Natural Hazards
and Earth System
Sciences

Quantitative comparison between two different methodologies to define rainfall thresholds for landslide forecasting

D. Lagomarsino, S. Segoni, A. Rosi, G. Rossi, A. Battistini, F. Catani, and N. Casagli

Earth Science Department, University of Firenze, Firenze, Italy

Correspondence to: S. Segoni (samuele.segoni@unifi.it)

Abstract. This work proposes a methodology to compare the forecasting effectiveness of different rainfall threshold models for landslide forecasting. We tested our methodology with two state-of-the-art models, one using intensity–duration thresholds and the other based on cumulative rainfall thresholds.

The first model identifies rainfall intensity–duration thresholds by means of a software program called MaCumBA (MAssive CUMulative Brisk Analyzer) (Segoni et al., 2014a) that analyzes rain gauge records, extracts intensity (I) and duration (D) of the rainstorms associated with the initiation of landslides, plots these values on a diagram and identifies the thresholds that define the lower bounds of the I–D values. A back analysis using data from past events is used to identify the threshold conditions associated with the least number of false alarms.

The second model (SIGMA) (Sistema Integrato Gestione Monitoraggio Allerta) (Martelloni et al., 2012) is based on the hypothesis that anomalous or extreme values of accumulated rainfall are responsible for landslide triggering: the statistical distribution of the rainfall series is analyzed, and multiples of the standard deviation (σ) are used as thresholds to discriminate between ordinary and extraordinary rainfall events. The name of the model, SIGMA, reflects the central role of the standard deviations.

To perform a quantitative and objective comparison, these two models were applied in two different areas, each time performing a site-specific calibration against available rainfall and landslide data. For each application, a validation procedure was carried out on an independent data set and a confusion matrix was built. The results of the confusion matrixes were combined to define a series of indexes commonly used to evaluate model performances in natural hazard assessment. The comparison of these indexes allowed to identify the most effective model in each case study and, consequently, which threshold should be used in the local early warning system in order to obtain the best possible risk management.

In our application, none of the two models prevailed absolutely over the other, since each model performed better in a test site and worse in the other one, depending on the characteristics of the area.

We conclude that, even if state-of-the-art threshold models can be exported from a test site to another, their employment in local early warning systems should be carefully evaluated: the effectiveness of a threshold model depends on the test site characteristics (including the quality and quantity of the input data), and a validation procedure and a comparison with alternative models should be performed before its implementation in operational early warning systems.

1 Introduction

One of the most common methodologies for the forecasting of landslide occurrence is the definition of rainfall thresholds. A rainfall threshold is an equation (based on two or more rainfall parameters) that discriminates between the rainfall conditions for which one or more landslides would or would not be triggered.

Since the pioneering works of Endo (1970), Campbell (1975), Lumb (1975), Guidicini and Iwasa (1977) and Caine (1980), the rainfall threshold approach has achieved great success, and many thresholds have been proposed based on a large variety of rainfall parameters (an exhaustive review can be found in Guzzetti et al., 2007). The thresh-

olds based on intensity and duration are probably the most common (Caine (1980), Guzzetti et al. (2008) and references therein); another very used threshold typology makes use of rainfall amount accumulated over given time periods (Wilson, 2000; Chleborad, 2003; Cardinali et al., 2006; Cannon et al., 2008, 2011) or variable time windows (Lagomarsino et al., 2013).

Independently from the rainfall parameters used to characterize the triggering conditions, every study that made use of both rainfall events that triggered and that did not triggered landslides highlighted that it is impossible to perfectly divide the diagram into a 100 % landslide field and a 100 % non-landslide field (Berti et al., 2012; Staley et al., 2013). This brings the necessity of taking a fundamental conceptual decision when defining a threshold: a conservative threshold that would encompass all future landslides should be defined or the best trade-off between identified landslides and missed alarms should be researched? It does not exist a universally valid response, as the right answer depends on the objective of the threshold. Indeed, it is important to highlight that in the existing literature, some thresholds have been used to identify the minimum rainfall conditions possibly leading to landsliding, while others have been specifically designed to be operated in warning systems for civil protection purposes.

The first kind of threshold (minimum thresholds henceforth) is commonly defined as the lower bound to a data set of rainfall conditions that in the past were associated to landslide triggering (Caine, 1980; Larsen and Simon, 1993; Cannon et al., 2008; Brunetti et al., 2010; Berti et al., 2012): it is expected that any future landslide will fall above the thresholds. Since minimum thresholds are very conservative, a high number of false alarms is usually expected, because the lower the threshold, the lower the possibility of missing a landslide and the higher the possibility of committing false alarms.

The second kind of thresholds (early warning thresholds henceforth) usually aims to obtain the best possible compromise between effectiveness in recognizing triggering conditions (for which a low threshold would be preferable) and effectiveness in committing a low number of false alarms (for which a high threshold would be preferable) (Martelloni et al., 2012; Staley et al., 2013; Segoni et al., 2014a, 2015a). In other words, the task of a warning system is to avoid both missed alarms and false alarms as much as possible. Both kinds of errors are considered dangerous, as missed alarms may expose society to unrecognized hazards, while false alarms, especially when recurring, may lead to a misperception of risk and to a distrust in the warning system itself (Staley et al., 2013).

The errors committed by a threshold can be recognized and evaluated only after a validation procedure is carried out, but despite rainfall thresholds for the occurrence of landslides being a long-debated research topic, only a small number of works completes the presentation of a new threshold with a quantitative validation of its performances (Martelloni et al., 2012; Staley et al., 2013; Lagomarsino et al., 2013;

Segoni et al., 2014a, b; Gariano et al., 2015) or with a comparison with an independent data set of landslide and rainfall data (Giannecchini et al., 2012). This leads to an additional limitation when a comparison between different thresholds is needed. In fact, while many studies on rainfall thresholds contain a comparison between different literature thresholds (Guzzetti et al., 2007, 2008; Rosi et al., 2012; Chen and Wang, 2014), in most cases this is just a visual comparison of the threshold equations. This comparison is interesting from many scientific points of view (e.g., the influence of meteorological regime, landslide typology or other physical features on the threshold equations), but thresholds are very site-specific (Segoni et al., 2014b) and when a comparison is needed to decide which threshold should be used in a warning system, it is of limited usefulness to compare a threshold obtained using a given methodology in a test site with the threshold obtained using a different methodology in another test site. Moreover, a comparison would be more useful if it were based on quantitative indexes describing the performances of the thresholds.

