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 #2

I have revised the manuscript "Estimating global landslide susceptibility and its uncertainty through ensemble modelling" by Anne Felsberg and co-authors. The procedure is very interesting and it would be interesting to apply it to a smaller area where you can control better the training and the validation landslide and thematic data. The manuscript is well written and well organized but there are few major thinks nor really convincing.

We thank the reviewer for the detailed feedback. We extended and improved explanations for many points that were mentioned as unclear and hope that this is able to increase readability and understanding.

1. It is possible to select the hydrologically triggered landslides from the Global Landslide Catalogue?

Please see response to comment 1 of reviewer #1.

2. Is not clear how you have used the road network density

Thank you for pointing this out. We will extend the mentioning of the road network density in section 2.1 to the following: *"To address the remaining landslide presence bias originating from more landslide reporting in frequently accessed areas, we use stratified data on the [average road network density] (including highways and all types of roads, ranging from primary to local roads) provided by the Global Roads Inventory Project (GRIP) (Meijer et al., 2018) as a random effect, explained in sect. 3.1." (Line 112-115)*

3. In the analysis you have mentioned model uncertainty and input uncertainty. I think the uncertainty associated to a 36-km spatial resolution grid is so large that the entire analysis is not relevant. As you mention at line 51 "*coarser input data might be less representative for local events, such as landslides*". If data are not representative the entire modelling is not representative.

Thank you for voicing your concern. There are indeed limitations of a susceptibility analysis at 36-km resolution for specific hillslopes and we used some unfortunate phrasing to address the spatial representativeness issue. We removed the confusing statement and the text is updated as follows: *"Input uncertainty principally results from errors in the environmental data. To assess how input uncertainty propagates into the total prediction uncertainty, ensemble simulations can be used." (Line 58-61)*

Furthermore, we will follow up on this and acknowledge the effect of spatial representativity and how that error is caught in the optimization of the "input" uncertainty. Please see our answer to comments 5 and 6 below. Yet, at this stage, it may already be relevant to point out that the native resolution of many of our input layers (e.g. slope) was much finer than 36 km and the variability within these large grid cells is (at least partially) taken into account. As explained in the introduction, the main reason for constructing a model at such a coarse resolution (and assess the associated uncertainties) is to make a

product that can be used in combination with currently existing soil moisture satellite observations (see below).

4. Line 20 *LSS maps derived from environmental conditions are a fundamental tool for informing local population, city planners and decision makers both on the immanent landslide likelihood, but also about secondary effects such as major sediment sources* → this is true but very difficult to be applicable at the resolution of your analysis.

We agree. This sentence was intended as a first introduction into the topic of landslide susceptibility. As we state in line 39-43, the susceptibility assessment in this study aims at a subsequent combination with satellite soil moisture products for spatio-temporal hazard assessment. This motivated the choice of a 36-km spatial resolution which is common for these satellite products. To prevent any misinterpretation, we will however update the sentence as follows: *"Regional high-resolution LSS maps derived from environmental conditions are a fundamental tool for informing local population, city planners and decision makers both on the immanent landslide likelihood, but also about secondary effects such as major sediment sources (Crozier, 2013; Maeset al., 2017; Broeckx et al., 2020). Large scale low-resolution LSS maps can serve as background information to be downscaled for the above applications at the local scale, or they can be used in conjunction with large-scale satellite data to construct a spatio-temporal estimate of the likelihood for a landslide" (Line 23-28)*

We would also like to point out that although gridded landslide susceptibility is usually assessed at finer resolution, the end result is still frequently aggregated into communal or provincial units which often cover areas comparable to one of our grid cells.

5. (Line 60) The total uncertainty is estimated by comparing the predicted average LSS against the observed presence and absence of landslides → in your case the presence/absence of landslide is related to a too coarse grid resolution.

Thank you for your remark. Indeed, the evaluation suffers from representativeness error, and that uncertainty is now included in the LSS uncertainty. The text is updated as follows: *"One such important source of uncertainty is spatial representativeness error (Blöschl and Sivapalan, 1995; van Leeuwen, 2015), especially when evaluating spatially averaged grid cell LSS estimates using single landslide observations as reference data." (Line 72-74)*

6. *Due to their generalizing nature, LSS models are however prone to uncertainty.* → true but the uncertainty it is also highly related to the thematic data/landslide distribution/model used for the assessment. In your case the uncertainty associated to data and landslide distribution is more relevant than the entire modelling.

