Landslides (2012) 9:53–62
DOI 10.1007/s10346-011-0274-8
Received: 25 November 2010
Accepted: 24 May 2011
Published online: 9 June 2011
© Springer-Verlag 2011

Angelo Doglioni · Francesco Fiorillo · Francesco Guadagno · Vincenzo Simeone

Evolutionary polynomial regression to alert rainfall-triggered landslide reactivation

Abstract The derivation of an alert model for landslide risk management is a paramount problem for those sites which are affected by complex landslides involving strategic infrastructures as well as towns. This is a quite common scenario all over the world and then it is a primary problem for the management of geomorphological risk. Along the Adriatic Coast of south Italy, Petacciato landslide is peculiar, since it showed 11 reactivations between 1924 and 2009. It is a deep-seated landslide, and the history of its reactivations shows that even if generally related to quite abundant rainfall periods, there is no clear correlation between rainfall events and reactivations. For this reason, here, an analysis based on a data-driven evolutionary modeling technique is attempted, in order to identify an alert model based on cumulative rainfall heights. Modeling results are quite interesting and encouraging, since they are able to provide landslide forecasting whereas no false positive are ever returned. This work shows the results of this attempt as well as an analysis of the input to the modeling approach, in order to identify which are those cumulative rainfall heights which are physically sound with respect to the particular landslide.

Keywords Rainfall · Evolutionary polynomial regression · Landslide risk management

Introduction

Damages to infrastructures and sometimes killing and/or injuring people living nearby the landslide sites pushed stakeholders to look for effective and feasible instruments allowing for the prediction of landslides activation, in order to implement an alert system. Rainfall is an important factor for landslide triggering, therefore the determination of rainfall thresholds as well as the return periods of rainfall, which elicit landslides, is still a hot topic (Van Asch et al. 1999; Sengupta et al. 2010; Tatard et al. 2010; First Italian Workshop on Landslides 2009). Ideally, an optimal risk management would claim for a model able to provide a landslide occurrence probability or incipient condition related to a rainfall scenario. This would be of particular use, whereas a high number of people live at the adjoining areas.

The relation between rainfall and deep landslide is very difficult to be ascertained, since it is not easy to directly correlate rainfall events to landslide movements even for shallow landslide (Iverson 2000) where the relation cause–effect is generally clearer. For deep-seated landslide, the relation cause–effect can be very complex, as the activation is connected to longer rainwater infiltration time and can be controlled by progressive failure into the slope, besides further variables not easily identifiable. Besides, each reactivation of a deep-seated landslide can modify the boundary conditions of the slope, thus it may be difficult to compare different reactivations. The complexity of the hydrologic and mechanic characteristics complicates the relation between rainfall and deep-seated landslide, and cause a difficult to fix specific hydrologic thresholds. Therefore, classic deep-seated landslide analysis based on the integration of hydrologic infiltration models and slope stability analysis could not be definitive. In these cases data-driven models could be an effective tool to relate landslide reactivation to antecedent rainfall and to try to understand the relation between rainfall and landslide.

Past studies present some examples about modeling landslide-triggering phenomena and early warning systems based on data-driven modeling. Bai et al. (2009) present a landslide susceptibility map based on a data-driven bivariate analysis, whereas data are extracted from a GIS. Kawabata and Bandibas (2009), Melchiorre et al. (2008), and Lu and Rosenbaum (2003) presented modeling of landslide susceptibility based on the use of artificial neural networks (ANN). Flentje et al. (2007) attempt this analysis using data mining, while Chang and Chien (2007) present a genetic algorithm (GA)-based approach. Glade et al. (2000) introduce a physically based decay coefficient which they derive for each region from the recessional behavior of storm hydrographs and is used to produce an index for antecedent rainfall.

Here a data-driven hybrid evolutionary modeling technique, evolutionary polynomial regression (EPR, Giustolisi and Savic 2006; Giustolisi and Savic 2009), is used in order to predict reactivation occurrence of the deep-seated landslide of Petacciato along the Adriatic coast (south Italy, see Fig. 1; Guerricchio et al. 1996; Fiorillo 2003). It is a specific deep-seated landslide phenomenon along the Adriatic coast, between the Petacciato village and the sea, characterized by 11 reactivations in the period 1932–2009. Each reactivation causes small displacement along the slope, generally less than 1 m, vertically and horizontally, modifying only partially the geometry of the slope. Such characteristics together with the frequency of reactivations are quite unusual for a deep landslide and encourage the construction of a data-driven model aimed at the determination of potential reactivation of the landslide due to specific rainfall scenarios.

In particular, starting from the rainfall data, which are available as daily records since 1924, a data-driven model aimed at describing the reactivation as function of cumulative rainfall values is here identified. The particular mechanisms of the landslide as well as the exceptional data availability, in terms of reactivations, make it possible to cope with this system according to a data-driven approach.

Geological and geomorphological features of the investigated site

Petacciato landslide is one of the multiple landslides along the Adriatic coast (Fig. 1) of Italy (Cancelli et al. 1984; Guerricchio and Melidoro 1996; Cotecchia 2006), where a sequence of Plio-Pleistocenic marine blue-grey, silty clay deposits outcrop, filling the foredeep system of the Apennines (Casnedi et al 1981; Patacca and Scandone 2007). The marine sequence is characterized by overconsolidated clays interbedded with silty sandy layers, 1–10-cm thick. These deposits were subject to tectonic uplift, causing a typical regressive sequence, whereas sands and conglomerates constitute the topping layers of the marine sequence. The sandy conglomerate marine deposits form typical terraced areas, which are gently sloping towards the sea.

Petacciato landslide, as well as the other large landslides along the Adriatic coast, shows some common features. These features consist in large landslides on clayey slopes, characterized by broad and multiple scarps, on sand and conglomerate deposits of the topping layers of the marine sequence. It is also observed with a large ratio between width and length of the landslides, mainly due to the coalescence of multiple landslides. Further peculiar characteristics are: long valleys and watersheds parallel to the main scarps, sometimes located upstream to the main scarps, and involvement of the sea bottom, whereas uplift phenomena can be observed after each movement.

Along the coastal edge of Petacciato, the summit of the terrace lies at an elevation of 220 m, 2 km from the coastline, and a probable local retreatment occurred since late Pleistocene. The coastal slope is slightly sloping, and reflects the local monocline structure, dipping towards N–NE, with a slope ranging between 5° and 8°. A detailed description of Petacciato landslide and of the mechanism related to its activation and of the reactivation history is reported by Guerricchio et al. (1996) and Fiorillo (2003).

Petacciato landslide reactivated more than 10 times in the last 100 years. The latest reactivation was on 20 February 2009 (Fig. 2). Each reactivation of the landslide damaged Petacciato town, as well as the coastal railway and the motorway, which are national strategic infrastructures (Fig. 3).

Comparing the surveys made after the 1991 and 2009 reactivations (Fig. 2), it was observed that for both the cases, the maximum shift due to the landslide is approximately 1 m, both vertically and horizontally. Moreover the involved areas are very similar and no indicators of the incoming landslide were observed before it took place.

Along the beach, down to the landslide, grey-blue Pleistocenic clays were extruded as well as a general uplift of the bottom of the sea was observed (see Fig. 4). This is proved by a no sand sea bottom, whereas the irregular clayey sublayer outcrops. After the 2009 reactivation, a general uplift of the beach was observed close to the tower of Petacciato in particular, a long step related to a toe thrust of the landslide is evident (see Fig. 5).

