Landslides (2021) 18:3135–3147
DOI 10.1007/s10346-021-01704-7
Received: 9 November 2020
Accepted: 14 May 2021
Published online: 27 May 2021

David J. Peres · Antonino Cancelliere

Comparing methods for determining landslide early warning thresholds: potential use of non-triggering rainfall for locations with scarce landslide data availability

Abstract Rainfall intensity-duration landslide-triggering thresholds have become widespread for the development of landslide early warning systems. Thresholds can be in principle determined using rainfall event datasets of three types: (a) rainfall events associated with landslides (triggering rainfall) only, (b) rainfall events not associated with landslides (non-triggering rainfall) only, (c) both triggering and non-triggering rainfall. In this paper, through Monte Carlo simulation, we compare these three possible approaches based on the following statistical properties: robustness, sampling variation, and performance. It is found that methods based only on triggering rainfall can be the worst with respect to those three investigated properties. Methods based on both triggering and non-triggering rainfall perform the best, as they could be built to provide the best trade-off between correct and wrong predictions; they are also robust, but still require a quite large sample to sufficiently limit the sampling variation of the threshold parameters. On the other side, methods based on non-triggering rainfall only, which are mostly overlooked in the literature, imply good robustness and low sampling variation, and performances that can often be acceptable and better than thresholds derived from only triggering events. To use solely triggering rainfall—which is the most common practice in the literature—yields to thresholds with the worse statistical properties, except when there is a clear separation between triggering and non-triggering events. Based on these results, it can be stated that methods based only on non-triggering rainfall deserve wider attention. Methods for threshold identification based on only non-triggering rainfall may have the practical advantage that can be in principle used where limited information on landslide occurrence is available (newly instrumented areas). The fact that relatively large samples (about 200 landslides events) are needed for a sufficiently precise estimation of threshold parameters when using triggering rainfall suggests that threshold determination in future applications may start from identifying thresholds from non-triggering events only, and then move to methods considering also the triggering events as landslide information starts to become more available.

Keywords Early warning systems · Intensity-duration model · Shallow
landslides · Robustness · Performance · Landslide initiation

Introduction

Landslides induced by rainfall are quite frequent in mountainous areas and represent a threat to property and population (Froude and Petley 2018). Landslide early warning systems (LEWS) are a mitigation measure that, if coupled with adequate emergency plans, can save lives (Piciullo et al. 2018; Guzzetti et al. 2020). The analysis of the conditions, in terms of precursor characteristics, for which landslides are foreseen is essential for LEWS. It is customary to consider only precipitation as a precursor and to derive rainfall thresholds based on two characteristics: rainfall event duration (D) and (mean) intensity (I) or cumulated depth, also known as event rainfall ($E = I \times D$) (Caine 1980; Guzzetti et al. 2008). Most of the works consider a power-law threshold, derived from the analysis of rainfall events associated to past landslide occurrence (Brunetti et al. 2010; Segoni et al. 2018). Some others, more recent, consider both triggering and non-triggering events, in order to have a complete assessment of the performances of the derived threshold in terms of the so-called confusion matrix, i.e., the count of true/false positives/negatives (Berti et al. 2012; Staley et al. 2013; Peres and Cancelliere 2014; Postance et al. 2018; Marino et al. 2020). A third possibility, that is indeed seldom considered, is to use the non-triggering events only. To the knowledge of the authors, this method has been discussed only in the gray literature (http://www.cswcs.org.tw/AllDataPos/DownLoadPos/Root3/1.3-Glade_121106-print.pdf, last accessed 16 April 2021) or just as a way to provide an upper bound of the uncertainty range related to threshold identification (Zêzere et al. 2015). This method may be particularly useful when an area is potentially prone to landslides but only a few cases have been recorded. Prone areas with little or no recorded landslide events may come out also as a result of susceptibility analysis (Reichenbach et al. 2018). In fact, a landslide may occur in an area where adequate rainfall records started to be collected only a few decades ago, covering a period where no other landslides have been likely recorded. In this case, the only available information on landslide occurrence is that, presumably, none of the preceding recorded rainfall events has triggered landslides. Still, in this case, protecting the area with an early warning system may be useful as a landslide mitigation measure and/or be requested by who has the responsibility to manage landslide risk (e.g., civil protection agencies).

In this paper, we analyze the three mentioned methods and compare them with respect to three statistical properties: robustness, sampling variation, and performance in terms of the confusion matrix. Robustness measures how the determination of threshold is affected by the presence of errors. To this end, we consider errors induced from exchanging triggering events with non-triggering ones and vice versa, as this may occur for the incompleteness of landslide inventories, or an inappropriate attribution of a given rainfall event to a landslide (Gariano et al. 2015). It should be mentioned that many other sources of uncertainty do exist in threshold determination which have been the subject of other studies (Nikolopoulos et al. 2014, 2015; Marra et al. 2016, 2017; Destro et al. 2017; Rossi et al. 2017; Peres et al. 2018; Marra

2019; Gariano et al. 2020), but that the illustrated erroneous attribution may be considered not quantified so far.

Sampling variation here refers to the fact that threshold parameters are usually estimated from a limited sample and that the true values of the threshold parameters are unknown in the real world. Monte Carlo simulation allows simulation of repeated sampling of a given sampling size and to compute the sampling distribution. The standard deviation of the sampling distribution, known as the standard error, can be then plotted as a function of sample size, to investigate how sensitive a threshold is to the given sample available. In other terms, sampling variation provides the precision of a given estimation of the threshold parameters for a given sample size. This may help answer the question: how many landslide-triggering storms constitute a statistically significant sample for threshold determination?

Regarding performances, it is customary to use the confusion matrix or receiver operating characteristics (ROC) analysis, i.e., to evaluate the quality of the predictions related to a threshold in terms of the corresponding frequency of false/true positives/negatives, often summarized by metrics such as the true skill statistic (Peirce 1884).

Our analysis is based on the use of synthetic rainfall-landslide data generation through the Monte Carlo simulation approach, combining stochastic generation of rainfall events with physically based infiltration and slope stability models, presented in our previous work (Peres and Cancelliere 2014, 2018), calibrated and validated for typical hillslopes prone to landslides in the Peloritani Mountains area, Sicily, Italy.

Methods and data

Threshold estimation methods

We refer to the case of power-law intensity-duration thresholds:

$$I = aD^b \tag{1}$$

where I and D are rainfall event intensity and duration, respectively, while a and b are the two parameters to be estimated based on the data. In our scheme, as usual, we consider that an alarm for possible landslide occurrence is issued when $I \geq aD^b$.

The methods for estimation of parameters that we compare are the following (Fig. 1):

1. Frequentist P, i.e., the so-called *frequentist* method (Brunetti et al. 2010; Peruccacci et al. 2012) applied to rainfall events associated to landslides (triggering rainfall only—“positives”)
2. Frequentist N, the frequentist method applied to rainfall events not associated to landslides (non-triggering rainfall only— “negatives”)
3. Frequentist PN, namely methods using both triggering and non-triggering events (positives and negatives) and an optimization procedure applied to an indicator based on the confusion matrix

The Frequentist P method is widely described and used in previous works. It is the most commonly adopted method for threshold identification (Piciullo et al. 2017) and software has been made available to apply it to any suitable case study (Melillo et al. 2018). Here, we recall that it basically consists in applying the following procedure: (i) linear regression of the data in the double-logarithmic plane, where only triggering events are considered, and (ii) parallel translation of the regression line such that the probability of non-exceedance is a fixed (low) value. In performing the analysis, the residuals of the regression are considered to be normally distributed with zero mean (cf. Fig. 1). In the case of Frequentist P, it is customary to choose a 5% of probability of non-exceedance by triggering events. Conversely, for the Frequentist N, we consider the choice of a threshold with 5% of probability of exceedance by non-triggering events. For the third method, we search the threshold that maximizes a given function of the confusion matrix, representing the trade-off between correct and wrong predictions. For the purpose, we apply the particle swarm optimization algorithm implemented within MATLAB®, so to find the global maximum of the true skill statistic (Wilks 2006):

$$T_S = F\left(I > aD^b|P\right) - F\left(I > aD^b|N\right) \tag{2}$$

where $F(I > aD^b| P)$ is the true positive rate or ROC sensitivity, an estimation of the conditional probability to predict a landslide (a positive, P) given that it occurred, and $F(I > aD^b| N)$, the false positive rate, an estimation of the conditional probability that a landslide is predicted given that no landslide (a negative, N) has been observed. For a threshold that perfectly separates triggering and non-triggering rainfall events, $T_S = 1$, while $T_S = 0$ for a threshold performing no better than a random guess. Indeed, the three illustrated methods can all be framed in terms of the last one (Frequentist PN), as special cases where only a column of the confusion matrix is taken into account (see Fig. 2). This consideration provides already some insights on the possible performances of the methods: Frequentist P aims at obtaining a fixed (low) probability of false negatives (missed alarms), Frequentist N aims at obtaining a fixed (low) probability of false positives (false alarms). Hence, there is no control of the false positives (false alarms) and of the false negatives (missed alarms), respectively; so this research aims also at understanding which possible tendencies can be derived for the part of the confusion matrix that is not explicitly taken into account (see the “Results and discussion” section). On the other side, by construction, the threshold from Frequentist PN, built using all elements of the confusion matrix, allows taking into account the trade-off between false and true predictions. Another aspect that is noteworthy to be explicitly mentioned is that the number of negatives (non-triggering rainfall events) is generally of one order of magnitude greater than the number of positives (see later in Table 2). This has impacts on the investigated properties of the three analyzed methods.

