REFEREE'S COMMENTS ON egusphere-2025-1590

RC#2

GENERAL COMMENT

R: The contribution "Use of delayed ERA5-Land soil moisture products for improving landslide early warning" by N. Palazzolo and co-Authors is interesting and addresses questions relevant with the scope of NHESS. In this paper the researchers conducted a series of direct experiments. They aimed to determine how effective it is to use antecedent soil moisture data (ranging from 0 to 15 days old) from a global reanalysis model, combined with rainfall data, to predict landslide triggers using machine learning. As one might expect, the findings indicate that this historical soil moisture data enhances model performance, though its benefit diminishes as the time lag increases. Results, and discussion sections are short compared to amount of work done. They should be increased. The theoretical background is well-argued but not complete. Review of literature seems completed. The description of study area is sufficiently complete. The description of methodology and successive parts of paper are well organized but not complete. The readability of the whole paper is good with a good English. In general, the synthetic approach of the study is very clear, but I think that going into more detail on several aspects (later reported) could be useful for the readers. It can be published on NHESS journal only after minor revision.

A: The authors sincerely thank the reviewer for the detailed review and the insightful comments, which will help us improve the quality of our work. Based on your suggestions, we will expand the methodology section and enriched the results and discussion with additional analyses. Below, we provide detailed point-by-point responses to each comment.

SPECIFIC COMMENTS

R: 1 Introduction (Line 30-40)

My main comment regards the conceptual background of the comparing of the performance of different models. It would be necessary to better understand from the authors whether it is possible to directly compare the performance of the models with and without soil moisture (on the empirical rainfall thresholds) information based exclusively on the comparison with the TSS index or whether other strategies exist in this direction.

A: Thank you for the comments. First let us clarify that we compare the performances of ANN models with and without soil moisture, so the comparison of True Skill Statistic (TSS) is made under the same conditions and also has been the main topic of previous papers (Distefano et al., 2022). We acknowledge, however, that relying exclusively on a single performance index such as TSS may not fully capture all aspects of model performance. Although TSS is widely used and recommended in the literature for summarizing model skill in binary classification tasks (Frattini et al, 2010), it also corresponds to the performance closest to the top-left corner of the ROC space, which represents the ideal balance between true positive and false positive rates. For this reason, we selected TSS as the main metric to allow a direct and consistent comparison across models. Nevertheless, we agree that additional analyses can help refine the interpretation. In the revised version of the manuscript, we will therefore include a more comprehensive ROC-based analysis to better characterize the trade-offs between missed alarms and false alarms across the different models. Furthermore, we have adopted strategies for ensuring generalization capabilities of the developed ANNs, by considering training, validation and test datasets and the early stopping technique, for taking into account of the different complexity of models. We will however mention that other metrics such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), or adjusted performance scores can be useful when model complexity differs.

REFERENCE: P Frattini, G Crosta, A Carrara, Techniques for evaluating the performance of landslide susceptibility models, Engineering Geology, Volume 111, Issues 1–4, 2010, Pages 62-72, ISSN 0013-7952, https://doi.org/10.1016/j.enggeo.2009.12.004.

2 Material and methods

R: (Line 48-49) Have you tried using landslide data from other catalogues? (i.e. Peruccacci et al., 2023)

A: We thank the reviewer for the suggestion. At this stage, we have not yet tested other landslide catalogues. As a first step, we used the same landslide dataset as Distefano et al. (2023) to allow a consistent comparison and to evaluate how our methodology performs under similar conditions. However, we fully agree that testing different landslide inventories (Peruccacci et al., 2023), as well as alternative sources of soil moisture and rainfall data, may be a direction for future

research, in order to strengthen the validity of our findings. Hence, we will add a sentence about this in the Discussion section of the revised manuscript.

Reference: Peruccacci S., Gariano S.L., Melillo M., Solimano M., Guzzetti F., Brunetti M.T. (2023) The ITAlian rainfallinduced LandslIdes CAtalogue, an extensive and accurate spatio-temporal catalogue of rainfall-induced landslides in Italy. Earth Syst Sci Data, 15, 2863–2877, doi: 10.5194/essd-15-2863-2023

R: (Line 55-60) The Authors reconstruct triggering and non-triggering rainfall events.
I didn't understand the following list of things:
The rainfall event or landslide triggering condition (MPRC automatically reconstructed by the tool) normally represents a subset of the entire rainfall event while non-triggering rainfall events are always considered as a whole. Did the authors take this difference into account when creating the input data for the neural network?
A: Yes, we confirm that this is the case, even though in practice in most of the cases the entire rainfall events is attributed to the landslide (see next reply).

R: How many triggering conditions coincide with the entire rainfall event? If not, how many of them differ significantly in terms of duration and cumulative rainfall, and what justifies the choice of these two sets of triggered and non-triggered rainfall events? The authors should provide a better argument and commentary on the method used.