This paper explores the aforementioned issues and proposes a quantitative approach for comparing different methodologies for rainfall threshold definition, in order to assess which of them is the most effective for operational use in civil protection warning systems.

Two state-of-the-art models based on rainfall thresholds, namely SIGMA (Martelloni et al., 2012; Lagomarsino et al., 2013) and MaCumBA (Segoni et al., 2014a, b), are taken into account and are applied in two test sites. In each test site, each model undergoes a site-specific calibration to optimize its performance. A validation procedure is carried out on an independent data set and a confusion matrix is built. The results of the four confusion matrixes (true positives, true negatives, false positives and false negatives) are combined to define some indexes commonly used to evaluate model performances in hazard assessment (Begueria, 2005) and in rainfall thresholds (Martelloni et al., 2012; Gariano et al., 2015; Rosi et al., 2015). The comparison of these indexes assessed which model provides the best performance in each case study and, consequently, which threshold should be used in the local early warning system in order to obtain the best possible risk management.

2 Material and methods

2.1 SIGMA

SIGMA is the model used to define the thresholds for the Emilia Romagna regional landslide early warning system. It is explained in detail in Martelloni et al. (2012) and it is based on the concept that landslides occur in case of rainfall events that can be considered exceptional for either the duration or the rainfall amount. Its main feature is a statistical analysis of historical rainfall series considering differ-
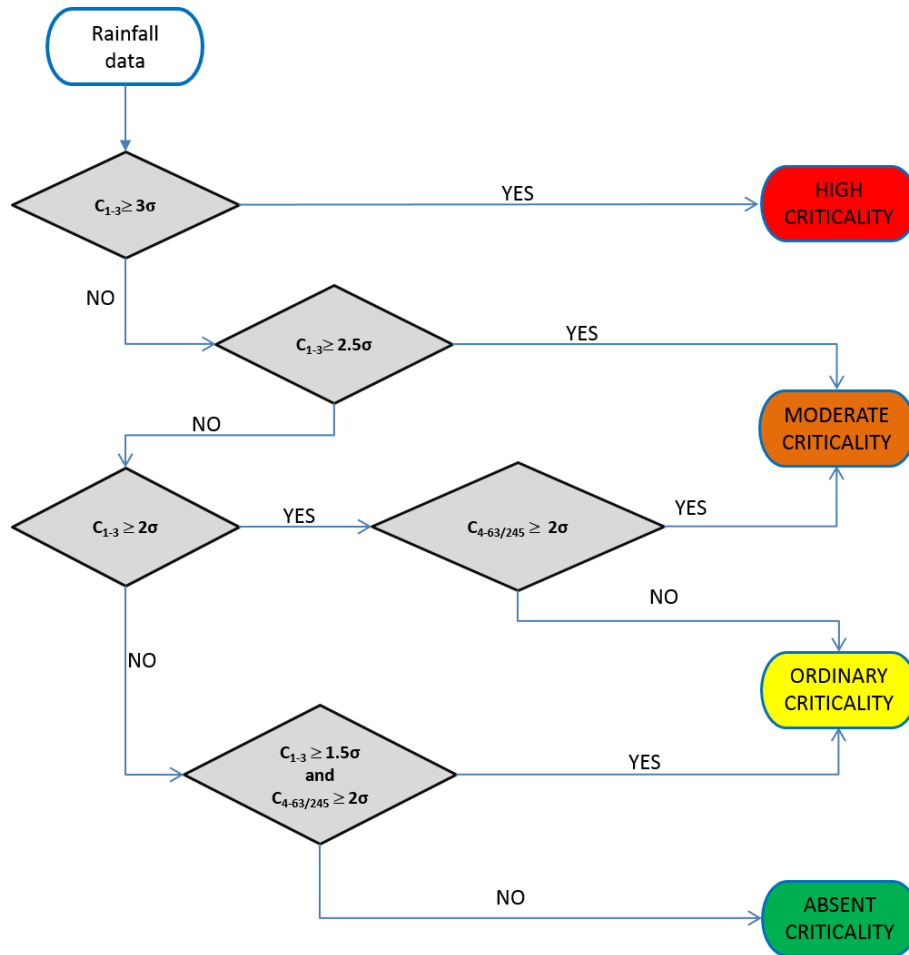
Figure 1. SIGMA algorithm (modified after Martelloni et al., 2012). C_{1-3} stands for the cumulate rainfall of the last 1, 2 or 3 days. $C_{4-63/245}$ stands for the rainfall values cumulated in the last 4 days, last 5 days and so on, up to the last 63 days during the dry season or 245 days during the wet season.

ent periods of accumulation: from 24 h up to 245 days, with daily step (Martelloni et al., 2012). These analyses allow the recognition of anomalous rain values, quantifying the value of the standard deviation of the distribution for each accumulation period. Considering different multiples of standard deviation, different thresholds are then defined (σ curves). An optimization algorithm compares the σ curves with the landslides contained in a calibration data set and identifies the σ curves that minimize the occurrence of false alarms (Martelloni et al., 2012). The selected σ curves are implemented in a warning system (named SIGMA, as the model) in which the measured and the forecasted rainfall is compared with these thresholds, according to the algorithm depicted in Fig. 1, to define the daily criticality level.

The entire territory of Emilia Romagna is subdivided into eight alert zones (AZ). For each of these, different rain gauges are selected, for a total of 25. Each rain gauge is representative of an area called the territorial unit (TU). The alerts calculated for each TU belonging to the same AZ are then combined to give a single alert for each AZ (Lagomarsino et al., 2013).

The Emilia Romagna regional early warning system is completed by a module that accounts for the effects of snowmelt and snow accumulation (Martelloni et al., 2012) and by a combination with purposely developed landslide susceptibility zonation that improves the spatial accuracy of the SIGMA model (Segoni et al., 2015b). However, these additional features are not considered in this work.