The large area of Petacciato slope, and its low gradient, together with the observed kinematics, seems to address a complex mechanism of rupture, which is difficult to be fully explained. Several surface features show a stratification which is parallel to the slope, typical of a dip/overdip slope (Fiorillo 2003). This seems to address the high hydraulic head through the silty and sandy layers at the bottom of the slope. In fact, after the 1991 reactivation, some mud extruded due to the high water pressure through the silty and sandy layers of the clayey sublayer sequence (Guerricchio et al. 1996). Some piezometers located along the landslide toe allowed for the surveillance of artesian conditions just before the 1996 reactivations (Fiorillo 2003). These are quite typical of layered slope, whereas permeability anisotropy exists, and where water pressure at the toe increases more than hydrostatically with the depth (Angeli 1985; Hodge and Freeze 1977). This scenario together with the observed kinematics depicts a sudden spreading mechanism (Fiorillo 2003).

Rainfall influence over the landslide-triggering dynamics

The relationship between the rainfall and landslide reactivations up to 1996 was investigated by Guerricchio et al. (1996) and by Fiorillo and Guadagno (2000), using hydrological and statistical

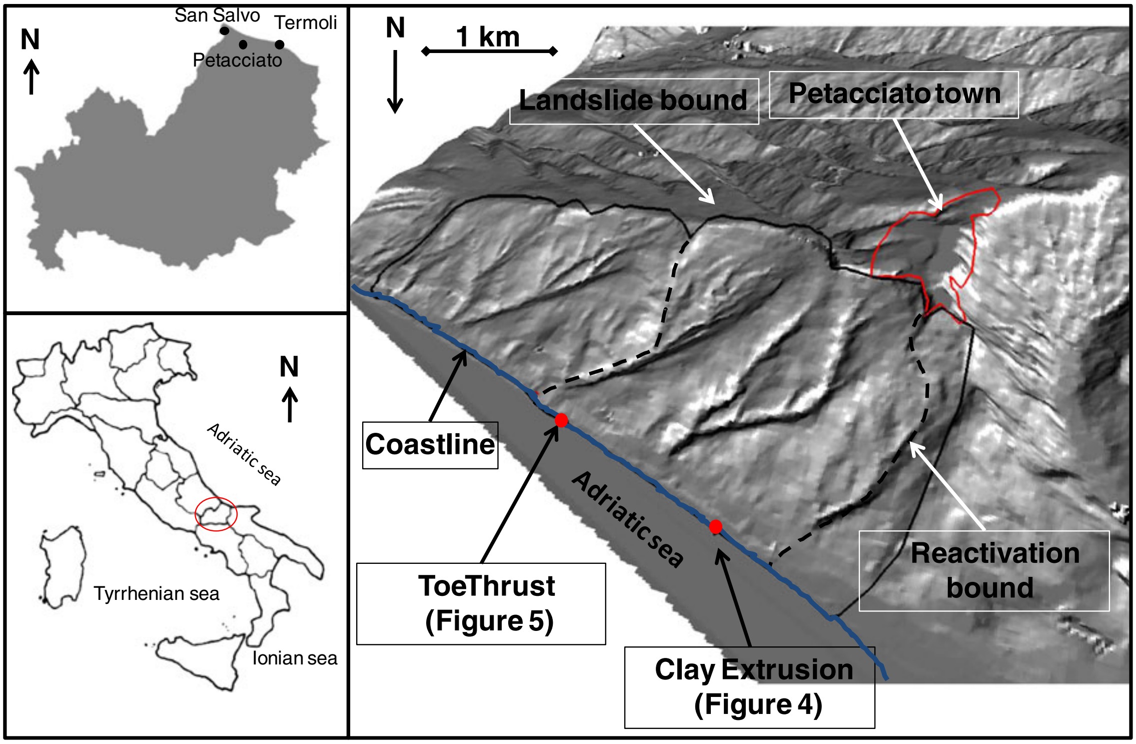


Fig. 1 Overview and location of Petacciato site

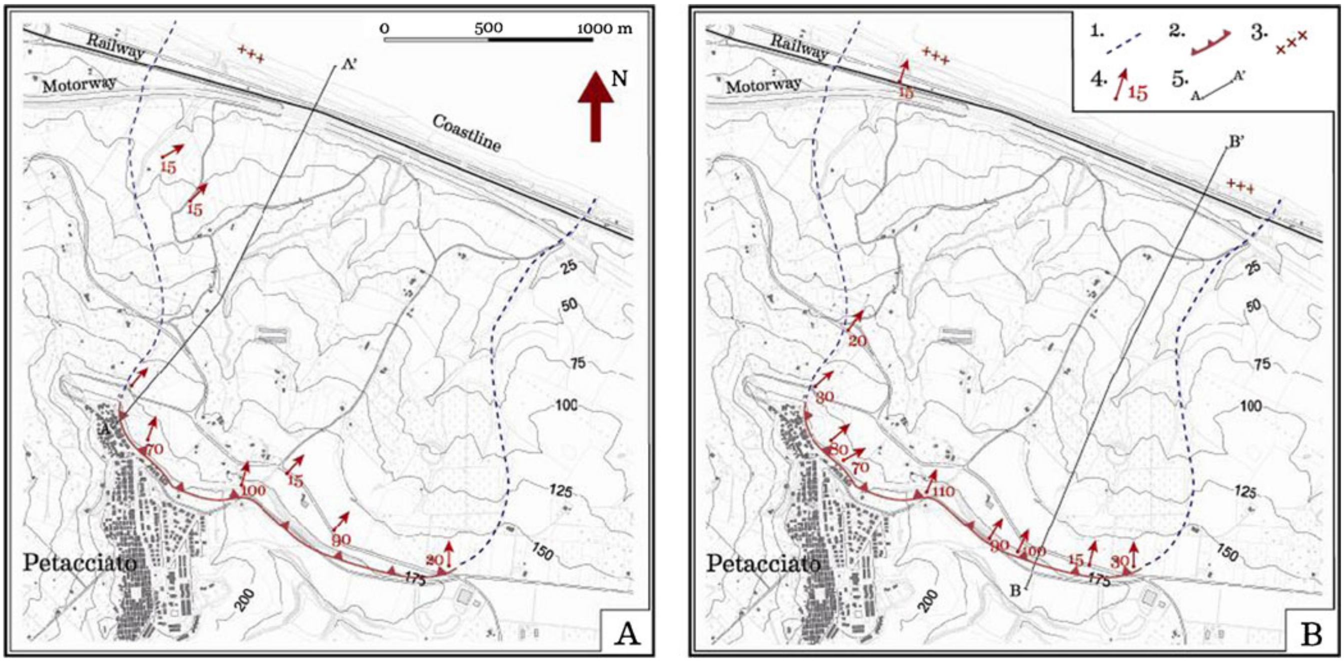
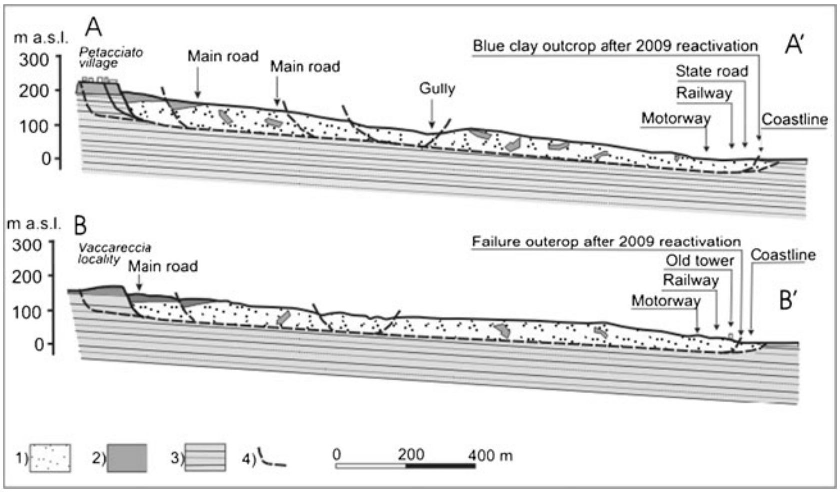