A: For most of the landslide events (103 out of 144) only the day of occurrence was available in the database. In this case we assumed that the triggering instant was at the end of the day. Hence, in many cases the entire rainfall is attributed to landslide triggering. However, we are aware that from a theoretical standpoint it may not be correct to attribute to a landslide the part of the triggering event that goes beyond the triggering event. We may add a comment on this point in the revised manuscript.

R: The number of reconstructed non-triggering rainfall events is at least two orders of magnitude greater than those that trigger landslides. What procedure was used to account for this substantial difference? Are the samples used for the training, validation and testing phases balanced or unbalanced?
A: Thank you for mentioning this issue. Indeed, the dataset used in this study is unbalanced, and the higher number of non-triggering rainfall events reflects the actual availability of observed landslides in the catalogue. However, this imbalance does not compromise the reliability of the modelling process. ANNs were trained using a cross-entropy loss function, which is widely adopted in classification tasks and known to be robust to class imbalance. Then, the network outputs are transformed into binary predictions through a thresholding procedure, where a landslide is predicted (i.e., output = 1) when the network output exceeds a certain threshold, and 0 otherwise. Within a study currently under preparation we have investigated the influence of different proportions of non-triggering and triggering events. The outcome of these tests suggests that it not necessary to balance the datasets with our model setting and that unbalanced datasets lead to more robust models, as, while using the entire information available, they reflect the actual likelihood of triggering and non-triggering events. A comment on this point will be added to the revised manuscript in the Methodology and Discussion sections.

R: What observations are used in this last case, and how is sampling performed? Have statistical tests been conducted to confirm whether these samples are representative of the population? Please specify the setting more precisely.
A: We further clarify that the entire dataset, comprising both triggering and non-triggering rainfall events, is randomly split into three subsets: 70% for training, 15% for validation, and 15% for testing. The subset allocation was random, according to a uniform distribution over the dataset indices. This means that each observation had the same probability of being assigned to any of the three subsets, ensuring an unbiased and representative sampling process. To account for the variability introduced by this random sampling, each neural network configuration was trained and evaluated 30 times, each with a different random split. This strategy allows us to assess the robustness of model performance and to simulate realistic operational conditions, where landslide-triggering events are relatively rare. The use of box plots in the results highlights this aspect: the width of the boxes reflects how sensitive the model performance is to the random selection of the subsets. Narrow boxes indicate stable results across different splits, while wider boxes reveal a stronger dependence

on the specific partitioning. Thus, the repeated randomization approach, together with the use of box plots to represent the results, was explicitly adopted to explore and quantify this variability. This point will be stressed in the revised manuscript.

R: 2.3 Artificial Neural Network models (ANNs)
(Line 80-90) ANNs Input variables/data
The reconstruction of the rainfall conditions that triggered landslides is performed starting from the assumption that in the place where the landslide occurs the rainfall is the same as that measured on a representative rain gauge (specific criteria dependent on a parameterized variable that considers the distance was used by CTRL-T tool). As regards the choice of cells that contains information associated with the soil moisture status, it seems that the cell that includes the representative rain gauge chosen automatically by the tool has been considered. Under this hypothesis (soil moisture associated with the cell that includes rain gauge) what considerations can be made from a physical point of view? Is it conceptually correct to look for a relationship between the saturation state of a soil different from that in which the triggering of a landslide occurs? If the cell that contains the landslide is considered, do the results change? Considering the spatial resolution of the ERA5-Land product (about 9km x 9km) probably in most cases the cell containing the landslide or the rain gauge is the same, but I believe that deepening this aspect by making descriptive statistics could be useful. Moreover, for an operational use, the forecast value returned as output by the model would describe the possibility that a landslide could happen in that specific cell for which the soil moisture value was provided as input. Therefore, in my opinion, it should perhaps also be used in the model's learning phase.

A: We agree that ideally the rain gauge, soil moisture information and landslide locations should be the same. Nevertheless, this is never possible in practice (with rain gauges) as the location of future landslides is unknown. What our model predicts is "a possible areal landslide event" within a fixed radius R_b around the rain gauge. The performance we have obtained prove that the use of soil moisture, even with this spatial resolution provides better performances respect to the use of precipitation only. This may be related to spatial correlation of soil moisture. A comment will be added on this point in the revised

R: 3 Study area and data
(Line 140-150) CTRL-T parameters setting
What is the length of the "warm" spring–summer period (CW) and "cold" autumn–winter period (CC). More in detail, what values were assigned to the variables sws (beginning of warm season) and ews (end of warm season), i.e. CW from May to October and CC from November to April correspond sws=5 (May), ews=10 (October). Please specify the setting more precisely. Add the *sws* and *ews* parameters in the Table 1 (cfr. comment refered to Line 140-150)
A: Thank you for the opportunity to provide further clarification. In our study, we defined the warm season (CW) as the period from April to October, and the cold season (CC) as the period from November to March. Accordingly, the parameters sws = 4 and ews = 10 will be added to Table 1 for clarity.