We agree about these different sources of uncertainty in landslide modelling. This is why later on in the introduction (line 44 onwards) we introduce the concept of input and model uncertainty and stress the need for assessing the input uncertainty, especially at a 36-km resolution. We would nevertheless like to highlight that most of the environmental data, and especially the ones building on topography, originally come from much finer resolution (see also the answer to comment 10) and by aggregating the information, the error is – in fact – reduced at the coarse resolution (aggregation of noise, purely statistically). We will include the original spatial resolutions in Table 1 (see below).

Furthermore, the uncertainty of the exact landslide locations used for training of the LSS models (location accuracy ranging from 1 km to up to 25 km in the GLC) actually becomes less of a problem when aggregating the information into a 36-km grid cell.

Table 1 Predictor variables used in this study

Predictor variables	Data source	Original spatial resolution	Aggregation method to or within EASEv2, 36 km grid cell
slope (mean, maximum) [°]	USGS: details in Verdin et al. (2007) based on SRTM DEM[a] and GTOPO30[b]	3" (SRTM DEM), 30" (GTOPO30)	mean and maximum
elevation (mean, standard deviation) [m a. s. l.]	CLSM parameters: details in Verdin (2013) based on SRTM DEM[a] and GMTED2010[c]	3" (SRTM DEM), 7.5" (GMTED2010)	mean and standard deviation
depth to bedrock [m]	CLSM parameters: details in De Lannoy et al. (2014) based on GSWP-2[d]	1°	spatial interpolation
percentage of gravel (0-30 cm) [vol%]	CLSM parameters details in De Lannoy et al. (2014) based on STATSGO2[e] and HWSD1.21[f]	30"	most representative 30" sample
percentage of clay (0-30 cm and 0-100 cm) [w%]			
percentage of sand (0-30 cm and 0-100 cm) [w%]			
porosity (0-30 cm and 0-100 cm) [m^3/m^3]			
wilting point divided by porosity (0-30 cm and 0-100 cm) [-]			
compound topographic index, CTI (mean, maximum) = ln(specific catchment area/tan(slope)) [log(m)]	CLSM parameters: details in Verdin (2013) based on SRTM DEM[a] and GMTED2010[c]	3" (SRTM DEM), 7.5" (GMTED2010)	mean and maximum
land fraction within grid cell	CLSM parameters: HYDRO1k based on GTOPO30, 1996 (EROS, 2018; Verdin, 2013)	10"	areal fraction
fraction covered by each of 13 lithological classes [-]: *metamorphic rocks, mixed sedimentary rocks, siliclastic sedimentary rocks, basic plutonic rocks, acid plutonic rocks, basic volcanic rocks, intermediate volcanic rocks, carbonate sedimentary rocks, unconsolidated sediments, intermediate plutonic rocks, pyroclastics, evaporites, acid volcanic rocks*	GLiM created by Hartmann and Moosdorf (2012)	polygons	areal fraction
peak ground acceleration, PGA [m/s^2] due to earthquakes expected with a return period of 475 years (i.e. 10% exceedance probability in 50 years)	GSHM[g] created by GSHAP[h] (Giardini et al., 2003)	1°	nearest neighbour
rainfall climatological statistics [mm]	MERRA-2 (Bosilovich, 2015)	0.625° lon x 0.5° lat	bilinear interpolation
surface soil moisture climatological statistics (0-5 cm) [m^3/m^3]	CLSM output	EASEv2, 36 km	-
root zone soil moisture climatological statistics (0-100 cm) [m^3/m^3]			
profile soil moisture climatological statistics (0-100 cm) [m^3/m^3]			
land surface temperature climatological statistics [K]			
runoff climatological statistics [mm]			
evaporation climatological statistics [mm]			
snow depth climatological statistics [mm]			

7. (line 89) When you aggregate landslide data in a landslide location, do you check that your aggregation is reliable?

Unfortunately, we do not fully understand this question. However, we have added some related text in response to other reviewer questions and hope this may help:

"Since LSS informs about the static environmental landslide likelihood, it is common practice to exclude the temporal aspect of landslide occurrence and instead work with landslide presence and absence locations. Multiple landslides within the same 36-km EASEv2 grid cell are therefore aggregated into one 'landslide presence grid cell', resulting in a total of N_{LS}=3757 (orange grid cells, Fig. A1). 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 104-111)

8. (Line 90) Multiple landslides within the same 36-km grid cell are aggregated into one 'landslide location --> The environmental condition selected in a 36 km grid cell can be completely inappropriate and not relevant to explain the landslide.