Fig. 2 Comparison between 1991 (**a**) and 2009 (**b**) reactivations. *1* landslide boundary, *2* main scarp, *3* clay extrusion, *4* towards and amount of the landslide displacement (in centimeters), *5* cross-section trace

analyses aimed at finding rain singularities. Those analyses were not able to comprehensively explain the relation between rainfall and landslide reactivations, as no singular characteristics in the rainfall patterns were found. Long cumulative rainfalls were identified as possible landslide-triggering factors, as well as antecedent dry summer, but no clear hydrological statistical model could be identified.

Here a brief standard hydrologic analysis looking at the comparison between reactivations and peculiar values of rainfall on multiple cumulative windows is undertaken to validate previous results. This analysis is mainly based on Termoli rain gauge data, located 11 km east side of Petacciato (Fig. 1). Annual average rainfall is 647.7 mm, the maximum value is 1016.5 mm, the minimum 361.6 mm, and the standard deviation is 133.1 mm. Rainfalls are quite variable during the year according to a typically Mediterranean regime: a minimum is usually reached during the summer period, July and August, while the maximum corresponds to the winter months, December and January. The analysis covers an 85-year period, ranging between 1 January 1924 and 28 February 2009. There is just one main discontinuity in data: it ranges between 1 January 1942 and 12 January 1951, which barely corresponds to 9 years of missing data. Another minor discontinuity (3 months between 1 January 1935 and 31 March 1935) was augmented by San Salvo rain gauge data (Fig. 1), located 11 km west side of Petacciato (Fig. 1). This rain gauge has an average annual rainfall of 632.4 mm and standard deviation of 136.1 mm, very close to that of Termoli rain gauge, confirming the uniform climatic regime along this stretch of the Adriatic coast which includes Petacciato slope. Cumulative rainfall heights for windows of 7, 15, 30, 60, 90, 120, 240, 365, and 500 days are

Fig. 3 Geological cross-section of *A*, *A′* (after Fiorillo 2003, modified) and *B*, *B′*. *1* landslide deposits, *2* sands and conglomerates (middle Pleistocene), *3* blue clay sequence, stratified with silty sandy levels (middle to lower Pleistocene), *4* likely slip surface



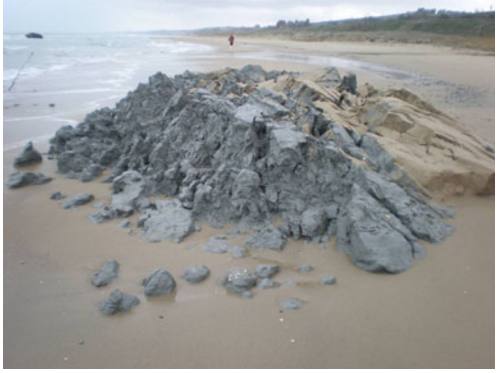


Fig. 4 Extrusion of grey-blue Pleistocenic clays, after 2009 reactivation (courtesy of Eng. Silvano Sgariglia — Petacciato Technical Bureau)

evaluated as a simple moving sum covering the assumed cumulative window.

Different probability distributions were tested to choose the one that could best fit the data. Gamma probability distribution seems to better fit data, thus it is chosen, whereas its parameters are estimated using the maximum likelihood estimates (Benjamin and Cornell 1960). Assuming a probability distribution for each cumulative interval, the return periods of the cumulative rainfall values evaluated in a range covering the 6 months preceding landslide occurrence are estimated.

Figure 6 shows the return periods for the aforementioned cumulative rainfall intervals of the maximum values registered during the 6 months preceding the landslide reactivations. It is noteworthy that reactivations follow rainy periods, which are not necessarily severe. Indeed just the three reactivations occurred during the periods 1953–1956 follow exceptional rainfalls. However those years were very rainy for all southern Italy, thus representing an exceptional scenario.

It is also envisaged that the largest part of the reactivations follow cumulative rains characterized by return periods lower than 10 years, which is consistent with the high number of the reactivation occurred: 11 in 85 years. However, this characteristic complicates the identification of the hydrological conditions inducing landslide reactivations as well as of those which do not imply any landslide.

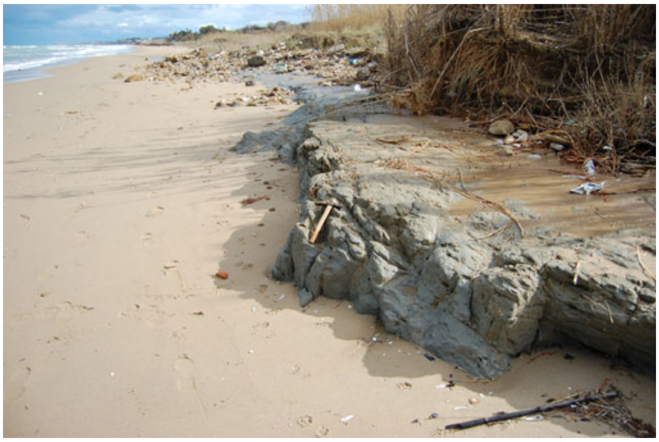


Fig. 5 Toe thrust of the landslide, after 2009 reactivation

This somehow addresses that there are further triggering factors as well as that single cumulative rainfall data cannot explain the landslide triggering. Moreover, this is consistent with the aforementioned studies of this landslide (Guerricchio et al. 1996; Fiorillo and Guadagno 2000), which did not identify a definitive rainfall component which is determinant for the landslide. Therefore, this seems to address a more complicated relation between rainfall and landslide reactivations, which may be related to a particular nonlinear combination of short-time and long-time cumulative rainfalls.

A further investigation about the influence of long rainy periods is here investigated in terms of relationship between landslide reactivations and deviation of the cumulative rainfall from the annual average value. In particular, starting from 1 September before the landslide reactivation, Fig. 7 shows the total amount of cumulative rainfall at the reactivation day. All the reactivation years were characterized by cumulative rainfall heights higher than the average value, but not exceptional. This roughly means that the reactivation years are not characterized by extreme precipitations even if abundant rainfalls occurred almost always during reactivation years.

The described scenario addresses that the relation between the cumulated amount of rainfall and landslide reactivation is quite complex and nonlinear. For this reason, a data-driven approach can be effective and is here undergone, in order to try to catch those nonlinear effects on the landslide due to rainfall, given a reasonably suitable dataset made of 11 reactivation events. For this purpose, an evolutionary modeling approach, as EPR (Giustolisi and Savic 2006; 2009), can be an effective tool in order to seek for the relation between rainfall and landslide. EPR can pursue the information contained in data, whereas this is not immediately evident.