Assessment of statistical properties

Triggering and non-triggering events are generated via the Monte Carlo simulation approach that we developed in our previous work, as cited in the “Introduction”—some more details are recalled in Section “Synthetic data generation”. What is relevant here is that Monte Carlo simulation allows generating a virtual sequence of rainfall events as long as desired, and then to dichotomize the events into triggering and non-triggering based on slope stability analysis. Then, the generated events can be analyzed according to the three methodologies illustrated in the previous section.

For each sample size M (from 2000 to 100,000 rainfall events), $n = 30$ realizations are drawn, and thresholds are determined for each sample, so to explore sampling variation. This variation consists in the dispersion of threshold parameter estimates as obtained from different samples of the same size (Everitt and Skrondal 2010). Then, sampling variation is visualized by plotting the thirty thresholds for each sample size, as well as the normalized standard errors of threshold intercept and slope as functions of sample size. The normalization is made with respect to the reference threshold parameter values obtained using the entire dataset of 3,000,000 generated rainfall events (see the "Synthetic data generation" section).

For what concerns robustness, it must be acknowledged that somewhat different meanings are used for this concept. Hence, it may be useful to recall a standard definition. According to the Oxford dictionary of Statistics (Upton and Cook 2008), robustness measures how "the outcome of a statistical procedure is affected by the presence of a small number of unusual or incorrect values." Here, to measure robustness, we consider a "polluted" dataset where the p % of the triggering rainfall events is randomly assigned to non-triggering events. In this way, we try to mimic the effect of missing landslide events in inventories. At the same time, an identical number of non-triggering rainfall events is randomly assigned to triggering events, to emulate inappropriate attribution of rainfall events to a landslide, which may occur mainly because the nearest rain gauge available records a rainfall event that differs from the one in the known location of the landslide, or for an imprecise knowledge of the triggering instant. Here, we consider the case of $p = 10\%$. This scheme keeps unvaried the number of triggering and non-triggering events for a given simulation. For the sake of clarity, it may be noteworthy to mention that it would have been unrealistic to take the p % of non-triggering rainfall events and attribute it to landslide events, as the number of the erroneous triggering rainfall events would generally exceed the number of true landslide events (cf. end of the "Threshold estimation methods" section and later Table 2).

Regarding performance of thresholds, as discussed again in the "Threshold estimation methods" section, some entries of the confusion matrix are assigned by construction from the methods (with respect to their own calibration sample). Nevertheless, for Frequentist P and Frequentist N methods, the false positive rates and true positive rates are respectively unassigned, so it is still useful to plot the true skill statistic T_S, which can be used as a single metric of performance combining the four entries of the confusion matrix. For a more rigorous assessment of this aspect, the T_S is computed using the thresholds estimated on each sample size against the reference complete dataset of 3,000,000 rainfall events. In fact, the value of the T_S computed in this way—given that the scatter plot of both triggering and non-triggering events remains the same—is a measure of how the threshold is placed within the I-D plane. This is particularly useful as a complement to the information given by the variation of the standard error of threshold parameters with sample size. It is known that two thresholds having different a and b may instead be located quite nearly in the I-D plane, as a lower value of a can be compensated by a lower value of b (less steep). Also, due to the logarithmic scale, thresholds with the intercept parameter that differ quite significantly can be instead very close in the log(I)-log(D) plane (see for instance Fig. 4 and Table 2, PN vs. P threshold for the case of $\tau = 0$).

Synthetic data generation

The stochastic rainfall and the infiltration and slope stability model are calibrated based on climate, geomorphological, hydrological, and geotechnical characteristics of hillslopes in the Peloritani Mountains area, Sicily, Italy (Fig. 3). The area has been described in many other studies (De Guidi and Scudero 2013; Schilirò et al. 2015; Stancanelli et al. 2017; Cama et al. 2017), and a detailed description on how the models have been calibrated and validated is presented in Peres and Cancelliere (2018). Hence, we illustrate only briefly the approach for data generation—the reader is referred to the cited study for more details.

The stochastic model generates rainfall events that have a duration D and a constant intensity $I = E/D$. Total event depth E and duration D are generated through a bivariate distribution obtained via a copula approach, while storm interarrivals U are generated through a separate independent probability distribution. Generated rainfall events are separated by at least 24 h of no rain. The generated rainfall data are inputted to the infiltration model, which is the extension of the Iverson (2000) diffusive model for a finite soil depth, based on a linearization of Richards' vertical infiltration equation. This solution is implemented within TRIGRS v.1 (Baum et al. 2010). We have made our own MATLAB® code to better couple the stochastic model with the infiltration model. The infiltration model provides the pressure head to compute the factor of safety according to the infinite slope equation. The initial conditions to each rainfall event are computed by a drainage model that takes the maximum pressure head computed for the previous rainfall event and makes it decay according to a negative exponential law, with time constant τ, resembling the linear reservoir model for sub-horizontal drainage. More complex models could have been applied, such as those that consider variably saturated conditions during infiltration. For instance, this is done by the more recent TRIGRS v.2 software and its parallel implementation (Alvioli and Baum 2016; Palazzolo et al. 2021). However, the diffusive model based on TRIGRS v.1 which we apply here has proven sufficiently reliable for many cases, including the area considered in this study. This allowed to run the high number of simulations required in this study in a reasonable computation time.

The rainfall stochastic model has been calibrated on the hourly rainfall data observed at Fiumedinisi for a period of about 9 years (from 5 January 2002 to 23 February 2011). This rain gauge is near the Giampilieri area where on 1 October 2009 a slide-to-flow regional event was occurred, killing 37 persons. Rainfall model calibration has been performed separately for three statistically homogeneous periods of the year (JFMA, MJJA, SOND), leading to the distributions and parameters reported in Table 2 of Peres and Cancelliere (2018). The average of the number of rainfall events per year is 45.36, which conversely means that 1000 events correspond to about 22 years of recordings.

Regarding the hydrologic and geomechanical properties for the landslide model, these are shown in Table 1.

The value of τ for the drainage model can also be varied, so to speculate on the possible relevance of antecedent rainfall memory

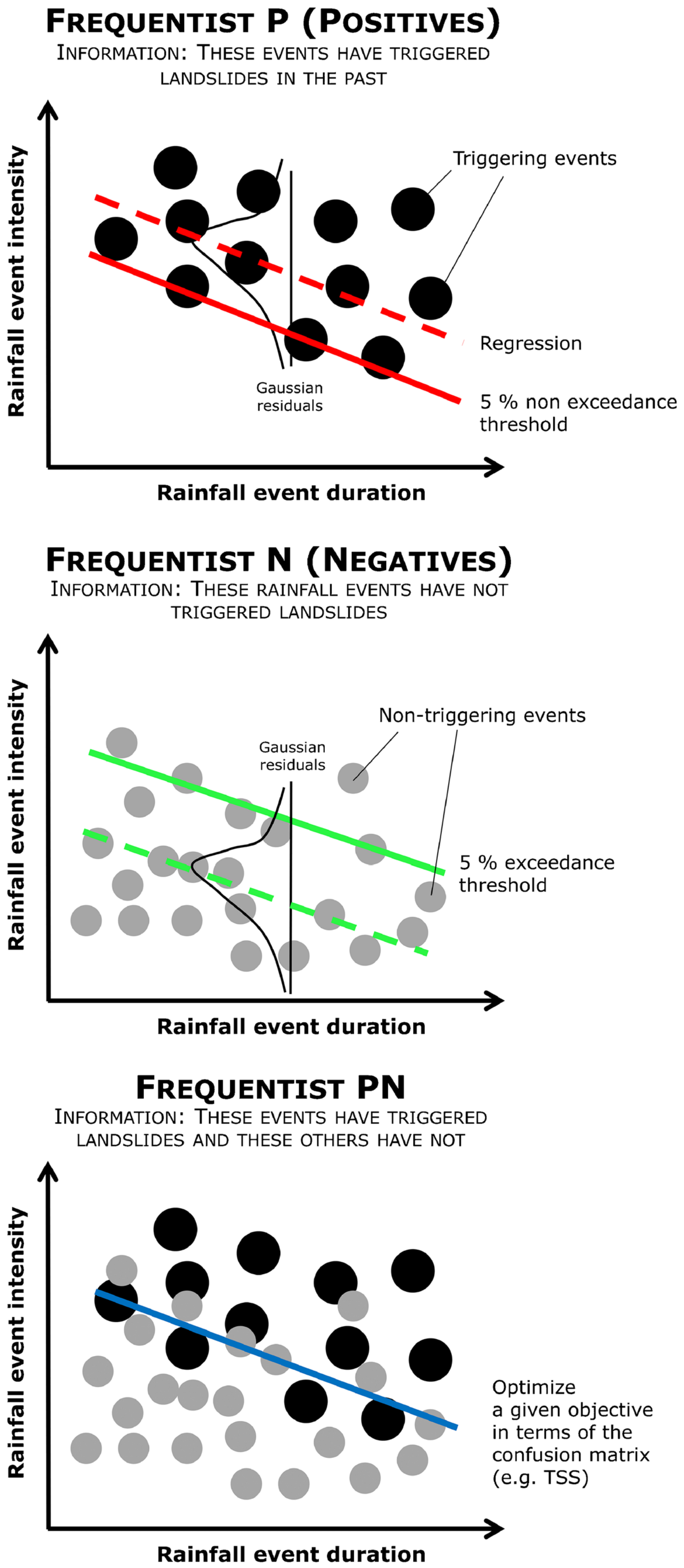
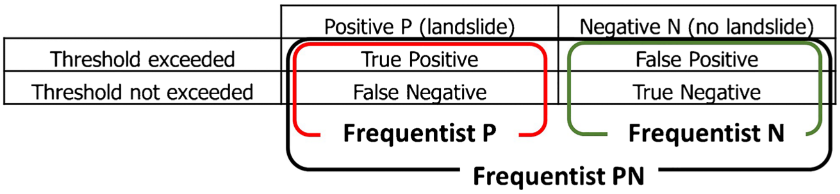