Thank you for pointing this out. We are aware of the fact that a resolution of 36-km allows for rather large variations within the grid cell. This is caught in the ensemble uncertainty, because it implicitly accounts for spatial representativeness error. We would like to refer to the answer to comment 3 that this study's intention was to retrieve the likelihood of landslide occurrence for an area (of one grid cell) rather than single slopes.

Of course, one slope that is very prone for landslides can be situated in an area that is generally not susceptible to landslides. This is also visible in our results, where we do find low predicted LSS for some landslide presence grid cells (see e.g. grid cell 18 in Figure 5). We will include the following sentence in the discussion, section 5.2: *"At the same time, landslide presence grid cell 18 also has a very wide LSS distribution with a rather low average. This could either indicate that a non-hydrological process caused the landslide (misclassification) or that specific unrepresented features are present within the grid cell area." (Line 402-404)*

9. (Line 111) *Absence grid cells are hence selected from grid cells 7 to 15 around a landslide occurrence*
→ How you can be sure that the selected conditions are not prone to landslides?

Thank you for this remark. In our study, we follow the philosophy that "the past is the key to the future", as commonly done for landslide susceptibility approaches. This entails that for simplicity, we assume that areas where landslides have occurred in the past will also be prone to landslides in the future. Areas without historical landslide observations, are hence assumed to be not prone for landslides. This of course generally calls for a reliable landslide absence reporting, which for the GLC is only the case to a certain degree: "*For large or remote areas, however, no reported landslide does not necessarily mean that the site never experienced [a landslide]." (Line 119-120)* ***The introduction of the buffer and maximum radius was intended to tackle some of this issue by excluding grid cells in the vicinity of known landslide locations to be selected as a landslide absence grid cell. To be 100% sure of landslide absence, we would need to be in the field or do intensive visual assessments based on google Earth as has for example been done by Depicker et al. (2021). The approach taken in this study can be regarded as our best educated guess.***

We will update the text as follows to reflect better on the issue: *"[…] it is still possible that an absence grid cell could experience a landslide, even if none has been reported in the GLC. A prominent example of this are absence grid cells 1 and 7, located in the East African Rift and India, respectively. Both grid cells have no reported landslide, but very wide LSS distributions, with relatively high LSS values. This discrepancy between prediction and observation could indicate the need to visit this location for landslide research. […] Overall, we find an average* $\overline{LSS}_{2500}$ *of 0.18 (0.82) for landslide absence (presence) grid cells (as displayed in Fig. A1) which makes us confident in our classification of these grid cells. "(Line 398-405)*

10. To compute the compound topographic index, you need the specific catchment area and the slope. How do you measure then in a 36 km grid?

Indeed, the CTI was originally calculated per catchment based on 3'' data from SRTM observation south of 60°N, and on 30'' data from GTOPO30 for the high northern latitudes (as described in Verdin 2013, see Table 1). The values that we use are the mean and maximum CTI per 36-km grid box (see Table 1). We have now added the resolution of the datasets that were used to generate CLSM model parameters in Table 1 (see comment 6).

11. Why do you consider the peak ground acceleration if you want to evaluate hydrologically triggered landslides?

Thank you for this question. While local lithology is an essential predictor for landslide susceptibility assessment, various studies have shown that lithological classes alone tell only part of the story. Even when accounting for topography and lithology, seismic proxies like PGA play a key role in explaining regional, continental and global patterns of landsliding. This is also the case across regions where seismicity is overall too weak to directly trigger landslides (e.g. Vanmaercke et al., 2017; Broeckx et al., 2018; Stanley et al., 2021). The most likely reason for this is that weak yet prolonged seismic and tectonic activity can have large effects on rock fracturation and, by extent, weathering and lithological strength (e.g. see discussion by Molnar et al., 2007). As such, peak ground acceleration can be a highly relevant preparatory factor that explains landslide susceptibility, even for hydrologically triggered landslides. We will extend this explanation to prevent confusion on this aspect: *"Peak ground acceleration (PGA) is the likely level of ground motion from earthquake (Giardini et al., 2003). Here, we do not use it as the likelihood of a seismic landslide trigger, but rather as a proxy for the fracturation and weakening that lithologies have undergone due to seismic and tectonic activity (Lin et al., 2017; Vanmaercke et al., 2017; Broeckx et al., 2018)" (Line 156-160)*

