Reliability and effectiveness of early warning systems for natural hazards: Concept and application to debris flow warning

Martina Sättele [a,*], Michael Bründl [a], Daniel Straub [b]

[a] WSL Institute for Snow and Avalanche Research SLF, Avalanche Dynamics and Risk Management, Flüelastrasse 11, 7260 Davos Dorf, Switzerland
[b] Technische Universität München, Engineering Risk Analysis Group, Arcisstrasse 21, 80 333 Munich, Germany

A B S T R A C T

Early Warning Systems (EWS) are increasingly applied to mitigate the risks posed by natural hazards. To compare the effect of EWS with alternative risk reduction measures and to optimize their design and operation, their reliability and effectiveness must be quantified. In the present contribution, a framework approach to the evaluation of threshold-based EWS for natural hazards is presented. The system reliability is classically represented by the Probability of Detection (POD) and Probability of False Alarms (PFA). We demonstrate how the EWS effectiveness, which is a measure of risk reduction, can be formulated as a function of POD and PFA. To model the EWS and compute the reliability, we develop a framework based on Bayesian Networks, which is further extended to a decision graph, facilitating the optimization of the warning system. In a case study, the framework is applied to the assessment of an existing debris flow EWS. The application demonstrates the potential of the framework for identifying the important factors influencing the effectiveness of the EWS and determining optimal warning strategies and system configurations.

1. Introduction

Early Warning Systems (EWS) are frequently applied as cost-effective risk mitigation measures against natural hazards, which provide timely information on future or ongoing events to reduce loss of life and damages [1]. In contrast to structural protection measures such as dams, galleries and rock fall nets, EWS are cheaper, have shorter installation time and have lower impact on the environment [2]. During the last decade, EWS have undergone a rapid technical development and are today frequently implemented as mitigation measures in an integrated risk management approach [3]. To compare the economic efficiency of mitigation measures and to identify the optimal risk reduction strategy, cost-benefit analyses are conducted. Following the standard convention, risk is defined as the expected value of adverse consequences [4]. The risk associated with an object i and scenario j is [5]

$$R_{ij} = p_j \times pe_{ij} \times v_{ij} \times A_i \qquad (1)$$

where p_j is the probability of occurrence of a scenario j, pe_{ij} is the presence probability of object i in scenario j, v_{ij} is the vulnerability of object i in scenario j and A_i the value of object i. The overall risk R is

evaluated by summing or integrating over all possible scenarios and exposed objects

$$R = \sum_{j=1}^{n_{scen}} \sum_{i=1}^{n_{obj}} R_{ij} \qquad (2)$$

Existing guidelines recommend that the benefits achieved due to reduced risk are compared against the costs induced to develop and maintain the measure [6,7]. Detailed guidelines for calculating the effectiveness, i.e. the achieved risk reduction, are available for structural risk mitigation measures [8]. However, for natural hazard EWS such guidelines and procedures for quantifying the effectiveness are lacking.

The reliability of EWS for natural hazards has been investigated in the past. It is generally accepted that an evaluation of EWS must include both the benefits of risk reduction and the negative consequences of missed events and false alarms [9–12]. A first approach for quantification of the reliability of a flood EWS was published by Krzysztofowicz et al. [13]. Following earlier work carried out in other areas e.g. [14], they quantify the reliability of a flood EWS following the concept of signal detection theory through the Probability of Detection (POD) and the Probability of False Alarm (PFA). In more recent case studies, the reliability of flood EWS and their forecasting performance are likewise expressed in terms of hits, missed events and false alarms for different thresholds [15,16]. Similar concepts are used for the

* Corresponding author. Tel.: +41 81 417 03 61.
  E-mail address: martina.saettele@slf.ch (M. Sättele).

assessment of EWS operated for other natural hazard processes. E.g., Simmons and Sutter [17] express the Tornado warning performance of the U.S. National Weather Service in terms of number of detected events and the ratio between false alarms and warnings, and Rheinberger [18] models the performance of avalanche warnings through POD, PFA and additional measures. As shown by Paté-Cornell [14], such an analysis is ideally based on detailed models of the response to false warnings, facilitating the identification of an optimal trade-off between POD and PFA. In addition to POD and PFA, the reliability of EWS depends on the probability of technical failures of system components. Bründl and Heil [19] assessed the technical reliability of the Swiss avalanche EWS in a case study. They conducted a fault tree analysis to identify the most critical system components but concluded that the method is not sufficient to cover the entire complexity of EWS. In a subsequent study, Sturny and Bründl [20] apply Bayesian Networks (BN) to assess the technical reliability of a glacier lake EWS. In Sättele et al. [21], we propose an enhanced BN to evaluate the reliability of a debris flow EWS, which computes POD and PFA including the technical reliability of the system components.

In this contribution, a first step towards a generic framework for quantifying the effectiveness of EWS for natural hazards is presented. EWS can be classified into alarm, warning and forecasting systems [22]. These classes differ in their degree of system automation. Alarm systems detect ongoing hazard events, have short lead times and include fully automated threshold-based decisions. Warning and forecasting systems monitor precursors to predict events and are only partly automated including model-based human decisions. In this contribution, we limit ourselves to alarm systems, and show how their effectiveness can be quantified from their POD and PFA using a BN. We first define the terms reliability and effectiveness in the context of alarm systems, before we propose a framework BN and an associated Decision Graph (DG). In a case study, we apply the framework on an existing debris flow threshold-based alarm system to find the optimal system configuration, to identify the main factors influencing the system effectiveness and to demonstrate the applicability of the novel framework approach.

2. Reliability of alarm systems for natural hazards

Following [23], reliability is defined as the "ability of an item to fulfill a required function under stated conditions for a stated period of time". An EWS for natural hazards fulfills its designated function if it detects all hazard events in a timely manner, transfers the warning to the effected persons and leads to measures that avoid damage and loss of life.

This requires (a) that the system and its components are available and work perfectly, and (b) that the monitoring and the data interpretation units are able to perfectly distinguish between hazard events and background noise. The requirement (a), to which we refer as *technical reliability*, can be quantified using the classical methods for assessing the reliability of technical systems, including fault trees, bow-tie models, failure mode and effective analysis [24,25]. More recently, BN have been applied as a flexible and powerful alternative to these models [26]. The requirement (b), to which we refer as *inherent system reliability*, is quantified through POD and PFA using the concepts of signal detection theory [27], which has found applications in many field including medical testing [28,29] and non-destructive testing of technical systems [30,31]. In the context of alarm systems, one can define the POD and the PFA as

$$POD = E\left[\frac{\text{number of detected events}}{\text{number of events}}\right] \tag{3}$$

$$PFA = E\left[\frac{\text{number of days with false alarms}}{\text{number of event free days}}\right] \tag{4}$$

where $E[\cdot]$ is the expectation operator.

Note that the PFA must be defined using a reference unit, which is chosen here as days, but other temporal or spatial references can be appropriate. To ensure comparability, it is important to use the same unit consistency throughout all studies. Unfortunately, this is often overlooked and many studies do not even state the reference unit of the PFA.

POD and PFA are both influenced by the interpretation of the monitoring data. This is illustrated in Fig. 1, which shows the basic concepts of signal detection theory. The measured signal can be either due to a hazard event H or due to noise N. The decision to issue a warning is based on the threshold t. If the measured signal is larger than t, a warning is issued. With $f_{S|H}(s)$ being the conditional probability density function (PDF) of the signal S given a hazard event H, and $f_{S|\overline{H}}(s)$ being the conditional PDF of S given no hazard event $\overline{H}$, it is [27,32]

$$POD(t) = \int_t^\infty f_{S|H}(s)\,\mathrm{d}s \tag{5}$$

$$PFA(t) = \int_t^\infty f_{S|\overline{H}}(s)\,\mathrm{d}s \tag{6}$$

With increasing threshold t, both the POD and the PFA decreases. This dependence between the two is graphically embodied in the Receiver Operator Characteristic (ROC) curve, see Fig. 1. ROC curves summarize the reliability of EWS for varying thresholds. They graphically represent the system reliability as a trade-off between POD and PFA.

