

A Tool for the Automatic Aggregation and Validation of the Results of Physically Based Distributed Slope Stability Models

Maria Alexandra Bulzinetti ¹, Samuele Segoni ^{1,*} , Giulio Pappafico ², Elena Benedetta Masi ¹ ,
Guglielmo Rossi ³ and Veronica Tofani ¹

¹ Department of Earth Sciences, University of Firenze, 50121 Firenze, Italy; mariaalexandra.bulzinetti@unifi.it (M.A.B.); elenabenedetta.masi@unifi.it (E.B.M.); veronica.tofani@unifi.it (V.T.)

² Department of Pure and Applied Sciences, University of Urbino, 61029 Urbino, Italy; giulio.pappafico@uniurb.it

³ Centre for the Civil Protection, University of Firenze, 50125 Firenze, Italy; guglielmo.rossi@unifi.it

* Correspondence: samuele.segoni@unifi.it; Tel.: +39-0552-755975

Abstract: Distributed physically based slope stability models usually provide outputs representing, on a pixel basis, the probability of failure of each cell. This kind of result, although scientifically sound, from an operational point of view has several limitations. First, the procedure of validation lacks standards. As instance, it is not straightforward to decide above which percentage of failure probability a pixel (or larger spatial units) should be considered unstable. Second, the validation procedure is a time-consuming task, usually requiring a long series of GIS operations to overlap landslide inventories and model outputs to extract statistically significant performance metrics. Finally, if model outputs are conceived to be used in the operational management of landslide hazard (e.g., early warning procedures), the pixelated probabilistic output is difficult to handle and a synthesis to characterize the hazard scenario over larger spatial units is usually required to issue warnings aimed at specific operational procedures. In this work, a tool is presented that automates the validation procedure for physically based distributed probabilistic slope stability models and translates the pixelated outputs in warnings released over larger spatial units like small watersheds. The tool is named DTVT (double-threshold validation tool) because it defines a warning criterion on the basis of two threshold values—the probability of failure above which a pixel should be considered stable (failure probability threshold, FPT) and the percentage of unstable pixels needed in each watershed to consider the hazard level widespread enough to justify the issuing of an alert (instability diffusion threshold, IDT). A series of GIS operations were organized in a model builder to reaggregate the raw instability maps from pixels to watershed; draw the warning maps; compare them with an existing landslide inventory; build a contingency matrix counting true positives, true negatives, false positive, and false negatives; and draw in a map the results of the validation. The DTVT tool was tested in an alert zone of the Aosta Valley (northern Italy) to investigate the high sensitivity of the results to the values selected for the two thresholds. Moreover, among 24 different configurations tested, we performed a quantitative comparison to identify which criterion (in the case of our study, there was an 85% or higher failure probability in 5% or more of the pixels of a watershed) produces the most reliable validation results, thus appearing as the most promising candidate to be used to issue alerts during civil protection warning activities.

Keywords: shallow landslide; hazard; early warning; validation; GIS; factor of safety



Citation: Bulzinetti, M.A.; Segoni, S.; Pappafico, G.; Masi, E.B.; Rossi, G.; Tofani, V. A Tool for the Automatic Aggregation and Validation of the Results of Physically Based Distributed Slope Stability Models. *Water* **2021**, *13*, 2313. <https://doi.org/10.3390/w13172313>

Academic Editor: Yu Huang

Received: 13 July 2021

Accepted: 21 August 2021

Published: 24 August 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

A landslide early warning system is defined as the set of capacities needed to generate and disseminate timely and meaningful warning information to enable individuals, communities, and organizations threatened by hazards to prepare and act appropriately

and in sufficient time to reduce the possibility of harm or loss [1]. Warning systems for landslides can be designed and employed at different reference scales.

Two categories of early warning systems can be defined on the basis of their scale of analysis—local systems for single slopes [2,3] and regional systems [4]. Regional early warning systems for shallow landslides can be developed following two approaches: (a) rainfall thresholds based on statistical analysis of rainfall and landslides, and (b) physically based models. While the first approach is currently extensively used at regional scale [4–14], the latter is more frequently applied at slope or catchment scale [15–29].

The physically based models would allow for the spatial and temporal prediction of the occurrence of landslides with high accuracy, producing dynamic and spatially accurate hazard maps that can be of help for landslide risk assessment and management [30–32]. However, poor knowledge of the spatial distribution of hydrological and geotechnical parameters, caused by the extreme heterogeneity and inherent variability of soil properties at large scale [33–38], hinders the physically-based model application at regional scale.

The uncertainty related to hydrological and geotechnical parameters (such as cohesion, internal friction angle, and hydraulic conductivity) can be overcome through the use of a probabilistic approach, supported by the combined use of Monte Carlo simulations [29,36,39]. In this case, the output is a distributed probability of failure, i.e., the probability of having the factor of safety below a defined threshold (usually one). The interpretation of the results is already troublesome since a calibration and validation procedure has to be taken into account in order to define the value of probability of failure to discriminate stable and unstable pixels and, in the perspective of possible application purposes, to convert the probabilistic output into alerts referred to wider spatial units (e.g., catchments or alert zones). Indeed, the pixelated model output at fine spatial and temporal resolution is difficult to manage and to interpret and, in order to provide reliable results to stakeholders involved in the hazard management, a reaggregation over spatial units (e.g., basins, slope units, and municipalities) and temporal units could be advisable and it is a strategy pursued by many works focused on operational applications of slope stability models [37,40–43]. In particular, for the spatial aggregation of probabilistic outputs, a calibration procedure must be performed in order to define how many pixels (over a value of failure probability that has to be defined) are necessary to consider a spatial unit unstable. At present, a standardized approach for model validation does not exist [41]. Rossi et al. [29] considered a watershed as potentially affected by a landslide if the probability of failure was higher than 80% in at least 1% of the pixels of the basin, whereas Kuriakose et al. [44] identified slope failures in locations with failure of probability higher than 60%. Salvatici et al. [45] used a methodology of verification of their model based on the count of the pixels with probability of failure higher than 75%, while Salciarini et al. [20] correlated the spatial distribution of the landslides with different classes of probability of failure (the higher one being >40%). Conversely, Ho and Lee [19] considered the mean value of the factor of safety in each catchment instead of a probabilistic value. Given the fact that validation is a time-consuming activity and that no standardized procedure has been established, in the recent scientific literature it is still common to make use of qualitative validation procedures based on visual comparison between model outputs (at the pixel scale) and landslide occurrences (usually mapped as polygons or points) [34,46]. In this work, we present a GIS-based tool (DTV, double-threshold validation tool) that automatically performs (i) the reaggregation of results from pixel to basins scale and (ii) a quantitative validation of the results, including the setting up of confusion matrixes. The tool has been applied to the simulation results obtained by Salvatici et al. (2018) [45] with the physically-based model HIRESSS (high-resolution slope stability simulator) [29] in a test site located in Valle d'Aosta region (northern Italy). The tool was used to explore the sensitivity of the validation results to some parametrical settings and to define how its configuration could be set to obtain an early warning system as effectively as possible.

2. Materials and Methods

2.1. Test Site

The test area was located in the eastern part of the Valle d'Aosta region, northern Italy, in a sub-area of alert zone B (Figure 1). The area stretches over approximately 195 km², in the south of the region and is characterized by two main valleys. The first is located on the Dora Baltea catchment, from Champ De Vigne to Pont-Saint-Martin and the second valley encompasses the Lys catchment, Dora Baltea's tributary, from Pont Trenta to Pont Saint Martin.

