface perpendicular to the direction of the maximum slope.	0.9094261	0.525836679
Profile curvature	The curvature of the surface in the direction of the maximum slope.	0.909605174	0.526032985
Categorical			
Litology	The original categories have been reclassified by expert criteria (Geoeuskadi).	0 *	0 *
Permeability	The original categories have been reclassified by expert criteria (Geoeuskadi).	$1.496\cdot10^{-33}$ *	$7.632\cdot10^{-72}$ *
Regolith thickness	The layer for the study area has been obtained from the Lithological Map (Geoeuskadi).	0 *	0 *
Land Use	The original categories have been reclassified by expert criteria (Geoeuskadi).	$5.14\cdot10^{-291}$	$1.42\cdot10^{-87}$
Vegetation	The original categories have been reclassified by expert criteria (Geoeuskadi).	0 *	$1.596\cdot10^{-173}$ *
Aspect	It represents the downslope direction measured in degrees classified in 9 classes.	0 *	0 *