2.2 MaCumBA

MaCumBA is the model used to define the thresholds for the Tuscany regional landslide early warning system, which is based on intensity–duration thresholds expressed in the form (Caine, 1980):

$$I = \alpha D^\beta,$$

where I is the rainfall intensity (mm h^{-1}), D is the rainfall duration (h), α (>0) and β (<0) are empirical parame-
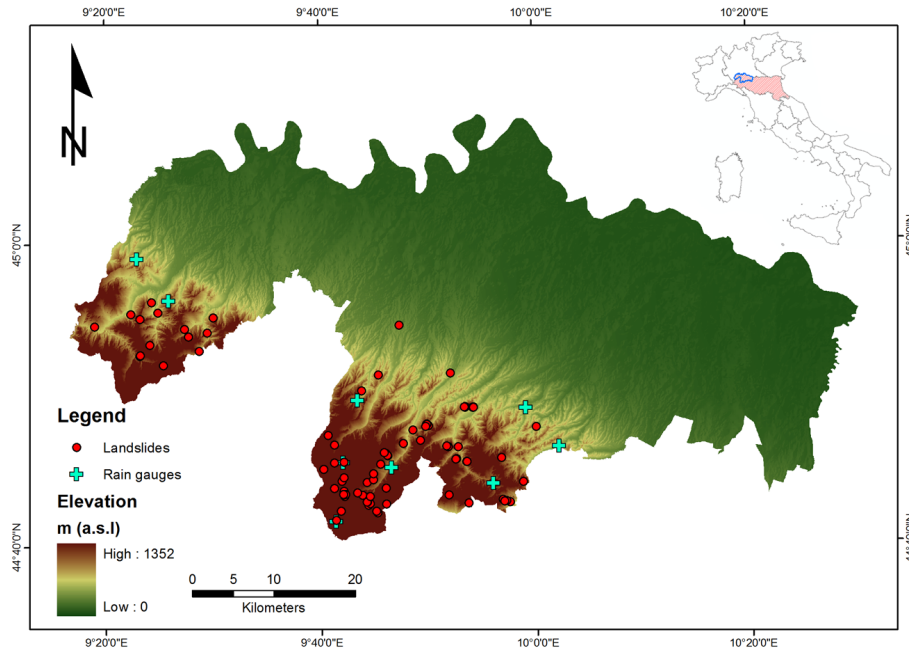
Figure 2. The test site in the Emilia Romagna region, with the location of rain gauges and landslides used in this study.

ters. One of the peculiarities of the MaCumBA model is that thresholds are characterized by a third parameter, called "no rain gap" (NRG). NRG is the number of consecutive hours without rain necessary to separate two rainfall events (Segoni et al., 2014a); this parameter is of fundamental importance to ensure the replicability of the analysis and to consistently employ the thresholds into an operational early warning system (Segoni et al., 2015a).

The procedure for parameters calculation is automated (Segoni et al., 2014a) and allows a large amount of data to be processed: starting from a landslide and a rainfall database, a software program analyzes each cumulated rainfall recorded in the vicinity of a landslide and the most critical rainfall conditions are identified and characterized in terms of I and D. Once the I and D parameters of every landslide are calculated, they are plotted in a I–D diagram and the lower bound threshold is automatically identified. The procedure is completed by a back analysis that identifies the NRG value that minimizes the occurrence of errors during a calibration period.

The model MaCumBA is explained in detail in Segoni et al. (2014a), while Segoni et al. (2014b) discusses its application to Tuscany, which was subdivided into 25 alert zones, each of them characterized by a specific threshold. Segoni et al. (2015a) described the integration of the thresholds into the Tuscany regional warning system, which compares the mosaic of thresholds defined by MaCumBA with rainfall forecasts and rainfall measurements from an automated network composed of about 300 rain gauges.

2.3 Similitudes and differences between SIGMA and MaCumBA

Both methods are presently used by regional civil protection agencies for landslide early warning systems at regional scale (over $20\,000\,\mathrm{km}^2$). SIGMA and MaCumBA operate in the Italian regions of Emilia Romagna and Tuscany, respectively. They provide automatic outputs, based on the comparison of rainfall thresholds with rainfall forecasts and real-time measurements from automated rain gauge networks. Both early warning systems are based on a mosaic of local-scale thresholds: the region is subdivided into smaller areas that are characterized by a site-specific threshold and that are monitored independently. This approach allows landslides of mixed typology to be accounted for and increases the spatial accuracy.

The main difference between the models lies in the calculation of the thresholds and in the input data required. While SIGMA thresholds are based on cumulative rainfall and consider variable time spans ranging from 1 to 245 days, MaCumBA is based on intensity–duration thresholds. SIGMA requires long rainfall recordings (50–60 years time series) but, on its basic implementation, thresholds can be defined even without landslide data. In turn, MaCumBA needs a complete landslide database to evaluate intensity–duration thresholds, but a shorter period of rainfall data (5–10 years) is required.

To quantitatively compare these two models, we applied MaCumBA in an Emilia Romagna alert zone (Fig. 2) and SIGMA in a Tuscany alert zone (Fig. 4).

The application to real case studies carries additional differences, as in the two test sites the rainfall and the land-

slide data sets present peculiar characteristics, which will be described in the next sections. For instance, the landslide data set in Tuscany extends from 2000 to 2009, while in Emilia Romagna from 2004 to 2010. However, a straight comparison between the models is guaranteed by adopting identical decisions during the validation and calibration of both models within the same test site. The application in Emilia Romagna follows the characteristics of Lagomarsino et al. (2013), while application in Tuscany is coherent with Segoni et al. (2014a, b). As a consequence, in Emilia Romagna, the data set was split into two independent subsets: 2004–2007 for calibration, and 2008–2010 for validation. In Tuscany, the data set from 2000 to 2007 is used for the calibration, and the data set from January 2008 to January 2009 is used for the validation.

