Nat. Hazards Earth Syst. Sci. Discuss., 2, C2549–C2564, 2014
www.nat-hazards-earth-syst-sci-discuss.net/2/C2549/2014/

Natural Hazards and Earth System Sciences
Discussions

Interactive comment on "Statistical modeling of rainfall-induced shallow landsliding using static predictors and numerical weather predictions: preliminary results" *by* V. Capecchi et al.

V. Capecchi et al.

capecchi@lamma.rete.toscana.it

Received and published: 25 November 2014

article

Author Comments to anonymous Referee #1

V. Capecchi, M. Perna and A. Crisci

25 November 2014

We thank the Referee for her/his relevant and constructive comments and useful corrections.

Conventions adopted in this document:
Bold is used for Referre's comments
Italic is used to cite parts of the manuscript
Plain text is used for the responses to the Referre's comments

According to the Referee's text, we divided our reply in four sections: major limitations, specific comments, technical corrections and table and figures.

MAJOR LIMITATIONS:

1. **[. . .] Perhaps the most significant methodological limitation is the apparent lack of a performance estimation on an independent test set or by using cross-validation.**

   We fully agree with the Referee that one of the major methodological shortcoming of the paper, in its first version, is the lack of the estimation of the model's performances

on an independent dataset. The reason for such choice was that the size of the event inventory maps is meager, especially for 18MAR2013 (only 127 unstable points) and thus models accuracy might be dramatically affected beyond its intrinsic value. However, in order to test the robustness of both models (GLM and RF), we decided to re-run them for both study cases (25OCT2011 and 18MAR2013) dividing randomly each event inventory dataset in two groups: a training set and a validation set. The first group is 70% of each event inventory map, whereas the validation set is the remaining 30%. Due to the small number of elements in the resulting training sets (in particular for 18MAR2013), overfitting may occur in predictive relationship of the models (GLM and RF). So we decided to re-run the models 100 times (for each study case), with 100 different random choice of the training and test sets. Results, figures and discussions were updated according to the new findings.

We think that the new findings do not change substantially the discussions and conclusions drawn in the first version of the paper. Nevertheless Results and Discussions sections has been rewritten as asked by the Referee (see below), in order to be more complete and to analyze in depth the new results.

2. **Furthermore, the relevance of the findings is limited by the fact that the models used here were trained using landslides that resulted from the rainfall predicted by the NWP, i.e. it is merely a hindcasting exercise rather than a forecasting attempt in which a model would be trained on earlier landslide and NWP data (e.g., trained using data from 2011 but applied to predict landslides in 2013 using NWP for the 2013 event).**

   In this preliminary stage, it was not possible to consider additional study cases, because it was not possible to collect enough observed landslides. In the near future, the idea is to "train" the models for some study cases (i.e. at least 10-15 rainfall events) and then apply the models to other independent cases (or on a daily basis, feed the models with NWP data and get the predictions for the following days). At this stage, we did not trained the models for 25OCT2011 and then applied them for 18MAR2013, because

the two study cases show different characteristics from a meteorological point of view (convective precipitation vs stratiform precipitation) and because the two events occurred in two different areas even if geographically contiguous. At this stage, the purpose of the work is just to test the possible benefits of NWP data into a statistical model of shallow landslides triggered by precipitations. Moreover we think that the work presented is not strictly a hindcast exercise since the NWP data are not a numerical description of the rainfall occurred (that is we did not use reanalysis data as initial and boundary conditions). On the contrary, NWP were obtained by numerical integration using global analysis and forecast.

3. **Overall, based on my assessment substantial additional analyses would be required. . .**

   Is the Referee referring to the need to test the results on an independent dataset (point 1 in MAJOR LIMITATIONS)?

4. **[. . .] as well as rewriting of important parts of the paper, including the Results and Discussion.**

   The Results and Discussions sections has been partially rewritten considering the new results and in particular the need, stated by the Referee, to be more complete and to analyze in depth the results. The authors hope that now the reading is complete and adequate.

   The Results section has been rewritten keeping in mind these guidelines: (i) objectively present the key findings in a logical sequence, (ii) initiate each sentence with the purpose of the results presented (i.e. "To evaluate", "To enable", etc. . .), (iii) summarize the outputs achieved in light of the new approach/methodology adopted.