The evolutionary polynomial regression

Evolutionary polynomial regression is a data-driven hybrid technique based on evolutionary computing. EPR integrates a GA (Goldberg 1989) with a least square (LS) approach, providing "transparent" and structured system identification (Giustolisi and Savic 2006, 2009). EPR is a two-stage method that (1) searches model structures based on a GA and (2) estimates their parameters based on the linear optimization.

EPR is an evolutionary modeling technique, therefore like all the population-based strategies for model identification, it can generate some doubts. Actually, the use of such strategies does not necessary imply a good result regardless of the quality of data. In fact, even if the procedure is mainly stochastic, it is driven by some objective functions, and, among these, by an objective which represents the fitness of model to data. Therefore, whereas no information is contained in data or no relationship exists between input and output data, the methodology does not return any significant result (Bäck et al. 1997).The structural search performed by the EPR procedure returns explicit equations, represented by a linear combination of nonlinear monomial terms. These can be used also to better understand the main variables of the phenomenon at stake.

EPR actually searches models having symbolic pseudo-polynomial structures, which are reported in Giustolisi and Savic (2006, 2009). The general structure adopted for the

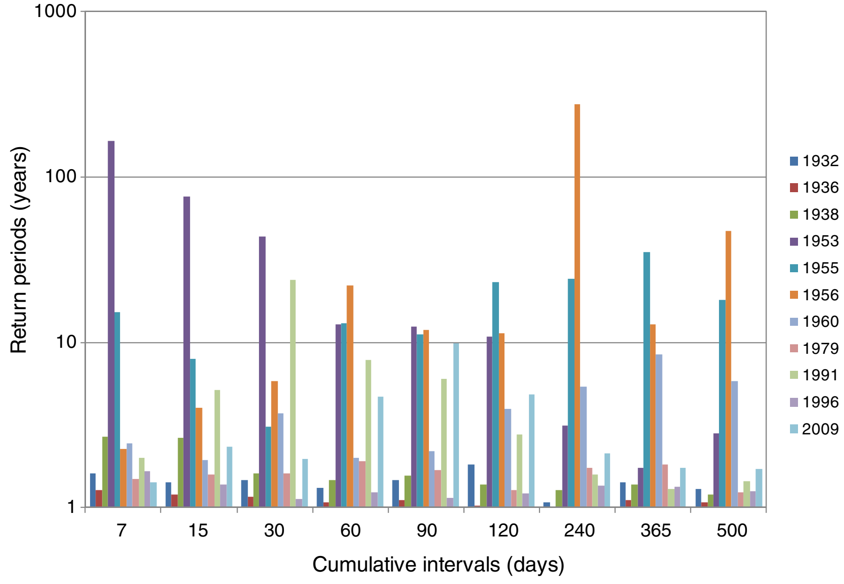


Fig. 6 Return periods for each cumulative interval at each reactivation

evolutionary search in the present case study is reported in the following equation:

$$Y = a_0 + \sum_{j=1}^{m} a_j \cdot (X_1)^{\mathrm{ES}(j,1)} \cdot \ldots \cdot (X_k)^{\mathrm{ES}(j,k)} \qquad (1)$$

where X_i are the vectors of candidate inputs, Y the simulated output of the system, ES is the matrix of exponents (coded as integers in the GA), a_j are constant parameters, and m is the length of the returned expressions, i.e., the number of terms of the polynomial structure returned by EPR. When a null element of the matrix ES is selected as an exponent for a variable in a monomial term, then the variable assumes the value of one and is therefore deselected. The constant parameters a_j, j=1,..., m, are estimated by a least squares method integrated in the EPR procedure. The LS guarantees a biunique correspondence between the structure and its constant values.

A multi-objective (MO) strategy (Giustolisi and Savic 2009) is implemented and integrated into EPR, which is named EPR-MOGA. This procedure tries to achieve a higher value of fitness to data while forcing the model structure to be parsimonious (Giustolisi and Savic 2009). An advantage is that, as a product of the MO approach, a set of models will be constructed and available for further judgment. Therefore, the goal of such optimization is to unearth a set of solutions which are acceptable for the analyst.

The Pareto dominance criterion is here used in order to pursue the MO search. In particular, a vector is said Pareto optimal if there exists no different vector which would decrease some criterion indicator without causing a simultaneous increase in at least one other criterion (Coello Coello 1999; Van Veldhuizen

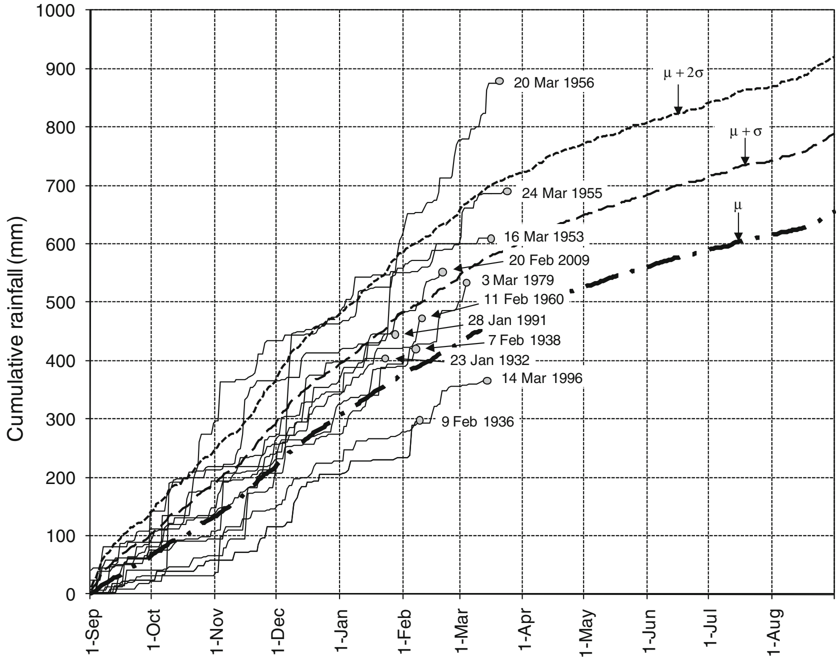


Fig. 7 Cumulative rainfall heights and statistics, • is the average value, σ is the standard deviation

and Lamont 2000). Moreover the MO strategy is fostered assuming one or more user-defined mathematical constraints. Pareto optimality almost never implies a single solution, but rather a set of solutions that constitute the so-called non-dominated set.

EPR-MOGA tries to minimize three objective functions: (a) the sum of squared errors (SSE), which addresses the performance of models in terms of fitness to data; (b) the number of constant values a_j; and (c) the total number of inputs involved in the symbolic expression. The latter two objective functions relate to the structural complexity of the models. Note that the total number of inputs corresponds to the number of times each input is involved in the symbolic expression. The user sets the maximum number of constant values, which poses an upper constraint on the length of the candidate expressions.

The GA used for the evolutionary stage of EPR-MOGA is OPTIMOGA (Giustolisi et al. 2004), which is employed to select the set of independent variables (X_k) that must form the model structure. Finally, Giustolisi and Savic (2006) accurately describe how OPTIMOGA is applied in EPR in order to conduct the structural identification of models.