12. (Line 161) → The mixed effects approach allows us to include a so-called 'random effect', here the random intercept α, for which we use the average road network density stratified into 6 groups (divided by the global quintile thresholds) → not clear

Thanks for pointing this out. For the logistic regression in this study, we decided to have the intercept vary with stratified road network density in order to prevent this bias in the landslide reporting from affecting the retrieved connection to the environmental predictors. We will update the text as follows: *"The mixed effects approach allows us to include a categorically scaled variable as a so-called 'random effect', here the random intercept α, for which we use the average road network density (RND) stratified into 6 classes. We summarize all land grid cells where average RND is negligible (< 1 m/km²) into the first class and use quantiles 20, 40, 60 and 80 of those grid cells with non-negligible RND to divide the rest into additional 5 classes. The mixed effects approach will then result in one global logistic regression equation that has the same β-factors for all grid cells, but different α values according to each grid cell's RND class. The 6 α values are assumed to come from a zero-mean normal distribution (Zuur, 2009)." (Line 184-190)*

13. (Line 185) We group the grid cells into a total of 100 blocks according to climatological conditions within 10 predefined regions (roughly two per continent), independent of landslide absence or presence → a) this means that each block has about 75 pixel?

This is approximately correct. Since we do not enforce each block to consist of the same number of grid cells (pixels), the number of grid cells per block varies, with the median being 55. We will mention this as part of the alterations in answer to comment 14.

b) What is the rational to select roughly two climatological conditions per continent? If this is not a consistent selection is not representative.

Thank you for your concern. The described method does not aim at concrete climatologic classification but rather at an appropriate process to delineate the aforementioned "blocks". The (sub-)continents were delineated based on our expert opinion, with the intention of a very broad first climatological stratification: North America west, North America east, South America west, South America east, Europe, Africa west, Africa east, Asia east, Asia west, Australia-Oceania. Within those (sub-) continents, the kmeans clustering according to average climatological conditions divided grid cells with similar climatological regimes into 10 groups that we refer to as "blocks". We assume the climatological regimes to also be representative of the landslide regimes. The so grouped grid cells of

one block do not necessarily have to be neighboring. (Please see our answer to comment 14 for some additional details.)

This grouping varies per CV model creation (with changing subsets used for training and testing), but we don't see this as an issue. Had we opted for a random selection of pixels, which is the most common approach in LSS modelling, there would also not have been any consistency in the assigning to different subsets.

We will add the following sentence: *"Note that the definition of the individual blocks varies between each repetition of absence grid cell sampling due to the kmeans clustering algorithm." (Line 233-234)*

14. (Line 183) One subset consists of 20 randomly sampled 'blocks', i.e. small groups, of the 7514 grid cells selected for model creation. We group the grid cells into a total of 100 blocks according to climatological conditions within 10 predefined regions (roughly two per continent), independent of landslide absence or presence. Within these regions, we mimic typical climatological zonations (for example that of Köppen) through k-means clustering (Lloyd, 1982) of 30-year average soil surface temperature and rainfall (see Table 1), dividing each region into 10 blocks. → Not clear the relation between the blocks and the 5 subset.

Thank you for this remark. In line with our reply to comment 13, we will alter the text as follows: *"We employ a blocked random CV (B-CV), as recommended by Roberts et al. (2017), which we found to indeed yield most realistic error estimates in comparison to random or spatial sampling (not shown). This means that instead of randomly sampling individual grid cells into the 5 subsets for training and testing the model as part of CV, we randomly sample small groups of grid cells with similar environmental conditions, so-called "blocks" (see Fig. 1). We expect that the environmental conditions are similar in neighboring pixels (for example same subcontinent) and for similar climate zones. We therefore derive blocks in 2 steps. First, the 7514 grid cells selected for model creation are divided according to 10 predefined (sub-) continents. Within each (sub-) continent, we then derive in a second step 10 blocks through kmeans clustering (Lloyd, 1982) of 30-year average soil surface temperature and rainfall (see Table 1). In total we retrieve 100 blocks comprising different numbers of grid cells (median: 55) that are not necessarily located next to each other. The 100 blocks are then randomly divided into the 5 subsets for model creation (20 each)." (Line 215-228)*