Frequentist P (Positives)
Information: these events have triggered landslides in the past
Triggering events
Regression
5 % non exceedance threshold
Gaussian residuals
Rainfall event intensity
Rainfall event duration
Frequentist N (Negatives)
Information: these rainfall events have not triggered landslides
Non-triggering events
Gaussian residuals
5 % exceedance threshold
Rainfall event intensity
Rainfall event duration
Frequentist PN
Information: these events have triggered landslides and these others have not
Optimize a given objective in terms of the confusion matrix (e.g. TSS)
Rainfall event intensity
Rainfall event duration

Fig. 1 Illustration of the three possible methods for determining landslide-triggering thresholds



	Positive P (landslide)	Negative N (no landslide)
Threshold exceeded	True Positive	False Positive
Threshold not exceeded	False Negative	True Negative

Fig. 2 Links among the three considered methodologies in terms of the confusion matrix

of the hillslope. The smaller the τ, the faster the soil drains during dry intervals between rainfall events, and the lesser antecedent rainfall is important for landslide occurrence. This means that in the limit case of τ = 0, antecedent rainfall has no importance and *I* and *D* are the only factors influencing landslide triggering (ideal situation). We have considered three cases $\tau = 0$, $\tau = \tau_0 = 2.75$ days and $\tau = 2\tau_0 = 5.49$ days, performing a sensitivity analysis with respect to the *I-D* model structural uncertainty—as this parameter controls the degree of confusion between triggering and non-triggering rainfall events described in terms of the considered predictors *I* and *D*. In the case of $\tau = 0$, there is a curve that can perfectly separate triggering and non-triggering events, though this differs slightly from a power-law (straight line in the log*D*-log*I* plane). Figure 4 compares the thresholds for a dataset of 3,000,000 rainfall events, which are assumed as the true thresholds for each method, useful for computing the normalized standard error of threshold parameters varying sample size (cf. "Methods and data" section). Table 2 shows, for each value of the recession constant τ, the ratio between the number of landslides and the number of rainfall events, and the thresholds corresponding to the entire datasets.

Results and discussion

Main case: intermediate structural uncertainty (Peloritani Mountains)

We present the results relative to the case of $\tau = \tau_0$, which represents the intermediate level of structural uncertainty, and the value that can be deemed representative for the case study of the Peloritani mountains. From Fig. 5, it can be seen that the Frequentist P method is the least robust. Errors tend to bring to an underestimation of the threshold parameters that would further increase the already high false alarm rates typical of this method. The Frequentist PN is more robust, but less than the Frequentist N. This happens because the former method is still affected by the errors in triggering rainfall data, while the latter is not significantly affected, as the impact of errors on non-triggering rainfall is lower in a relative sense. The same figure provides also some insight on sampling variation (width of the area spanned by the black lines). The low sampling variation of the Frequentist N is evident. However, for a better comparison of the Frequentist P and PN methods, it remains still useful to see how single threshold parameter estimation is affected by sample size (Fig. 6). As can be seen from the plots of normalized standard error varying sample size, for both parameters the Frequentist PN is generally way less variable than Frequentist N. Starting from a sample size of 30,000 rainfall events, both methods have very close normalized standard errors for the intercept parameter.

Regarding the performances, as already mentioned, the PN method performs the best, by construction (Fig. 7). What is interesting to see is that the Frequentist N could tendentially provide a performance closer to the best trade-off between correct and wrong predictions than the Frequentist P. Indeed, it may be true that this last method is cautionary as it reduces false negatives (potentially missed alarms) as much as possible, but it is also true that the false positive rate is too high to a point that it can induce a distrust in the early warning system eventually built on a so-determined threshold (cry-wolf syndrome). This plot gives also additional information on sampling variation of the thresholds;

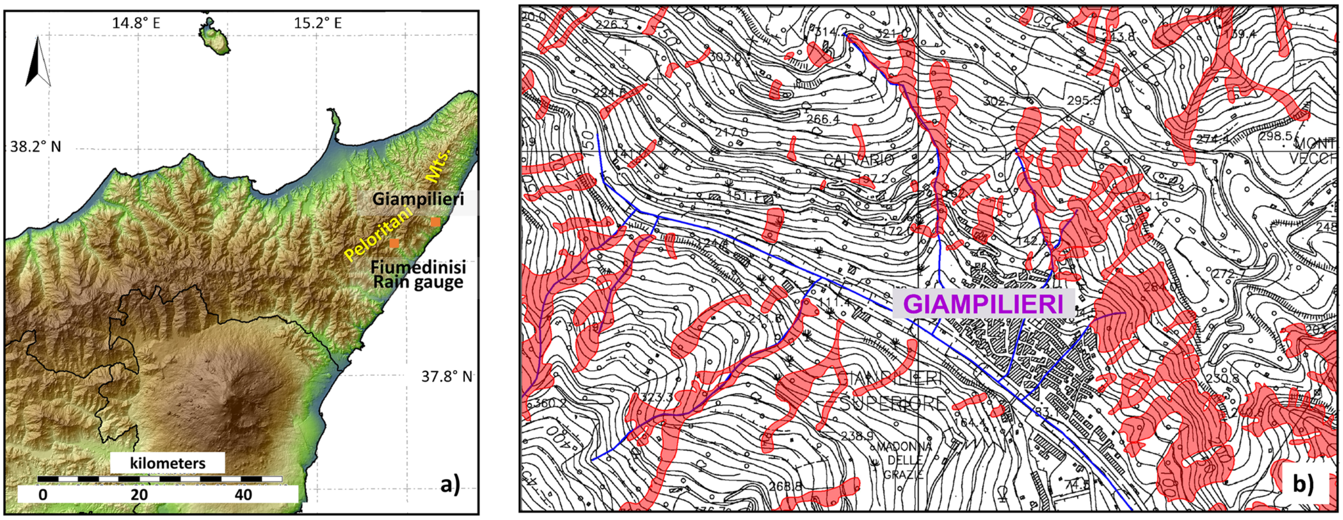


Fig. 3 Maps showing **a** the Peloritani Mountains area in Sicily, Italy, with the location of the rain gauge and of the town of Giampilieri and **b** the landslides occurring in the surroundings of the town on 1 October 2009 (red-colored areas)

Table 1 Hydrologic and geomechanical properties of a representative hillslope in the Peloritani Mountains area, Sicily, Italy

Property	Units	Value
Soil friction angle	°	39
Soil cohesion	kPa	4
Saturated soil unit weight	kN/m^3	19
Unit weight of water	kN/m^3	9.8
Soil saturated water content	-	0.35
Soil residual water content	-	0.045
Saturated hydraulic conductivity	m/s	2×10^{-5}
Saturated hydraulic diffusivity	m^2/s	5×10^{-5}
Soil depth	m	2
Terrain slope	°	40
Recession constant for drainage, τ	Days	2.75

with respect to Fig. 6, it reflects the variability of the position of the threshold taking into account both threshold parameters simultaneously. In fact, as already mentioned at the end of the "Assessment of statistical properties" section, for a given set of parameters, another threshold with a higher intercept may have a steeper (negative) slope that compensates the intercept and makes it quite similar to the given threshold. The T_S variability does not suffer from this issue, and reflects the actual position of the threshold in the *I-D* plane. It can be seen that performances get stable for the PN method starting from 15,000 rainfall events (about 100 landslide-triggering storms), while for the P method threshold position starts to reduce again for 30,000 rainfall events (about 200 landslide-triggering storms, cf. Table 2). These minimum data availability requisites—that, in addition, are to be deemed as referred to an area with homogenous landslide phenomenology—are not met by several past studies focussed on threshold determination: just to give an example, the Caine (1980) global threshold was based just on 73 landslide-triggering storms (non-triggering events were not considered).