The overall system reliability as a combination of the technical reliability and the inherent system reliability is also expressed in terms of ROC curves. To this end, we compute the POD and PFA as the conditional probability of a warning given a hazard event, including the probability of system component failures. This is achieved by modeling both the inherent and the technical reliability jointly in a Bayesian network, as described in Section 4.

3. Effectiveness of alarm systems for natural hazards

It is commonly accepted that the effectiveness of a mitigation measure equates to the relative reduction of the overall risk [5,8]. We propose to calculate the effectiveness of an EWS E_w from R being the overall risk without the EWS and $R^{(W)}$ the risk with the EWS system installed

$$E_w = 1 - \frac{R^{(W)}}{R} \tag{7}$$

Both R and $R^{(W)}$ are evaluated according to Eqs. (1) and (2). EWS aim to generate information before a hazard event causes damage; they reduce risk primarily by mitigating the exposure probability pe_{ij} of persons and mobile objects i in a hazard scenario j, following Eq. (1). If the EWS provides sufficient lead time, the risk can be additionally reduced through the implementation of supplementary intervention measures such as e.g. mobile flood protection. Consequently, the effectiveness of EWS E_w is primarily a result of the reduced exposure probability pe_{ij}, but can also be due to other factors such as the vulnerability of object i in scenario j in Eq. (1).

In the present contribution, we focus on alarm systems with limited lead time, during which the only possible action is to reduce the presence probability from a value pe_{ij} without warning to a value $pe_{ij}^{(W)}$. Combining Eq. (7) with Eqs. (1) and (2), the

effectiveness for this case becomes

$$E_W = 1 - \frac{\sum_{j=1}^{n_{scen}} \sum_{i=1}^{n_{obj}} p_j \times pe_{ij}^{(W)} \times v_{ij} \times A_i}{\sum_{j=1}^{n_{scen}} \sum_{i=1}^{n_{obj}} p_j \times pe_{ij} \times v_{ij} \times A_i} \tag{8}$$

Most alarm systems are installed primarily to warn people. Thus, n_{obj} is the number of exposed people and it is reasonable to assume that the exposure probability is the same for different i, i.e. $pe_{ij} = pe_j$. Finally, we limit ourselves to a situation with only one relevant scenario $j = 1$, and the warning effectiveness then reduces to

$$E_w = 1 - \frac{p_j \times pe_j^{(W)} \times \sum_{i=1}^{n_{obj}} v_{ij} \times A_i}{p_j \times pe_j \times \sum_{i=1}^{n_{obj}} v_{ij} \times A_i}$$

$$= 1 - \frac{pe_j^{(W)}}{pe_j} \tag{9}$$

The alarm system reduces the exposure probability to $pe_j^{(W)}$. This reduction is equal to the probability that a warning is issued, transferred to the target persons and that the affected people comply with the warning. The former corresponds to the POD, the latter to the Probability of Compliance (POC). Therefore,

$$pe_j^{(W)} = pe_j(1 - POD \times POC) \tag{10}$$

Inserting in Eq. (9), the effectiveness becomes

$$E_W = POD \times POC \tag{11}$$

The POC, i.e. the degree to which warnings are followed in practice, is strongly dependent on the PFA. A high number of false alarms reduces the POC to an issued warning, due to a loss of trust that is known as the cry-wolf syndrome [33,34]. We calculate POC as a result of a basic compliance probability POC_0 and a compliance reduction factor due to false alarms RF(PFA)

$$POC = POC_0 \times RF(PFA) \tag{12}$$

For the case study, we estimate the general compliance rate $POC_0 = 0.95$ from traffic analyses [35,36]. One analysis investigated the behavior of pedestrians towards red lights and revealed that 5% ignore red-lights. The second analysis considered the behavior of cyclists, where about 7% ignore red lights. To estimate the compliance reduction factor due to false alarms RF(PFA) we adopt results from a case study that assessed the compliance frequency of students as a function of false alarms [37]. The resulting compliance frequencies (corresponding to our RF) at different levels of the False Alarm Ratio (FAR) are shown in Fig. 2, together with a fitted quadratic function

$$RF(FAR) = -0.34 \, FAR^2 - 0.66 \, FAR + 1 \tag{13}$$

To incorporate the effect of decreasing compliance for a given number of false alarms in the effectiveness of the alarm system, the FAR, which is defined as the ratio of false to correct alarms, is related to the PFA by
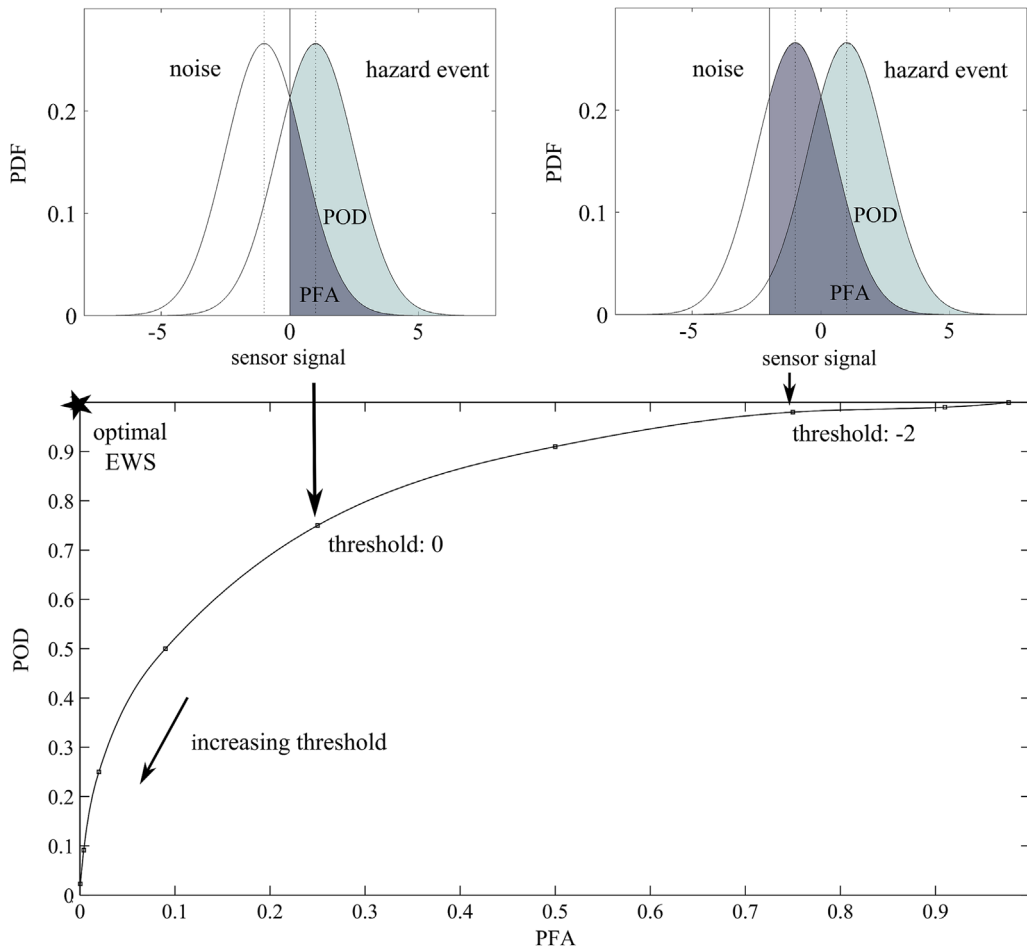
$$FAR = PFA \frac{Pr(\overline{H})}{Pr(A)} \tag{14}$$