   The Discussions section has been rewritten keeping in mind these guidelines: (i) answer the issue posed in the introduction (evaluate the relative importance of NWP variable in rainfall-induced shallow landslide susceptibility mapping, (ii) discuss how the results achieved concur with those of others and contribute to the wider research, (iii) explain

the limitations of the study. Recommendations for further research and possible future developments are relegated in a separate section.

SPECIFIC COMMENTS:

1. **P4992L25 and P4994L7 Landslides were mapped based on what kind of image source? Indicate sensor, date, resolution?**

   For 25OCT2011, landslides were mapped both by local technical authorities post event surveys and using Rapid-Eye images (multispectral 5-bands with 5-meter resolution) pre/post event (13OCT2011 and 29OCT2011 for 25OCT2011 case) semi-automatic detection with field check. For 18MAR2013, due to time constraints, it was not possible to integrate field surveys observations with pre/post event images analyses and thus commission/omission errors are very likely to occur. This information was added to the text.

2. **How complete is the inventory?**

   A few considerations have been added, in sections 'Study case 25th October 2011' and 'Study case 18th March 2013'. Information added regards mainly the inventory map for 25OCT2011.

3. **Provide summary statistics of landslide size.**

   It was possible to collect summary statistics only for 25OCT2011 inventory map. This information was added to the text.

4. **Based on this information and the information on P5002L20 it is unclear if all landslide-affected grid cells (including the landslide deposit) were used as landslide observations for modeling purposes (i.e., value of response variable = 1), or, for example, only their scarps or a point ("initiation point") at the center of the scarp. The interpretation of the hazard map, model performances and variable importance will vary accordingly.**

   Initiation points at the center of the scarp were used as landslide observations to determine landslide-affected grid cells. This information was added to the text.

5. **Also, how many non-landslide locations were selected for modeling fitting, and how?**

   The number of non-landslide locations equals the number of landslide locations. The non-landslide locations were selected randomly in the areas of interest. This information was added to the text.

6. **P5000 L9 - 45 seems to be an unusually large number of land cover classes, and an impractical one for the purpose of landslide modeling. In the context of RF, for example, there are more than 10 trillion ways of creating a binary split based on a 45-class predictor variable (see Strobl et al., 2007, in BMC Bioinformatics); in the GLM, 44 coefficients need to be estimated when including this categorical predictor in the model.**

   That was an error in the first version of the paper. We incorrectly indicated 45 as the number of land-cover classes, but this is the number of the original CORINE dataset. In the two areas of our interest only 8 classes belong to 25OCT2011 area and only 10 classes belong to 18MAR2013 area. This information was corrected and a list of the classes was inserted in the text.

7. **P5002 L22-23 NWP data was not downscaled from 3 km x 3 km resolution to 30m x 30m resolution. Please provide a justification and discuss (in the Discussion) possible implications due to finer-scale variability, considering that high-intensity rainfall events are often spatially very variable.**

   A few discussions about possible possible implications due to finer-scale variability has been added to the text. Anyway, at this stage, we wanted to preserve the information content provided by NWP output and, as stated in the text, evaluate possible benefits in integrating the meso-γ scale ($\simeq$2-20 km the typical resolution of nowadays NWP products) information, with the micro-γ scale ($\leq$ 20 m of spatial resolution) information.

8. **P5003 L4 provide reference for application of RF in landslide modeling or geomorphology. However, RF may also overfit to spatial data, as seen in other geomorphological applications.**

   References for application of RF in landslide modeling were added. Some discussions on its use in landslide modeling, past results achieved and possible overfitting are provided.

9. **P5003 L24ff - While RF is often praised based on these general characteristics, in the present context it may be appropriate to think about the characteristics of both RF and GLM that make them adequate choices for landslide modeling. In this study, for example, there don't seem to be missing values that would need to be handled by RF (c). Also, (b) GLM doesn't assume a "formal distribution" of predictors either (this is a common misbelief). Likewise, GLM can handle categorical predictors in addition to quantitative ones (a), perform automatic variable selection (in this study, using the AIC) (c), has "little need to fine-tune parameters" (e). GLM is also less in need of cross-validation (d) since it doesn't tend to overfit less than RF. Variable importance (P5004L4) can also be assessed in the GLM based on model coefficients, which have a fairly straightforward interpretation, while on the other hand the permutation-based approach can also be applied to GLM, although this is not very common. Finally, interaction terms can be incorporated into GLMs (less conveniently than in RF though), and nonlinear extensions such as the generalized additive model (GAM) are known in the landslide literature (e.g., Goetz et al., 2011 in Geomorphology).**

   Overall, rather than praising random forests and announcing "excellent performances" (L29), a more balanced account of the relative advantages and disadvantages of RF and GLM should therefore be presented.