Sensitivity analysis for different levels of structural uncertainty

The results presented vary with the level of structural uncertainty, which can be shown from analogous plots relative to $\tau_M = 0$, and $\tau = 2\tau_0$. In the first case (Figs. 8, 9, and 10), which is an "ideal" situation, it corresponds to a negligible structural uncertainty for the *I-D* model; both the Frequentist PN and Frequentist N methods are very robust, despite the Frequentist P, which is revealed again to be too sensitive to the presence of errors in the dataset (Fig. 8). These errors induce once more a very high underestimation of the threshold that may induce a dramatic increase in false positives. Regarding the sampling variation of the parameters, Frequentist PN and Frequentist P have a low sampling variation (Fig. 9). However, when looking at the performances (Fig. 10), clearly due to the low structural uncertainty of the *I-D* model in this case, the Frequentist P method performs well as the Frequentist PN, while the Frequentist N has slightly lower performances. This last method still has good performances, and represents a cautionary option as it does not uselessly leave any triggering events below the threshold (no false negatives), which in the case of the Frequentist P still remain, by construction, the 5% of total positives.

In the case of an increased level of structural uncertainty (recession constant $\tau = 2\tau_0$), robustness improves for both Frequentist P and PN methods (Fig. 11). This occurs when, as in this case, the impact of these errors may be comparable with the level of structural uncertainty. On the other side, and differently from the other two cases analyzed above, sampling variation of the threshold intercept for the Frequentist P method is lower than that for the Frequentist PN, while for the threshold slope, normalized standard errors have similar values for these two methods (Fig. 12). Nevertheless, when looking at the plot of the true skill statistic vs. sample size (Fig. 13), it can be seen that for the PN method, the dispersion of the T_S is way lower than that for the P method. This again means that, while single parameters may vary widely for the PN method, the position of the thresholds remains quite stable in this case, and thus again sampling variation of the threshold as a whole is dramatically lower for the PN and N methods with respect to the P method. Looking at the same plot, it can be seen that performances of this last method are dramatically lower than those of the other two methods.

Conclusions

In this paper, we have analyzed some relevant statistical properties of rainfall intensity-duration thresholds for landslide early warning determined by three different methods, i.e., based on the analysis of (i) triggering rainfall events only (Frequentist P), (ii)

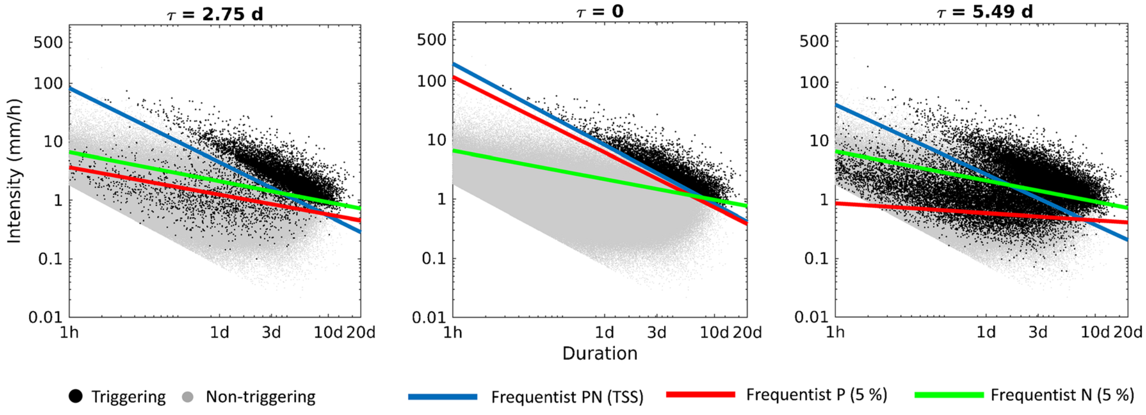


Fig. 4 Reference thresholds obtained using the whole virtual dataset of 3,000,000 rainfall events, for three values of the recession constant τ

Table 2 Number of landslides per thousand rainfall events (P/(P + N) rate) and reference thresholds for the three methods, varying the value of the recession constant τ

τ (days)	P/(P+N) rate ‰	PN threshold	P threshold	N threshold
2.75	7.7	$I = 82.86\ D^{-0.92}$	$I = 3.65\ D^{-0.34}$	$I = 6.61\ D^{-0.36}$
0	5.6	$I = 197.80\ D^{-1.00}$	$I = 118.23\ D^{-0.93}$	$I = 6.61\ D^{-0.36}$
5.49	13.5	$I = 41.42\ D^{-0.86}$	$I = 0.86\ D^{-0.12}$	$I = 6.61\ D^{-0.36}$

non-triggering rainfall events only (Frequentist N), and (iii) both triggering and non-triggering events (Frequentist PN). The first approach is the most commonly applied in the literature, while the last method has been mentioned only by a few scholars, mainly just as a way to provide an upper bound of the uncertainty range related to threshold identification. The third approach has been applied in more recent studies, and it presents the advantage of allowing the estimation of all entries of the confusion matrix, and thus to optimize performances in terms of both true positive and false positive ratios, which makes this approach the one that performs best in terms of the skill in separating triggering from non-triggering events. The results show that it is also a quite robust method. On the other side, Frequentist P method is poorly robust, can potentially lead to a high number of false alarms, and seems quite unstable with respect to the specific sample at hand. For a reasonable application of the Frequentist P method, at least 200 landslide-triggering storms should be available for a given geomorphologically homogeneous area, as otherwise the estimation of threshold parameters would be too imprecise (high sampling variation). For the frequentist PN method, the position of the threshold can be deemed stable starting from a sample with one hundred triggering events (half than that for the Frequentist P). Regarding the Frequentist N method, the performed analysis shows that it is the most robust and least variable. It also seems to have a greater potential than the Frequentist P method in providing performances that are a better compromise between false alarms and missed alarms. All these considerations lead to the conclusion that Frequentist N has not received the attention that it must deserve as a method for determining thresholds in its own right. A practical implication of the good properties of the Frequentist N method is that it could be a valid method to derive rainfall thresholds in locations where only few landslides have been recorded. Another point is that given the low robustness of the Frequentist P method, very low data points in the *I-D* plane

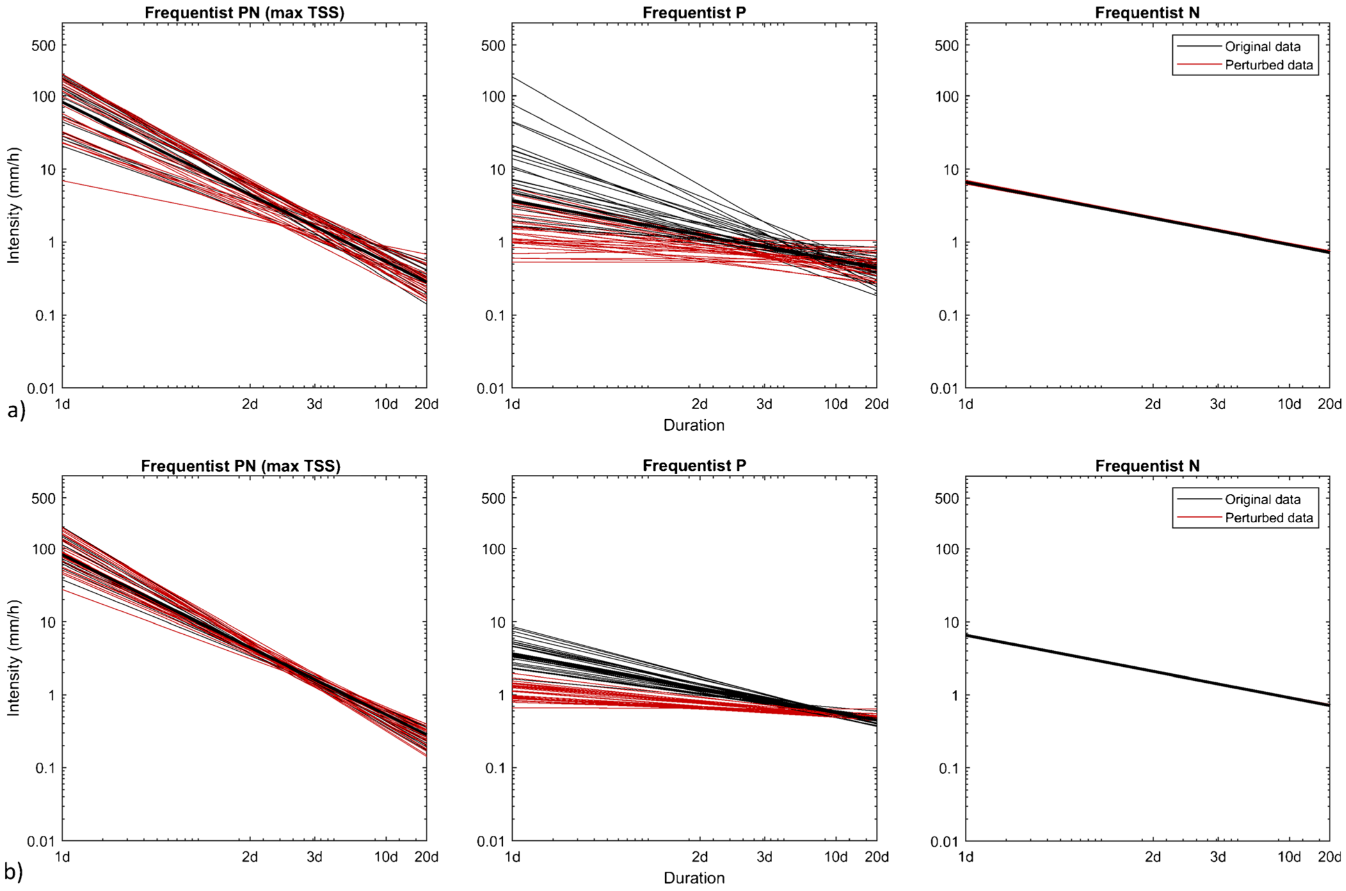


Fig. 5 Robustness test: for the three methods investigated, black lines represent thresholds for the error-free dataset while the red lines represent those derived from a polluted dataset ($p = 10\%$ of triggering events are exchanged with non-triggering, and vice versa): sample sizes 5000 (**a**) and 50,000 (**b**). The different thresholds for the Frequentist N method vary little and are not visible in the plot