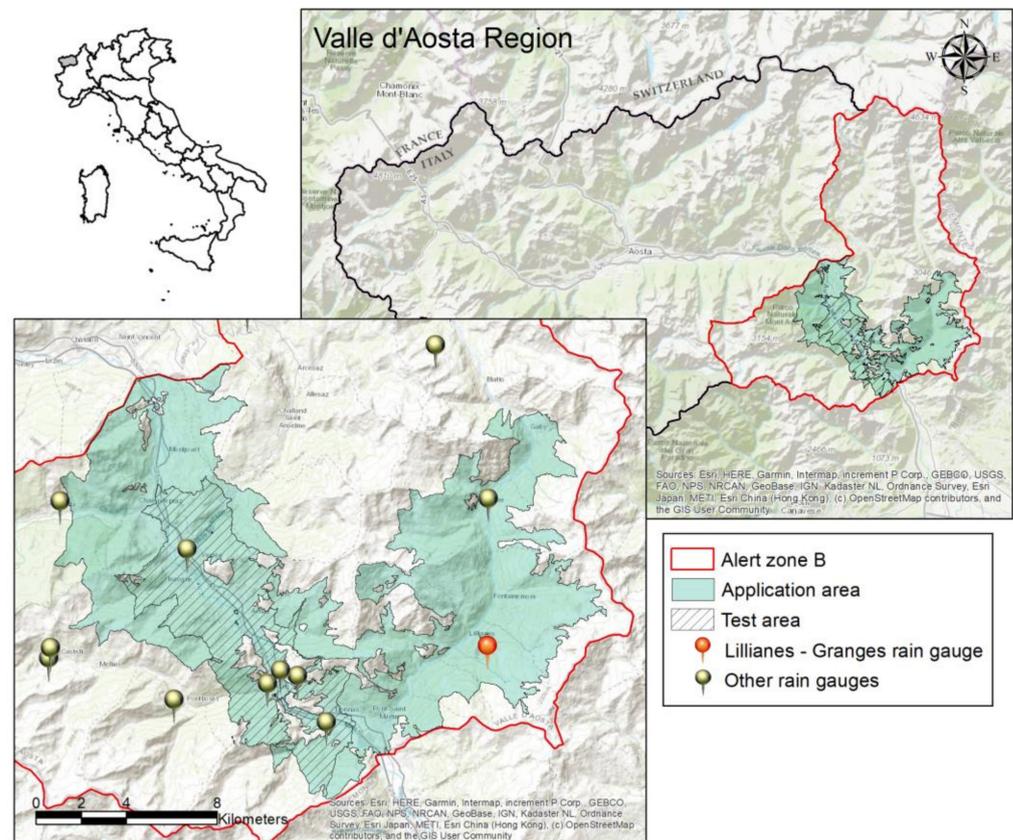


Figure 1. Map showing the location of the test site and the areas used to apply and test the DTVT (double threshold validation tool).

The Valle d'Aosta region is located in the Italian alpine area and the main geological units stretch from north-west to south-east, through two main geological alignments—one downstream near the city of Borgofranco d'Ivrea, (the Canavese line) and the other upstream (the Pennidico front). From a lithological point of view the study area is mainly characterized by intrusive and metamorphic rocks, in particular granites, metagranites, schists, and serpentinite [45].

The geomorphology of the study area is characterized by steep slopes, which predispose the area to the triggering of landslides. The landscape shows valleys shaped by glaciers, mainly during the Pleistocene glacial period, with elevations ranging from 400 m a.s.l. on the Dora Baltea River floodplain to about 3000 m a.s.l. on the crests. The climate of the region is characterized by high variability that is strongly influenced by altitude, with a continental climate in the valley floors and an alpine climate at high altitudes [45].

Different types of landslides such as rockfalls, deep seated gravitational slope deformations (DSGSD), rocks avalanches, debris avalanches, debris flows, and debris slides (Catasto dei Dissesti Regionale—from Val d'Aosta Regional Authorities) are quite common. These landslide typologies have very different triggering mechanisms and predisposing factors, which should be accounted for by different models. In this work, we model the

triggering conditions of shallow landslides, i.e., soil slips and translational slides and we do not take other types of movement into account.

The GIS-based tool was applied to the HIRESSES simulation results of two selected events during which several landslides had occurred in the study area and a good quality landslide inventory was available. These events took place in April 2009 and June 2010. The event of April 2009 affected the southeastern part of the Aosta Valley region from 26 April to 28 April, with the highest total precipitation (about 268 mm) recorded at the Lillianes-Granges station, mainly concentrated on 27 April (daily amount of 208 mm and peak hourly rainfall of 17.4 mm); in the same day several shallow landslides were triggered. The second event started on 8 June 2010 and ended on 17 June 2010. The maximum total precipitation for the entire event was recorded at Issime and Lillianes-Granges stations, where 190 mm and 230 mm were measured, respectively. The days with the highest precipitation were the 15 and 16 June. The maximum daily and hourly precipitation (78.0 mm and 10.4 mm, respectively), were both recorded at the Lillianes-Granges station on 16 June. This event triggered several debris flows during the last days of the storm.

The former rainfall event was used to investigate the sensitivity of the model outputs to the calibration strategies adopted, while the latter was used to verify a possible strategy for issuing alerts. The areal extension of the simulations was limited only in the sectors of the alert zone B where the information about the landslides was deemed complete and reliable (Figure 1).

2.2. Previous Slope Stability Modeling in the Test Site

The DTVT tool presented in this work can be applied to the outputs of any distributed probabilistic slope stability model; it can therefore be applied to failure probability maps, i.e., raster maps in which pixel values describe the factor of safety in probabilistic terms (i.e., the probability of the factor of safety being lower than one) for points of a certain area and time-step (e.g., 1h or 1 day). In this work, we present an application of the DTVT tool to the failure probability maps obtained using HIRESSES [29] by Salvatici et al. [45], but we wish to stress that the DTVT tool and HIRESSES are independent of each other.

Details about the HIRESSES model and the Salvatici et al. study are present in references [28,29,45] and for the convenience of the reader there follows some information about HIRESSES. HIRESSES is a physically-based distributed slope stability simulator developed to analyze shallow landslide-triggering conditions on large scale at high spatial and temporal resolution using a parallel calculation method [29,45]. The model is composed of two different modules—hydrological and geotechnical. The hydrological module is based on the hydraulic diffusivity concept, using an analytical solution of an approximated form of the Richards equation under wet condition [47] with rainfall as input data, whilst geotechnical stability analysis is based on an infinite slope stability model, taking into account the effect of strength and cohesion increase due to matric suction. The HIRESSES model needs the following spatially distributed input data: slope gradient, effective cohesion (c'), root cohesion (c_r), friction angle (ϕ'), dry unit weight (γ_d), soil thickness, hydraulic conductivity (k_s), initial soil saturation (S), pore size index (l), bubbling pressure (h_s), porosity (n), residual water content (θ_r), and rainfall intensity [29,43]. HIRESSES computes the factor of safety at each selected time step (and not only at the end of the rainfall event) and at different depths within the soil layer and it can operate at any spatial resolution. A Monte Carlo simulation is implemented in the model to manage the uncertainty of the input data, and consequently the output is the probability of failure, i.e., the probability to have a value of factor of safety below 1. Detailed information about HIRESSES can be found in Rossi et al. and Mercogliano et al. [28,29] showing past applications of HIRESSES in early warning systems. All details about the application of HIRESSES to this test site can be found in Salvatici et al. [45]; as mentioned, the present work builds on those results to develop and test the DTVT tool.

2.3. Input Data Needed

The following list has a twofold objective: (i) to explain in general terms the input data needed to apply the DTVT tool to any test site where a distributed slope stability model has been previously applied and (ii) to describe the datasets used for the present application.

- Pixel-based failure probability raster maps: raster maps describe the failure probability distribution of a certain area (in a specific time) obtained using a distributed slope stability model (i.e., raster maps, in which every pixel reports the probability to have factor of safety below one) (Figure 2a). In this work, we used raster outputs of HIRESSS [45] with 10 m resolution, consisting in a summarizing raster map computed by the model in which each pixel was associated with the highest probability value calculated during each hour of that day.
- Reaggregation units: a polygonal shapefile of territorial units identifying the elements in which the user needs to spatially aggregate the pixel-based information, to be used for hazard management. In our case, as reaggregation units, we used sub-watersheds (Figure 2b) with 0.37 km² average areal extension and extracted starting from a DEM of the study area using ArcMap. Other possible reaggregation units could be municipal areas, alert areas, and sub-areas used by civil protection centers.
- Landslide dataset: this is the inventory of the landslides that occurred in the modeled events. In this case study, it is represented by a shapefile of landslides including all landslides triggered by the meteorological event described in the previous section, as reported by the local authorities to the regional Civil Protection offices (Figure 2b, “observed unstable areas”). Most of these landslides were originally mapped as polygons. In this case, they were just imported in the shapefile. Other landslides were registered less accurately: only the approximated location (e.g., name of the locality, road trait hit by the slide, and address of a damaged infrastructure) was noted and thus they were represented only as points. The points were transformed into circles with a radius corresponding to the associated uncertainty in the location. Only landslides with a small spatial uncertainty that would allow the triggering point into a precise watershed to be located (a total of 10 landslides) were included. Moreover, on the basis of the reports and of a field survey, all landslide types not compatible with the triggering mechanism encompassed by HIRESSS (e.g., rockfalls) were discarded.