Modeling assumptions

One of the advantages of EPR is that this approach does not need a strict preliminary selection of input. Before starting the procedure, a set of potential input is selected, afterwards during the evolutionary stage, the methodology selects those input which constitute the models corresponding to the best trade-off among the three objective functions. Theoretically, a large and complex set of input can be assumed, but a preliminary reduction of this set is guessed, in order to avoid a potentially huge space of search, which can bias the efficiency of the methodology. In this case the efficiency is meant both in terms of run time of the algorithm and in terms of probability of selecting good models. For the case under investigation, the preliminary choice is based on the hydrological investigation made by other authors (Fiorillo and Guadagno 2000) for the same case study. However, it is noteworthy that analysis based on hydrological investigations never retuned models able to pragmatically predict a reactivation, or a threshold scenario after which the probability of a landslide reactivation is high.

In order to set a pragmatic output to the model, able to discriminate between landslide occurrence or not, it is assumed that in case there is a landslide reactivation, the output of the model is 1, if no reactivation is observed, the output is 0. This implies a threshold between reactivation or not, which is 0.5. In fact, the actual output of the candidate models is brutally rounded to 1 if equal or higher than 0.5, while it is rounded to zero if lower than 0.5. This assumption is actually strong and somehow questionable, since it associates different output with the same meaning. However, an analysis of the output in terms of distribution of the values between 0 and 1 will be later introduced and discussed to show how reliable results can be.

The total investigated period, in terms of rainfall, ranges between 1924 and 2009. Two different analyses are undergone on the aforementioned time window. These two analyses differ in terms of hypotheses on the input to the methodology. For both the attempts, data were split into two different sets: rainfall and reactivation data ranging between 1924 and 1985 are used as training set, i.e., the set of data used in order to identify the models. Data ranging between 1986 and 2009 are used to test the models, i.e., these data are never used for model construction, they are just used to test the already identify models, in order to evaluate their generalization abilities. Splitting data assuming a training set comprising two third of the dataset and the test set comprising one third of the data set is a common practice in system identification procedure (Ljung 1987). It is noteworthy that eight reactivations occur in the training set, while three are in the test set.

The earlier attempt involves cumulative rainfall as follows: 15, 60, 120, and 500 days evaluated at 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, and 120 days before the reactivation occurrence. Shorter cumulative intervals are neglected, since Petacciato landslide is unlikely influenced by short-time cumulative rainfall. All the reactivation took place during the period January to March, therefore if no reactivation occurs, the aforementioned 15, 60, 120, and 500 days cumulative rainfall are evaluated at 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, and 120 days before April 1. This date is chosen since no landslide activations were ever observed after April 1. This attempt aims at investigating if long antecedent cumulative rainfall heights can affect the landslide-triggering process.

In the latter attempt, the input to the model is accounted for lag times before the potential reactivation which are shorter than for the earlier case. In particular, cumulative 20, 30, 45, 60, 75, 90, 120, and 150-day rainfall heights are considered at 0, 5, 10, 20, and 30-day lag before the reactivation date or before April 1. This case is supposed to investigate if the potential reactivations can be related to cumulative rainfall up to 1 month before the event.

It is noteworthy that several other attempts were made in order to test further input combinations, both in terms of lag time before the landslide and in terms of cumulative rainfall heights. However, none but the presented returned acceptable or interesting results, therefore they are herein omitted.

Further assumptions made on EPR-MOGA follow. The set of candidate exponents of Eq. 1 is given as {−2, −1, −0.5, 0, 0.5, 1, 2}. This set is chosen according to the criterion that exponent 0 is useful in order to allow the methodology to unselect a nonuseful input; while the exponent 0.5, 1, and 2 provide in the order a square root-dumping effect, a linear effect, and a power 2 amplification effect to the selected input. If these exponents are negative, they provide the same effect on the inverse of the input.

The choice of involving negative exponents can be questionable, since it can potentially return more complex equations, which are characterized by a good interpolation attitude but poor physical meaning. However, during the several attempts made in order to avoid their use, it was realized that if not involved, the quality of results in terms of fitness of prediction was extremely poor. This may be due to the presence of further unknown triggering input which conditions the landslide. These unknown inputs are likely to be described by using negative exponents for cumulative rainfall values. In order to round the results to 1, if landslide occurs, or to 0, if no landslide occurs, Eq. 1 was assumed as follows:

$$Y = \text{round}\left(a_0 + \sum_{j=1}^{m} a_j \cdot (X_1)^{\text{ES}(j,1)} \cdot \ldots \cdot (X_k)^{\text{ES}(j,k)}\right) \quad (2)$$

where the rounding is made by the introduced *round* operator, while the expression which is argument of *round* will be rounded assuming the 0.5 value as threshold, according to the aforementioned criterion.

Finally, based on the experience made on EPR-MOGA, the maximum number of generations of the GA was assumed 54,600 and 140,000 respectively for the earlier and later attempts. Such number of generations is related to the number of potential input and combinations, and then the latter attempt characterized by a higher number of combinations needed for a higher number of generations. It was also observed that for both cases, increasing the number of generations did not return any reasonable benefit in terms of results. Further minor assumptions are the maximum number of monomial terms of Eq. 1 which was set to 3 for the earlier attempt and 5 for the latter, and for both cases the bias term a_0 (see Eq. 1) is a priori discarded. Also the last assumption was consequence of further attempts involving a_0, which did not give any benefit.

Results and discussion

Both the two described attempts returned a Pareto set (Van Veldhuizen and Lamont 2000) of models which is made of six models each. A common dimension of the Pareto front is here casual, however such dimension allows for making cross comparison of models on the same Pareto set. Moreover, it may be attempted a cross comparison of models between the two Pareto sets.

The earlier attempt returned the following set of equations, which are here ordered according to their decreasing structural complexity as returned by EPR-MOGA.

$$R = \mathrm{round}\left(76.8456 \cdot P15_{60}^{-0.5} P60_{120}^{2} P60_{80}^{-0.5} P60_{10} P120_{100}^{-2} P120_{30}^{-0.5} P120^{2} P500^{-2}\right) \tag{3}$$

$$R = \mathrm{round}\left(-2.8574 \cdot P60_{70}^{-0.5} + 46.2043 \cdot P60^{2} P60_{80}^{-0.5} P60_{10}^{0.5} P120_{100}^{-2} P500_{40}^{-0.5}\right) \tag{4}$$

$$R = \mathrm{round}\left(0.82143 \cdot P60_{120}^{2} P60_{80}^{-1} P60_{10} P120_{80}^{-2}\right) \tag{5}$$

$$R = \mathrm{round}\left(0.061909 \cdot P15_{10}^{2} P60 P120_{100}^{2}\right) \tag{6}$$

$$R = \mathrm{round}\left(0.0003585 \cdot P60_{10}^{2} P500_{110}^{-0.5}\right) \tag{7}$$

$$R = \mathrm{round}\left(1.1322 \cdot 10^{-5} \cdot P60_{10}^{2}\right) \tag{8}$$

The latter attempt returned the following Pareto set of models.

$$R = \mathrm{round}\left(2.3712 \cdot P20_{30}^{0.5} P30^{2} P30_{30}^{2} P45_{5}^{-2} P45_{20}^{-1} P60^{-0.5} P60_{10} P90_{10}^{-2} P90_{20} P90_{30}^{2}\right) \tag{9}$$

$$R = \mathrm{round}\left(1.2715 \cdot P20_{30}^{0.5} P30^{2} P30_{30} P45_{5}^{-2} P45_{20}^{-1} P90_{10}^{-1} P90_{20} P90_{30}^{2}\right) \tag{10}$$

$$R = \mathit{round}\left(2.378 \cdot P20_{30}^{0.5} P30^{2} P30_{30}^{2} P45_{5}^{-2} P45_{20}^{-1} P90_{20}^{2}\right) \tag{11}$$

$$R = \mathrm{round}\left(0.90432 \cdot P30^{2} P45_{5}^{-2} P45_{20}^{2} P90_{30}\right) \tag{12}$$

$$R = \mathrm{round}\left(1.0818 \cdot P45_{10}^{0.5} P60_{10}^{2}\right) \tag{13}$$

$$R = \mathrm{round}\left(1.404 \cdot P30_{30}^{2}\right) \tag{14}$$

where R means reactivation or not, and it could be 0 or 1, and Pxx_{yy} represents the rainfall cumulated on xx-day interval evaluated yy days before the reactivation or before April 1, when no subscript shows, it means yy equals 0.