Fig. 1. Conditional Probability Density Functions (PDFs) representing noise and a hazard event and the Probability of Detection (POD) and Probability of False Alarms (PFA) for two different thresholds (upper part). Correlated Receiver Operator Characteristics (ROC) curve for varying thresholds and the optimal performance of an EWS (lower part).
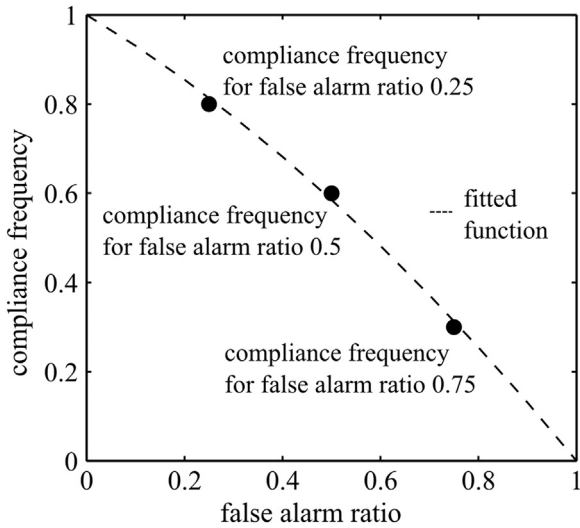
Fig. 2. Compliance frequency at different levels of False Alarm Ratio (FAR), according to [37].
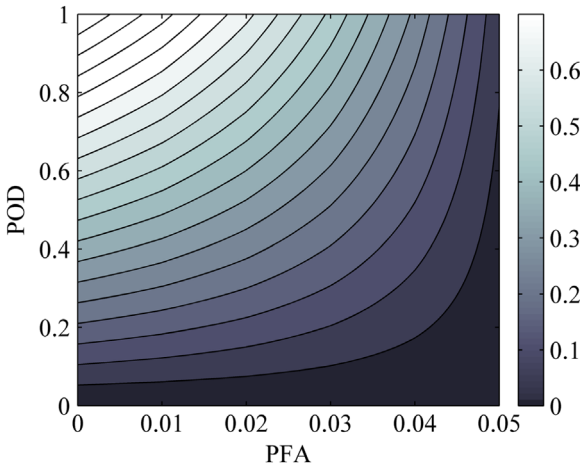

Fig. 3. Effectiveness as a function of Probability of Detection (POD) and Probability of False Alarms (PFA) for the case study.

where $\Pr(\overline{H})$ is the probability of no hazard event and $\Pr(A)$ is the probability of an alarm (both correct and false) on a given day. For the case study considered here, it is approximately $\Pr(\overline{H}) \approx 95\%$ and $\Pr(A) \approx 5\%$, therefore $FAR \approx 19\,PFA$. Combining Eqs. (11)–(14), we obtain the effectiveness as a function of POD and PFA, see also Fig. 3

$$E_w = \text{POD} \times 0.95 \left(-0.34\,\text{FAR}^2 - 0.66\,\text{FAR} + 1 \right)$$

$$= \text{POD} \times \left(0.95 - 116\,\text{PFA}^2 - 11.9\,\text{PFA} \right), \quad \text{PFA} \leq \frac{1}{19} \tag{15}$$

4. Bayesian network to quantify the system reliability and maximize the effectiveness

To probabilistically model the system reliability for varying thresholds of an existing debris flow alarm system and to identify the threshold combination that implies the optimal effectiveness, we design a BN and an associated DG. A BN is a graphical probabilistic model, in which each node represents a random variable and the arcs among the nodes characterize the stochastic dependence among these [38,39]. In many instances, the arcs can be constructed following the causal relations between the random variables [40]. To each node, a Conditional Probability Table (CPT) is attached, specifying the probability of the random variable conditional on its parent nodes. The BN facilitates the computation of the probability of any set of nodes conditional on observations of other nodes. BNs can be extended to DGs for decision making under uncertainty, whereby the strategy that maximizes the expected utility is sought [38,41]. DGs are essentially BNs augmented with decision and utility nodes, wherein the latter describe the preferences of the decision maker.

BN allow the incorporation of expert knowledge, can deal with rare data and are based on an intuitive modeling approach. In recent years BN have been applied frequently for environmental modeling and for the evaluation of natural hazard risks [42–44]. Applications of BN for modeling EWS are presented by Medina-Cetina and Nadim [45], who present a BN of a landslide EWS and apply it to determine optimal thresholds, and by Blaser et al. [46], who use BN to assess a Tsunami EWS in Sumatra.

Our framework BN to model the reliability of alarm systems for natural hazards is designed according to three main units of a EWS [47]: monitoring, data interpretation and information dissemination (Fig. 4). The monitoring unit is equipped with sensors, which continuously monitor the environment. In the data interpretation unit, the measured data are analyzed to detect irregularities and make the final warning decision. The information dissemination unit conveys the warning information to responsible authorities and finally to endangered persons and responsible authorities. This BN describes the causal chain from the event to the warning. Component failure nodes are included to model the technical reliability of the system dependent on the failure probabilities of different system components. The node "event indicated" represents the inherent system reliability as a function of the selected threshold. To compute the POD, the top node is set to the state "hazard event=true" and the BN is evaluated; the POD is then obtained as the probability of a warning. Likewise, the PFA is obtained by setting the top node to "hazard event=false". By varying the threshold, different combinations of POD and PFA are obtained, allowing the construction of the ROC curve.

By adding a utility node, the BN is extended to a DG, which can automatically identify the optimal warning threshold (Fig. 5). This is of particular use when multiple sensors are installed. In this case, thresholds must be set for all sensors and combination rules (logic operators) must be defined, e.g. that a warning is issued only if more than x sensors have a signal above their threshold. This
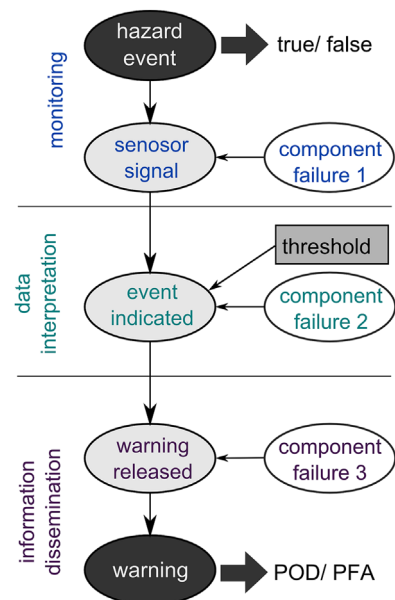

Fig. 4. Schematic framework of a Bayesian Network (BN) to model the reliability in terms of Probability of Detection (POD) and Probability of False Alarms (PFA) for alarm systems.

leads to a high-dimensional optimization problem, which can be effectively solved with the DG.

5. Case study: the Illgraben debris flow alarm system

The system under investigation is located at the Illgraben catchment in the western part of the Swiss Alps. The catchment ranges in elevation from 610 m a.s.l. to 2716 m a.s.l. and half of the catchment area ($\sim$4 km^2) is covered by bedrock and debris deposits. Due to the geological conditions there is a remarkably high occurrence rate of debris flows. A debris flow is a spontaneous fast-flowing mixture of water and solid particles, which typically consists of surges . In 2006, the Swiss Federal Institute for Forest, Snow and Landscape Research WSL designed an alarm system to protect local residents and tourists frequently crossing the catchment (Fig. 6). In the present case study, we assess the reliability and effectiveness of the existing Illgraben system, which is a typical fully-automated threshold-based system [48].