Figure 3. Intensity–duration threshold calculated by MaCumBA for the Emilia Romagna test site. Since some of the landslides occurred on the same day and at nearby locations, a single I–D point in the graph can be representative of more than one landslide.

3   Application to the Emilia Romagna test site

The region of Emilia Romagna (northern Italy) is dominated in the south by the Apennines. The hilly and mountainous sector extends from the Apennine ridge, in the SW of the region, to the Pede-Apennine margin, in the NE. The chosen alert zone, denoted H, lies in the northwestern part of the region (Fig. 2), and consists of a hilly and mountainous zone, with a maximum elevation of about 1300 m.

The application of SIGMA in the test site is already published (Martelloni et al., 2012; Lagomarsino et al., 2013) and considered the time span 2004–2007 as the calibration period, and the time span 2008–2010 as the validation period. The calibration data set consists of data of 71 landslides, triggered during 17 distinct rainfall events, while for the validation, the data of 39 landslides triggered during 18 rainfall events were available (Fig. 2).

Flysch is the lithology most frequently associated to landslides (about 70 %), while 26 % occurred on hillslopes made up of soft or incoherent rocks (pelagic limestone, claystone and chaotic complex), which are usually covered with cohesive terrains.

The landslide database does not include complete information on the landslide typology, as in most cases (54 %) it is not specified. A total of 11 and 15 % of the occurrences can be comprehended in the "shallow landslide" and "deep-seated" category, respectively, while for 19 % of the landslides, flow is the prevailing mechanism. This information seems to be in accordance with the landslide characteristics commonly reported by the existing literature, which states that the most frequent phenomena are deep-seated landslides (mainly rotational–translational slides, slow earth flows and complex movements) (Bertolini and Pellegrini, 2001; Bianchi and Catani, 2002; Trigila et al., 2010) and that rapid shallow landslides, although less recurrent, have increased their frequency in the last few years (Martina et al., 2010; Montrasio et al., 2011).
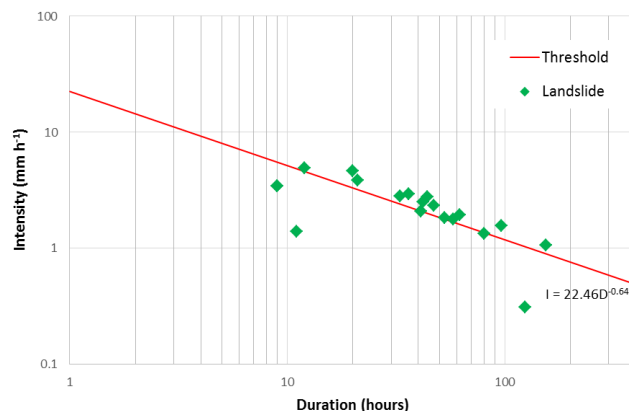
While the SIGMA model makes use of only two reference rain gauges (one for the western sector and one for the eastern sector of the alert zone), to apply MaCumBA at its full potential, all nine automated rain gauges installed in the alert zone were used (Fig. 2). For all of them, we extracted hourly rainfall measurements pertaining to the calibration and validation period and we applied the procedure described in Segoni et al. (2014a) and summarized in Sect. 2.2.

The application of MaCumBA to this case study resulted in a threshold represented by the equation:

$$I = 22.46D^{-0.64}.$$

This threshold is reported in Fig. 3, where the events used for its calibration are also represented. Since some of the landslides occurred on the same day and at nearby locations, a single I–D point in the graph can be representative of more than one landslide. In particular, the three points below the thresholds are each representative of a single landslide. Consequently, the threshold encompasses 68 out of 71 landslides of the calibration data set, which is within the 95 % confidence level selected for the threshold analysis (as in Segoni et al., 2014a).

4   Application to the Tuscany test site

Tuscany is located in central Italy and is characterized by a mainly hilly and mountainous territory. The alert zone (AZ) chosen as test site corresponds with the Serchio Basin (Fig. 4) and includes part of the Northern Apennines, a fold-and-thrust post-collisional belt. This area is mainly mountainous and shows two different geological settings (Rossi et al., 2013): in the western sector, mountain tops are mainly made up of carbonaceous rocks and have very steep flanks. The summits are typically connected to the lower parts of
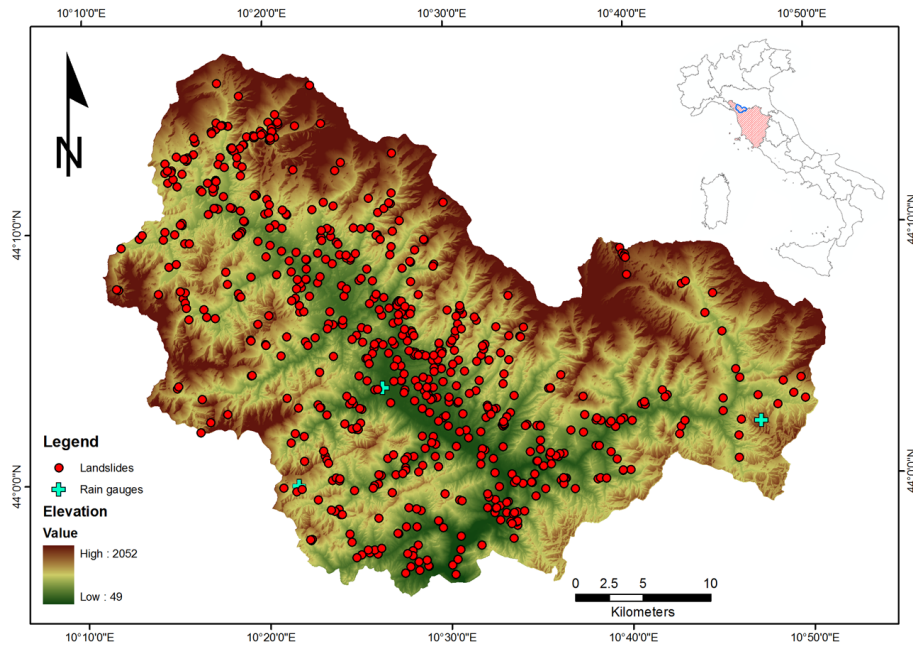
Figure 4. The test site in the Tuscany region, with the location of rain gauges and landslides used in this study.

the slopes, composed of metamorphic sandstone and phyllitic schist and covered by talus and scree deposits. The eastern sector shows a more uniform geological condition with the prevalence of flysch rocks.