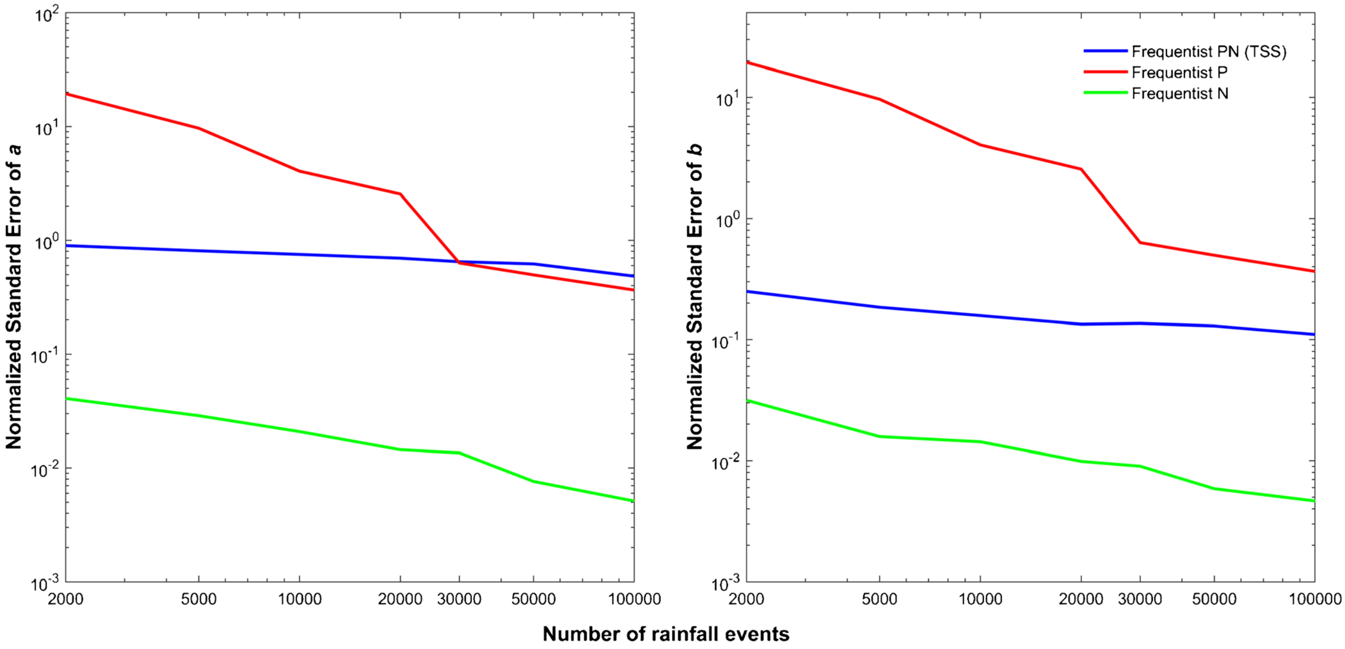


Fig. 6 Variability of threshold parameters against the sample size in terms of the normalized standard error. In this case (recession constant $\tau = \tau_0 = 2.75$ days), there are 7.7 landslide events per 1000 rainfall events

must be carefully checked for errors and thus possibly removed. If they are not errors in the dataset, the presence of low-intensity triggering points is a sign that only intensity and duration may not be sufficient to obtain a threshold with the necessary reliability, and that approaches taking into account predisposing hydrological factors must be investigated (Bogaard and Greco 2016, 2018; Thomas et al. 2019, 2020; Marino et al. 2020). As an overall conclusion, the analysis we have carried out suggests that future studies for threshold determination for newly investigated areas may start using the Frequentist N method first, and then move to the Frequentist PN as more data becomes available, as this method delivers globally the best of the techniques explored herein.

Acknowledgements

The handling editor, two anonymous referees, and Ben Mirus (USGS) provided constructive comments on earlier versions of this work. The research has been partially conducted within the following projects: BLUE-GREEN funded through the program PIACERI 2020-2022 of University of Catania, LIFE 17 CCA/IT/000115 SimetoRES funded by the EASME (now CINEA) of the

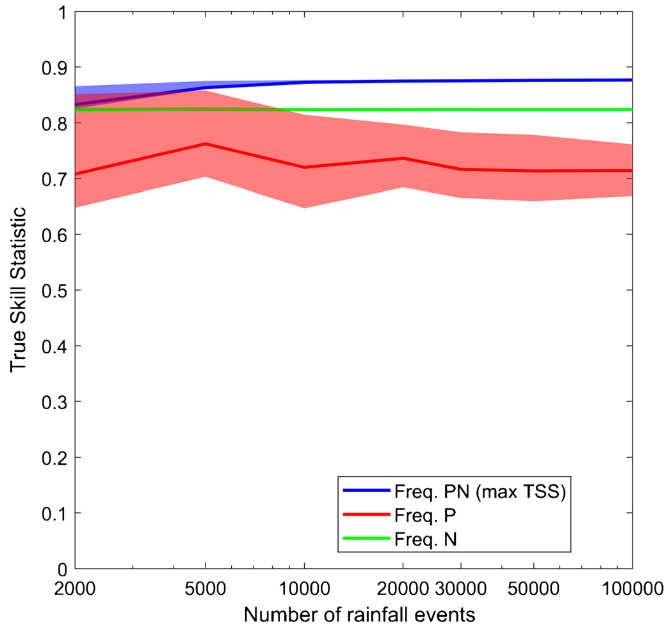


Fig. 7 True skill statistic of partial datasets varying sample size. The central lines represent the average of the T_S within the 30 samples of each size, while the bands are the interquartile ranges. In this case (recession constant $\tau = \tau_0 = 2.75$ days), there are 7.7 landslide events per 1000 rainfall events