The monitoring unit includes five sensors that are located close to the release area to detect events in real-time. In the upper catchment, one single sensor, Geophone 1 (G1), continuously monitors ground vibrations. Further down in the catchment, some hundred meters below, two geophones, geophone 2 (G2) and geophone 3 (G3), measure ground vibrations and two radar devices, radar 1 (R1) and radar 2 (R2), measure the flow depth
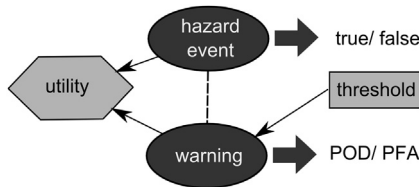
Fig. 5. Schematic Decision Graph (DG) to identify the optimal threshold combination that maximizes the alarm system effectiveness.

in the river bed. The upper G1 is controlled by one logger and the remaining four sensors are controlled by a second logger. The power at these remote locations is supplied via solar panels and batteries. The loggers build an interface between the monitoring unit and the data interpretation unit. If predefined threshold values in the data loggers are exceeded, a warning call is automatically activated via modem and transmitted to the valley. The incoming warning calls are forwarded via two communication devices to the information dissemination unit. To release the warning information to endangered persons in the catchment, three alarm stations are located close to three crossings of the streambed. Each station consists of an audible signal and a red light. The lead time of the system is determined by the velocity of the debris flow and the runtime between the lower sensor units and the upper crossing and is in the range between 5 and 15 min.

6. BN to model the reliability of the Illgraben alarm system

By applying the BN framework (Fig. 4) to the system sketch of the alarm system (Fig. 6), the BN depicted in Fig. 7 is obtained. The oval gray nodes in the BN represent the causal chain from the event to the warning. This chain can also be interpreted as the information flow. For each sensor, a local interpretation is made in node "event indicated", which is in state "true" only if the sensor signal exceeds the corresponding threshold. The information from sensors in the lower catchment (G2, G3, R1, R2) is merged in the node "warning issued 2", where it is decided whether or not to issue a warning, following the selected criterion defined in the node "decision criteria". The node "warning transmitted" is in state "true" if either of the two warnings is issued (OR connection). If the warning is transmitted, a warning is released at each of the three stations, given that no component failures occur. Therefore, the final node "warning" should in principle have four states 0,1,2,3, corresponding to the number of stations where warnings are released. However, to comply with the binary definition of POD
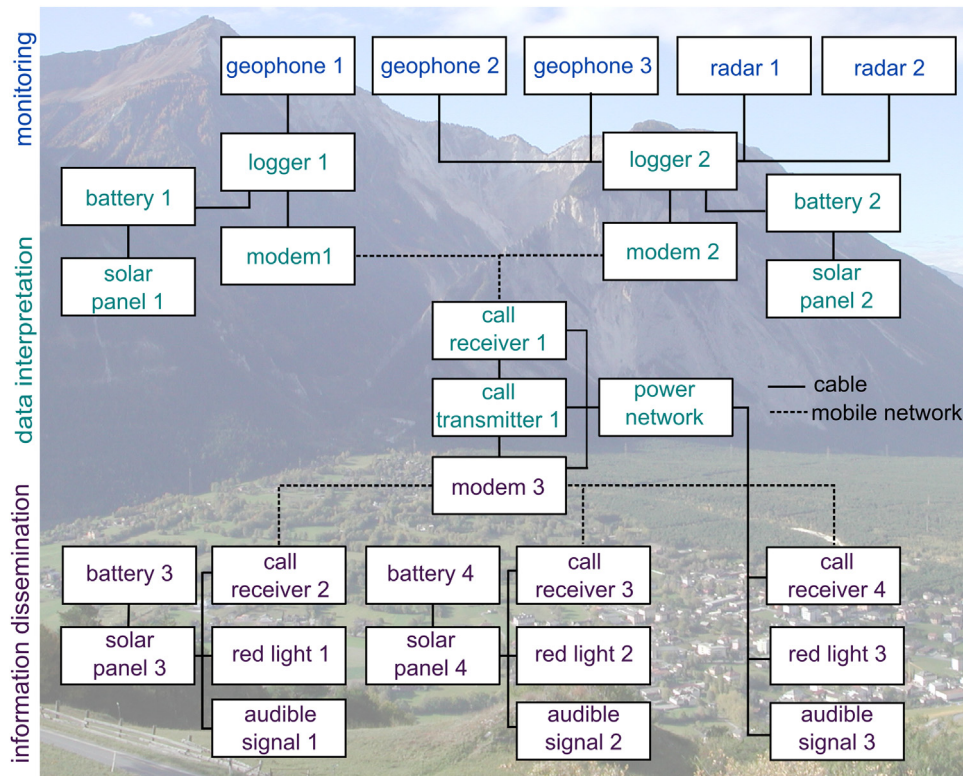
Fig. 6. The components of the Illgraben debris flow alarm system can be described in three main units: monitoring, data interpretation and information dissemination.
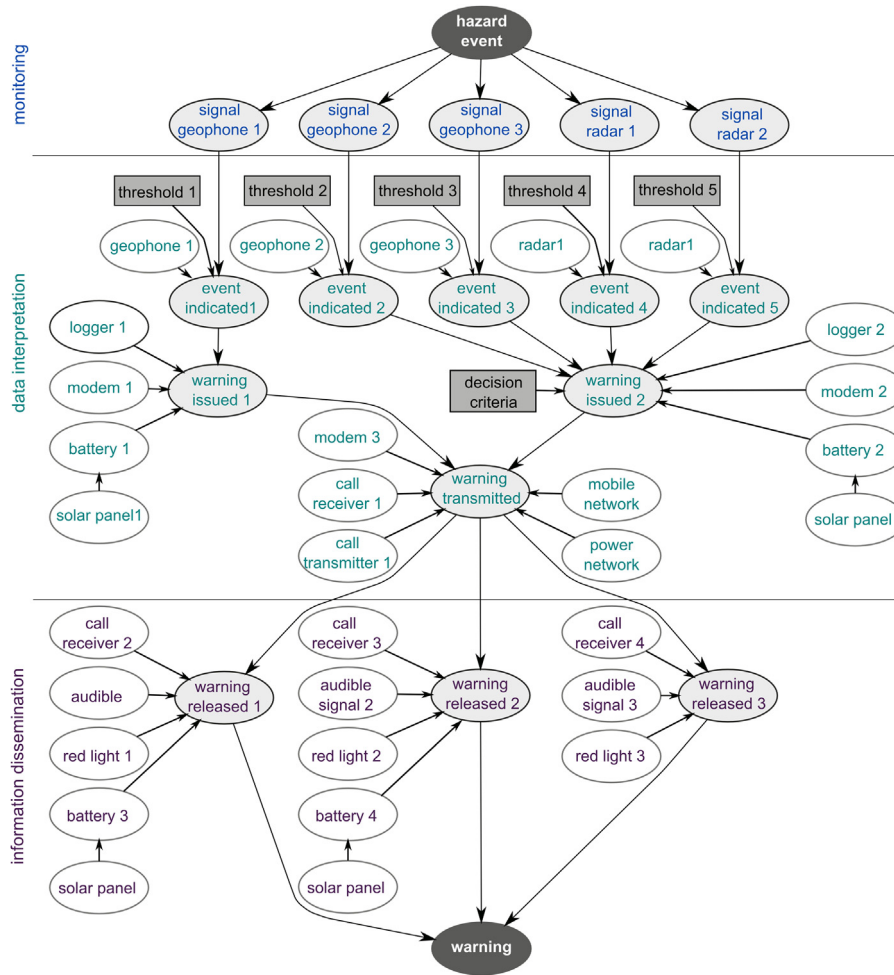
Fig. 7. Tailored BN to model the reliability of the Illgraben debris flow alarm system.

and PFA, this node has only two states "yes" and "no". To account for the number of warnings released, the conditional probability of "warning=yes" is 0.33 if two stations release a warning respectively 0.67 if only one station releases a warning. The BN is implemented with the free GeNIe software [49].