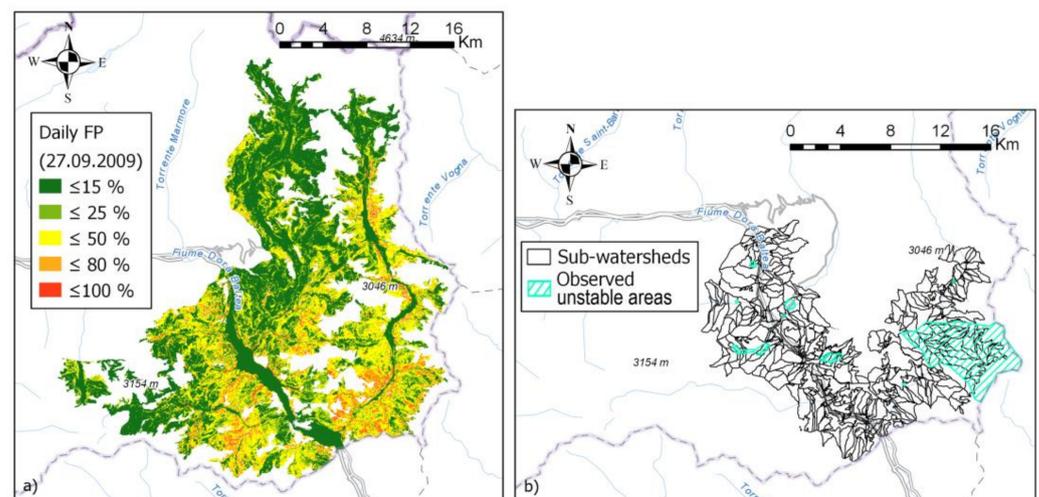


Figure 2. Input data: (a) pixel-based failure probability (FP) raster map and (b) reaggregation units (sub-watersheds) and landslide database (observed unstable areas).

2.4. Instability Diffusion Threshold—Tool Development

The input data were combined into a GIS system (ESRI ArcMap 10.6) and the “model builder” application was used to get a procedure for the automatic validation of the slope

stability results aggregated on small watersheds. The procedure, named double-threshold validation tool (DTVT) is based on the definition of watershed to be warned and on two threshold values.

The first threshold operates at the pixel scale and is used to discriminate unstable pixel from stable pixels. Probabilistic distributed models produce a raster map in which each pixel is associated with its probability of slope failure calculated by the model. These values range from 0 to 1 and the failure probability threshold (FPT) is used to define the probability value above which pixels should be considered unstable.

The second threshold operates at the small watershed scale and is used to discriminate watersheds with a relevant triggering hazard from watersheds that could be considered relatively safe. Indeed, one could observe that an unstable pixel is not enough to consider a whole watershed as unstable—issuing a warning that may involve serious countermeasures (e.g., evacuation or road traffic closure) requires that the instability is quite widespread. Therefore, we considered an instability diffusion threshold (IDT), which expresses the percentage of the watershed area that needs to be occupied by unstable pixels (as defined by the FPT) to issue a warning in the whole watershed. This approach is quite well established in landslide hazard management [37,40].

The two thresholds (FPT and IDT) operate jointly and the outcomes may vary greatly according to the threshold values used. Indeed, FPT and IDT can be considered parameters that can be adjusted and calibrated during a testing phase (e.g., with a trial and error procedure or empirically calibrated through case studies), to define the best arrangement for forecasting purposes. This issue will be explored in the second part of the manuscript, while in this section a detailed explanation of the steps performed by the tool is reported.

The procedure performed by the DTVT combines 32 ArcMap commands into an automated tool: it is shown in Figure 3 and it is described hereafter, divided into five main steps.

First, by means of a simple graphic user interface (Figure 3), all input data listed in Section 2.3 must be loaded, the user-defined threshold values (FPT and IDT) have to be set, and the directory where the results have to be saved has to be defined. Afterwards, the model can be run.

Step 1: Set up the Failure Probability Threshold (FPT)

Using the raster maps of the probability of landslide-triggering produced by the distributed slope stability model, a default expression of the Raster Calculator tool identifies all the raster cells with probability values greater than the user-defined FPT value. The result is a binary raster map, where the cells with probability of instability higher than FPT are assigned the value 1; all the other cells are assigned the value 0.

Afterwards, the tool Tabulate Area overlays the aforementioned binary raster with the watershed shapefile and defines how unstable raster cells are located within each watershed.

A field is added to the table and the R Index, which represents the ratio R, is calculated and converted to a percentage, between the number of cells with a value of 1 and the total number of cells included in the watershed. The R index is calculated for each watershed.

Step 2: Set up the Instability Diffusion Threshold (IDT)

Another field is added to the table and therein a field calculator is launched to compare the R index of each basin with the user-defined IDT. As a result, table elements with a spatial diffusion of instability greater than IDT are assigned a value 1, otherwise they are assigned value 0.

A join is added to connect the modified table to the watershed shapefile. In this way, watersheds modeled as stable (thus, considered safe) are distinguished by those modeled as unstable (thus, to be warned).

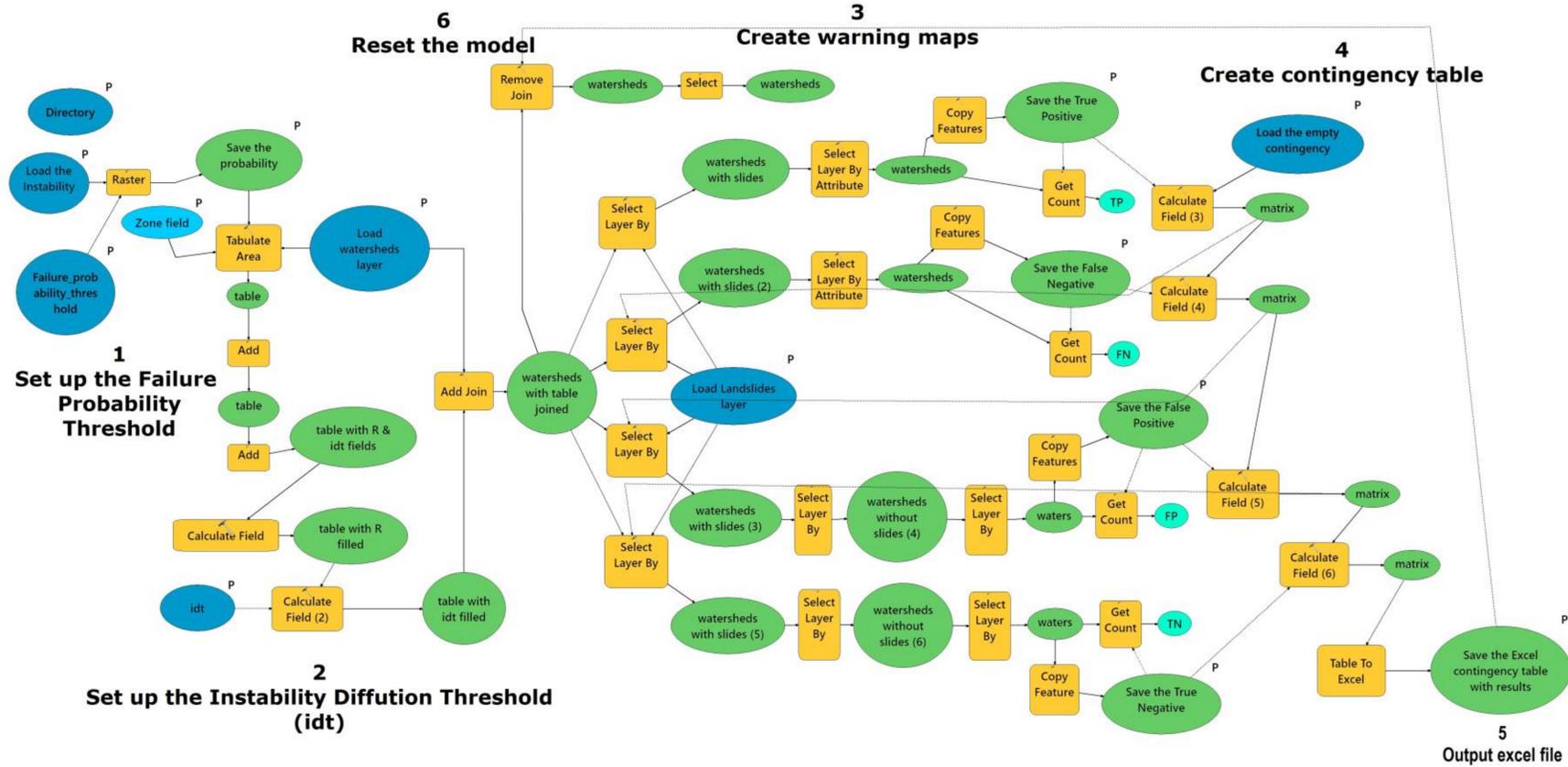


Figure 3. Model builder of the double-threshold validation tool (DTVT).