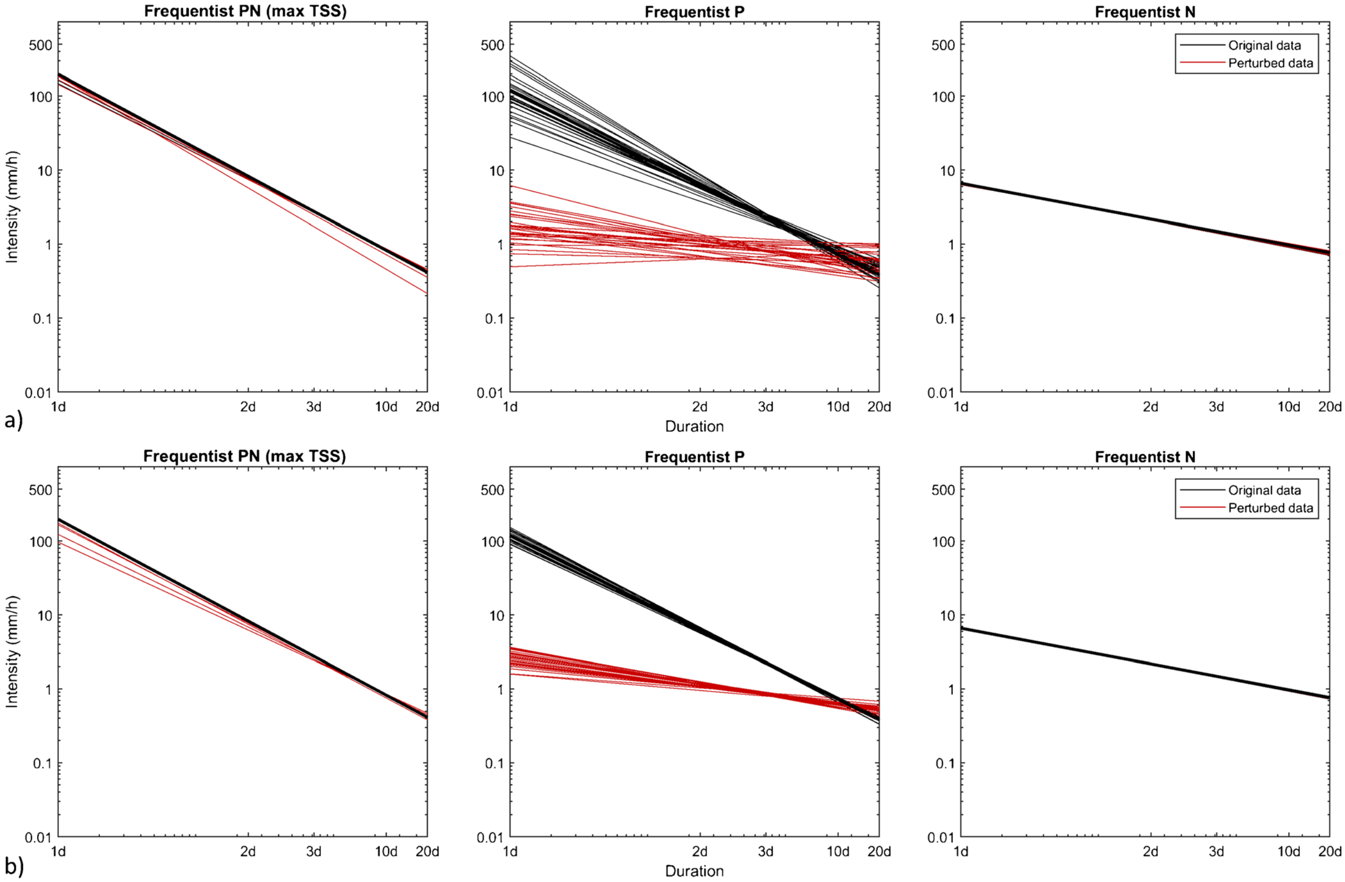


Fig. 8 Robustness test for the case of negligible structural uncertainty (recession constant τ = 0): for the three methods investigated, black lines represent thresholds for the error-free dataset while the red lines represent those derived from a polluted dataset (p = 10% of triggering events are exchanged with non-triggering, and vice versa): sample sizes 5000 (**a**) and 50,000 (**b**). The different thresholds for the Frequentist N method vary little and are not visible in the plot

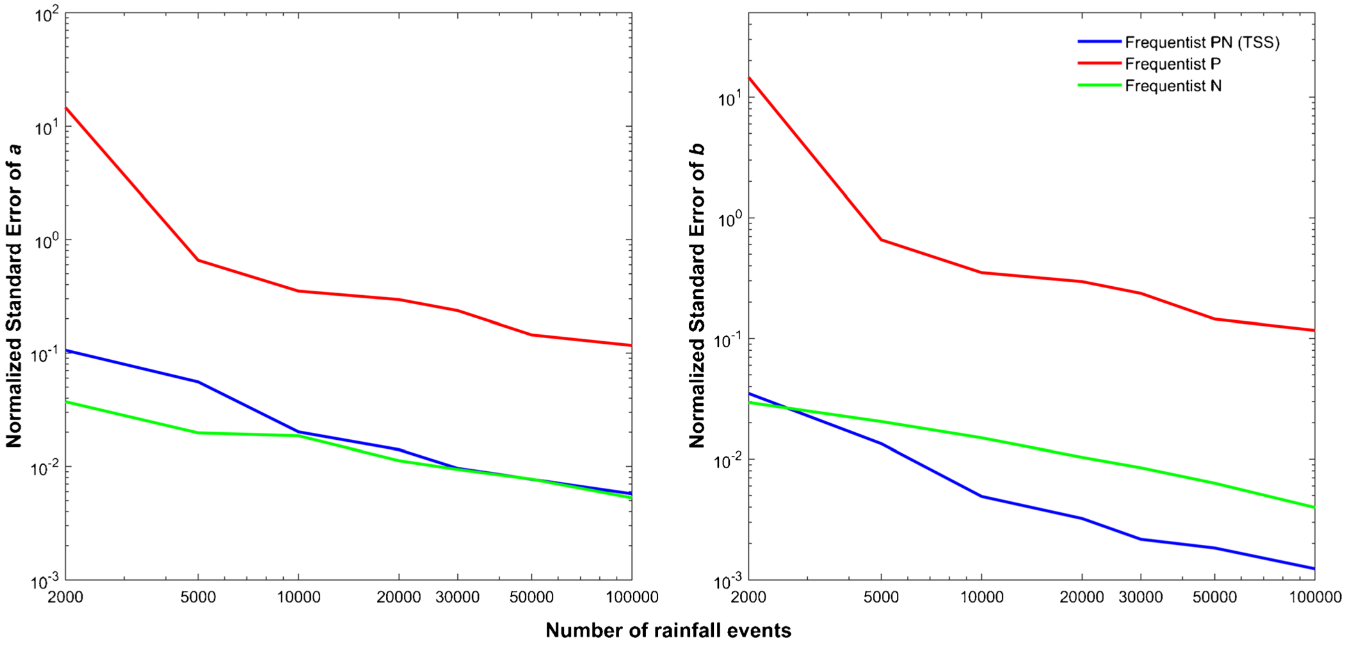


Fig. 9 Variability of threshold parameters against sample size in terms of the normalized standard error, for the case of negligible structural uncertainty (recession constant τ = 0). In this case, there are 5.6 landslide events per 1000 rainfall events

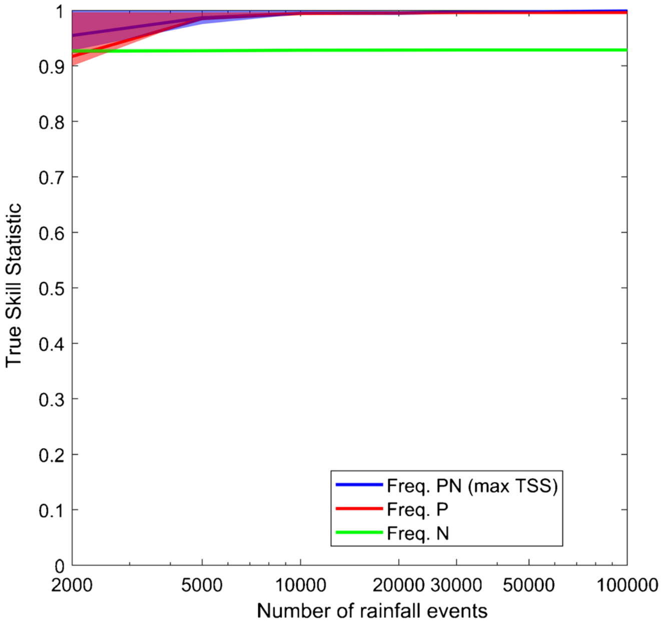


Fig. 10 True skill statistic of partial datasets varying sample size. The central lines represent the average of the T_S within the 30 samples of each size, while the bands are the interquartile ranges. Case of negligible structural uncertainty (recession constant τ = 0). In this case, there are 5.6 landslide events per 1000 rainfall events

