Author Comments Manuscript “Explainable Artificial Intelligence for deriving 3D dynamic rainfall thresholds for landslide triggering using Kolmogorov-Arnold Networks” (EGUSPHERE-2026-1796)

RC1: 'Comment on egusphere-2026-1796', Anonymous Referee #1

Referee #1 Comment 01

> *The study addresses an important and timely problem by combining XAI with dynamic rainfall threshold modelling for landslide triggering, and the topic is clearly relevant to today’s needs. Although the study appears promising in scope and ambition, the current version would benefit from a more rigorous methodological validation framework, clearer justification of modelling choices, stronger uncertainty analysis, and deeper demonstration that the learned thresholds are both physically meaningful and operationally robust. I have some recommendations/suggestions regarding methodological improvements for the authors.*

We sincerely thank the anonymous Referee 1 for the recommendations for improvements and constructive comments on the manuscript. We greatly appreciate the time and effort invested in the review process, as well as the helpful suggestions provided. We agree that the suggested revisions will improve the quality and clarity of our work. We believe that a revised version of the manuscript will adequately address the points raised.

Referee #1 Comment 02

> *Please demonstrate strict spatiotemporal leakage control, because rainfall samples are drawn across neighboring grid cells and long-time sequences, which may otherwise inflate the reported generalization skill. I suggest reading a studies on physics-informed rainfall forecasting with XAI. The authors may also read the following study: “https://doi.org/10.1007/s44163-026-00833-z”*

Leakage control has been addressed in the original dataset. The dataset as published contains a 30-day precipitation time series and the correct output class (failure or non-failure) per sample. The training is performed on random samples from this dataset meaning that there is no temporal link between the samples as the model receives the classification samples as inputs, as opposed to time series forecasting. Furthermore, all samples are provided as precipitation time series alone, without spatial identifier. The model does thus not learn spatial relations from the data and no spatial modelling is used to build the model. However, future

work could address this in more detail by, for example, employing a spatio-temporal model, including spatial modelling of the data using Graph-Neural Networks to capture spatial variations in the data. Following this comment, we will add a more thorough explanation of the dataset construction to the manuscript to ensure the reviewers' concerns are properly addressed.

Referee #1 Comment 03

> *Please replace the single random train-test split with blocked temporal and spatial cross-validation to verify that the KAN advantage remains stable across years, regions, and event densities.*

The study area has been divided into non-overlapping cells where the cell size was based on the grid-size for the available GPM-IMERG rainfall data. We agree however that a temporal split into training (first 80% of available time) and test/validation (remaining 20% of time) could prevent temporal leakage. We will address this issue in the revised manuscript by re-training all the models on this new temporal split.

Referee #1 Comment 04

> *Please justify the asymmetric comparison between a two-output KAN and a one-output MLP or re-run all classifiers under matched output parameterization and tuning budgets for a fair benchmark.*

We agree that this would further improve the comparability of the models. The two-output KAN has originally been chosen to being able to derive the Dynamic Rainfall Thresholds. As this adds more activation functions to learn the model might be able to better perform with given training than a single output KAN of the same dimensions, but this is offset by the fact that we train different depths and widths of networks for all architectures. As such, the comparison is still inherently fair as the hyperparameters cover different depts and widths.

Finetuning of the hyperparameters themselves has been implemented as a grid search with hyperparameter ranges based on typical sensible ranges. For all models, this included manual attempts at tuning the model best as possible first and then ensuring a more thorough grid-search is conducted. Following this useful comment, we will add clarification to the manuscript regarding the hyperparameter search and discuss the comparability with special regards to the difference in the last layer of the KAN network.

Referee #1 Comment 05

> *Please quantify calibration uncertainty in addition to discrimination, since early warning usefulness cannot be established from ROC-AUC and TSS alone on an imbalanced dataset.*

The comment of the reviewer is well justified, as there is no common practice on reporting performance scores. We opt to report the most used scores Accuracy, ROC-AUC and TSS together with True Positive Rate (TPR) False Positive Rate (FPR). Calibration can in this case not really be achieved as we do not predict probabilities and rather focus on a binary classification task. What is more, the KAN model does not feature a sigmoid activation function in the last layer as is common to predict probabilities. This results in the output being non-confined to the internal [0,1]. The threshold visualisation is solely based on a comparison of the outputs as well where the larger of the two outputs decides the class. We will add a clarifying paragraph to the revised manuscript as well discussing the future possibilities of adapting the method to include better ways to approach calibration.

Referee #1 Comment 06

> *Please test whether the strong contribution of day of year reflects genuine process seasonality or inventory and sampling bias by comparing cyclic encodings, debiased formulations, and ablation results.*

The dataset has been assembled to include an event distribution where all days are represented equally. However, the positive samples have a distribution with a clear peak in the autumn months. The final dataset aims at drowning the positive samples by adding many negative samples such that the positive samples make up only a small percentage of the total samples for all days. The model seems to pick up the distribution of positive sample nonetheless and the day of year is thus able to contribute significantly to the prediction. Following this comment, we will include a more detailed description of the dataset, including the temporal distribution of the positive samples as well as all samples to clarify this point.

Referee #1 Comment 07

Please provide sensitivity and robustness analyses for the dynamic rainfall thresholds, including prototype function choice, fit acceptance threshold, stochastic initialization, and the stability of the resulting threshold surfaces.

The prototypes have been chosen according to the available functions in the KAN implementation pykan, while the fit acceptance threshold has been set as high as possible while allowing. Initialisation of the parameters follows the default KAN implementation (in pykan) which samples the base functions from a normal distribution with random noise. The threshold surfaces largely depend on the grid size used to visualise the threshold points in three dimensions. Depending on the resolution, an interval for the maximum difference between the function values at the points can be chosen. Coarser resolutions will thus result in worse representations of the surface for the same fit acceptance thresholds. The reviewer highlights an important point here which we will add to both the method description of the DRTs as well as to the discussion. We firmly believe that thoroughly discussing this point will improve the manuscript significantly.

Referee #1 Comment 08

Please assess the physical and operational validity of the learned thresholds against established intensity duration or soil moisture-informed baselines, because interpretability claims require process-consistent agreement beyond visual plausibility. For this, the study suggested in Comment No. 01 can be utilized as well.

We first and foremost propose a model that outperforms recently proposed deep-learning models for predicting rainfall-induced landslides based on precipitation time-series. The existing literature does in this context not always compare to Intensity-Duration baselines. We agree that a thorough comparison is of scientific interest. We argue however that this is beyond the scope of this machine learning-based manuscript which aims at motivating the geoscientific community to employ interpretable machine learning models rather than well-established and hard to interpret models such as the commonly used LSTM and Random Forest based approaches. As we agree that this point is important and of interest to the scientific community, we will address the reviewers' concerns in the discussion and outlook sections of the revised manuscript.