The authors thank the reviewers for their constructive comments. The comments are shown in regular fonts, ***our responses are in bold italic, blue fonts***. *Changes made in the manuscript are printed in italic, underlined, blue fonts.* ***Our line references refer to the updated manuscript with track-changes.***

Reviewer #1

I must first apologize to the authors and to the editor for this late review. As an author, I have myself experienced how irritating it can be to wait for an overdue review. I can only mention heavy load of work as an excuse for my delayed response.

Not being myself an expert on landslide causes and occurrences, I am not in a position to comment on the parts of the paper that specifically deal with those aspects, nor actually to evaluate the paper in comparison with what has already been done on estimation of landslide susceptibility and uncertainty. But I have comments on the methodological aspects of the paper, which may be useful for all readers.

I must say I have had difficulties to clearly understand what the authors have done, as concerns both their methodological approach and the validation of the results they have obtained. I will limit myself to what are the most important points I want to stress.

We thank the reviewer for the feedback on the methods. We improved the readability of the paper for people in various research fields, because we hope to bridge multiple disciplines in this paper and appreciate feedback from outside the landslide community.

1. My first question has actually to do with landslides. The authors focus their study to *hydrologically triggered landslides* (l. 68). This means that they ignore, for instance, landslides triggered by earthquakes. What is the reason for that restriction ? How can the distinction be made between different kinds of landslides, once they have occurred ? And does the Global Landslide Catalog report only hydrologically triggered landslides ? These questions may look naïve to specialists, but some appropriate information (and references) may be useful to outsiders.

Thank you for this valid question. As mentioned in lines 39-43, the susceptibility assessment carried out in this study is intended to be used with satellite soil moisture observations. Since the likelihood of earthquake triggered landslides is primarily dependent on the presence of an earthquake and its magnitude, the soil saturation is not a reliable indicator for occurrence of these landslides. For hydrologically triggered landslides, on the other hand, increased soil water content is the actual underlying condition that leads to slope failure by i) decreasing the shear strength and ii) increasing the shear stress.

Landslide inventories (such as the GLC) based on media reports usually take the information on the trigger from the report itself. The GLC was designed mainly for the purpose of collecting information on hydrologically triggered landslides, namely triggered by "continuous rain", "downpour", "monsoon", "flooding", "rain" and "tropical cyclone" (GLC classifiers), but also contains a small number of landslides from other triggers (less than 5%) and for about 16% triggers are unknown. We will update section 2.1 as follows: *"The GLC is a landslide inventory that contains information about location, date and trigger. It is originally based on media reports (Kirschbaum et al., 2010, 2015) but has recently been supplemented with the citizen science-based Landslide Reporter Catalog (LRC) data (Juang* *et al., 2019), see Stanley et al. (2021) for details. […] For this study, we use 12515 hydrologically triggered landslides (GLC classifiers “continuous rain”, “downpour”, “monsoon”, “flooding”, “rain” and “tropical cyclone”) reported mainly between January 2007 and November 2020.” (Line 95-104)*

2. The first purpose of the paper is to derive *LSS model equations* (ll. 66-67). The authors do not actually show any equation that is explicitly identified as such. The model equations must be equations of form (1), where the quantity P(Y = 1) is what is called LSS elsewhere. Equations of form (1), which are defined as a form of logistic regression, are used on appropriate training sets for determining, through MELR and Cross Validation, values of the parameters α and βi (i=1,…,n). This raises a number of questions.

a. What is the rationale for the logistic form of equation (1) ? What are the advantages of that specific approach ? It seems to me to be a rather arbitrary choice. In their conclusion, the authors mention the choice of the statistical model (l. 365) as one possible source of uncertainty in the whole estimation process. Do they refer there to Eq. (1) ? The authors give references concerning logistic regression, but some basic explanation would be useful.