Step 3: Create Validation Maps (Output #1)

The model overlays the landslides' shapefile with the watershed and a series of “select by location” operations is performed. The two layers are intersected, and the modeled instability is validated by comparison with the known landslide inventory. As a result, the watersheds are classified as TP (true positive—watersheds modeled as unstable where landslides have been reported), FN (false negative—watersheds modeled as stable where landslides have been reported), TN (true negative—watersheds modeled as stable where landslides have not been reported), or FP (false positive—watersheds modeled as unstable where landslides have not been reported) (Figure 4a). The four instances are saved in separate shapefiles, which in turn are used to draw the resulting validation map (Figure 4b).

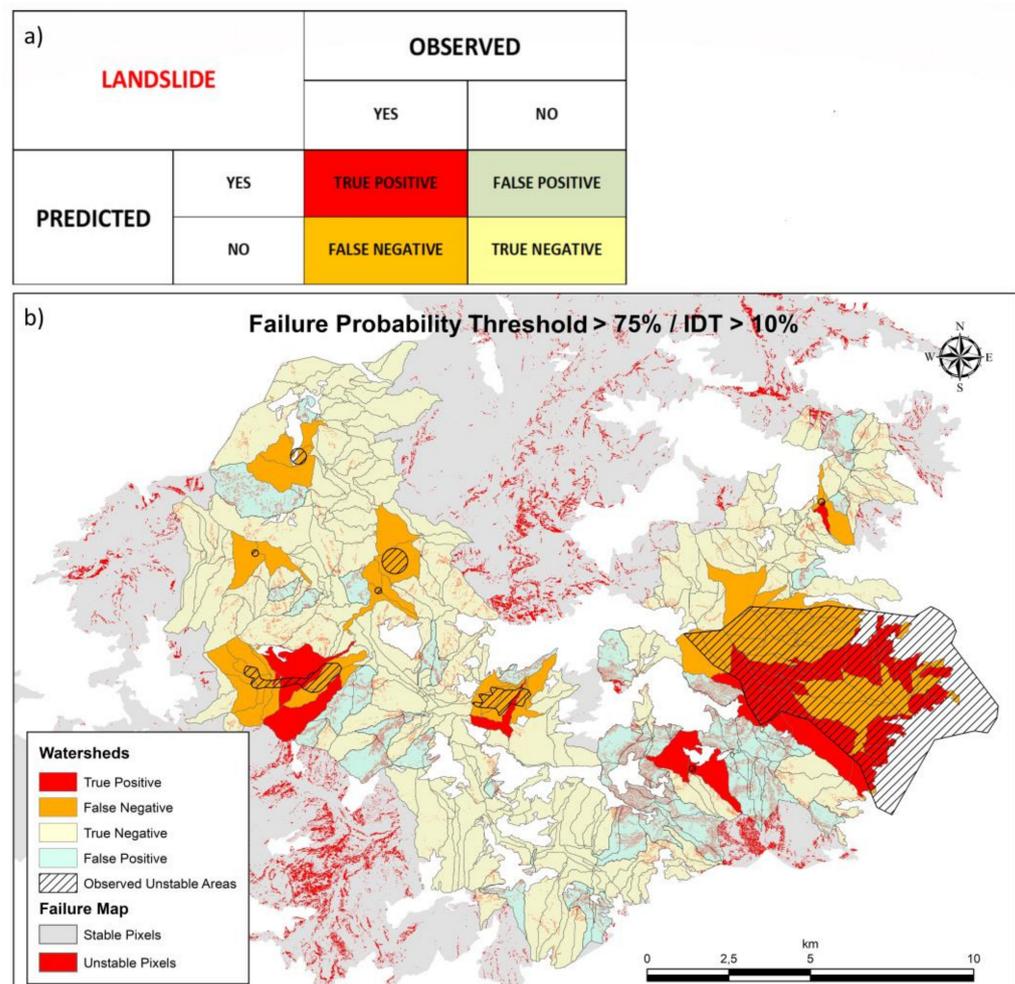


Figure 4. (a) Contingency matrix defining true positives, false positives, true negatives, and false negatives and (b) graphical representation, with the same set of colors, of the validation outcomes obtained setting FPT = 75% and IDT = 10%.

Step 4: Create Contingency Table (Output #2)

With a series of “get count” and “calculate field” operations, the TP, TN, FP, and FN instances are counted, and a contingency table is built and saved as an .xls file. The results of the contingency table can be easily imported in a spreadsheet or other software to design graphics and to calculate state-of-the-art skill scores, which are commonly used to quantitatively assess the performance of a model after a validation procedure.

Step 5: Reset the Model

Joins are removed and all active selections are cleared. The model is ready to be used again with new input data and/or with a different couple of threshold values.

In summary, the system automatically performs a validation of the probabilistic outputs generated by distributed slope stability models by comparison with an inventory of reported landslides. The validation output is twofold and includes a map and a contingency table. Thanks to the implementation of the procedure in the ArcGIS Model Builder tool (Copyright © 1995-2017 ESRI), the workflow can be tested repeatedly and quickly after changing the input data or the user-defined settings.

2.5. Tool Application

In this work, the DTVT tool was also applied in a series of tests with a twofold objective: (i) to explore the sensitivity of the model outputs to the validation criteria and (ii) to define some interpretation criteria of the probabilistic outputs to be used in operational application to issue warnings.

Using the same set of input data, the DTVT tool was used repeatedly to perform several validation procedures, each time changing the criteria (i.e., threshold values) used to discriminate stable and unstable watershed. The modeled event was the one occurred in April 2009, because this meteorological event produced a severe and representative impact in terms of triggered landslides. The following FPT and IDT values were tested—60, 70, 75, 80, 85, and 90% and 3, 5, 10, and 15%. Each pair of FPT-IDT values defines a specific configuration of the DTVT, which was used to automatically validate the results of the landslide modeling performed by HIRESSES. For each of the 24 DTVT configurations, the validation maps were saved, and the contingency matrixes were imported into a spreadsheet to calculate and compare the skill scores that are commonly used as quantitative indicators of model performance [48].

3. Results

3.1. DTVT Outputs and Sensitivity Analysis

The results of the validation were visually rendered with a series of shapefiles that were automatically loaded in the opened ArcMap project and “dressed” according to a default symbology (Figures 4 and 5).

True-positive watersheds are colored in red, to stress the high hazard level correctly captured by the model (in a hypothetical warning system, those watersheds should have been warned and the warning would have been successful because a landslide occurred there). False negative watersheds are colored in orange—a high hazard level was present there (a landslide was reported), but the slope stability model output, reaggregated with the DTVT at hand, has not adequately captured it. The watersheds where no landslides have been reported are colored in light colors—light yellow if the model correctly returned a safe condition (true negatives) and light blue if the model overestimated the hazard level and returned a false positive.

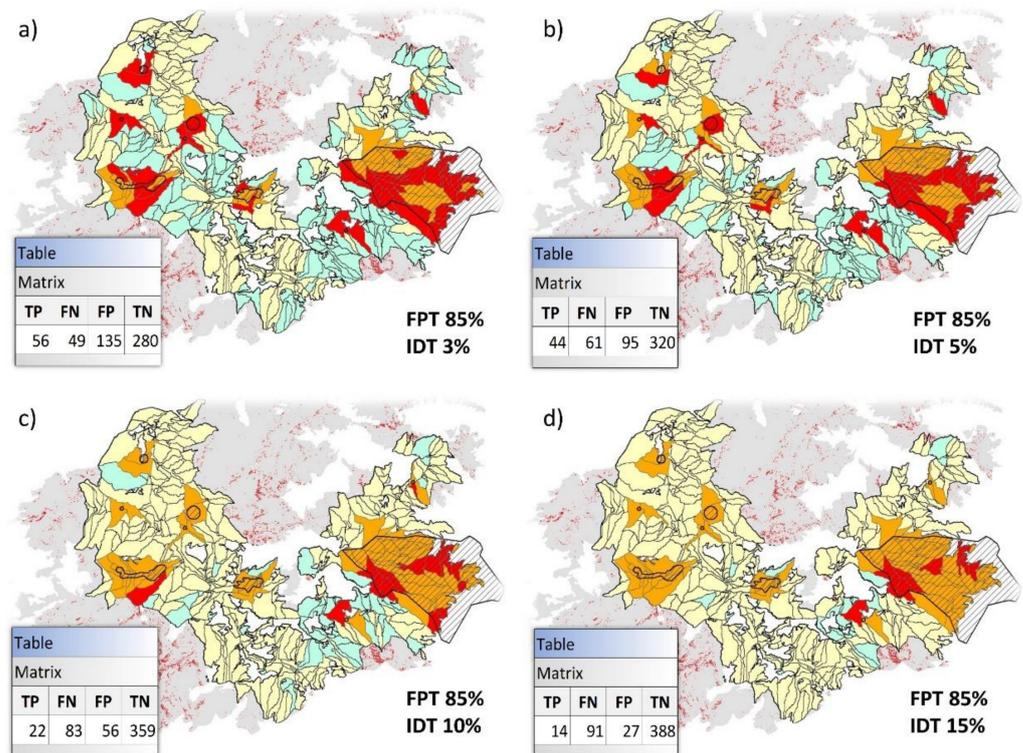


Figure 5. Example of outputs of the DTVT tool, according to four different configurations: (a) 85%/3%, (b) 85%/5%, (c) 85%/10%, and (d) 85%/15%; pixels and watersheds are considered unstable (colors are explained in Figure 4) and the resulting contingency matrixes are shown to highlight the sensitivity of the results to the approach used.