The application of MaCumBA in Tuscany and in the Serchio alert zone is already published (Segoni et al., 2014a, b) and considered the time span 2000–2007 as the calibration period and the time span from 1 January 2008 to 31 January 2009 as the validation period. The calibration data set counts 719 landslides, related to 79 distinct rainfall events, while the validation data set counts 272 landslides, related to seven distinct rainfall events (Segoni et al., 2014a, b). Among these, debris flows and shallow landslides are the largely prevailing typologies (89 % of the landslides with known typology). The lithologies most affected by landslides are flysch (60 % of the occurrences), limestone and marble (22 %), clayey rocks (8 %) and granular terrain (7 %).

Using the calibration data set, the SIGMA model has been applied to the Serchio AZ (Fig. 4). Concerning rainfall data, the 37 automated rain gauges used for MaCumBA were analyzed; however, most of these instruments were installed in recent times, and only three of them have the characteristics (time series between 60 and 70 years) to be used for the statistical analyses needed in SIGMA (Fig. 4). One of the three rain gauges is located in the center of the alert zone, while the other two are close to the eastern and southwestern borders (Fig. 4).

As demonstrated by Lagomarsino et al. (2013), it is not straightforward to decide how many and which rain gauges have to be used in SIGMA to obtain the best possible landslide prediction. According to Lagomarsino et al., (2013), the

application of SIGMA comprehended some tests to identify the optimal configuration of the model. We tested all possible configurations: the alert zone subdivided into three territorial units, each with one of the three instruments as the reference rain gauge; three different configurations in which the alert zone was not partitioned and the three rain gauges were selected each time as the only reference rain gauge for the whole of the alert zone and three possible combinations using two rain gauges as reference for an alert zone split into two distinct territorial units. We verified that the best outcomes were obtained using the central rain gauge as the unique reference rain gauge for the entire alert zone. This result is only partially surprising, as when Lagomarsino et al. (2013) tuned SIGMA to optimize the results, an identical circumstance was found in one of the eight Emilia Romagna alert zones.

Using the calibration procedure reported in Martelloni et al. (2012) and summarized in Sect. 2.1, the thresholds shown in Fig. 5 were selected as the optimal ones for the Serchio alert zone.

5   Results

The first step to evaluate the performances of the models consisted of simulating their response to past events that are independent from those used in the threshold calibration process. For the Emilia Romagna test site, the independent validation data set spans from 1 January 2008 to 31 December 2010, while for the Tuscany test site, it spans from 1 January 2008 to 31 January 2009. The models were run using
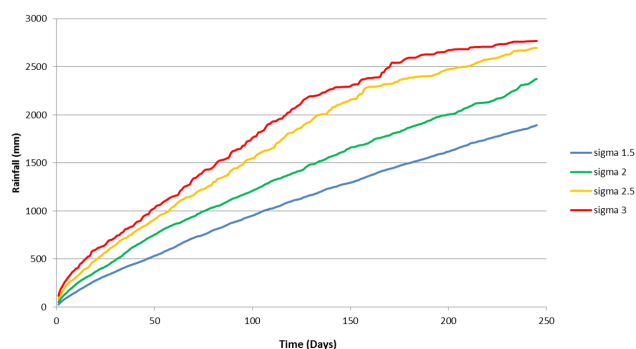
Figure 5. Rainfall thresholds obtained with the SIGMA model in the Serchio AZ; please note that the thresholds are defined for a maximum accumulation period of 245 days, since longer periods of accumulation are not used in the decisional algorithm of the model (Fig. 1).

rainfall data from the validation data set. The simulated daily outputs of each model were compared to the landslides which occurred during the validation period, so as to count

- true positives (TP), which are days with landslides correctly detected by the model (the model raised an alarm and it was verified that a landslide occurred);

- true negatives (TN), which are days without landslides in which the model did not raise an alarm;

- false positives (FP), which are days in which the model raised an alarm but no landslides occurred (false alarms or "errors of commission");

- false negatives (FN), which are days in which at least one landslide occurred, but the model did not raise an alarm (missed alarms or "errors of omission").

In each case study, these occurrences were combined to define some indexes commonly used to evaluate model performances in hazard assessment (Begueria, 2005) and in rainfall thresholds (Martelloni et al., 2012). The following indexes quantify the forecasting effectiveness of the models in the different test sites and allow for a rigorous comparison of the performances.

- Positive predictive power (PPP) is the proportion of positive results that are true positives: $\mathrm{PPP} = (\mathrm{TP})/(\mathrm{FP} + \mathrm{TP})$.

- Negative predictive power (NPP) is the proportion of negative results that are true negatives: $\mathrm{NPP} = (\mathrm{TN})/(\mathrm{FN} + \mathrm{TN})$.

- Sensitivity (Se, also called the true positive rate) measures the proportion of positive occurrences (landslides) which are correctly identified as such: $\mathrm{Se} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$.

Table 1. Contingency matrix displaying the results of the validation of MaCumBA in the Emilia Romagna test site. In this test site, the validation data set spans from 2004 to 2007. TP denotes true positives, FP false positive errors, FN false negative errors and TN true negatives.

Emilia Romagna test site		Observed truth	
MaCumBA model		Landslide	No landslide
Prediction	Landslide	6 (TP)	7 (FP)
	No landslide	12 (FN)	1071 (TN)

- Specificity (Sp, also called the true negative rate) measures the proportion of negative occurrences (days without landslides) which are correctly identified as such: $\mathrm{Sp} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP})$.

- Likelihood ratio (LR) evaluates both the sensitivity and the specificity of a model in a single parameter: $\mathrm{LR} = \mathrm{Se}/(1 - \mathrm{Sp}) = [\mathrm{TP}/(\mathrm{TP} + \mathrm{FN})]/[1 - \mathrm{TN}/(\mathrm{TN} + \mathrm{FP})]$.