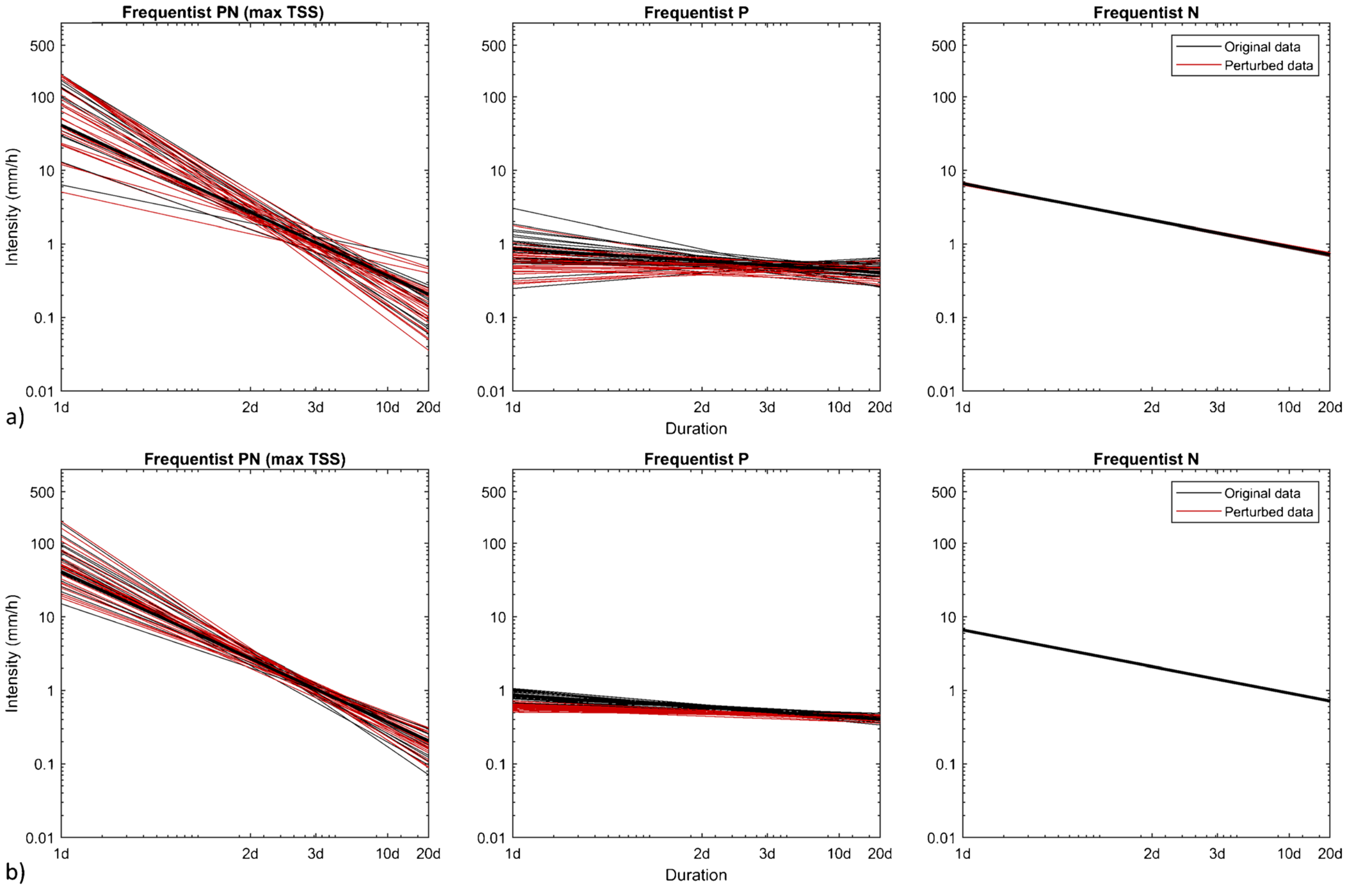


Fig. 11 Robustness test for the case of high structural uncertainty (recession constant τ = 5.49 days—double respect to the reference case): for the three methods investigated, black lines represent thresholds for the error-free dataset while the red lines represent those derived from a polluted dataset (p = 10% of triggering events are exchanged with non-triggering, and vice versa): sample sizes 5000 (**a**) and 50,000 (**b**). The different thresholds for the Frequentist N method vary little and are not visible in the plot

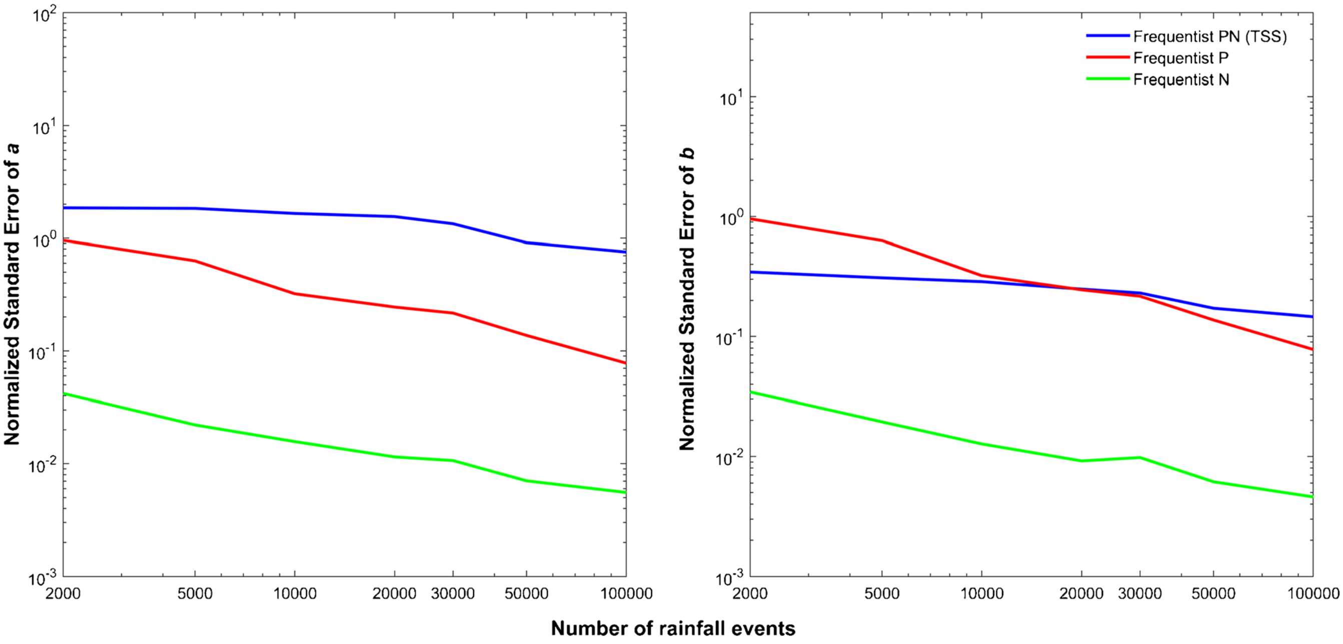


Fig. 12 Variability of threshold parameters against sample size in terms of the normalized standard error, for the case of high structural uncertainty (recession constant τ = 5.49 days). In this case, there are 13.5 landslide events per 1000 rainfall events

European Commission, and the Programma Operativo Nazionale Governance e Capacità Istituzionale 2014-2020 - Programma per il supporto al rafforzamento della governance in materia di riduzione del rischio ai fini di protezione civile CUP (Program to support the strengthening of governance in the field of risk reduction for civil protection purposes). David J. Peres was supported by the post-doctoral grant on "Sviluppo di modelli per la valutazione di strategie innovative di gestione delle risorse idriche in un contesto di cambiamenti climatici" (Development of models for the evaluation of new strategies for water resources management in a changing climate).

Availability of data and material Rainfall data used in this paper can be downloaded by request at http://www.sias.regione.sicilia.it/.

Code availability By request to authors.

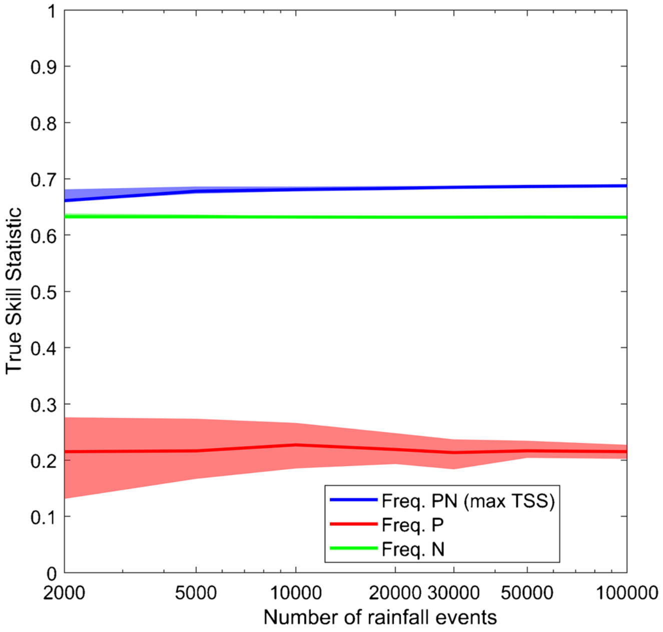


Fig. 13 True skill statistic of partial datasets varying sample size. The central lines represent the average of the T_S within the 30 samples of each size, while the bands are the interquartile ranges. Case of high structural uncertainty (recession constant τ = 5.49 days). In this case, there are 13.5 landslide events per 1000 rainfall events

Funding

Open access funding provided by Università degli Studi di Catania within the CRUI-CARE Agreement.

Declarations

Conflict of interest The authors declare no competing interests.

References

Alvioli M, Baum RL (2016) Parallelization of the TRIGRS model for rainfall-induced landslides using the message passing interface. Environ Model Softw 81:122–135. https://doi.org/10.1016/J.ENVSOFT.2016.04.002

Baum RL, Godt JW, Savage WZ (2010) Estimating the timing and location of shallow rainfall-induced landslides using a model for transient, unsaturated infiltration. J Geophys Res 115:F03013. https://doi.org/10.1029/2009JF001321

Berti M, Martina MLV, Franceschini S, Pignone S, Simoni A, Pizziolo M (2012) Probabilistic rainfall thresholds for landslide occurrence using a Bayesian approach. J Geophys Res Earth Surf 117:1–20. https://doi.org/10.1029/2012JF002367

Bogaard TA, Greco R (2016) Landslide hydrology: from hydrology to pore pressure. Wiley Interdiscip Rev Water 3:439–459. https://doi.org/10.1002/wat2.1126

Bogaard T, Greco R (2018) Invited perspectives: hydrological perspectives on precipitation intensity-duration thresholds for landslide initiation: proposing hydro-meteorological thresholds. Nat Hazards Earth Syst Sci 18:31–39. https://doi.org/10.5194/nhess-18-31-2018

Brunetti MT, Peruccacci S, Rossi M, Luciani S, Valigi D, Guzzetti F (2010) Rainfall thresholds for the possible occurrence of landslides in Italy. Nat Hazards Earth Syst Sci 10:447–458. https://doi.org/10.5194/nhess-10-447-2010

Caine N (1980) The rainfall intensity-duration control of shallow landslides and debris flows. Soc Swed Annaler Geografiska Geogr Phys 62:23–27