Thank you for your question. Indeed, the choice of logistic regression as statistical model was in a way arbitrary, but based on literature review and expert opinions. Many smaller scale LSS studies compare multiple statistical models and subsequently use the best performing one (for example Steger et al. 2015, Zêzere et al. 2017 or Depicker et al. 2020). Larger scale studies usually choose a priori one statistical model to limit computational time (for example Stanley and Kirschbaum 2017, Lin et al. 2017 or Broeckx et al. 2018). The findings of Zêzere et al. 2017 also show that the choice of statistical model has less impact on the final LSS results than the choice of mapping unit. We opted for a logistic regression because it is a very simple approach and one of the earliest statistical, data-driven models used for LSS (Reichenbach et al. 2018). The advantage lies in its rather robust nature, that is not as prone to overfitting as certain machine learning algorithms. As such, logistic regression is also the most commonly used technique (Reichenbach et al. 2018). We will extend the manuscript as follows:

“Logistic regression is the most commonly used approach for statistical LSS mapping (Reichenbach et al., 2018). It is associated with strong generalizing capabilities (Brenning, 2005), which is a necessity when working at the global scale, and it has proven to be reliable in continental to global LSS assessments (Broeckx et al., 2018; Lin et al., 2017).” (Line 164-167)

“Because Zêzere et al. (2017) found that the choice of spatial mapping unit influences LSS estimates stronger than the choice of statistical model, we do not expect that our results would fundamentally change for approaches other than MELR.” (Line 450-452)

b. In MELR, what is the criterion for quality ? Given a tentative set of values (α, β_i,i=1,…,n), by which measure is the corresponding fit to P(Y = 1) evaluated ? A simple quadratic fit, or what ?

For this study we used the lme4 package in R, specifically the glmer function, in which the fit is optimized based on maximum likelihood estimation, which involves a minimization of squared deviances with penalty terms. Subsequently, we evaluate the model fitting performance of each fitted equation by means of AUC on the test data, but this information is not used for optimization. We will extend section 3.1 by the following statement in hope for clarification: *“We use the glmer function from the lme4 package (Bates et al. 2015) to create MELR models in R version 4.0.3 (R Core Team, 2020) where the best fitting parameters are obtained by maximum likelihood estimation.” (Line 201-203).*

In addition, note that: *“a measure that is proportional to the sum of squared errors […]” (L.177-178)* ***is used to identify the most useful predictor variables.***

c. I understand that the values P(Y = 1) in the training sets are taken in the data set built in subsection 2.1 from the GLC catalog, so that these values are restricted to 0 and 1 (absence or presence of landslides). It would a priori seem more appropriate to consider a quantity such as the frequency of occurrence of landslides (that would not be impossible from GLC since the latter mentions more than one landslide for a number of individual grid cells). That may not be practically possible, but it would in my mind be appropriate to mention, and preferably briefly discuss, that alternative approach.

That is a valid point that also crossed our mind. A posterior analysis of average LSS ($\overline{LSS}_{2500}$) against the number of landslide reports per grid cell showed indeed mostly an increased median $\overline{LSS}_{2500}$ for higher reporting frequency (see Figure 1 below).

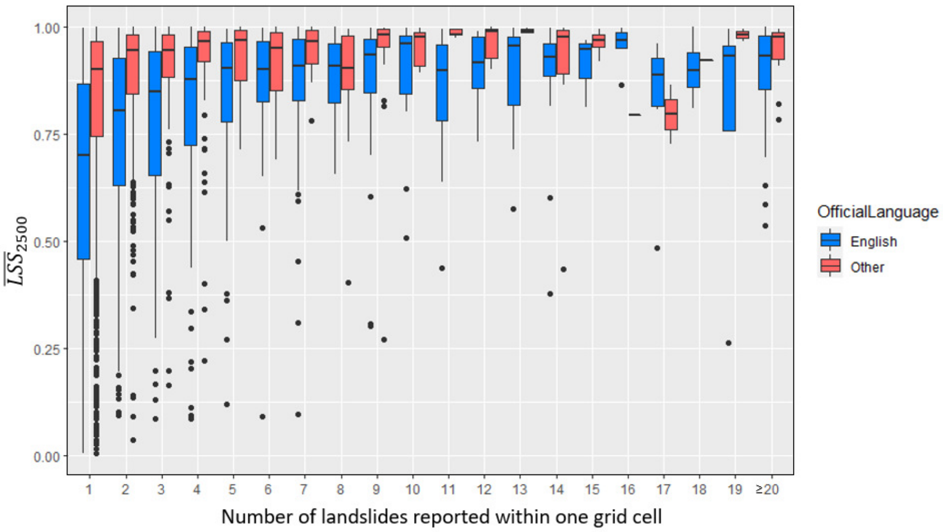

Figure 1 Ensemble average LSS ($\overline{LSS}_{2500}$) for grid cells with different number of reported landslides in the GLC, stratified for the official language denoted by the United Nations Group of Experts on Geographical Names.