- Efficiency (Ef) is an index that evaluates the overall performance of a model, measuring the proportion of correct predictions with respect to the total: $\mathrm{Ef} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{FP} + \mathrm{FN} + \mathrm{TP} + \mathrm{TN})$.

A perfect predictor would be 100 % sensitive and 100 % specific and would have a PPP and NPP equal to 1. In a warning system, the best possible trade-off between sensitivity and specificity is usually researched. Two indexes that help to evaluate this trade-off and thus the overall performance of the model are the efficiency and likelihood ratios. However, when used in circumstances where TN are 1 or 2 orders of magnitude higher than all other occurrences, as in landslide early warning systems, efficiency values can be very close to 1 (optimal value): this strongly reduces the weight of TN occurrences in assessing the final value and prevents a proper comparison between efficiency values, which are very close to each other. This drawback does not affect the likelihood ratio, which evaluates both the sensitivity and the specificity in a single parameter: the higher its value, the better the model.

For the Emilia Romagna test site, the validation results are summarized in contingency matrixes (Tables 1 and 2) and can be quantitatively compared in Table 3.

For the Tuscany test site, the validation results are summarized in Tables 4 and 5 and can be quantitatively compared in Table 6.

6 Discussion

The performance of the models can be quantitatively evaluated by comparing the validation indexes and the contingency tables presented in the previous section. A comparison

Table 2. Contingency matrix displaying the results of the validation of SIGMA in the Emilia Romagna test site. In this test site, the validation data set spans from 2004 to 2007. TP denotes true positives, FP false positive errors, FN false negative errors and TN true negatives.

Emilia Romagna test site		Observed truth	
SIGMA model		Landslide	No landslide
Prediction	Landslide	18 (TP)	12 (FP)
	No landslide	0 (FN)	1066 (TN)

was performed separately for each test site, to assess which model would perform better in a landslide warning system.

In the Emilia Romagna test site, MaCumBA identified only six out of 18 landslide events, while SIGMA correctly identified all of them (Tables 1 and 2). However, SIGMA committed a higher number of false positives (12 against seven committed by MaCumBA). Looking at validation statistics (Table 3), it can be seen that SIGMA indexes are higher than MaCumBA ones, especially in the case of the positive predictive power and sensitivity. Considering both efficiency (that balances positive and negative predictive power) and likelihood ratio (that balances sensitivity and specificity), SIGMA performs better than MaCumBA (Table 3).

In the Tuscany test site, MaCumBA and SIGMA identified a similar number of landslide events (18 and 19 out of 21, respectively), while a relevant difference exists in the number of false alarms: 12 for SIGMA and only one for MaCumBA (Tables 4 and 5). Consequently, MaCumBA has higher positive predictive power and specificity, but lower negative predictive power and sensitivity than SIGMA (Table 6). To assess which model has the best overall performance, we compared efficiency and likelihood ratio: both indexes are higher for MaCumBA (0.98 against 0.93, and 158.6 against 13.9, respectively).

This comparison revealed that none of the two models can be considered better than the other: SIGMA performed better than MaCumBA in the Emilia Romagna test site, while MaCumBA prevailed in the Tuscan test site. Indeed, the performances of a model can vary substantially from one application to another. It is evident that in each test site, the best results were obtained with the model specifically conceived for the characteristics of the case study. Among these characteristics, the different landslide typologies could be put in relation with the performance of the models: MaCumBA, which is based on intensity–duration thresholds, prevails in the Serchio Valley that is affected mainly by shallow landslides; SIGMA is based on a more complex decisional algorithm conceived to account for both shallow and deep-seated

landslides, and it prevails in the Emilia Romagna test site, which is affected by both typologies of landslides.

Another feature that can greatly influence the performance of a model from one application to another is the quantity and quality of the rainfall data available. For instance, it is well established (Staley et al., 2013; Vessia et al., 2014) that the $I-D$ threshold provides the best results when rainfall is measured at hourly or even smaller time steps, while the SIGMA model is specifically conceived to be applied on rainfall data with a daily time step. However, in this work, rain gauges provide hourly rainfall data, and the larger flexibility of SIGMA is not fully exploited. A feature that could have had a relevant impact in the results is the spatial density of the rainfall measurements. In the Emilia Romagna test site, fewer rain gauges are available, but they all have long rainfall series. This is an optimal condition to apply SIGMA, which needs only a limited number of rain gauges, since each territorial unit is analyzed and monitored by a single reference rain gauge. Conversely, this condition is a strong limitation for the employment of an $I-D$ threshold model like MaCumBA: the very longer time series do not provide an additional value, and the lower number of points of measurements constrains the accurate characterization of the landslides in terms of intensity and duration of the triggering rainfall. The Tuscany test site has opposite conditions: the rain gauge network is very dense, but only very few instruments (namely, three) have long enough time series to implement SIGMA. Taking into consideration the three rain gauges that could serve as reference, the calibration procedure of SIGMA allows the best possible model configuration to be defined, but it is the best option among a few options available. Moreover, the calibration results (a single rain gauge used as a reference for the whole of the area) highlight that large sectors of the area could not be fully represented by the rain gauges available. On the contrary, MaCumBA can be successfully applied in these conditions, as the short time series are not a handicap (provided they cover the same time period of the landslide inventory) and the high network density allows the triggering rainfall intensity to be identified with sufficient approximation.

It should be noted that in this study we decided to give the same weight to errors of omission (FN) and errors of commission (FP). In other applications, it could be decided to give different weights to one (or more than one) of the occurrences of the contingency table and to recalculate a modified contingency table and a series of modified performance indexes. The weights should be decided in advance, depending on the objectives of the research or the local civil protection procedures. For instance, in case of a comparison between two or more "minimum thresholds", false alarms could be tolerated, while missed alarms should receive a relevant weight, because the aim of these thresholds is to point out the minimum rainfall conditions potentially responsible for landslides. Concerning the evaluation of warning systems, the balance between false alarms and missed alarms

Table 3. Validation statistics and comparison of the performances of the two models in the Emilia Romagna test site.