   This part has been partially rewritten. A more balanced assessment of the relative advantages of GLM and RF is presented. A reference reporting overfitting of RF is reported.

   Possible impact of RF overfitting in the present context is discussed (in Discussion section).

10. **P5003 L15 Please check if stepwise forward variable selection was performed instead of backward, as stated. Fitting a full GLM with $>$90 predictor degrees of freedom (see Table 4) as a starting point for backward variable selection seems problematic in the case of the 18 March 2013 event with only 127 landslide observations.**

    That was an error in the first version of the paper. Stepwise forward variable selection was performed. Text was corrected.

11. **P5004 L7-23 can be shortened substantially and refocused on the relative utility of RF in these applications. E.g., how well did RF perform in studies that compared performances of various techniques in a landslide or geomorphological context? Was overfitting an issue in any of these studies?**

    This part was rewritten and focused on the relative utility of RF in geomorphological context.

12. **The performance measure (AUC) and the procedure used for its estimation are not mentioned in the Results section. Since there is no specific mention of setting a holdout data set aside, one might get the impression that the training sample was also used for the estimation of the AUC. The result would be an over-optimistic AUC estimate probably a highly overoptimistic one in the case of RF**

    This point has been overcome by the new procedure adopted in this second version of the paper for the estimation of model performances (see point 1 in "MAJOR LIMITATIONS" of Author Comments)

13. **P5005L3-4 and Table 4: The use of slope aspect as a predictor variable (without any transformation) is problematic since this is a directional vari-**

able, i.e. 0 degrees = 360 degrees, and e.g. 359 degrees minus 0 degrees = 1 degree. One way to fix this is to use the cosine and sine of slope aspect instead of aspect itself.

In this last version of the manuscript the sine of slope aspect was used. This information was added to the text.

14. **P5005L18-20, Figure 5, Tables 2 and 3: Insert a scatterplot showing the correlation between model predictions and observations of 24-h rainfall, in addition to scatterplots of modeled and observed 1-h rainfall. To the tables please add summary statistics of the differences between observed and modeled, since this would provide information on model bias and precision. Discuss these results and performance measures in Section 3.1**

To give an idea of model bias, we added, in the text, the indication of the multiplicative bias i.e. $\frac{\frac{1}{N}\sum F_i}{\frac{1}{N}\sum O_i}$, where N is the total number of observed/forecasted values, F_i are the forecasted values, O_i are the observed values. It measures the average forecast magnitude compare to the average observed magnitude. Perfect skill score is 1, even if it is possible to get a perfect score for a bad forecast if there are compensating errors.
In addition the correlation coefficients were added.
We evaluated that descriptive statistics (which surrogate box plots), RMSE, POD, FAR, multiplicative bias and correlation coefficients give an idea of the model forecasting skills. Adding scatterplots would result in new figures in the paper and perhaps plots on POD and FAR should be removed. Nevertheless, these latter plots add information on the underestimation of the model's data and allow discussions as done in section "Discussion".

15. **P5008 The text refers to predictors as being "classified" as being important. Increased note purity is a quantitative measure, please indicate in the Methods how you classified based on this measure. Is there a rationale for using this particular importance measure rather than decrease in accuracy or AUROC? In fact decrease in AUROC would seem to be the most natu-**

ral measure of predictive variable importance in this context since AUROC was chosen as the performance measure and has an actual interpretation, while increase in node purity is a very abstract measure that can only be interpreted in relative terms

In this second version of the paper the mean decrease in accuracy instead of the mean decrease in node impurity was used as the measure of variable's importance.