The maps are completed by the landslide catalog and by the binary raster map created in the first steps of the DTVT procedure, which visualizes in red the pixels considered unstable by the FPT of the current configuration. In this way, a complete overview of the validation result can be visually acquired and used to understand and interpret the results.

The contingency matrixes consist of a simple count of the TP, FP, TN, and FN instances. This raw output of the DTVT can be visualized as a table in the ArcMap project, together with the maps (Figure 5).

Figure 6 reports a summary of all the contingency matrixes obtained in our application, grouped by IDT values.

The contingency matrixes clearly show that depending on the criteria used for validation, the same model may get very different results. For instance, in the configurations tested, the count of true positives ranged from 68 to 105 as the IDT parameter grew. At the same time, the false positive count showed an inverse trend, diminishing from 182 to 21. This result shows the large sensitivity of the output of the contingency matrixes to the criteria used to validate the results of a slope stability model. Moreover, if the stability model has to be used for operational purposes in an EWS, the decision about which configuration produces the most desirable forecasts is not trivial and involves the selection of a tradeoff between errors of omission (false negatives) and errors of commission (false positives). In our application, we investigated two different possible criteria to select the optimal DTVT configuration to be implemented and used in a warning system to reaggregate and manage the slope stability model outputs from a pixel scale to a watershed scale.

| IDT = 3% | | Observed | | FPT |
|-----------|-----|----------|-----|-----|
| | | Yes | No | |
| Predicted | Yes | 105 | 182 | 60% |
| | No | 0 | 233 | |
| | Yes | 105 | 170 | 70% |
| | No | 0 | 245 | |
| | Yes | 105 | 158 | 75% |
| | No | 0 | 257 | |
| | Yes | 105 | 143 | 80% |
| | No | 0 | 272 | |
| | Yes | 105 | 135 | 85% |
| | No | 0 | 280 | |
| | Yes | 105 | 121 | 90% |
| | No | 0 | 294 | |

| IDT = 5% | | Observed | | FPT |
|-----------|-----|----------|-----|-----|
| | | Yes | No | |
| Predicted | Yes | 105 | 152 | 60% |
| | No | 0 | 263 | |
| | Yes | 105 | 138 | 70% |
| | No | 0 | 277 | |
| | Yes | 105 | 128 | 75% |
| | No | 0 | 287 | |
| | Yes | 105 | 110 | 80% |
| | No | 0 | 305 | |
| | Yes | 105 | 95 | 85% |
| | No | 0 | 320 | |
| | Yes | 101 | 88 | 90% |
| | No | 4 | 327 | |

| IDT = 10% | | Observed | | FPT |
|-----------|-----|----------|-----|-----|
| | | Yes | No | |
| Predicted | Yes | 100 | 97 | 60% |
| | No | 5 | 318 | |
| | Yes | 97 | 85 | 70% |
| | No | 8 | 330 | |
| | Yes | 95 | 75 | 75% |
| | No | 10 | 340 | |
| | Yes | 95 | 67 | 80% |
| | No | 10 | 348 | |
| | Yes | 87 | 56 | 85% |
| | No | 18 | 359 | |
| | Yes | 81 | 42 | 90% |
| | No | 24 | 373 | |

| IDT = 15% | | Observed | | FPT |
|-----------|-----|----------|-----|-----|
| | | Yes | No | |
| Predicted | Yes | 95 | 68 | 60% |
| | No | 10 | 347 | |
| | Yes | 95 | 55 | 70% |
| | No | 10 | 360 | |
| | Yes | 81 | 43 | 75% |
| | No | 24 | 372 | |
| | Yes | 81 | 33 | 80% |
| | No | 24 | 382 | |
| | Yes | 68 | 27 | 85% |
| | No | 37 | 388 | |
| | Yes | 68 | 21 | 90% |
| | No | 37 | 394 | |

Figure 6. Contingency matrixes resulting from all the DTVT configuration tested.

3.2. Identification of the Most Balanced Prediction

The contingency matrix tables were imported in a spreadsheet to calculate some skill scores that are commonly used as quantitative indicators of model performances [48,49]. The complete set of skill scores can be found in the Supplementary Materials. Here, for reasons of space, we present as an example the calculation of the Efficiency, defined as

$$\text{Efficiency} = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

Among all possible skill scores, we considered the *efficiency* because it can be conveniently used to describe the overall behavior of a model and measure its balance between errors of omission (false negatives) and errors of commission (false positives) [48,49]; it accounts for all four instances of the contingency matrix and it is easily understandable as it is expressed as a percentage ranging from 0 (no correct predictions obtained) to 1 (perfect prediction in all occurrences). By comparing the efficiency score reported by the 24 DTVT configurations tested (Table 1), the 80%–15% configuration can be considered as the most desirable one, because it led to the prediction outcomes that were most balanced between false alarms (FP) and missed alarms (FN). Of course, this is just an example of application and other criteria could be based on a different skill score (or on a combination of two or more skill scores), as can be seen on the next section.

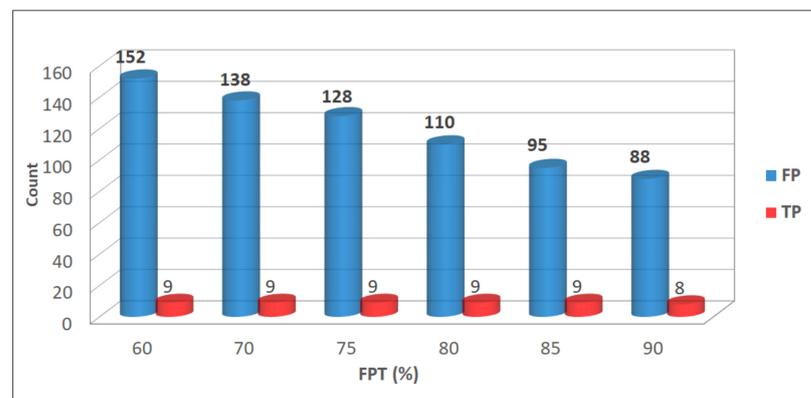
Table 1. Comparison of the efficiency skill score of all the 24 DTVT configurations tested and identification of the configuration providing the most balanced outcomes.

| Efficiency | FPT 60% | FPT 70% | FPT 75% | FPT 80% | FPT 85% | FPT 90% |
|------------|---------|---------|---------|--------------|---------|---------|
| IDT 3% | 0.650 | 0.673 | 0.696 | 0.725 | 0.740 | 0.767 |
| IDT 5% | 0.708 | 0.735 | 0.754 | 0.788 | 0.817 | 0.823 |
| IDT 10% | 0.804 | 0.821 | 0.837 | 0.852 | 0.858 | 0.873 |
| IDT 15% | 0.850 | 0.875 | 0.871 | 0.890 | 0.877 | 0.888 |

3.3. Identification of the Operational Warning Criterion

Since one of the objectives of this work was to identify a DTVT configuration to be operated in a warning system for the interpretation and reaggregation of the probabilistic outputs of HIRESSS, we performed another analysis to identify an “operational warning” criterion. In this case, the criterion used to identify the optimal configuration was driven by legal, operational, social, and political constraints which, in countries like Italy, tend to consider a missed alarm much worrying than a false alarm [11,50,51]. Consequently, the operational warning criterion was identified with a double-step procedure: first, the DTVT configurations that minimize missed alarms (FN) and thus, at the same time, maximize correct warnings (TP) should be identified. Afterwards, among them, the configuration leading to the lowest number of false alarms (FP) is selected as the operational warning criterion.