Emilia Romagna test site	Efficiency	Positive predictive power	Negative predictive power	Sensitivity	Specificity	Likelihood ratio
MaCumBA	0.98	0.46	0.99	0.33	0.99	51.3
SIGMA	0.99	0.60	1	1	0.99	89.8

Table 4. Contingency matrix displaying the results of the validation of SIGMA in the Tuscany test site. In the Tuscany test site, the validation data set spans from 1 January 2008 to 31 January 2009. TP denotes true positives, FP false positive errors, FN false negative errors and TN true negatives.

Tuscany test site		Observed truth	
MaCumBA model		Landslide	No landslide
Prediction	Landslide	18 (TP)	1 (FP)
	No landslide	3 (FN)	184 (TN)

Table 5. Contingency matrix displaying the results of the validation of SIGMA in the Tuscany test site. In the Tuscany test site, the validation data set spans from 1 January 2008 to 31 January 2009. TP denotes true positives, FP false positive errors, FN false negative errors and TN true negatives.

Tuscany test site		Observed truth	
SIGMA model		Landslide	No landslide
Prediction	Landslide	19	12
	No landslide	2	173

is usually desirable and the weights could be assigned with a political decision. The impact of the countermeasures to be taken in response to alarms may lead to different levels of acceptance of false alarms, which in turn could lead to different weights.

7   Conclusions

Rainfall thresholds are widely used in landslide forecasting and they often constitute the core of civil protection warning systems. However, most of the rainfall thresholds presented in the literature were not subject to a rigorous validation procedure. Moreover, no publication exists that quantitatively compares two or more different rainfall threshold models with the aim of choosing the one with the best forecasting effectiveness.

This paper proposes a methodology to compare different rainfall threshold models and to assess which of them would constitute the most effective warning system.

The proposed methodology goes beyond the commonly adopted visual comparison of literature thresholds and consists of the application of the models to a common case study to define site-specific thresholds, performing a calibration and a validation procedure against independent data sets, building a confusion matrix and using it to derive a series of statistical indexes. These indexes can be considered as indicators of the performance of the thresholds and can provide an objective basis for the quantitative comparison of the effectiveness of the threshold models. We propose, in particular, taking the likelihood ratio and efficiency into consid-

eration, as they can estimate the overall performance of the models with a single value.

We tested two different models, namely SIGMA (Martelloni et al., 2012) and MaCumBA (Segoni et al., 2014a), which have already been used for the regional landslide early warning systems operated in Emilia Romagna and Tuscany, respectively. To compare these two models, each of them was applied in a part of the region in which the other is already active. This work demonstrated the technical feasibility of exporting each model in test sites different from those where they have been conceived, however the performance of the models varied substantially, depending on the characteristics of the test site and on the quality and quantity of the rainfall measurements. In the test site affected by shallow landslides and equipped with a dense rain gauge network, the intensity–duration thresholds of MaCumBA provided the best outcomes. In the test site affected by both shallow and deep-seated landslides and equipped with a limited number of rain gauges with long time series, the best results were obtained using SIGMA, which is based on a more complex decisional algorithm based on rainfall time series aggregated over variable time windows.

We conclude that even if state-of-the-art threshold models can be exported from a test site to another, their employment in local early warning systems should be carefully evaluated: the effectiveness of a threshold model depends on the test site characteristics (including the quality and quantity of the input data), and a validation procedure and a comparison with alternative models should be performed before its implementation in operational early warning systems.

Table 6. Validation statistics and comparison of the performances of the two models in the Tuscany test site.

Tuscany test site	Efficiency	Positive predictive power	Negative predictive power	Sensitivity	Specificity	Likelihood
MaCumBA	0.98	0.95	0.98	0.86	0.99	158.6
SIGMA	0.93	0.61	0.99	0.90	0.94	13.9

References

Begueria, S.: Validation and evaluation of predictive models in hazard assessment and risk management, Nat Hazards, 37, 315–329, 2005.

Berti, M., Martina, M. L. V., Franceschini, S., Pignone, S., Simoni, A., and Pizziolo, M.: Probabilistic rainfall thresholds for landslide occurrence using a Bayesian approach, J. Geophys. Res., 117, F04006, doi:10.1029/2012JF002367, 2012.

Bertolini, G. and Pellegrini, M.: The landslides of the Emilia Apennines (Northern Italy) with reference to those which resumed activity in the 1994–1999 period and required civil protection interventions, Quaderni di Geologia Applicata, 8, 27–74, 2001.

Bianchi, F. and Catani, F.: Landscape dynamics risk management in Northern Apennines (Italy), Environ. Stud., 7, 319–328, 2002.

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

Caine, N.: The rainfall intensity–duration control of shallow landslides and debris flows, Geogr. Ann., 62, 23–27, 1980.

Campbell, R. H.: Soil slips, debris flows, and rainstorms in the Santa Monica Mountains and vicinity, southern California, US Geological Survey Professional Paper 851, US Government Printing Office, Washington, D.C., 51 pp., 1975.

Cannon, S. H., Gartner, J. E., Wilson, R., Bowers, J., and Laber, J.: Storm rainfall conditions for floods and debris flows from recently burned areas in southwestern Colorado and southern California, Geomorphology, 96, 250–269, 2008.

Cannon, S. H., Boldt, E., Laber, J., Kean, J., and Staley, D.: Rainfall intensit–duration thresholds for postfire debris-flow emergency-response planning, Nat. Hazards, 59, 209–236, 2011.

Cardinali, M., Galli, M., Guzzetti, F., Ardizzone, F., Reichenbach, P., and Bartoccini, P.: Rainfall induced landslides in December 2004 in south-western Umbria, central Italy: types, extent, damage and risk assessment, Nat. Hazards Earth Syst. Sci., 6, 237–260, doi:10.5194/nhess-6-237-2006, 2006.

Chen, H. X. and Wang, J. D.: Regression analyses for the minimum intensity-duration conditions of continuous rainfall for mudflows triggering in Yan'an, northern Shaanxi (China), Bull. Eng. Geol. Environ., 73, 917–928, doi:10.1007/s10064-013-0567-3, 2014.