Among the models returned by the two attempts, Eqs. 3 and 9 returned the best results, both on the training set, 1926–1986, and on the test set, 1986–2009. In particular looking at the test set, which is constituted by 23 years of data never involved during the model identification stage, Eqs. 3 and 9 never return false-positive reactivation. Moreover, both the models are able to correctly forecast 1991 and 2009 reactivations. Looking at the training set, which is made up of 60 years of data, both the models perform reasonably well. Both of them never return false positive, while Eq. 3 fails to forecast 1953 reactivation. Equation 9 is able to forecast all the reactivations of the training set.

It is also very interesting to observe that 1996 reactivation, part of the test set, is never forecasted by both models. This is quite interesting, since it occurred mid-March, which is later than the other events with respect to the rainy season, and after a nonexceptionally rainy period. In fact it can be considered an anomalous reactivation with small displacement and damages. In summary, Eq. 3 was able to forecast all the reactivations but 1996 and 1953, while it is really remarkable that it never returned a false positive. Equation 9 was able to forecast all the reactivations events, but 1996, even in this case no false positive are returned. From a structural point of view, it is interesting to observe that both the chosen models are monomial, even if they involve several inputs. This somehow confirms the complex relation existing between rainfall amount and reactivation for Petacciato landslide.

A discussion about the involved input is not easy and can be physically questionable (Giustolisi et al. 2008). Indeed, some inputs can actually be related to the physics of the problem, while others are purely interpolative. However, the availability of multiple models can somehow hint which terms are likely to have a physical meaning. In this case, *P*60 is common to almost all the models of both the Pareto fronts, this may indicate that cumulative value has some role for landslide reactivations. Similarly, *P*30 and *P*45 are common to the largest part of the models returned by the latter attempt. The structure of the equations can be hardly related to a physical interpretation. Indeed, the complexity of the structure indicates a likely complex relationship between rainfall and landslide. However, it is also interesting to observe that EPR selected just one polynomial structure, made of two terms (see Eq. 4), even if it was allowed to search for structures up to five terms. This means that input terms of those equations may be quite significative for the studied phenomenon. The negative exponents of the rainfall variables are a further critical point, since it is quite hard to provide them with a physical meaning. However, it is interesting that when a different set of exponents, i.e., just positive exponent, was assumed, the quality of results dramatically decreased. A further interesting point is the lag *yy* of the cumulative rainfall, accounted by the model. These lags may be related to a potential delay of response to rainfall.

In summary, a physical interpretation of the models is very challenging even if the main rainfall components *P*60, *P*30, and *P*45 have been identified as relevant. On the other hand, here the main focus of this model is to provide a relatively simple equation which can reasonably provide reactivation forecasting, when rainfall scenarios are assumed.

Analysis of the expression to be rounded

Due to the pragmatic criterion used for rounding, it is quite important to assess how uncertain can be the reactivation/no reactivation prediction. This problem is here faced evaluating how far the results of the expression to be rounded of Eqs. 3 and 9 are from 0.5. In word, a result which is close to 0.5 is quite uncertain, while values close to 0 for no reactivation or to 1 for reactivation are assumed reasonably certain. This analysis is solely undergone for the aforementioned better performing equations.

Figure 8 presents the analysis of results for Eq. 3. It is envisaged that for those years when no reactivation occurs, the values of the expression to be rounded to 0 are almost always lower than 0.2 and just 5 years exceed 0.4, being anyway lower than 0.45. This result is interesting since the risk of forecasting a reactivation when cumulative rainfall values are not exceptional is quite unlikely. For those years characterized by a reactivation, the situation is not as sharp as for no reactivation. Indeed, just four values out of nine are higher than 0.7, note that 1953 and 1996 are excluded since they are not forecasted by the model. The remaining five values range between 0.5 and 0.6, whereas 1960 reactivation is characterized by a value 0.51 and should be really uncertain. Anyway, being the no reactivation years very low, this implies that a sharp separation between forecasted reactivation and no reactivation exists. Therefore even those reactivation events characterized by values which are close to 0.5 can be assumed reasonably certain. It is also interesting to observe that 1991 and 2009 events, which belong to the test set of data, are characterized by values close to 0.8 and then very well separated by the no reactivation values.

Figure 9 shows the analysis of results for Eq. 9. Also this case presents very low values when no reactivation is forecasted, rarely exceeding 0.25. This should reasonably prevent from forecasting a false positive. Looking those values corresponding to reactivations, here they are 7 times out of 10 higher than 0.6, while just 1932, 1936 and 1953 reactivations range between 0.53 and 0.54. However, similar to Eq. 3, also this case shows a sharp separation between no reactivations and reactivations. It is also interesting to observe that Eq. 9 has a good performance on the test set, whereas 1991 and 2009 events are predicted with values higher than 0.65.

It is noteworthy that comparing the values of Eqs. 3 and 9, the same reactivations are denoted by different values. For instance, 1960 reactivation is characterized by a value 0.51 for Eq. 3, while Eq. 9 returns 0.86. In general, Eq. 9 tends to return values that are averagely closer to 1, for reactivations, than Eq. 3. This kind of behavior is not easily interpretable, in fact on the one hand it can be related to a better choice both of the input and of

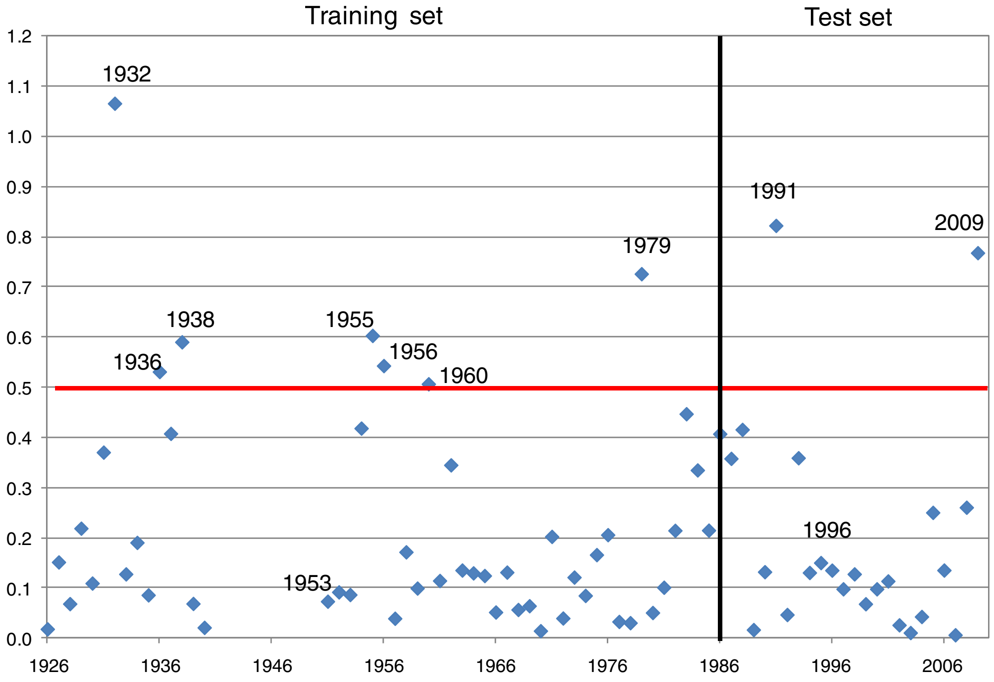


Fig. 8 Analysis of results of Eq. 3 before rounding, the vertical black line denotes the bound between training and test set, while the horizontal red line denotes the bound between reactivation and no-reactivation