6.1. Technical reliability

The technical reliability of the system describes the probability that failures of technical system components (TSC) lead to a malfunctioning of the alarm system. The TSC are the white nodes in the BN of Fig. 7. They are modeled by binary random variables, with states "functioning" and "failed". Failures occur following a Poisson process, i.e. they occur randomly in time and independently of each other. The probability of a TSC failure at time $\Pr(F(t))$ is calculated as [50]

$$\Pr(F(t)) \approx \lambda \times E[T_r] \tag{16}$$

where λ is the failure rate of the TSC and $E[T_r]$ is the expected time it takes to detect and repair a failure. The approximation holds for small values of λ, i.e. for $\lambda \ll 1/E[T_r]$. In the Illgraben system, $E[T_r]$ is one day for all TSCs, because diagnosis tools are incorporated into the system to ensure that failures are detected within one day. If the failures cannot be repaired immediately, additional operational measures are taken to ensure detection of an event.

The failure rate λ of TSC includes both internal failures, with corresponding rate λ_{IF}, and failures caused by external influences, with rate λ_{EF}

$$\lambda = \lambda_{IF} + \lambda_{EF} \tag{17}$$

The internal failure rate λ_{IF} is directly derived from the Mean Time To Failure (MTTF) or, for repairable parts, from the Mean Time Between Failure (MTBF), as specified by the suppliers. As an example, for radar devices the MTTF is 60 years and the corresponding internal failure rate is $\lambda_{IF} = 4.5 \times 10^{-5}$ per day. If MTTF or MTBF are not specified by the supplier, expert judgment is used to estimate λ_{IF}.

Failures probabilities due to external causes λ_{EF} are more difficult to quantify. EWS are primarily installed in remote areas in alpine regions, close to rivers and glaciers, in high altitudes, steep catchments and are thus prone to numerous external failure causes. Lightning, humidity, storm and extreme temperatures are the most frequent external factors that cause failures on system components. Rock falls, snow avalanches and snow load, ice blocks, flood, vegetation, mud, dust and fog are site or system specific causes that can lead to failures of TSC. Additional potential failure causes such as construction, vandalism and animals must also be considered. To estimate λ_{EF}, we consult experts and evaluate historical data from repair records. Since the installation of the Illgraben debris flow system in 2006, one solar panel was destroyed by a rock fall. In the Illgraben, rock fall is common and we assume the failure rate to be 3×10^{-4} d^{-1}, which corresponds to a return period of 10 years. System failure due to extreme floods, lightning, animals, vandalism and extreme temperatures have not occurred yet, but should be considered as possible failure causes. We assume a failure rate of

3.0×10^{-5} d^{-1} for each external failure, which corresponds to a return period of 100 years. Summing up these rates, we receive an overall $\lambda_{EF} = 4.5 \times 10^{-4}$ d^{-1} for all TSC. This is in good agreement with available repair records.

To quantify the effect that technical failures have on the overall system reliability, we incorporate technical failure rates λ for all TSC in the BN. In doing so, the maximum POD (achieved with the optimal thresholds described later in the paper) is decreased by 0.34%. Thereof, 0.12% are due to internal failures (λ_{IF}) and 0.22% are due to external failures (λ_{EF}).

6.2. Inherent system reliability

The inherent reliability of the Illgraben system, as expressed through POD and PFA, depends on the selected threshold for each sensor signal. To analyze the influence of these thresholds, decision nodes representing varying thresholds are included in the BN/DG (Fig. 7). In addition, a decision node "decision criteria" allows various criteria to be analyzed for issuing warnings based on the indications from the individual sensors, e.g. a warning is issued if at least two sensors indicate an event.

Each of the five signal nodes in the monitoring unit are described by the conditional PDF of maximum measured signal during a day, conditional on whether or not a debris flow event occurs during that day. These conditional probability distributions correspond to those of the signal detection theory as illustrated in Fig. 1. To estimate them, recorded sensor data from the period between 1st of May 2008 and 24th September 2012 were used. During this period, 44 debris flow events were recorded on 883 days. For each of the five sensors, a probability distribution is fitted to the observed signals for days with and for days without events, as displayed in Fig. 8 for geophone G2. For inclusion in the BN, the signal is discretized in 10 classes, as exemplarily shown in Table 1 for G2.

To quantify the inherent reliability of individual sensors, POD and the PFA are evaluated from the conditional distributions of the signal following Eqs. (5)–(6). The resulting ROC curves that represent the reliability of individual sensors for varying thresholds are presented in Fig. 9. They indicate that the inherent reliability of the individual sensors varies strongly. Geophone G1 performs best and reaches a reliability close to the optimum with POD = 0.992 and PFA = 10^{-4}, whereas the remaining sensors have much lower inherent reliability. The difference among the reliabilities of the sensors is mainly due to the positioning of the sensors in the field, which influences their ability to detect

hazardous debris flow events and the amount of external disturbances, e.g. from animals, humans, rock falls.

6.3. Decision graph to identify optimal threshold combinations

With five sensors, and all the signals discretized in 10 classes, there are $9^5 = 59 \times 10^3$ possible threshold combinations, each of which leads to a POD and a PFA. Furthermore, for combing the individual sensor results different decision criteria can be defined, which further increase the number of possible warning strategies. For the Illgraben case study, two such decision criteria are considered. Either one individual sensor can issue a warning individually or a warning is issued when geophone 1 or at least one geophone and one radar device in the lower catchment indicate an event. The optimal warning criterion in all instances for the Illgraben case study is the second criterion.

Most of the possible warning strategies will be sub-optimal. Of interest are only the Pareto optimal warning strategies, for which it holds that no other strategy exists with simultaneously higher POD and lower PFA. To identify the Pareto optimal solutions, we employ the DG of Fig. 5. In the utility node, we modify the ratio of cost of false alarm to cost of a missed event. The costs of false alarm include spendings for activating the alarm units and are typically low compared to the expected cost of a missed event. Lattest, involve costs that are caused through damage and loss of life. The DG is used to identify the optimal threshold combination and decision criterion for each utility ratio. In this way, we obtain a set of Pareto optimal solutions, which allow the construction of

Table 1
Discretized probability distribution of signals measured by geophone 2 on days with and without event.

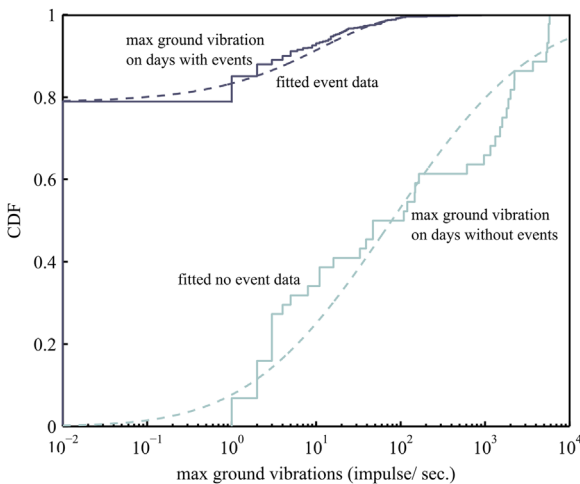| Class	Impulses/s		No event	Event
1	≤ 1		0.8332	0.0767
2	> 1	≤ 5	0.0512	0.1071
3	> 5	≤ 10	0.0295	0.0663
4	> 10	≤ 20	0.0305	0.0772
5	> 20	≤ 30	0.0163	0.0492
6	> 30	≤ 40	0.0102	0.0362
7	> 40	≤ 50	0.0069	0.0286
8	> 50	≤ 200	0.0208	0.1789
9	> 200	≤ 500	0.0014	0.1075
10	> 500		0.0001	0.2713

Fig. 8. Cumulative distribution function (CDF) of the signal of geophone 2. Solid lines represent observed data and dashed lines the fitted probability distributions.
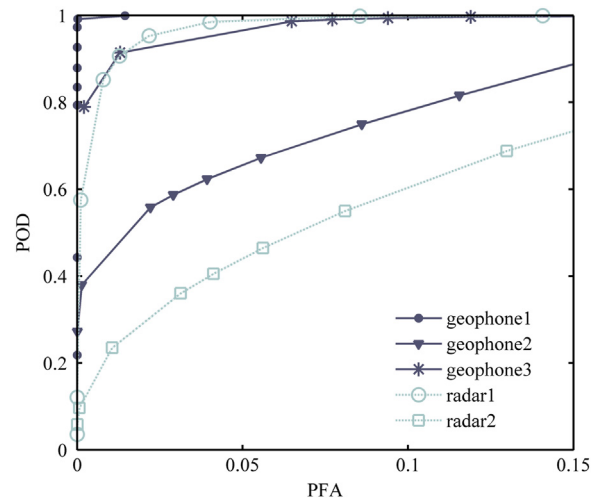

Fig. 9. Receiver Operator Characteristic ROC curves illustrate the reliability of sensors for nine predefined thresholds. The highest threshold is represented by operation points left. Geophone 1 shows the best performance.

the system ROC curve. In Table 2, the optimal threshold combinations for 20 utility ratios are presented, together with the corresponding POD, PFA and effectiveness, as computed with Eq. (15). The results are also graphically illustrated in Fig. 10. Here, the technical reliability is already included, i.e. the results show the overall sytem reliability and effectiveness.