The result of the data processing showed that the IDT value of 5% returned the best cases. Figure 7 shows that an IDT value of 5% with FPT thresholds from 60% to 85%, allowed the forecast of all nine landslides. These configurations were considered for the subsequent step (while the 90% threshold was discarded because it generated a missed alarm). Among them, the operational warning criterion was selected by identifying DTVT 5–85% as the configuration with the lowest possible number of false alarms.

**Figure 7.** Double-step procedure applied to the IDT 5% outcomes to identify the optimal operational warning criterion: identification of the configurations maximizing true positives (TP) (red cylinders) and, among them, selection of the configuration minimizing false positives (FP) (blue cylinders).

It is important to remark that with respect to the criterion proposed in Section 3.2, the operational warning criterion has two main differences. First, true positives are based on the count of the landslides instead of on the number of the watersheds affected by landslides. Second, the definition of the “optimal” configuration is not based on the balance of the prediction between missed alarms and false alarms, but it is biased towards the priority of avoiding as many missed alarms as possible. False alarms are considered as well, but their reduction is, in this case, of secondary importance.

4. Discussion

The DTVT tool proved effective in automating a long series of operations (32) to get a complete validation of the results of a probabilistic distributed slope stability model. These operations are traditionally performed manually by operators in GIS programs, therefore the DTVT allows for saving a relevant amount of time and efforts.

The proposed tool is highly replicable. It should be stressed that it can be applied to any distributed slope stability model that provides probabilistic outputs at the pixel scale that need to be aggregated on wider spatial units (e.g., watersheds or municipal areas). The specific application presented here involves the HIRESSS model and sub-watersheds as spatial aggregation units, but any other slope stability model with a similar output could be used, and the spatial aggregation of the results could be performed based on any polygonal shapefile provided in the input data (wider or smaller watersheds, municipalities, slope units, and so on). To ensure the widest possible applicability, the tool was created in the form of a model builder in ArcGIS ESRI, which is the most widespread GIS software globally. The tool can be obtained upon request by contacting the corresponding author and contains a “help” section to guide the user.

The results of the tests also showed that the outcomes of the stability model can change greatly depending on how the probabilistic values are interpreted and reaggregated. As FPT and IDT threshold values are changed, watersheds can become stable or unstable, changing the count of correct and wrong predictions and, ultimately, the validation scores. As the thresholds are raised, the modeled instability is reduced and the count of missed alarms may consequently drop, but usually this happens at the cost of a lower amount of correctly predicted conditions of instability. Conversely, if thresholds are lowered, instability conditions are met more frequently, thus producing a higher count of correct hits and false alarms as well. It is also interesting to highlight that, since the validation procedure is based on two thresholds, it is possible to get intermediate results as one threshold is raised while the other is lowered, opening a wide range of possible intermediate outcomes.

This feature highlights once more the high sensitivity of the results to the key used to interpret and reaggregate them but could be also used as a point of strength to fill an existing gap in the landslide literature. No standard has been established yet to validate the results of probabilistic distributed slope stability models; thus, the tool can be used to investigate all the possible options with contained efforts. In our application, we tested 24 configurations obtaining quantitative performance indicators. These indicators can be used to define the best approach to interpret the stability model outcomes. For instance, if a balanced result is desired, the model configuration with highest efficiency could be selected. In our case, that would correspond to the 85%/15% configuration. Conversely, in preparation to an operative application in a warning system, a more precautionary configuration could be desirable (e.g., the minimization of missed alarms could be considered more important than the minimization of false alarms) and, in our case study, that would correspond to the 85%/5% configuration. It should be stressed that these values are expected to be highly site-specific, depending on the slope stability model used and on the characteristics of the case of study. However, these terms being equal, the operational warning configuration has proven to be quite stable. Indeed, it was further verified against another landslide event following the rainstorm occurred in June 2010, in the warning zone B. During this event, a single landslide event was reported in the warning zone B, in a location slightly different from the one used to calibrate the DTVT tool (Figures 1 and 8).

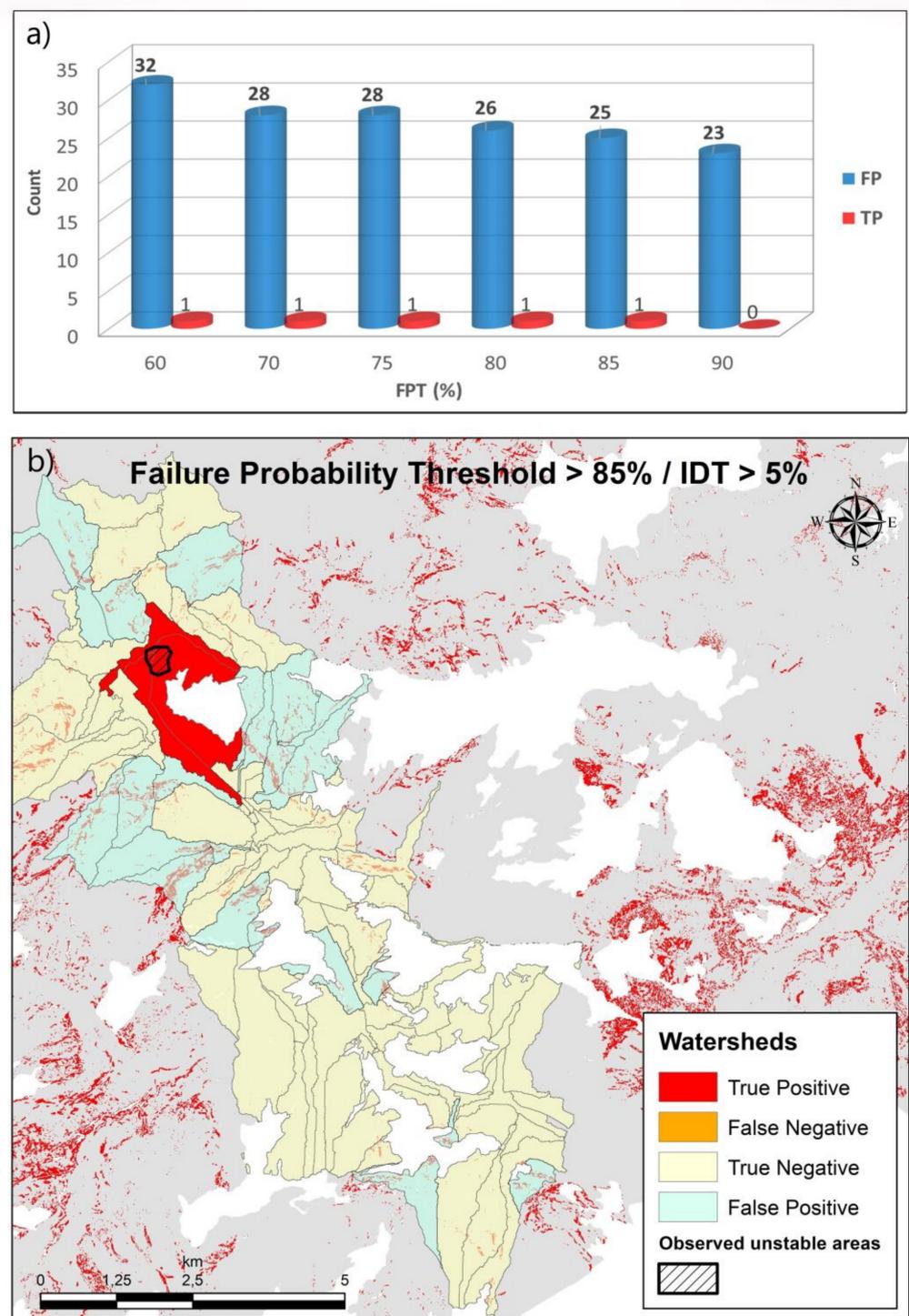


Figure 8. (a) Double-step procedure applied to the IDT 5% outcomes to verify (against the event of June 2010) the optimal operational warning criterion. The 85%/5% DTVT configuration was confirmed as the most effective one as, true positives being equal, false positives were minimized and (b) the resulting map.

The outputs of the HIRESSS model were processed using the DTVT tool until contingency matrixes were obtained for all configurations. Figure 8 shows that, as in the first series of tests, the optimal DTVT configuration for the operational warning was the one composed by a FPT of 85%, combined with an IDT of 5%. This combination allowed the

prediction of all landslides, returning the lowest number of false alarms (Figure 8a). The resulting validation map is portrayed in Figure 8b.