A final distribution of the 5 subsets might then look like this:

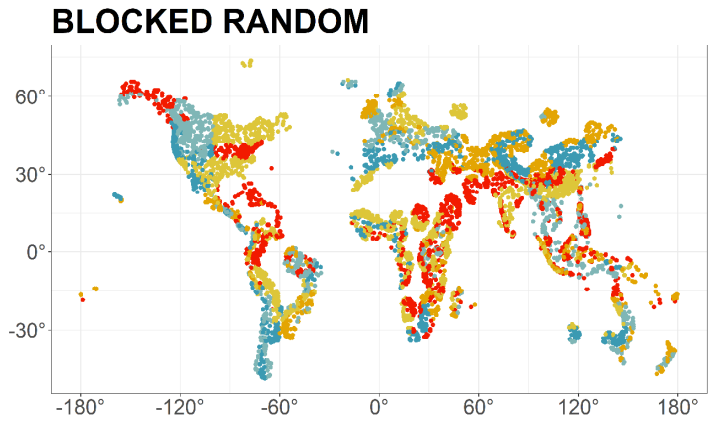

Figure 1 Spatial distribution of landslide presence and absence grid cells selected for LSS model creation. Colors indicate the subset they were sampled into, based on blocked random selection. One point on the map indicates the center of a grid cell.

15. (Line 224) Aggregated data vs observations → Do you really think is reliable to aggregate original observations? In Italy for example you have aggregated 5438 observations in 309 points. I think the two data are completely different and infact you get very low ROC curve.

Indeed, the aggregation has a quite extreme effect for Italy as compared to the other validation data sets of Russia and Africa. Please note, however, that the aggregation transforms the original landslide observation points into landslide presence grid cells. Essentially, what we learn is that in most areas of Italy, you can find landslides and should expect a high landslide susceptibility. The reason for the low AUC value was actually introduced by a mistake on our side that we now discovered: For the ROC analysis, we had selected a box around the reference validation data set region rather than really cutting at the country or continental borders. For Italy that meant that a large number of very high susceptibility grid cells in the Alps joined the analysis as false positives, while we actually did not have validation data for these grid cells. This mistake has been corrected, and we now obtain great AUC values of 0.91 for Italy, 0.84 for Africa and 0.92 for Russia.

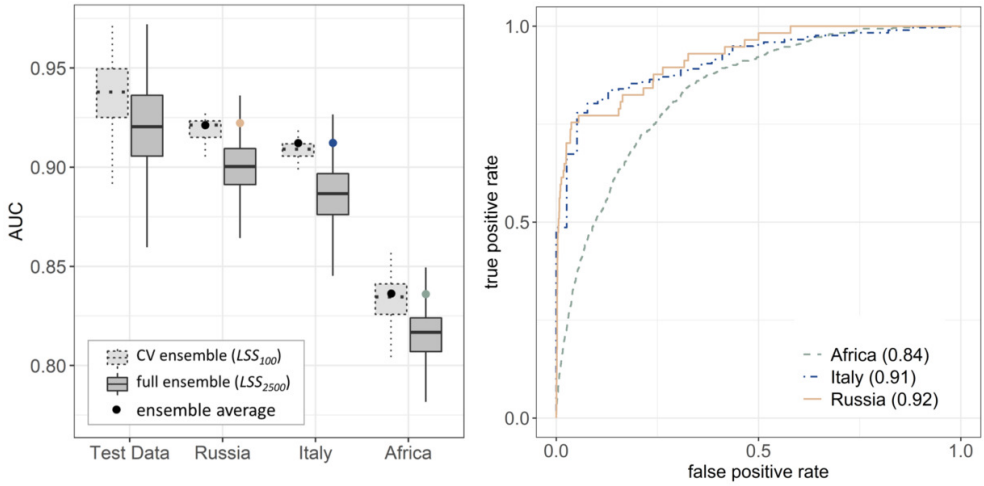

Figure 2 Corrected ROC curves and AUC values for the validation landslide inventories, see Figures 6 and 7 in manuscript

In addition, we will add the following thought in the text: *"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)*

16. (Line 236) Values of the intercept, which is part of all models, vary with road network density as part of the MELR and mostly have and average close to zero (not shown) → can you explain better this statement?