However, this connection significantly differed between grid cells located in English-speaking countries and those of other official languages. One reported landslide in an English-speaking country is associated with lower LSS than in a non-English-speaking country. Hence, keeping the possible response values restricted to 0 or 1 also contributes to mitigating observation biases from the GLC. This posterior analysis nevertheless gives confidence that locations with a higher frequency of landslide reports have also been assigned a higher LSS by the statistical method of MELR. We will add the following short discussions of this:

"While we acknowledge that grid cells with more frequent landslide reporting can in general be expected to have a higher LSS, we found that the information about the frequency of landslide occurrence within a grid cell strongly mirrors biases in the landslide inventory, e.g. more landslides are reported in English-speaking countries. The aggregation, on the contrary, reduces the landslide presence reporting bias of the GLC." (Line 108-111)

"Future research could explore the additional information, such as landslide sizes, types or the frequency of occurrence per grid cell instead of reducing the data to landslide presence and absence. For the latter, one would need to find ways to counteract the English-language and economic bias of the GLC which is more pronounced when using the actual number of reports instead of the presence-absence method chosen in this study. " (Line 452-456)

3. a. Concerning the quality of the uncertainty of their LSS estimates, the authors use as diagnostic the Receiver Operation Characteristic (ROC) and the associated area under the ROC curve (AUC). They mention AUC values for individual members of ensembles, i.e. LSS maps (ll. 214-217, l. 275 and Fig. 7), as

well as for global ensembles (ensembles of maps). The latter are all right for me, but I do not understand what AUC values for individual maps can be. ROC curves (https://en.wikipedia.org/wiki/Receiver operating characteristic) are parameterized by a threshold T, each point on the curve corresponding to a value of T. The corresponding coordinates are relative to the circumstances when the value of a given parameter is larger than T (in the context of the present paper, that parameter must be LSS). Unless all grid cells are lumped together, which does not seem to be reasonable, it does not make sense to consider the situation LSS > T on a single map. That makes sense only on an ensemble of maps, with grid cells being considered independently of each other. I may of course be mistaken as to what the authors have exactly done, but clarification is necessary.

Thank you for this valid question. All ROC curves and connected AUC values mentioned in this study were derived for individual LSS maps, i.e. "all grid cells lumped together", or a subset of these grid cells. In case of the ensemble, this means one AUC value is computed for each member map (or parts of it, such as Africa, Russia and Italy). The ROC allows to validate a continuous probability value against a discrete outcome (landslide presence and absence) by means of applying, as you correctly state, different threshold values T. For grid cells below (above) T, the prediction is assumed to be landslide absence (presence), and comparison against the validation data in form of a confusion matrix allows to compute the according true and false positive rate. The pairs of these rates are collected for different thresholds T and displayed in the ROC curve. "Lumping" all grid boxes of one map together, allows to evaluate the spatial accuracy and is in this way common practice for landslide susceptibility: Reichenbach et al. (2018) report that it was used as an accuracy measure in more than 20% of the 565 susceptibility studies they reviewed. We will alter section 3.3 as follows:

"To quantify how well a predicted LSS map represents observed landslide presences and absences, a BS can be used (see Equation 2). Alternatively, the Receiver Operating Characteristic (ROC) is commonly used as evaluation tool for categorical response values such as landslide presence and absence (Reichenbach et al., 2018). For the ROC, the true positive rate of one LSS map is displayed against its false positive rate for different possible thresholds in the continuous probability (here: LSS) that is predicted. The area under the ROC curve (AUC) is 1 for a perfect representation of the spatial LSS distribution, whereas an AUC value of 0.5 indicates that the model does not perform better than a uniform distribution.

Depending on the reference landslide data, the ROC analysis can be conducted for specific grid cells from a CV subset (independent data not used in the training), or from other independent landslide inventories. Here, we use landslide presence and absence information from the grid cells of the fifth CV subset (test subset, see Figure 1) to assess the model fitting performance for each LSS ensemble member map "on the go". To evaluate the final prediction performance of the complete ensemble averages and the corresponding ensemble members, we use 3 independent landslide inventories." (Line 252-272)

b. And the reference to one fully deterministic reference MELR equation (based on neither CV nor input perturbations) (ll. 220-221) is confusing. Does it mean you have performed the validation on other outputs than the ones obtained from CV ? I have a similar question about the one deterministic MELR equation mentioned on ll. 355-356.