This test performed against another meteorological event produced further evidence about the promising potential of the operational warning criterion. However, before being implemented in an early warning system and used to reaggregate probabilistic outputs to get warnings, more tests on a larger number of landslide events are necessary. From the point of view of operational use, the outcomes presented in this work represent a preliminary result; nevertheless, they are a first important step to automate, standardize, and validate the interpretation of the results of probabilistic distributed physically based models.

5. Conclusions

This work introduces a tool that automates the validation procedure for physically based distributed probabilistic slope stability models and translates the pixelated outputs in warnings released over larger spatial units. The tool is named DTVT (double-threshold validation tool) because it uses a warning criterion on the basis of two threshold values: the probability of failure above which a pixel should be considered stable (failure probability threshold, FPT), and the percentage of unstable pixels needed in each spatial unit to consider the hazard level enough widespread to justify the issuing of an alert (instability diffusion threshold, IDT). The tool is based on a series of GIS operations organized in a model builder.

To test its efficiency, the DTVT tool was applied to the results obtained by Salvatici et al. [45] with the distributed physically based model HIRESSES [29] in two past landslide events that occurred in the Valle d'Aosta region (northern Italy), for which a reliable landslide inventory is available. The spatial units used to consider the IDT are small watersheds. A series of tests demonstrated that the results of the validation procedure are very sensitive to the criterion used to consider a watershed stable or unstable—24 configurations were tested, and a series of skill scores commonly used in the literature were automatically calculated to quantify the effectiveness of the approaches used. We discovered that the value of the skill scores was very sensitive to the two thresholds used to consider a watershed unstable.

A sensitivity study was also carried out to define the most effective DTVT configuration to be used in operational procedures to post-process raw pixel-based failure probability maps to release warnings on a watershed level—in the selected test site the 85%/5% double threshold minimized both missed alarms and false alarms.

The application showed the promising potential of DTVT in terms of (i) computational time, since the tool was able to automatically carry out in limited time (less than one minute) a series of 32 GIS operations usually carried out manually by researchers and (ii) potentiality of applicability in other test areas—the DTVT tool was conceived to be applied rapidly and widely, thus ensuring the possibility of easily getting a proper application in other test sites, tuned to the user needs, to automate the validation procedures of distributed slope stability models or to calibrate the derived warning models.

Supplementary Materials: The following are available online at <https://www.mdpi.com/article/10.3390/w13172313/s1>, Table S1: Skill scores resulting from all the DTVT configurations tested.

Author Contributions: Conceptualization, S.S.; methodology, S.S. and M.A.B.; software, G.P. and M.A.B.; validation, M.A.B. and G.P.; formal analysis, V.T., E.B.M. and G.R.; investigation, M.A.B.; resources, G.R.; data curation, E.B.M.; writing—original draft preparation, M.A.B. and S.S.; writing—review and editing, E.B.M., S.S., G.P. and V.T.; supervision, S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by “Dipartimento della Protezione Civile–Presidenza del Consiglio dei Ministri” (Presidency of the Council of Ministers—Department of Civil Protection); this publication, however, does not reflect the position or official policies of the Department. The work was supported also by Florence University, in the framework of the project SEGONISAMUELERICATEN19.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The DTVT tool presented in this work is available upon request by contacting the corresponding author. In this work, the tool has been tested on already published HIRESSS results, which are also available on request due to restrictions because they were obtained in the framework of former research agreement between the Department of Earth Sciences of the University of Firenze and the Centro Funzionale, Regione Autonoma Valle d’Aosta.

Acknowledgments: The methodological approach has been tested on HIRESSS results obtained in the framework of former research agreement between the Department of Earth Sciences of the University of Firenze, and the Centro Funzionale, Regione Autonoma Valle d’Aosta.

Conflicts of Interest: The authors declare no conflict of interest.

References

- 2009 UNISDR Terminology on Disaster Risk Reduction | UNDRR. Available online: <https://www.undrr.org/publication/2009-unisdr-terminology-disaster-risk-reduction> (accessed on 7 April 2021).
- Intrieri, E.; Gigli, G.; Casagli, N.; Nadim, F. Landslide Early Warning System: Toolbox and general concepts. *Hazards Earth Syst. Sci.* **2013**, *13*, 85–90. [CrossRef]
- Michoud, C.; Bazin, S.; Blikra, L.H.; Derron, M.H.; Jaboyedoff, M. Experiences from site-specific landslide early warning systems. *Nat. Hazards Earth Syst. Sci.* **2013**, *13*, 2659–2673. [CrossRef]
- Piciullo, L.; Calvello, M.; Cepeda, J.M. Territorial early warning systems for rainfall-induced landslides. *Earth-Sci. Rev.* **2018**, *179*, 228–247. [CrossRef]
- Baum, R.L.; Godt, J.W.; Savage, W.Z. Estimating the timing and location of shallow rainfall-induced landslides using a model for transient, unsaturated infiltration. *J. Geophys. Res.* **2010**, *115*, F03013. [CrossRef]
- Rosi, A.; Segoni, S.; Canavesi, V.; Monni, A.; Gallucci, A.; Casagli, N. Definition of 3D rainfall thresholds to increase operative landslide early warning system performances. *Landslides* **2021**, *18*, 1045–1057. [CrossRef]
- Aleotti, P. A warning system for rainfall-induced shallow failures. *Eng. Geol.* **2004**, *73*, 247–265. [CrossRef]
- Lagomarsino, D.; Segoni, S.; Fanti, R.; Catani, F. Updating and tuning a regional-scale landslide early warning system. *Landslides* **2013**, *10*, 91–97. [CrossRef]
- Devoli, G.; Tiranti, D.; Cremonini, R.; Sund, M.; Boje, S. Comparison of landslide forecasting services in Piedmont (Italy) and Norway, illustrated by events in late spring 2013. *Nat. Hazards Earth Syst. Sci.* **2018**, *18*, 1351–1372. [CrossRef]
- Segoni, S.; Piciullo, L.; Gariano, S.L. A review of the recent literature on rainfall thresholds for landslide occurrence. *Landslides* **2018**, *15*, 1483–1501. [CrossRef]
- Gariano, S.L.; Sarkar, R.; Dikshit, A.; Dorji, K.; Brunetti, M.T.; Peruccacci, S.; Melillo, M. Automatic calculation of rainfall thresholds for landslide occurrence in Chukha Dzongkhag, Bhutan. *Bull. Eng. Geol. Environ.* **2019**, *78*, 4325–4332. [CrossRef]
- Tiranti, D.; Nicolò, G.; Gaeta, A.R. Shallow landslides predisposing and triggering factors in developing a regional early warning system. *Landslides* **2019**, *16*, 235–251. [CrossRef]
- Yang, H.; Wei, F.; Ma, Z.; Guo, H.; Su, P.; Zhang, S. Rainfall threshold for landslide activity in Dazhou, southwest China. *Landslides* **2020**, *17*, 61–77. [CrossRef]
- Dikshit, A.; Sarkar, R.; Pradhan, B.; Segoni, S.; Alamri, A.M. Rainfall induced landslide studies in Indian Himalayan region: A critical review. *Appl. Sci.* **2020**, *10*, 2466. [CrossRef]
- Bordoni, M.; Corradini, B.; Lucchelli, L.; Meisina, C. Preliminary results on the comparison between empirical and physically-based rainfall thresholds for shallow landslides occurrence. *Ital. J. Eng. Geol. Environ.* **2019**, *1*, 5–10.
- Dietrich, W.M.; Montgomery, D.R. *Shalstab: A Digital Terrain Model for Mapping Shallow Landslide Potential*; Technical Report; University of California: Berkeley, CA, USA, 1998.
- Aristizábal, E.; Vélez, J.I.; Martínez, H.E.; Jaboyedoff, M. SHIA_Landslide: A distributed conceptual and physically based model to forecast the temporal and spatial occurrence of shallow landslides triggered by rainfall in tropical and mountainous basins. *Landslides* **2016**, *13*, 497–517. [CrossRef]
- Chae, B.-G.; Park, H.-J.; Catani, F.; Simoni, A.; Berti, M. Landslide prediction, Monitoring and early warning: A concise review of state-of-the-art. *Geosci. J.* **2017**, *21*, 1033–1070. [CrossRef]
- Ho, J.Y.; Lee, K.T. Performance evaluation of a physically based model for shallow landslide prediction. *Landslides* **2017**, *14*, 961–980. [CrossRef]
- Salciarini, D.; Fanelli, G.; Tamagnini, C. A probabilistic model for rainfall-induced shallow landslide prediction at the regional scale. *Landslides* **2017**, *14*, 1731–1746. [CrossRef]
- Reder, A.; Rianna, G.; Pagano, L. Physically based approaches incorporating evaporation for early warning predictions of rainfall-induced landslides. *Hazards Earth Syst. Sci.* **2018**, *18*, 613–631. [CrossRef]
- Pack, R.T.; Tarboton, D.G.; Goodwin, C.N. Assessing Terrain Stability in a GIS using SINMAP. In Proceedings of the 8th Congress of the International Association of Engineering Geology, Vancouver, BC, Canada, 21–25 September 1998.