6.4. Reliability and effectiveness of the Illgraben alarm system

The POD and PFA of the Pareto optimal warning strategies for the Illgraben alarm system are summarized in Table 2 and Fig. 10. Using these values, the overall ROC curve of the system is constructed, as depicted in Fig. 11. This ROC curve is overlaid with the system effectiveness calculated as a function of POD and PFA, following Fig. 3.

Overall, the reliability of the Illgraben debris flow alarm system is high, and so is its efficiency. According to Table 2, the warning strategy that maximizes the effectiveness of the system is the one found with utility ratios 0.7/0.8 and 0.9. This warning strategy has low thresholds for sensors G1 and R1, whereas the thresholds of

the remaining three sensors G2, G3 and R2, are set to their maximum. Geophone G3 still has a POD of 0.79 even with the largest threshold (see also Fig. 9). For G2 and R2, these optimal maximum thresholds indicate that these sensors do not contribute to the system reliability and may even decrease the overall effectiveness of the Illgraben debris flow system.

To assess the influence of individual technical system components (TSCs) on the overall system reliability and the resulting effectiveness, a sensitivity analysis is conducted. For each TSC i, the system effectiveness with the optimal warning strategy is recalculated once by assuming that the TSC i failed and once by assuming that the TSC i is perfectly reliable. This is done by simply setting the node of TSC i to "functioning" or "failure" respectively. The difference in effectiveness between the system with the perfectly reliable TSC i and the original system is called Effectiveness Achievement Worth, as it corresponds to the Risk Achievement Worth importance measure [51]. Accordingly, the difference

Table 2
Pareto optimal solutions for varying utility ratios.

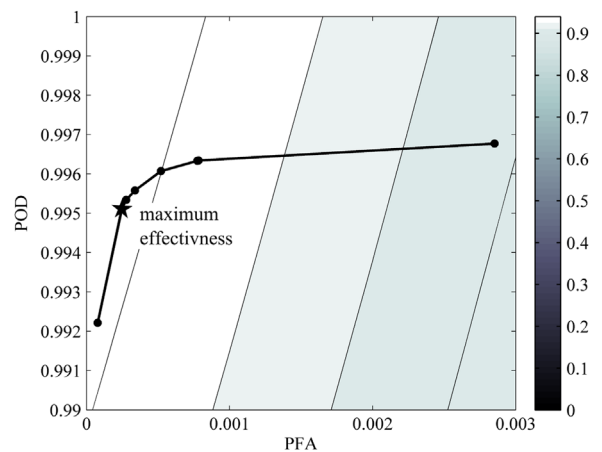| Utility ratio=cost of false alarm/cost of miss	Threshold					POD	PFA	Effectiveness
	G1	G2	G3	R1	R2			
0.009/0.01	2	8	7	3	9	0.996772	0.002851	0.912171
0.02	2	8	8	3	8	0.996342	0.000783	0.937166
0.03/0.04/0.05	2	8	8	3	9	0.996336	0.000775	0.937260
0.06/0.07/0.08/ 0.09/ 0.1	2	8	8	4	8	0.996072	0.000520	0.940068
0.2	2	8	9	3	9	0.995582	0.000339	0.941772
0.3	2	8	9	4	8	0.995339	0.000277	0.942281
0.4/ 0.5/ 0.6	2	9	9	4	8	0.995125	0.000248	0.942423
0.7/ 0.8/ 0.9	2	9	9	4	9	0.995110	0.000247	0.942424
1	3	8	9	4	8	0.992215	0.000078	0.941680

Fig. 11. The resulting Receiver Operator Characteristic ROC curve of the Illgraben alarm system, overlaid on the effectiveness.
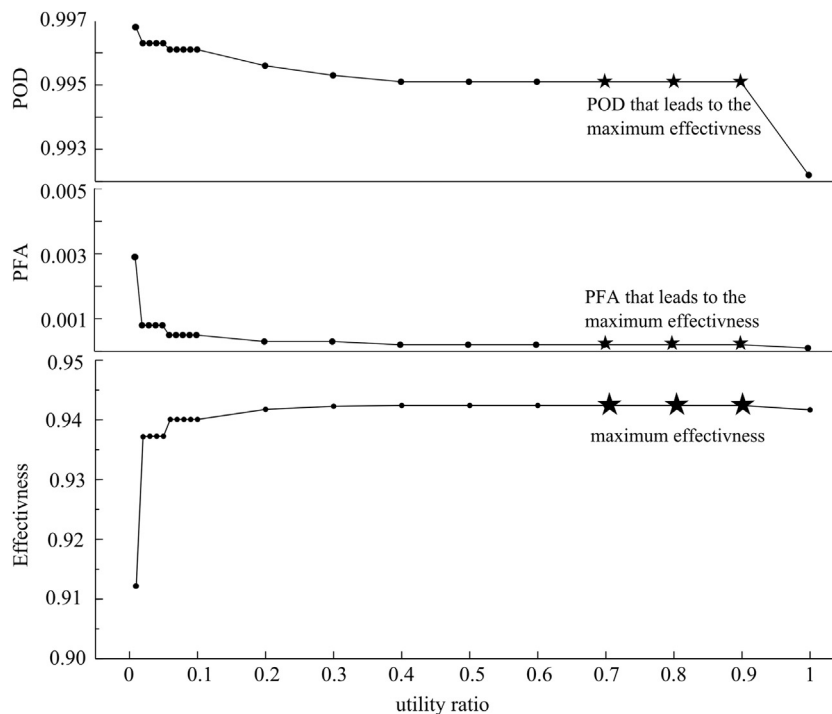

Fig. 10. Reliability and effectiveness of Pareto optimal warning strategies, as shown in Table 2.

in effectiveness between the original system and the one with TSC *i* failed is called Effectiveness Reduction Worth, corresponding to the Risk Reduction Worth importance measure. The results are summarized in Table 3, where TSCs are ordered according to their importance. Overall, the Effectiveness Achievement Worth of all TSCs is small; indicating that little can be gained from improving the reliability of individual TSCs. On the other hand, the Effectiveness Reduction Worth of the TSCs that are responsible for the data transmitting within the Illgraben system (modem 3, call receiver 1, call transmitter 1, mobile network or power supply), is large (9.42×10^{-1}). Upon failure of any of these TSCs, the system will not work, which is a consequence of the missing redundancy. Redundantly constructed data transmitting devices would therefore improve the system reliability and so its effectiveness considerably. However, as the Effectiveness Achievement Worth shows, the effect would be limited. For a further analysis of possible modifications in the system configuration, a cost analysis should be conducted.

The non-redundant TSCs in the information dissemination unit (call receiver 2/3/4, battery 3/4) are among the most critical TSCs and so their Effectiveness Reduction Worth (3.10×10^{-1}) is significant. Because all three alarm stations are equipped with redundant release devices, an audible signal and a red light, these two devices are less critical (1.82×10^{-4}).