Chleborad, A. F.: Preliminary evaluation of a precipitation threshold for anticipating the occurrence of landslides in the Seattle, Washington Area, US Geological Survey Open-File Report 03, US Geological Survey, 463 pp., 2003.

Endo, T.: Probable distribution of the amount of rainfall causing landslides, Annual report, Hokkaido Branch, Govern. Forest Experiment Station, Sapporo, 123–136, 1970.

Gariano, S. L., Brunetti, M. T., Iovine, G., Melillo, M., Peruccacci, S., Terranova, O., Vennari, C., and Guzzetti, F.: Calibration and validation of rainfall thresholds for shallow landslide forecasting in Sicily, southern Italy, Geomorphology, 228, 653–665, 2015.

Giannecchini, R., Galanti, Y., and D'Amato Avanzi, G.: Critical rainfall thresholds for triggering shallow landslides in the Serchio River Valley (Tuscany, Italy), Nat. Hazards Earth Syst. Sci., 12, 829-842, doi:10.5194/nhess-12-829-2012, 2012.

Guidicini, G. and Iwasa, O. Y.: Tentative correlation between rainfall and landslides in a humid tropical environment, Bull. Int. Ass. Eng. Geol., 16, 13–20, 1977.

Guzzetti, F., Peruccacci, S., Rossi, M., and Stark, C. P.: Rainfall thresholds for the initiation of landslides in central and southern Europe, Meteorol. Atmos. Phys., 98, 239–267, 2007.

Guzzetti, F., Peruccacci, S., Rossi, M., and Stark, C. P.: The rainfall intensity–duration control of shallow landslides and debris flows: An update, Landslides, 5, 3–17, 2008.

Lagomarsino, D., Segoni, S., Fanti, R., and Catani, F.: Updating and tuning a regional-scale landslide early warning system, Landslides, 10, 91–97, 2013.

Larsen, M. C. and Simon, A.: A rainfall intensity–duration threshold for landslides in a humid-tropical environment, Puerto Rico, Geograf. Ann. A, 75, 13–23, 1993.

Lumb, P.: Slope failure in Hong Kong, Q. J. Eng. Geol., 8, 31–65, 1975.

Martina, M. L. V., Berti, M., Simoni, A., Todini, E., Pignone, S.: Un approccio bayesiano per individuare le soglie di innesco delle frane, in: Rainfall-induced landslides: mechanisms, monitoring techniques and nowcasting models for early warning systems, vol. 2, edited by: Picarelli, L., Tommasi, P., Urcioli, G., and Versace, P., CIRAM, Naples, 2010.

Martelloni, G., Segoni, S., Fanti, R., and Catani, F.: Rainfall thresholds for the forecasting of landslide occurrence at regional scale, Landslides, 9, 485–495, 2012.

Montrasio, L., Valentino, R., and Losi, G. L.: Towards a real-time susceptibility assessment of rainfall-induced shallow landslides on a regional scale, Nat. Hazards Earth Syst. Sci., 11, 1927–1947, doi:10.5194/nhess-11-1927-2011, 2011.

Rosi, A., Segoni, S., Catani, F., and Casagli, N.: Statistical and environmental analyses for the definition of a regional rainfall threshold system for landslide triggering in Tuscany (Italy), J. Geogr. Sci., 22, 617–629, 2012.

Rosi, A., Lagomarsino, D., Rossi, G., Segoni, S., Battistini, A., and Casagli, N.: Updating EWS rainfall thresholds for the triggering of landslides, Nat. Hazards, 78, 297–308, doi:10.1007/s11069-015-1717-7, 2015.

Rossi, G., Catani, F., Leoni, L., Segoni, S., and Tofani, V.: HIRESSS: a physically based slope stability simulator for HPC applications, Nat. Hazards Earth Syst. Sci., 13, 151–166, doi:10.5194/nhess-13-151-2013, 2013.

Segoni, S., Rossi, G., Rosi, A., and Catani, F.: Landslides triggered by rainfall: a semiautomated procedure to define consistent intensity-duration thresholds, Comput. Geosci., 30, 123–131, 2014a.

Segoni, S., Rosi, A., Rossi, G., Catani, F., and Casagli, N.: Analysing the relationship between rainfalls and landslides to define a mosaic of triggering thresholds for regional-scale warning systems, Nat. Hazards Earth Syst. Sci., 14, 2637–2648, doi:10.5194/nhess-14-2637-2014, 2014b.

Segoni, S., Battistini, A., Rossi, G., Rosi, A., Lagomarsino, D., Catani, F., Moretti, S., and Casagli, N.: Technical Note: An operational landslide early warning system at regional scale based on space–time-variable rainfall thresholds, Nat. Hazards Earth Syst. Sci., 15, 853–861, doi:10.5194/nhess-15-853-2015, 2015a.

Segoni, S., Lagomarsino, D., Fanti, R., Moretti, S., and Casagli, N.: Integration of rainfall thresholds and susceptibility maps in the Emilia Romagna (Italy) regional-scale landslide warning system, Landslides, 12, 773–785, 2015b.

Staley, D. M., Kean, J. W., Cannon, S. H., Schmidt, K. M., and Laber, J. L.: Objective definition of rainfall intensity–duration thresholds for the initiation of post-fire debris flows in southern California, Landslides, 10, 547–562, doi:10.1007/s10346-012-0341-9, 2013.

Trigila, A., Iadanza, C., and Spizzichino, D.: Quality assessment of the Italian Landslide Inventory using GIS processing, Landslides, 7, 455–470, 2010.

Vessia, G., Parise, M., Brunetti, M. T., Peruccacci, S., Rossi, M., Vennari, C., and Guzzetti, F.: Automated reconstruction of rainfall events responsible for shallow landslides, Nat. Hazards Earth Syst. Sci., 14, 2399–2408, doi:10.5194/nhess-14-2399-2014, 2014.

Wilson, R. C.: Climatic variations in rainfall thresholds for debris-flows activity, in: Proc. 1st Plinius Conf. on Mediterranean Storms, edited by: Claps, P. and Siccardi, F., Maratea, 415–424, 2000.