23. Baum, R.L.; Savage, W.Z.; Godt, J.W. *TRIGRS-A Fortran Program for Transient Rainfall Infiltration and Grid-Based Regional Slope-Stability Analysis*; US Geological Survey: Reston, VA, USA, 1988.
24. Baum Jonathan, W.; Godt, R.L. Early warning of rainfall-induced shallow landslides and debris flows in the USA. *Landslides* **2010**, *7*, 259–272. [[CrossRef](#)]
25. Simoni, S.; Zanotti, F.; Bertoldi, G.; Rigon, R. Modelling the probability of occurrence of shallow landslides and channelized debris flows using GEOtop-FS. *Hydrol. Process.* **2008**, *22*, 532–545. [[CrossRef](#)]
26. Ren, D.; Leslie, L.M.; Fu, R.; Dickinson, R.E.; Xin, X. A storm-triggered landslide monitoring and prediction system: Formulation and case study. *Earth Interact.* **2010**, *14*, 1–24. [[CrossRef](#)]
27. Arnone, E.; Noto, L.V.; Lepore, C.; Bras, R.L. Physically-based and distributed approach to analyze rainfall-triggered landslides at watershed scale. *Geomorphology* **2011**, *133*, 121–131. [[CrossRef](#)]
28. Mercogliano, P.; Segoni, S.; Rossi, G.; Sikorsky, B.; Tofani, V.; Schiano, P.; Catani, F.; Casagli, N. Brief communication A prototype forecasting chain for rainfall induced shallow landslides. *Nat. Hazards Earth Syst. Sci.* **2013**, *13*, 771–777. [[CrossRef](#)]
29. Rossi, G.; Catani, F.; Leoni, L.; Segoni, S.; Tofani, V. HIRESSS: A physically based slope stability simulator for HPC applications. *Nat. Hazards Earth Syst. Sci.* **2013**, *13*, 151–166. [[CrossRef](#)]
30. Fusco, F.; De Vita, P.; Mirus, B.B.; Baum, R.L.; Allocca, V.; Tufano, R.; Clemente, E.D.; Calcaterra, D. Physically based estimation of rainfall thresholds triggering shallow landslides in volcanic slopes of Southern Italy. *Water* **2019**, *11*, 1915. [[CrossRef](#)]
31. Palazzolo, N.; Peres, D.J.; Bordonni, M.; Meisina, C.; Creaco, E.; Cancelliere, A. Improving spatial landslide prediction with 3d slope stability analysis and genetic algorithm optimization: Application to the oltrepò pavese. *Water* **2021**, *13*, 801. [[CrossRef](#)]
32. Crosta, G.B.; Frattini, P. Distributed modelling of shallow landslides triggered by intense rainfall. *Nat. Hazards Earth Syst. Sci.* **2003**, *3*, 81–93. [[CrossRef](#)]
33. Fusco, F.; Mirus, B.B.; Baum, R.L.; Calcaterra, D.; De Vita, P. Incorporating the effects of complex soil layering and thickness local variability into distributed landslide susceptibility assessments. *Water* **2021**, *13*, 713. [[CrossRef](#)]
34. Tofani, V.; Bicocchi, G.; Rossi, G.; Segoni, S.; D'Ambrosio, M.; Casagli, N.; Catani, F. Soil characterization for shallow landslides modeling: A case study in the Northern Apennines (Central Italy). *Landslides* **2017**, *14*, 755–770. [[CrossRef](#)]
35. Bicocchi, G.; Tofani, V.; Tacconi-Stefanelli, C.; Vannocci, P.; Casagli, N.; Lavorini, G.; Trevisani, M.; Catani, F. Geotechnical and hydrological characterization of hillslope deposits for regional landslide prediction modeling. *Bull. Eng. Geol. Environ.* **2019**, *78*, 4875–4891. [[CrossRef](#)]
36. Schmaltz, E.M.; Van Beek, L.P.H.; Bogaard, T.A.; Kraushaar, S.; Steger, S.; Glade, T. Strategies to improve the explanatory power of a dynamic slope stability model by enhancing land cover parameterisation and model complexity. *Earth Surf. Process. Landf.* **2019**, *44*, 1259–1273. [[CrossRef](#)]
37. Bregoli, F.; Medina, V.; Chevalier, G.; Hürlimann, M.; Bateman, A. Debris-flow susceptibility assessment at regional scale: Validation on an alpine environment. *Landslides* **2015**, *12*, 437–454. [[CrossRef](#)]
38. Vasseur, J.; Wadsworth, F.B.; Lavallée, Y.; Bell, A.F.; Main, I.G.; Dingwell, D.B. Heterogeneity: The key to failure forecasting. *Sci. Rep.* **2015**, *5*, 1–7. [[CrossRef](#)] [[PubMed](#)]
39. Zhou, G.L.; Xu, T.; Heap, M.J.; Meredith, P.G.; Mitchell, T.M.; Sesnic, A.S.Y.; Yuan, Y. A three-dimensional numerical meso-approach to modeling time-independent deformation and fracturing of brittle rocks. *Comput. Geotech.* **2020**, *117*, 103274. [[CrossRef](#)]
40. Canavesi, V.; Segoni, S.; Rosi, A.; Ting, X.; Nery, T.; Catani, F.; Casagli, N. Different approaches to use morphometric attributes in landslide susceptibility mapping based on meso-scale spatial units: A case study in Rio de Janeiro (Brazil). *Remote Sens.* **2020**, *12*, 1826. [[CrossRef](#)]
41. Domènech, G.; Alvioli, M.; Corominas, J. Preparing first-time slope failures hazard maps: From pixel-based to slope unit-based. *Landslides* **2020**, *17*, 249. [[CrossRef](#)]
42. Palau, R.M.; Hürlimann, M.; Berenguer, M.; Sempere-Torres, D. Influence of the mapping unit for regional landslide early warning systems: Comparison between pixels and polygons in Catalonia (NE Spain). *Landslides* **2020**, *17*, 2067–2083. [[CrossRef](#)]
43. Crosta, G.B.; Frattini, P. Rainfall-induced landslides and debris flows. *Hydrol. Process.* **2008**, *22*, 473–477. [[CrossRef](#)]
44. Kuriakose, S.L.; van Beek, L.P.H.; van Westen, C.J. Parameterizing a physically based shallow landslide model in a data poor region. *Earth Surf. Process. Landf.* **2009**, *34*, 867–881. [[CrossRef](#)]
45. Salvatici, T.; Tofani, V.; Rossi, G.; D'Ambrosio, M.; Tacconi Stefanelli, C.; Benedetta Masi, E.; Rosi, A.; Pazzi, V.; Vannocci, P.; Petrolo, M.; et al. Application of a physically based model to forecast shallow landslides at a regional scale. *Nat. Hazards Earth Syst. Sci.* **2018**, *18*, 1919–1935. [[CrossRef](#)]
46. Canli, E.; Mergili, M.; Thiebes, B.; Glade, T. Probabilistic landslide ensemble prediction systems: Lessons to be learned from hydrology. *Nat. Hazards Earth Syst. Sci.* **2018**, *18*, 2183–2202. [[CrossRef](#)]
47. Richards, L.A. Capillary conduction of liquids through porous mediums. *J. Appl. Phys.* **1931**, *1*, 318–333. [[CrossRef](#)]
48. Beguería, S. Validation and evaluation of predictive models in hazard assessment and risk management. *Nat. Hazards* **2006**, *37*, 315–329. [[CrossRef](#)]
49. Martelloni, G.; Segoni, I.S.; Fanti, I.R.; Catani, I.F. Rainfall thresholds for the forecasting of landslide occurrence at regional scale. *Landslides* **2012**, *9*, 485–495. [[CrossRef](#)]

-
50. Segoni, S.; Tofani, V.; Rosi, A.; Catani, F.; Casagli, N. Combination of rainfall thresholds and susceptibility maps for dynamic landslide hazard assessment at regional scale. *Front. Earth Sci.* **2018**, *6*, 85. [[CrossRef](#)]
 51. Abraham, M.T.; Satyam, N.; Bulzinetti, M.A.; Pradhan, B.; Pham, B.T.; Segoni, S. Using field-based monitoring to enhance the performance of rainfall thresholds for landslide warning. *Water* **2020**, *12*, 1–21. [[CrossRef](#)]