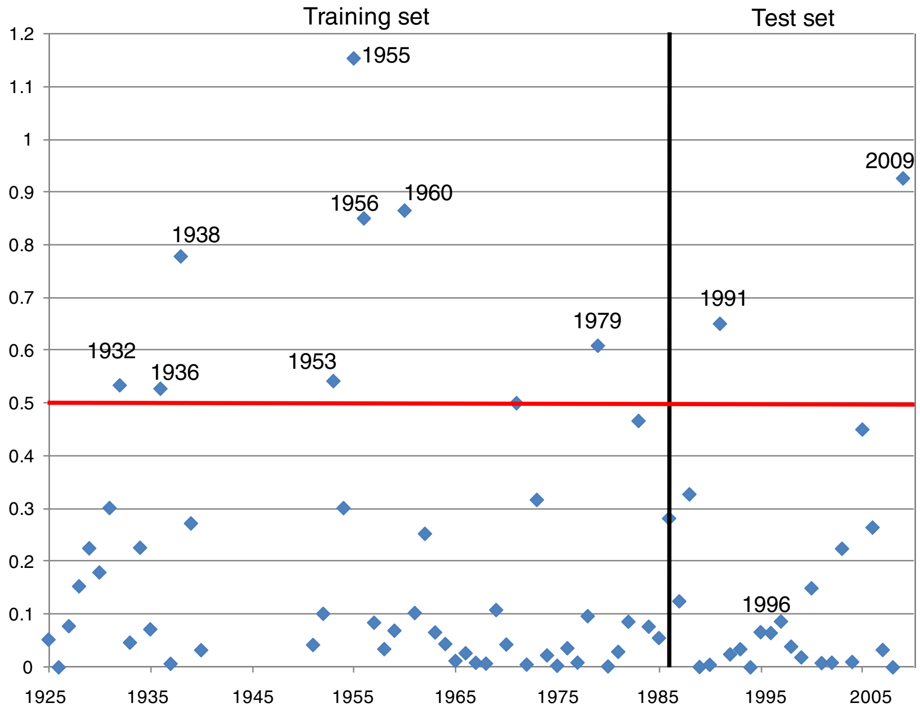


Fig. 9 Analysis of results of Eq. 9 before rounding, the vertical black line denotes the bound between training and test set, while the horizontal red line denotes the bound between reactivation and no-reactivation

the equation structure. On the other hand, this behavior may be due to a better interpolation capacity of Eq. 9 than Eq. 3. However, these interpretations can be speculative since both the models are purely data driven.

Conclusions

An analysis of landslide reactivation for the site of Petacciato, south Italy is here undergone based on a data-driven hydrological analysis. The site of Petacciato is characterized by the presence of a deep-seated landslide, which periodically reactivates, damaging national road and railway infrastructures as well as part of the town of Petacciato. In particular, the landslide reactivated 11 times between 1932 and 2009. During the same period and earlier, the time series of daily rainfall record is available from the site of Termoli, which is close to Petacciato slope. For this reason, it was attempted a data-driven analysis of the relationship between cumulative rainfall values and landslide reactivations aimed at identifying a reasonably simple and pragmatic alert model.

The approach herein presented is based on the evolutionary polynomial regression technique, and it returned promising results. Indeed, starting from slightly different assumptions, two attempts of model identifications are presented. Both of them returned encouraging results, whereas no false-positive reactivations are ever forecasted, while almost all the reactivation events were forecasted by the chosen models.

EPR shows the advantages of using an evolutionary data-driven approach, which typically consists in the identification of an explicit equation starting from measured data, which is of particular use for management purposes even for practitioners. Moreover, EPR was able to identify a relationship between rainfall data and landslide reactivation with a reasonable accuracy, whereas a classic hydrological analysis did not return a well-defined relationship. This is due to the nonlinear relation between rainfall and reactivations and to the influence of multiple variables conditioning the reactivations. However, EPR, like all the evolutionary data-driven algorithms, return an equation which cannot be generalized to other landslides and the accuracy of results can be related to the quality of data and to their quantity.

Anyway, the two chosen equations allow for a prediction of landslide reactivation based just on cumulative rainfall, which are easily and cheaply reachable data. In fact, both the models allow for alerting of a potential landslide reactivation starting from rainfall measures. In this case, the rainfall time series are measured by a rain gauge of the former National Hydrographic Bureau and are freely available on the internet at the address http://www.annali.apat.it/site/it-IT/. It is also important to emphasize that the identified models are relatively structurally simple and allow to be used with forecasted or hypothetic rainfall scenarios.

Moreover, both the models allow for the construction of hypothetical scenarios based on forecasted rainfall heights. The two equations differ for the assumption on the input: while the earlier involves long-term rainfall height lagged up to 120 days before the reactivation event, the latter involves cumulative rainfall up to 90 days, lagged 30 days before the event at most. It is not possible to state that Eq. 3 behave definitively better than Eq. 9 or vice versa. However it is possible to make a cross comparison between the models aimed at understanding if there are common features among the inputs.

A discussion about the physical interpretation of the results is partially questionable, since some input can be related to the physics of the problem, while others are purely interpolative. This is probably due to the complexity of the system, a really large deep-seated landslide, whereas the identification of the main variables is challenging. Therefore the interpolative terms, like those one having negative exponents, are likely representative of some extra unknown input. On the one hand, this is one of the major drawbacks of this kind of approaches, on the other hand we have to emphasize that such models are particularly useful for an alerting system scenario. However, the main focus of this model is to provide a relatively simple equation which can reasonably forecast reactivation occurrence, assuming rainfall scenarios. The identified models are primarily aimed at forecasting a potential reactivation of the landside based on simulation of rainfall scenarios or on actually measured data. The meaning of the input and the structure of the equation can be commented and analyzed, but very carefully. For instance negative exponents are hardly interpretable, under the umbrella of physics. In word, those input which are common through the equations returned by the methodology are likely to have some physical meanings, since they are related to reasonably good results, when the structural complexity of the models varies. On this premise, cumulative rainfall of 30, 45, and 60 days are assumed as key rainfall variables of the models.

Finally, the applied data-driven approach proved very effective at the identification of an alert model, since the particular investigated landslide is not easily correlated to particular rainfall events or sequence of the events. This is a deep-seated landslide, which reactivates without appreciably modifying the slope, thus allowing for a direct comparison of the hydrological conditions which trigger the landslide. Therefore it is here shown how a data-driven evolutionary technique can support the management of geomorphological risk of complicated scenarios.