16. **A table with model coefficients from the fitted GLM should be included in the article, and these coefficients should be interpreted, as far as possible, in terms of odds ratios, especially as far as NWP variables are concerned**

So far it was not possible, due to time constraints, to update the paper with further analysis regarding the interpretation of GLM coefficients in terms of odds ratio. Work is currently under development for the computation of odds ratio. In particular, additional calibrations need to be addressed for the determination of the threshold for dichotomizing outputs according to the predicted probability, since results highly depend on this parameter. Because of the lack of any possible and exhaustive discussions on this topic, the table with GLM coefficients was not inserted in the text.

17. **Discussion on P5009 - The text on this page is largely a summary of study objectives and procedures without actually discussing the present findings in the context of the literature, in terms of their broader relevance or with regards to their limitations. The text starting in L23 lists a number of papers that also utilize NWP data in a natural hazards context, without discussing the present results or methods in the context of these studies**

The present findings have been discussed (in Discussion section) with reference to similar papers (list of papers starting in L23).

18. **P5010 L13ff - Are AUROC values really comparable among studies in different study areas? The problem with any comparison is that study areas that**

contain larger "easy-to-predict" portions (such as flat valley floors or less steep forelands) will result in higher AUROC values. In other words, the exact, often arbitrary definition of a study area determines the AUROC to a large extent. Overfitting and the performance estimation technique further influence the reported AUROC values.

AUC values are compared only to one paper which has characteristics similar to those here under exam (i.e. similar and contiguous areas of interest, same performance estimation technique).

19. **A discussion of model results (e.g., model coefficients, variable importance, relationships between predictors and response in RF) in geomorphological or hydroclimatological terms is missing.**

    A few considerations on model results of the static predictors (variable importance mainly) have been added. These considerations were kept limited in the text because analyzing the impact of static predictors in LSM by using RF was already assessed in recent works and it was not the aim of the paper. We don't report any innovative finding on this issue.

20. **In the context of landslide prediction, what is the relevance of the present results for actual landslide forecasting (see my General Comment above) - i.e., fitting a model to a landslide inventory and NWP data, and applying it to new NWP data that comes in in near-real time in order to forecast future landslides.**

    A few considerations have been added in the Discussions section regarding these issues (see also point 2 in MAJOR LIMITATIONS).

TECHNICAL CORRECTIONS:

1. **Throughout the paper, the authors refer to the statistical models as 'indirect' models. There seems to be no justification for using this attribute, it**

should therefore be omitted.

We use the diciture 'indirect models.' as defined by Guzzetti et al (1999), i.e. "*Indirect methods for landslide hazard assessment are essentially stepwise. They require first the recognition and mapping of landslides over a target region or a subset of it (training area). It follows the identification and mapping of a group of physical factors which are directly or indirectly correlated with slope instability (instability factors). They then involve an estimate of the relative contribution of the instability factors in generating slope-failures, and the classification of the land surface into domains of different hazard degree (hazard zoning)*".

2. **P4988 In the Abstract, briefly mention AUROC (area under the ROC curve) values (L9) and variable importance of numerical forecast (L15).**

   AUC values and variable importance of numerical forecast were reported in the abstract.

3. **P4988 L2 'a ...modeling' - rephrase L5 'model's forecast' $->$ 'model forecasts; L5 'combine together': omit 'together'**

   Part of the abstract was re-written and the above words were omitted or replaces.

4. **P4999 L27 and throughout the paper: 'returning period' $->$ 'return period'**

   'returning period' was replaced by 'return period' here and elsewhere in the text.

5. **P5000 L11-28 Unfortunately it is not clear from this text how the EVI variable was derived from 'raw' (EVI?) data for the years 2000-2013. Phenology is mentioned at some point (L14) but there is no mention of specific measures of phenology. Precisely what EVI was used, e.g. the overall maximum value of these 14 years, or some multiannual average, or the date of seasonal maximum EVI, etc?**

   We used a layer derived from the temporal climatology (i.e. an average) of the EVI index, using all the available satellite imagery for the time series 2000-2013. Text was modified and we hope it is now more explanatory.

6. **P5002 L22 - 'to nudge' - reword L26 (and elsewhere) 'In bibliography' $->$ 'In the literature' - provide reference L27 (and possibly elsewhere) 'forecasting' - This is not forecasting since the models were trained based on the outcomes that are to be predicted. Prediction is a more general term that would apply here.**

   'to nudge' was replaced by 'to downscale, geo-statistically,'. 'forecasting' was replaced by 'prediction'. 'bibliography' was replaced by 'literature' here and elsewhere. References were provided.