Cama M, Lombardo L, Conoscenti C, Rotigliano E (2017) Improving transferability strategies for debris flow susceptibility assessment: application to the Saponara and Itala catchments (Messina, Italy). Geomorphology 288:52–65. https://doi.org/10.1016/J.GEOMORPH.2017.03.025

De Guidi G, Scudero S (2013) Landslide susceptibility assessment in the Peloritani Mts. (Sicily, Italy) and clues for tectonic control of relief processes. Nat Hazards Earth Syst Sci 13:949–963. https://doi.org/10.5194/nhess-13-949-2013

Destro E, Marra F, Nikolopoulos EI, Zoccatelli D, Creutin JD, Borga M (2017) Spatial estimation of debris flows-triggering rainfall and its dependence on rainfall return period. Geomorphology 278:269–279. https://doi.org/10.1016/j.geomorph.2016.11.019

Everitt BS, Skrondal A (2010) Cambridge dictionary of Statistics, 4th edn. Cambridge University Press, Cambridge

Froude MJ, Petley DN (2018) Global fatal landslide occurrence from 2004 to 2016. Nat Hazards Earth Syst Sci 18:2161–2181. https://doi.org/10.5194/nhess-18-2161-2018

Gariano SL, Brunetti MT, Iovine G, Melillo M, Peruccacci S, Terranova O, Vennari C, Guzzetti F (2015) Calibration and validation of rainfall thresholds for shallow landslide forecasting in Sicily, southern Italy. Geomorphology 228:653–665. https://doi.org/10.1016/j.geomorph.2014.10.019

Gariano SL, Melillo M, Peruccacci S, Brunetti MT (2020) How much does the rainfall temporal resolution affect rainfall thresholds for landslide triggering? Nat Hazards 100:655–670. https://doi.org/10.1007/s11069-019-03830-x

Guzzetti F, Peruccacci S, Rossi M, Stark CP (2008) The rainfall intensity-duration control of shallow landslides and debris flows: an update. Landslides 5:3–17. https://doi.org/10.1007/s10346-007-0112-1

Guzzetti F, Gariano SL, Peruccacci S, Brunetti MT, Marchesini I, Rossi M, Melillo M (2020) Geographical landslide early warning systems. Earth Sci Rev 200:102973

Iverson RM (2000) Landslide triggering by rain infiltration. Water Resour Res 36:1897–1910. https://doi.org/10.1029/2000WR900090

Marino P, Peres DJ, Cancelliere A, Greco R, Bogaard TA (2020) Soil moisture information can improve shallow landslide forecasting using the hydrometeorological threshold approach. Landslides 17:2041–2054. https://doi.org/10.1007/s10346-020-01420-8

Marra F (2019) Rainfall thresholds for landslide occurrence: systematic underestimation using coarse temporal resolution data. Nat Hazards 95:883–890. https://doi.org/10.1007/s11069-018-3508-4

Marra F, Nikolopoulos EI, Creutin JD, Borga M (2016) Space-time organization of debris flows-triggering rainfall and its effect on the identification of the rainfall threshold relationship. J Hydrol 541:246–255. https://doi.org/10.1016/j.jhydrol.2015.10.010

Marra F, Destro E, Nikolopoulos EI, Zoccatelli D, Creutin JD, Guzzetti F, Borga M (2017) Impact of rainfall spatial aggregation on the identification of debris flow occurrence thresholds. Hydrol Earth Syst Sci 21:4525–4532. https://doi.org/10.5194/hess-21-4525-2017

Melillo M, Brunetti MT, Peruccacci S, Gariano SL, Roccati A, Guzzetti F (2018) A tool for the automatic calculation of rainfall thresholds for landslide occurrence. Environ Model Softw 105:230–243. https://doi.org/10.1016/J.ENVSOFT.2018.03.024

Nikolopoulos EI, Crema S, Marchi L, Marra F, Guzzetti F, Borga M (2014) Impact of uncertainty in rainfall estimation on the identification of rainfall thresholds for debris flow occurrence. Geomorphology 221:286–297. https://doi.org/10.1016/J.GEOMORPH.2014.06.015

Nikolopoulos EI, Borga M, Creutin JD, Marra F (2015) Estimation of debris flow triggering rainfall: influence of rain gauge density and interpolation methods. Geomorphology 243:40–50. https://doi.org/10.1016/j.geomorph.2015.04.028

Palazzolo N, Peres DJ, Bordoni M, Meisina C, Creaco E, Cancelliere A (2021) Improving spatial landslide prediction with 3D slope stability analysis and genetic algorithm optimization: application to the Oltrepò Pavese. Water 13:801. https://doi.org/10.3390/w13060801

Peirce CS (1884) The numerical measure of the success of predictions. Science 4:453–454

Peres DJ, Cancelliere A (2014) Derivation and evaluation of landslide-triggering thresholds by a Monte Carlo approach. Hydrol Earth Syst Sci 18:4913–4931. https://doi.org/10.5194/hess-18-4913-2014

Peres DJ, Cancelliere A (2018) Modeling impacts of climate change on return period of landslide triggering. J Hydrol 567:420–434. https://doi.org/10.1016/J.JHYDROL.2018.10.036

Peres DJ, Cancelliere A, Greco R, Bogaard TA (2018) Influence of uncertain identification of triggering rainfall on the assessment of landslide early warning thresholds. Nat Hazards Earth Syst Sci 18:633–646. https://doi.org/10.5194/nhess-18-633-2018

Peruccacci S, Brunetti MT, Luciani S, Vennari C, Guzzetti F (2012) Lithological and seasonal control on rainfall thresholds for the possible initiation of landslides in central Italy. Geomorphology 139–140:79–90. https://doi.org/10.1016/j.geomorph.2011.10.005

Piciullo L, Gariano SL, Melillo M, Brunetti MT, Peruccacci S, Guzzetti F, Calvello M (2017) Definition and performance of a threshold-based regional early warning model for rainfall-induced landslides. Landslides 14:995–1008. https://doi.org/10.1007/s10346-016-0750-2

Piciullo L, Calvello M, Cepeda JM (2018) Territorial early warning systems for rainfall-induced landslides. Earth Sci Rev 179:228–247. https://doi.org/10.1016/J.EARSCIREV.2018.02.013

Postance B, Hillier J, Dijkstra T, Dixon N (2018) Comparing threshold definition techniques for rainfall-induced landslides: a national assessment using radar rainfall. Earth Surf Process Landf 43:553–560. https://doi.org/10.1002/esp.4202

Reichenbach P, Rossi M, Malamud BD, Mihir M, Guzzetti F (2018) A review of statistically-based landslide susceptibility models. Earth Sci Rev 180:60–91. https://doi.org/10.1016/j.earscirev.2018.03.001

Rossi M, Luciani S, Valigi D, Kirschbaum D, Brunetti MT, Peruccacci S, Guzzetti F (2017) Statistical approaches for the definition of landslide rainfall thresholds and their uncertainty using rain gauge and satellite data. Geomorphology 285:16–27. https://doi.org/10.1016/j.geomorph.2017.02.001

Schilirò L, De Blasio FV, Esposito C, Scarascia Mugnozza G (2015) Reconstruction of a destructive debris-flow event via numerical modeling: the role of valley geometry on flow dynamics. Earth Surf Process Landf 40:1847–1861. https://doi.org/10.1002/esp.3762

Segoni S, Piciullo L, Gariano SL (2018) A review of the recent literature on rainfall thresholds for landslide occurrence. Landslides 15:1483–1501

Staley DM, Kean JW, Cannon SH, Schmidt KM, Laber JL (2013) Objective definition of rainfall intensity-duration thresholds for the initiation of post-fire debris flows in southern California. Landslides 10:547–562. https://doi.org/10.1007/s10346-012-0341-9

Stancanelli LM, Peres DJ, Cancelliere A, Foti E (2017) A combined triggering-propagation modeling approach for the assessment of rainfall induced debris flow susceptibility. J Hydrol 550:143. https://doi.org/10.1016/j.jhydrol.2017.04.038

Thomas MA, Collins BD, Mirus BB (2019) Assessing the feasibility of satellite-based thresholds for hydrologically driven landsliding. Water Resour Res 55:9023. https://doi.org/10.1029/2019WR025577

Thomas MA, Mirus BB, Smith JB (2020) Hillslopes in humid-tropical climates aren't always wet: implications for hydrologic response and landslide initiation in Puerto Rico. Hydrol Process 34:4307–4318. https://doi.org/10.1002/hyp.13885

Upton G, Cook I (2008) Dictionary of statistics - Oxford Reference, 2 revised. Oxford University Press, Oxford. https://doi.org/10.1093/acref/9780199541454.001.0001

Wilks DS (2006) Statistical methods in the atmospheric sciences, 2nd edn. Academic Press, London

Zêzere JL, Vaz T, Pereira S, Oliveira SC, Marques R, Garcia RAC (2015) Rainfall thresholds for landslide activity in Portugal: a state of the art. Environ Earth Sci 73:2917–2936. https://doi.org/10.1007/s12665-014-3672-0

D. J. Peres (✉) · **A. Cancelliere**
Department of Civil Engineering and Architecture,
University of Catania,
95123, Catania, Italy
Email: djperes@dica.unict.it