Acknowledgments

Rainfall data have been provided by (ex) Servizio Idrografico e Mareografico Nazionale of Pescara (1924–1998 period) and Centro Funzionale del Servizio per la Protezione Civile of Regione Molise (1999–2009 period).

References

Angeli M.G. (1985) The role of anisotropic permeability on slopes instability conditions. In Proc. XI Int. Conf. of Int. Soc. Soil Mech. Found. Eng. S. Francisco, 4:2059–2063

Bäck T, Hammel U, Schwefel HP (1997) Evolutionary computation: comments on the history and current state. IEEE Trans Evol Comput 1(1):3–17. doi:10.1109/4235.585888

Bai SB, Wang J, Lu GN, Zhou PG, Hou SS, Xu SN (2009) GIS-based and data-driven bivariate landslide-susceptibility mapping in the Three Gorges Area, China. Pedosphere 19(1):14–20. doi:10.1016/S1002-0160(08)60079-X

Benjamin JR, Cornell CA (1960) Probability, statistics and decisions for civil engineering. McGraw-Hill, New York

Cancelli A, Pellegrini M, Tonnetti G (1984) Geological features of landslides along the Adriatic coast (Central Italy). In Proc. Int. Symp. on Landslides, Toronto, 2:7–17

Casnedi R, Crescenti U, D'Amato C, Mostardini F, Rossi U (1981) Il Plio-Pleistocene del sottosuolo molisano (Plio-Pleistocene and Molise underground). Geol. Rom. XX:1–42

Chang TC, Chien YH (2007) The application of genetic algorithm in debris flow prediction. Environ Geol 53:339–347. doi:10.1007/s00254-007-0649-2, 2007

Coello Coello CA (1999) A comprehensive survey of evolutionary-based multiobjective optimization techniques. Knowl Inform Syst 1(3):269–308

Cotecchia V (2006) The Second Hans Cloos Lecture. Experience drawn from the great Ancona landslide of 1982. Bull Eng Geol Environ 45:1–41. doi:10.1007/s10064-005-0024-z

Fiorillo F (2003) Geological features and landslide mechanisms of an unstable coastal slope (Petacciato, Italy). Eng Geol 67(3–4):255–267. doi:10.1016/S0013-7952(02)00184-9

Fiorillo F, Guadagno FM (2000) Analysis of rainfall patterns triggering reactivations of a large landslide in Pleistocene clay in Molise (Italy). In Proc. of the 8th Int. Symp. on Landslides, Thomas Telford, 2, 553–557

First Italian Workshop on Landslides (2009) Rainfall-induced landslides: mechanisms, monitoring techniques and nowcasting models for early warning systems. Naples, 8–10 June 2009

Flentje P, Stirling D, Chowdhury R (2007) Landslide susceptibility and hazard derived from a landslide inventory using data mining—an Australian case study. In: Proceedings of the First North American Landslide Conference, Landslides and Society: Integrated Science, Engineering, Management and Mitigation, Vail, Colorado 3–8 June 2007

Giustolisi O, Savic DA (2006) A symbolic data-driven technique based on evolutionary polynomial regression. J Hydroinform 3(8):207–222. doi:10.2166/hydro.2006.020

Giustolisi O, Savic DA (2009) Advances in data-driven analyses and modelling using EPR-MOGA. J Hydroinform 11(3–4):225–236. doi:10.2166/hydro.2009.017

Giustolisi O, Doglioni A, Laucelli D, Savic DA (2004) A proposal for an effective multiobjective non-dominated genetic algorithm: the OPTimised Multi-Objective Genetic Algorithm, OPTIMOGA, Report 2004/07, School of Engineering Computer Science and Mathematics, Centre for Water Systems, University of Exeter

Giustolisi O, Doglioni A, Savic DA, di Pierro F (2008) An evolutionary multi-objective strategy for the effective management of groundwater resources. Water Resour Res 44:W01403. doi:10.1029/2006WR005359

Glade T, Crozier M, Smith P (2000) Applying probability determination to refine landslide-triggering rainfall thresholds using an empirical "Antecedent Daily Rainfall Model". Pure Appl Geophys 157(6–8):1059–1079. doi:10.1007/s000240050017

Goldberg DE (1989) Genetic algorithms in search, optimization and machine learning. Addison-Wesley, Reading

Guerricchio A, Melidoro G (1996) Rischi da grandi frane nella fascia costiera adriatica (Large Landslide Hazard along the Adriatic coastline). In Proc. of Prevention of Hydrogeological Hazard: the role of scientific research, Italian National Research Council (CNR,) 5–7 November 1996, Alba (Italy), 1:317–330

Guerricchio A, Melidoro G, Simeone V (1996) Le grandi frane di Petacciato sul versante costiero adriatico (The large landslides of Petacciato on the Adriatic coastline). Mem Soc Geol Ital 51(2):607–632

Hodge RA, Freeze RA (1977) Groundwater flow system and slope stability. Can Geotech J 14(4):466–477. doi:10.1139/t77-049

Iverson RM (2000) Landslide triggering by rain infiltration. Water Resour Res 36 (7):1897–1910. doi:10.1029/2000WR900090

Kawabata D, Bandibas J (2009) Landslide susceptibility mapping using geological data, a DEM from ASTER images and an Artificial Neural Network (ANN). Geomorphology 113(1–2):97–109. doi:10.1016/j.geomorph.2009.06.006

Ljung L (1987) System identification: theory for the user. Prentice-Hall, Upper Saddle River, p 672

Lu P, Rosenbaum MS (2003) Artificial neural networks and grey systems for the prediction of slope stability. Nat Hazards 30(3):383–398. doi:10.1023/B:NHAZ.0000007168.00673.27

Melchiorre C, Matteucci M, Azzoni A, Zanchi A (2008) Artificial neural networks and cluster analysis in landslide susceptibility zonation. Geomorphology 94(3–4):379–400. doi:10.1016/j.geomorph.2006.10.035

Patacca E, Scandone P (2007) Geology of the Southern Apennines. In: Mazzotti A, Patacca E, Scandone P (eds) Results of the CROP Project, Sub-project CROP-04 So, Spec. Issue 7(2007), 75–119

Sengupta A, Gupta S, Anbarasu K (2010) Rainfall thresholds for the initiation of landslide at Lanta Khola in north Sikkim, India. Nat Hazards 52(1):31–42. doi:10.1007/s11069-009-9352-9

Tatard L, Grasso JR, Helmstetter A, Garambois S (2010) Characterization and comparison of landslide triggering in different tectonic and climatic settings. J Geophys Res 115: F04040. doi:10.1029/2009JF001624

Van Asch ThWJ, Buma J, Van Beek LPH (1999) A view on some hydrological triggering systems in landslides. Geomorphology 30(1–2):25–32. doi:10.1016/S0169-555X(99)00042-2

Van Veldhuizen DA, Lamont GB (2000) Multiobjective evolutionary algorithms analyzing the state-of-the-art. Evol Comput 8(2):125–144. doi:10.1162/106365600568158

A. Doglioni (✉) · V. Simeone
Engineering Faculty of Taranto,
Technical University of Bari,
V. le del Turismo n.8, 74123 Taranto, Italy
e-mail: a.doglioni@poliba.it

F. Fiorillo · F. M. Guadagno
Dept. of Environmental and Geological Studies,
University of Sannio,
Via dei Mulini 59/A, 82100 Benevento, Italy