R: 4 Results and discussion
(Line 160-165)
Figure 4 shows box plots of the results for the autocorrelation conditions. The text provides a good description of the behaviour of the variations in terms of lag k. However, a visual analysis of the (k) values (k = 1, 2, ..., 15 days) in the different layer depth cases (b) 7–28 cm, (c) 28–100 cm, and (d) 100–239 cm reveals that the values are indistinguishable (perhaps due to the scale). Consequently, it is unclear why layer 2 was selected for subsequent analysis (b). Please explain better.
A: Thank you for this comment. The purpose of Figure 4 is to explore overall the autocorrelation structure of soil moisture by comparing its lagged values (k = 1 to 15 days) with the reference value at lag 0. The box plots are intended to show how the autocorrelation decreases with increasing lag, rather than to directly compare the performance of models obtained from different soil layers' data. As for the choice of Layer 2 (7–28 cm) for subsequent analyses, this was motivated by our previous findings reported in Di Stefano et al. (2023), where layer 2 consistently yielded the best performance in landslide prediction tasks. Based on this, we decided to focus our experiments on the configurations presented within the manuscript. We will further clarify this in our revised manuscript.

R: Line 175-190

How does performance change for all combinations of layer use? For example, what is the maximum TSS (mean) value for D-E-S1, D-E-S3, D-E-S4, D-E-S all-1, and so on? Has a systematic analysis been conducted to better understand the effect of each input data component introduced in network training? If it is not very time-consuming, I would suggest trying this approach.

A: We thank the reviewer for the suggestion. A systematic analysis of the effect of different combinations of input layers (including D, E, S1, S3, S4, and their combinations) on model performance has already been conducted in Distefano et al. (2023), which our work takes as a starting point. For this reason, we did not replicate the same analysis here, as we believe it would not significantly add to the substance of the present study. However, we will better clarify this aspect in our revised manuscript.

R: Figures description

Description of Figure 2

In the structural diagram of the ANN, the symbol for dichotomy on the third layer is different from those for the three preceding nodes on layer 2.

A: We thank the reviewer for the observation. The difference in the symbol is intentional, as two different activation functions have been used: a tan-sigmoid function f(n) for the hidden layer, with output in the range (-1, 1), and a log-sigmoid function g(n) for the output layer, with output in the range (0, 1). Additional details will be added in the revise manuscript to make it clearer.

R: I suggest to:
remove "---and---" because it's implicit in the figure
replace "depth" with "cumulated rainfall" (also extended to the text)
***A**: Will be done*
R: Specify the terms S (ERA5-Land depth)
***A**: Will be done*
R: Specify which is the input, hidden and output level.
***A**: Will be done*

R: Description of Figure 3

Add geographical grid and coordinate labels.
***A**: Will be done*
R: I suggest changing the colour filling of the peninsular part of Italy (i.e. the monochrome palette) because the scale is different respect to the study area.
***A**: Will be done*
R: I suggest increasing the size of the dots and changing the starting colour of palette from "blue" to "green" for greater contrast.
***A**: Will be done*
R: Add the description of the symbols in the caption.
***A**: Will be done*

R: Description of Figure 4

Formatting borders, label text, and ticks in black instead of grey
***A**: Will be done*
R: Add the description of the continuous red and the dashed grey curves in the caption (also extended to the text).
Remove the labels and headings from the x-axis for (a) and (b) and from the y-axis for (b) and (d). This should make the figure clearer by increasing its size.
***A**: Will be done*

R: Description of Figure 5

Formatting borders, label text, and ticks in black instead of grey
***A**: Will be done*
R: Change dimensions of '(a)' and '(b)' as in figure 4.

***A**: Will be done*

TECHNICAL CORRECTIONS

R: Below a list of some more detailed comments and suggestions referred to specific parts of the text.
(Line 54): replace "(*D*, in hours)" with "*D* (h)"
***A:** Will be done*
R: (Line 55): replace "(*E*, in millimeters)" with "*E* (mm)"
***A**: Will be done*
R: (Line 116): remove the space in "non-triggering"
***A**: Will be done*
R: (Line 120): remove the space in "(missed alarms)"
***A**: Will be done*
R: (Line 123): Change the 'T' term format
***A**: Will be done*
R: (Line 126): enter the comma in the range [0,1]
***A**: Will be done*
R: (Line 129): replace "southern Italy" with "Southern Italy"
***A**: Will be done*
R: (Line 31): replace "700" with "700 mm"
***A**: Will be done*
REFERENCE: Peruccacci S., Gariano S.L., Melillo M., Solimano M., Guzzetti F., Brunetti M.T. (2023) The ITAlian rainfall-induced LandslIdes CAtalogue, an extensive and accurate spatio-temporal catalogue of rainfall-induced landslides in Italy. Earth Syst Sci Data, 15, 2863–2877, doi: 10.5194/essd-15-2863-2023