The overall high system effectiveness is mainly a consequence of the high reliability of geophone G1 in the upper catchment. If that single geophone G1 or the TSCs essential for its functioning (logger 1, modem 1, battery 1) fail, the loss in effectiveness is large (1.77×10^{-1}). The influence of this individual sensor exceeds the joint influence of all four sensors in the lower catchment. The latter is quantified through the influence of logger 2, modem 2 or battery 2, whose failures would render all four sensors in the lower catchment useless. The influence of the individual sensors varies drastically. While G3 and R1 have a considerable effect on the effectiveness of the Illgraben system, G2 and in particular R2 are assumed to be sensors with minor significance. Nevertheless, the positioning of the four sensors in the lower catchment is limited. The position is chosen to detect debris flow events that could enter the main channel below the upper geophone at the earliest possible moment.

7. Discussion

EWS for natural hazards are safety-critical systems, whose reliability and effectiveness depends on the technical reliability of its components and the inherent ability of the system to correctly identify the hazard events. The framework proposed in this paper combines these two aspects into a single model, using a Bayesian network (BN). The BN is constructed to calculate the reliability of automated alarm systems in terms of Probability of Detection (POD) and Probability of False Alarms (PFA) as a function of the thresholds set for all sensors. The reliability is defined probabilistically, in agreement with existing concepts of risk management for natural hazards. The flexibility of the BN makes it straightforward to include potential technical failures of system components into this analysis. By extending the BN to a decision graph (DG), we furthermore find an effective way to solve the multi-dimensional optimization problem of identifying the optimal warning strategy with multiple sensors, i.e. the determination of the optimal combination of signal thresholds at the individual sensors.

We define effectiveness as a relative measure of the achieved risk reduction. We show that if a EWS is installed primarily for limiting the presence of people in the endangered area, the effectiveness can be assessed as a function of the POD and PFA alone. A crucial point is the quantification of the effect of false alarms on people's compliance with warnings. It is well known that false alarms deteriorate compliance, the so-called cry wolf syndrome, but studies quantifying this effect for natural hazard warnings are lacking, and assumptions have to be made on a relatively weak basis. Even when owners and operators of EWS do not intend to quantify the effectiveness, they must still understand the effect of false alarms, in determining the maximum acceptable PFA of a system.

The approach presented is applicable to automated alarm systems with limited lead times. Modifications will be necessary for EWS that offer larger lead times, such as in the case of slow rock movement process or flood hazards in the lower catchment area. For such events, the system reliability will be a function of the lead time. It will then be necessary to find a trade-off not only between POD and PFA, as in the current study, but also between

Table 3
Influence of individual technical system components (TSC) on effectiveness.

Ranking	TSC	Effectiveness TSC = "functioning"	Effectiveness achievement worth	Effectiveness TSC = "failure"	Effectiveness reduction worth
1	Modem 3 Call receiver 1 Call transmitter Power network Mobile network	0.943741	1.32×10^{-3}	0.0	9.42×10^{-1}
2	Battery 3/4 Call receiver 2/3/4	0.942581	1.57×10^{-4}	0.632004	3.10×10^{-1}
3	Geophone 1 Logger 1 Modem 1 Battery 1	0.942509	8.45×10^{-5}	0.764853	1.77×10^{-1}
4	Modem 2 Logger 2 Battery 2	0.942427	3.30×10^{-6}	0.935417	7.01×10^{-3}
5	Radar 1	0.942427	3.24×10^{-6}	0.935876	6.55×10^{-3}
6	Geophone 3	0.942426	2.19×10^{-6}	0.937827	4.60×10^{-3}
7	Solar panel 3/4	0.942426	1.45×10^{-6}	0.939547	2.88×10^{-3}
8	Solar panel 1	0.942425	8.31×10^{-7}	0.940781	1.64×10^{-3}
9	Geophone 2	0.942424	2.34×10^{-7}	0.942048	3.77×10^{-4}
10	Red light 1/2/3 Audible signal 1/2/3	0.942424	1.04×10^{-7}	0.942242	1.82×10^{-4}
11	Solar panel 2	0.942424	3.51×10^{-8}	0.942359	6.50×10^{-5}
12	Radar 2	0.942424	8.16×10^{-9}	0.942405	1.91×10^{-5}

POD, PFA and the lead time. For warning and forecasting EWS, decisions are not fully automated. For these systems, a comprehensive evaluation of EWS reliability and effectiveness has to include human-decision making and the quality of predictive models. Nevertheless, parts of the framework introduced in this paper are applicable also to such EWS, as we demonstrate in a case study on a rockslide warning system [52].

For the investigated debris flow alarm system, the final sensitivity analysis showed that most of the individual technical components have little impact on the reliability, with the exception of the non-redundant communication system and the most important sensors. For these components, additional protection measures such as fences, rockfall nets and standardized double-wall box may be installed to protect them. The case study furthermore revealed that besides the types and numbers of sensors, mainly their positioning in the field is crucial for the inherent system reliability. Finally, we find that the combination of multiple sensors can increase the POD while keeping PFA low, but only to a certain level.

The approach presented can be applied to optimize existing alarm systems, but can also assist in the design phase when a new system configuration is developed. In the latter case, cost-benefit analyses should be conducted to assess and compare different system configurations. In contrast to the case study presented, data availability is a problem for most applications. The fact that the sensitivity of sensors to a hazard event is often strongly site-specific—as shown by our case study—makes it difficult to transfer models describing the inherent reliability of individual sensors among different applications. During a test phase, site-specific models may be developed, following the procedure presented in this paper. However, during the design phase or for locations with only few hazard events such an approach is not feasible. Models for the inherent reliability of sensors must then be developed based on expert opinions or using detailed physical models, e.g. as developed in structural reliability applications.

8. Conclusion

We propose a framework to quantify the reliability of alarm systems for natural hazards based on Bayesian network (BN), accounting for both technical failures and the inherent system ability. The reliability is expressed in terms of the Probability of Detection (POD) and the Probability of False Alarms (PFA). To find a warning strategy that offers an optimal trade-off between these two, we define the system effectiveness as a function of POD and PFA as a measure of risk reduction. The optimal warning strategy is the one maximizing the system effectiveness. We show that by enhancing the BN to a decision graph, one is able to automatically identify an optimal warning strategy for systems with multiple sensors, where the decision on whether or not to issue an alarm is based on a combination of signals from all these sensors. By implementing the framework for a debris flow alarm system, we are able to demonstrate the applicability and usefulness of the framework for real alarm systems installed in practice.

References

[1] UNISDR. Terminology. Geneva: The United Nations Office for Disaster Risk Reduction. ⟨http://www.unisdr.org/we/inform/terminology⟩, last retrieved in October; 2014 2007.

[2] Hattenberger D, Wöllik A. (Naturgefahren-) Mess-und Frühwarnsysteme: Einzelne rechtliche Aspekte. Baur Bl 2008;11:89–101.

[3] FOCP., Leitfaden KATAPLAN -. Kantonale Gefährdungsanalyse und Vorsorge. Bern: FOCP Swiss Federal Office for Civil Protection; 2013.

[4] UNDRO. Natural disasters and vulnerability analysis - Report of Expert Meeting. Geneva: Office of the United Nations Disaster Relief Co-ordinator UNDRO; 1980.

[5] Bründl M, Romang H, Bischof N, Rheinberger C. The risk concept and its application in natural hazard risk management in Switzerland. Nat Hazards Earth Syst Sci 2009;9:801–13.

[6] SafeLand. Quantitative risk-cost-benefit analysis of selected mitigation options for two case studies. Deliverable 5.3: SafeLand Project - Living with landslide risk in Europe: Seventh Framework Programme for research and technological development (FP7) of the European Comission; 2012.

[7] Penning-Rowsell EJC, Tunstall S, Tapsell S, Morris J, Chatterton J, Green C. The benefits of flood and coastal risk management: a handbook of assessment techniques. London: Middlesex University Press; 2005.