7. **P5003 L1 'GLM model' - omit 'model' since the 'M' in GLM stands for 'model' (also in L7 and possibly elsewhere). This sentence can be simplified since the models used 'are' the GLM and RF (not 'based on') L3 'see below for references' - please insert references here instead - 'below' could be anywhere in the article L5-6 'No interactions...' - I suggest to omit this sentence since it letting different prediction methods interact seems uncommon. L24 needs to be rephrased**

   P5003 L1: 'GLM model' was replaced by 'GLM' everywhere in the text. L3 references were inserted. L5-6 the sentence was removed. L24: this sentence was rewritten.

8. **P5004 L24-29 can be omitted**

   These lines were removed.

9. **P5005 L1 'R translation' $->$ 'R implementation'**

   'R translation' was replaced by 'R implementation'

10. **P5005 L20-25 Avoid sentences that either just report a method in the Results section (methods should be introduced in the Methods section), or that simply point to a figure or table without providing an actual summary or interpretation of the information that the reader is referred to. (Please**

also check the rest of the Results for such sentences, e.g. the first and third sentence on P5007 does not say anything substantial. Same for P5007L14-16.

Both sections 'Evaluation of the forecasting skills of NWP outputs' and 'Evaluation of landslide hazard maps' have been consistently rewritten. In general the 'Results' section has been rewritten keeping in mind these guidelines: (i) objectively present the key findings in a logical sequence, (ii) initiate each sentence with the purpose of the results presented (i.e. "To evaluate", "To enable", etc…), (iii) summarize the outputs achieved in light of the new approach/methodology adopted.

11. **P5005L21 It would appear that scatterplots would provide more direct evidence of the relationship and difference between modeled and observed rainfall amounts compared to contingency tables and Fig. 6. I suggest to omit Fig. 6.)**

    Instead of scatterplots, a set of skills are used to evaluate model's accuracy. See Author Comment number 14 in 'SPECIFIC COMMENTS'.

12. **P5006 L7-8 This 'shift' should be referred to as a 'model bias'.**

    The word 'shift' was removed and a new sentence was inserted to account for the model bias.

13. **P5007 L10-13 can be omitted.**

    We kept and moved these lines at the end of the section, since they allow a few considerations (in the Discussion section) on the feasibility of a early warning system based on the proposed modeling chain (considerations added in section Discussion). For the same reason, we added a single sentence on the CPU time required by the WRF simulations at the end of section "Evaluation of the forecasting skills of NWP outputs".

14. **P5007 L20-22 AUC values how estimated? Use more precise terminology, e.g. 'on the training set' or 'out-of-bag estimate of', as applicable.**

The estimate of AUC values is reported at the end of the Method section and briefly recalled in section 'Evaluation of landslide hazard maps'

15. **L23-L2 on following page: Move to Methods**

    This sentence was moved to Methods

16. **P5008 'All these layers are highlighted...' (two occurrence): Please omit this information from the text, it should be in the figure caption.**

    This sentence was removed

17. **P5008L24 Please rephrase, adding a particular set of predictor variables does not seem to constitute the development of a 'statistical framework'.**

    This sentence was rewritten and in general 'develop' was replaced by 'implement'.

18. **P5009L4 '...no interactions...' Please omit, this seems obvious and not relevant for the Discussion.**

    This sentence was rewritten.

19. **P5012L1 Avoid repetition from P5011L23-24.**

    This sentence was rewritten.

TABLE AND FIGURES:

1. **Table 4 should mention the 45 classes of the land cover variable**

   The list of the classes of land cover for both study areas has been inserted in the text.

2. **What unit is 'h' in 'hm-1' for curvature? (Certainly not 'hour'.)**

   Here we use 'hm' as the SI symbol for hectometer.

3. **Figure 2 should be integrated into Figure 1.**

   Figure 2 was removed and the extent of the WRF simulations has been integrated in Figure 1.

4. **Figure 3 could be omitted since this is a standard procedure that is not greatly modified by adding NWP data.**

   The figure was removed.

5. **Figure 4: improve readability by using different (larger?) symbols for rain gauges.**

   Figure 4 was modified.