We are sorry to have caused confusion with this. We wanted to evaluate whether the general introduction of an ensemble in contrast to deterministic predictions improves the accuracy of the LSS map, as found for hydrological and meteorological modelling (Kalnay et al. 2006). For this reason we created – in addition to the full (LSS_{2500}) and CV ensemble (LSS_{100}) – this one fully deterministic MELR equation. The AUC values retrieved for the resulting deterministic LSS map for Russia, Italy and Africa were then compared against the AUC values of the ensemble averages ($\overline{LSS}_{2500}$ and $\overline{LSS}_{100}$).

However, we realize that this is confusing for the reader and to some extent an unfair comparison because this fully deterministic MELR (without CV and predictor perturbations) is also trained on the complete set of landslide presence and absence grid cells (in contrast to the model training on 4 out of 5 subsets as part of CV). Since this additional comparison does not add much to the results of our study, we have decided to remove any mention of it from the manuscript:

We will alter section 3.3 as follows and hope that the procedure becomes more clear: *"The AUC and BS metrics can be computed for individual ensemble members (of the CV ensemble LSS_{100}, or the full ensemble LSS_{2500}, yielding a distribution of metrics) or for ensemble averages ($\overline{LSS}_{100}$ and $\overline{LSS}_{2500}$). It will be assessed whether i) an ensemble average outperforms an individual member LSS realization and whether ii) the full ensemble average with ensemble input perturbations ($\overline{LSS}_{2500}$) outperforms the CV ensemble average which does not include input perturbations ($\overline{LSS}_{100}$). This would be in line with the expectations for hydrological or meteorological models (Kalnay et al., 2006)." (Line 282-286)*

4. The authors write (ll. 357-359) *The finding of Kalnay et al. (2006) (show) that the introduction of ensembles increases the accuracy of the prediction does not hold for our LSS modelling. This is probably due to the non-linear characteristics of logistic regression and LSS being static.* I understand the authors mean that the accuracy of the mean of the output ensemble is higher than the accuracy of an individual deterministic estimate (at least statistically). From what I understand, non-linearity cannot be the problem here. Consider a process F(x) where there is uncertainty of the input x. Let $\{x_i\}$ a sample of independent realizations of the probability distribution for x. The ensemble $\{F(x_i)\}$ is a sample of independent realizations of the probability distribution for F(x). As such, the mean of that ensemble is the best estimate of F(x), at least in a least-square sense. That is true whether the process F is linear or not.

Thank you for this remark. Indeed you are right in that for one grid cell the mean of an ensemble $\{F(x_i)\}$ should remain the best estimate for the process F(x), independent of the linearity of the process, and the quoted finding was an error from our side, that will be corrected.

Our naïve inference was that the ensemble average should also improve the spatial performance (assessed by AUC, as explained for comment 3) over that of an unperturbed realization. This, however, we did not find to be the case. A more correct assessment of the influence of the ensemble on the performance would be based on the accuracy per grid cell. A measure for this is the ensemble skill (ensk, see Equation A2), which is essentially the squared difference between (ensemble average) LSS and the observation. When averaged over a number of grid cells, ensk turns into the Brier Score (BS, see Equation 2). We do find that the BS based on intervals of CV ensemble ensk ($ensk_{LSS100}$) is improved through predictor variable perturbation (full ensemble) where $ensk_{LSS100}$ is not already close to its optimum value of 0:

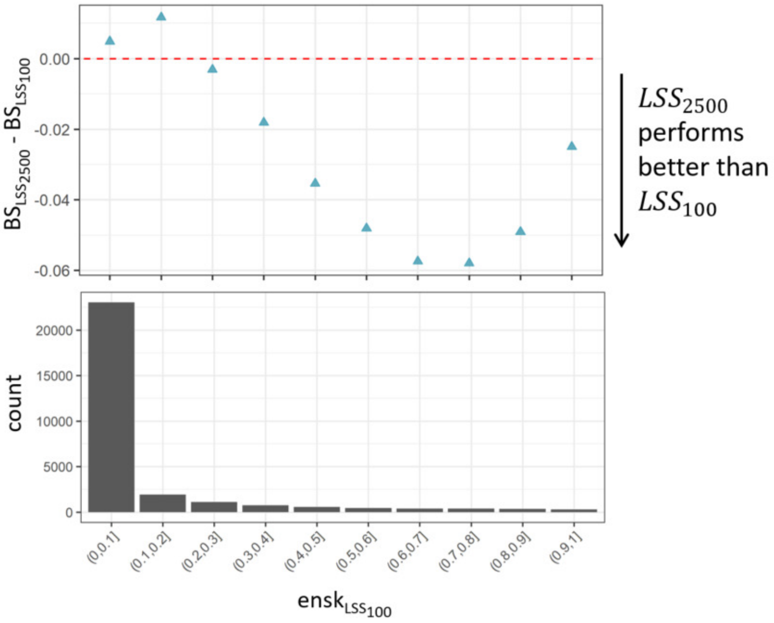

Figure 2 (Top) Difference in Brier Score (BS) between the full ensemble (LSS_{2500}) and the CV ensemble (LSS_{100}), for intervals of ensemble skill of the latter ($ensk_{LSS100}$). (Bottom) Number of grid cells within the $ensk_{LSS100}$ interval.

These grid cells with very small $ensk_{LSS100}$, i.e. where the predictor perturbations do not improve the performance, comprise by far the largest proportion of all land grid cells. They mostly have very low LSS values close to 0, where the predictor variable perturbation was found to introduce a small bias towards higher LSS values: *"Slightly increased (decreased) LSS_{2500} at the lower (upper) limits can be attributed to the resampling of predictor variable values if they exceed the definition interval of rescaled predictor variables (0,1)." (Line 421-423)* ***For landslide absence grid cells, this slightly decreases the accuracy and results in a slightly increased BS value. In the spatial accuracy assessment, it is these grid cells that strongly dominate the resulting AUC values due to their large number.***

We will alter section 5.3 as follows: *"The AUC analysis (Fig. 7) shows that the ensemble averages perform much better than individual ensemble members, and that $\overline{LSS}_{2500}$ and $\overline{LSS}_{100}$ perform equally well. Not shown is that the BS (Equation 2) decreases (i.e. improves) for LSS_{2500} in comparison to LSS_{100} where LSS is not very close to the observation already (landslide presence and absence). This effect is, however, not visible in the AUC comparison (spatial accuracy) for the validation data in Russia, Africa and Italy because the grid cells with BS improvement only make up for ~8%, ~9% and ~18% respectively. The AUC values of ensemble averages remain practically the same, and an LSS model without predictor perturbations would hence suffice for a general insight in the global spatial LSS pattern." (Line 425-431)*

5. The authors write in their conclusion (ll. 373-374) … *predictor variable perturbations results in a reliable assessment of the associated total prediction uncertainty*. It is of course more difficult to assess the uncertainty on an estimate than to obtain the estimate itself. But the authors' statement seems to be a bit of an exaggeration. The AUC values given in the paper do show some reliability in the assessment of the uncertainty, but no more. Actually, the amplitude of the predictor variable perturbations has been evaluated on the same data set as the LSS values. The whole process is therefore subject to some form of inbreeding, the impact of which is difficult to assess. And the authors write themselves *A comparison of $\sigma_{LSS2500}$ with independent global estimates is currently not possible for lack of uncertainty estimates* (ll. 340-341). I suggest the authors soften down their concluding statement.

Thank you for this valid remark and we will soften the concluding statement. The “inbreeding” is, however, only partially true, because we also evaluate AUC values for the independent validation inventories (Russia, Africa, Italy). These were not part of the GLC, which we used for the tuning of the perturbations. We find the same tendencies in AUC spread for these validation inventories and the test data (not used for training) from the GLC (see Figure 7). We do agree though, that our assumption of reliability is strongly rooted in and connected to our trust in the tuning of the perturbations. We agree to change the text as follows: *“The novel method of combining blocked random CV (B-CV) and predictor variable perturbations results in a reasonable assessment of the associated total prediction uncertainty.” (Line 459-461)*

I would have also comments on editing aspects of the paper, but I think they are of lesser importance at this stage.