[8] Margreth S, Romang H. Effectiveness of mitigation measures against natural hazards. Cold Reg Sci Technol 2010;64:199–207.

[9] Rogers D, Tsirkunov V. Costs and Benefits of Early Warning Systems. Global Assessment Report on Disaster Risk Reduction, 16pp, Published by ISDR and World Bank; 2011.

[10] Schröter K, IHWB MOT-D, Darmstadt GDS, Velasco C, GRAHI-UPC B, Nachtnebel SH-P, et al. Effectiveness and efficiency of early warning systems for flash floods. www document, http://www ihwb tu-darmstadt de/media/fachgebiet_ihwb/literatur/lit_ostrowski/paper78_ost pdf. 2006.

[11] UNEP. Early Warning Systems: a state of the art analysis and future directions. Nairobi: United Nations Environment Programme (UNEP), Division of Early Warning and Assessment (DEWA); 2012. p. 70.

[12] Carsell KM, Pingel ND, Ford DT. Quantifying the benefit of a flood warning system. Nat Hazards Rev 2004;5:131–40.

[13] Krzysztofowicz R, Kelly K, Long D. Reliability of flood warning systems. J Water Resour Plan Manag 1994;120:906–26.

[14] Paté-Cornell ME. Warning systems in risk management. Risk Anal 1986;6:223–34.

[15] Alfieri L, Thielen J, Pappenberger F. Ensemble hydro-meteorological simulation for flash flood early detection in southern Switzerland. J Hydrol 2012;424:143–53.

[16] Montesarchio V, Lombardo F, Napolitano F. Rainfall thresholds and flood warning: an operative case study. Nat Hazards Earth Syst Sci 2009;9:135–44.

[17] Simmons KM, Sutter D, Alarms False. Tornado warnings, and Tornado casualties. Weather Climate Soc 2009;1:38–53.

[18] Rheinberger CM. Learning from the past: statistical performance measures for avalanche warning services. Nat Hazards 2013;65:1519–33.

[19] Bründl M, Heil B. Reliability analysis of the Swiss avalanche warning system. In: Faber M, Köhler J, Nishijima K, editors. Proceedings of the 11th international conference on applications of statistics and probability in civil engineering. Zürich: CRC Press an imprint of the Taylor & Francis Group; 2011. p. 881–87.

[20] Sturny RA, Bründl M., Bayesian networks for assessing the reliability of a Glacier Lake warning system in Switzerland. In: Fujita M. et al., editors. Proceedings of the Interpraevent 2014 in the Pacific Rim - Natural Disasters Mitigation to Establish Society with the Resilience. Nara, Japan.

[21] Sättele M, Bründl M, Straub D. Bayesian Networks to Quantify the reliability of a debris flow alarm system. In: Deodatis E, Frangopol, editors. Saftey, Reliability, Risk and Life-Cycle Performance of Structures & Infrastructures. London: Taylor & Francis Group; 2013. p. 928.

[22] Sättele M, Bründl M, Straub D. A classification of warning system for natural hazards. In: Moormann C, Huber M, Proske D, editors. Proceedings of the 10th international probabilistic workshop. Stuttgart: Institut für Geotechnik der Universität Stuttgart; 2012. p. 257–70.

[23] IEEE.. IEEE Standard Framework for Reliability Prediction of Hardware, IEEE Std 1413. NY: Institute of Electrical and Electronics Engineers, Inc; 2010.

[24] Rouvroye J, Brombacher A. New quantitative safety standards: different techniques, different results? Reliab Eng Syst Saf 1999;66:121–5.

[25] Ericson CA. Hazard analysis techniques for system safety: Wiley. com; 2005.

[26] Langseth H, Portinale L. Bayesian networks in reliability. Reliab Eng Syst Saf 2007;92:92–108.

[27] Swets JA. Signal detection theory and ROC analysis in psychology and diagnostics: collected papers. New York: Lawrence Erlbaum Associates, Inc.; 1996.

[28] Hanely J, McNeil B. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.

[29] Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993;39:561–77.

[30] Müller C, Elaguine M, Gaal M, Scharmach M, Ewert U, Osterloh K. Determination of reliability in deminig and NDT. NDT–Competence & Safety, MATEST, Zagreb; 2004.

[31] Schoefs F, Clement A, Nouy A. Assessment of ROC curves for inspection of random fields. Struct Safety 2009;31:409–19.

[32] Peterson W, Birdsall T, Fox W. The theory of signal detectability. IRE Prof Group Inf Theory 1954;4:171–212.

[33] Breznitz S. Cry wolf: the psychology of false alarms. Hillsdale, NJ: Erlbaum; 1989.

[34] Dejoy DM, Cameron KA, Lindsay JD. Postexposure evaluation of warning Effectivness: a review of field studies and population-based research. In: Wogalter S, M, editors. Handbook of Warnings (Human Factors & Ergonomics). Mahwah, N.J: Lawrence Erlbaum Associates; 2006. p. 35–48.

[35] Johnson M, Newstead S, Charlton J, Oxley J. Riding through red lights: the rate, characteristics and risk factors of non-compliant urban commuter cyclists. Accid Anal Prev 2011;43:323–8.

[36] Rosenbloom T. Crossing at a red light: behaviour of individuals and groups. Transp Res Part F 2009;12:389–94.

[37] Bliss JP, Gilson RD, Deaton JE. Human probability matching behaviour in response to alarms of varying reliability. Ergonomics 1995;38:2300–12.

[38] Jensen FV, Nielsen TD. Bayesian networks and decision graphs. 2nd ed.. New York: Springer Science+Business Media; 2007.

[39] Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. San Francisco: Morgan Kaufmann; 1988.

[40] Straub D, Der Kiureghian A. Bayesian network enhanced with structural reliability methods: methodology. J Eng Mech-Asce 2010;136:1248–58.

[41] Shachter RD. Evaluating influence diagrams. Oper Res 1986;34:871–82.

[42] Straub D.. Natural hazards risk assessment using Bayesian networks. In: Augusti G, Schuëller GI, Ciampoli M, editors. Proceedings of the 9th international conference on structural safety and reliability ICOSSAR'05. Rome: Millpress; 2005.

[43] Aguilera PA, Fernández A, Fernández R, Rumí R, Salmerón A. Bayesian networks in environmental modelling. Environ Model Softw 2011;26:1376–88.

[44] Grêt-Regamey A, Straub D. Spatially explicit avalanche risk assessment linking Bayesian networks to a GIS. Nat Hazards Earth Syst Sci 2006;6:911–26.

[45] Medina-Cetina Z, Nadim F. Stochastic design of an early warning system. Georisk: Assess Manag Risk Eng Syst Geohazards 2008;2:223–36.

[46] Blaser L, Ohrnberger M, Riggelsen C, Babeyko A, Scherbaum F. Bayesian networks for tsunami early warning. Geophys J Int 2011;185:1431–43.

[47] Sättele M, Meier L. Elektronisch warnen. TEC21 2013;31–32:16–21.

[48] Badoux A, Graf C, Rhyner J, Kuntner R, McArdell BW. A debris-flow alarm system for the Alpine Illgraben catchment: design and performance. Nat Hazards 2009;49:517–39.

[49] DSL GeNIe & SMILE. DSL, Decision Systems Laboratory, University of Pittsburgh; 2013. ⟨http://genie.sis.pitt.edu/index.php/about⟩; last retrieved in October; 2013.

[50] Straub D. Lecture notes in engineering risk analysis. Munich, Germany: Technische Universität München, Engineering Risk Analysis Group; 2012.

[51] Vesely WE, Davis TC, Denning RS, Saltos N. Measures of risk importance and their applications. Ohio: NUREG/CR-3385, US Nuclear Regulatory Commission; 1983.

[52] Sättele M, Krautblatter M, Bründl M, Straub D. Forecasting rock slope failure: How reliable and effective are warning systems? Landslides. under review.