Thank you for your question. Hopefully, we were already able to provide some clarification concerning the varying intercepts in MELR in answer to comment 12 . For each average road network density ($\overline{RND}$) group, we retrieve one intercept (α). These are both positive and negative, i.e. increase or decrease the susceptibility resulting from the predictor variables depending on $\overline{RND}$ within the grid cell. That they have an average of zero is actually an initial condition, which we mention now already in section 3.1 (please see our answer to comment 12 for more details): "*The 6 α values are assumed to come from a zero-mean normal distribution (Zuur, 2009)." (Line 190)*

For low (high) $\overline{RND}$, α takes negative (positive) values. This means we move the connection between $\overline{RND}$ and landslide presence into the intercept instead of using $\overline{RND}$ as a predictor variable itself. We will alter the sentence as follows: *"The values of the intercept α take negative values for low RND and positive values for high RND (by design, not shown)." (Line 294)*

17. Fig. 3 → how much all this complex analysis improves/enhances at worldwide scale a simple regression model applied to obtain an LLS?

Thank you for this question. For individual grid cells where the accuracy was low, i.e. where the difference between predicted LSS and observed absence (0) or presence (1) is large, the ensemble approach was able to improve the accuracy in comparison to the CV ensemble (see answer to comment 3 from reviewer #1). If the interest lies only in average LSS assessment (without uncertainty) to retrieve information on the general global patterns, it is not much of a drawback to use a simple regression model.

We will expand the changes made in answer to comment 4 from reviewer #1: *"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 429-431)*

The key advantage of this study is that we obtain uncertainty estimates that can be used in conjunction with satellite data in a Bayesian framework, which is now better clarified in the text: *"A reliable uncertainty assessment of global LSS estimates is moreover crucial when subsequently combining them in a statistically optimal way with, for example, satellite soil moisture products from Soil Moisture Ocean Salinity (SMOS) or Soil Moisture Active Passive (SMAP) as used by Felsberg et al. (2021)." (Line 39-43)*

18. (Line 265) The LSS2500 map hence performs very well over Russia and Africa, while showing some difficulties to capture the patterns for Italy→ This is the situation where you have modelled 309 points aggregated from 5438 observation. Are you sure that the aggregated points are representative of the failure distribution around the world?

In answer to this, we would like to point to the answers of comments 3, 8, and 15. The aggregation of 5438 observation points into 309 landslide presence grid cells of 36-km resolution is reasonable when interested in an area's likelihood of landslides. We are convinced that the aggregated landslide presence grid cells for Africa, Russia and Italy paint a realistic pattern of areas that are prone to landslides within the continent or country.

Concerning the failure distribution around the world, Figure A.1 shows the 3757 landslide presence grid cells as aggregated from 12515 landslides reported in the GLC (see lines 103-108). The landslide presence grid cells cover all prominent landslide hot spots, as for example also visible in the original GLC point observations (Kirschbaum et al. 2015) or found by Froude and Petley (2018).

19. The procedure is quite complex and the real meaning between the LSS2500 and LSS100 is not very easy to understand

The LSS_{100} is an ensemble of 100 LSS maps that is the direct result of MELR with blocked random cross validation (5 subsets) and 20 times repetition of the absence grid cell subsampling. In addition to that, LSS_{2500} includes predictor variable perturbations, so that per 1 map in LSS_{100}, you have 25 maps in LSS_{2500}. This is introduced in lines 229-238:
"[…] the MELR […] results in 5 different model equations and corresponding LSS maps. By repeating the absence sampling 20 times, we obtain a total of 100 LSS maps (referred to as CV ensemble or LSS_{100}, see Fig. 1)[…]

For the input ensemble perturbations, we apply one fitted model equation to a slightly perturbed set of its predictor variable values. In total, 25 repetitions of this process are conducted […], this results in a total amount of 2500 LSS maps (referred to as full ensemble or LSS_{2500}) […]"

We will nevertheless update the text in a later part of the manuscript to be clearer about both types of ensembles in section 4.3: *"The above discussion of the full ensemble LSS_{2500} includes perturbations to the predictor variables on top of the CV ensemble LSS_{100} obtained by the CV techniques alone." (Line 333-335)*