PMT: new analytical framework for automated evaluation of geo-environmental modelling approaches

Omid Rahmati [1,2*], Aiding Kornejady [3], Mahmood Samadi [4], Ravinesh C. Deo [5], Christian Conoscenti [6], Luigi Lombardo [7], Kavina Dayal [8], Ruhollah Taghizadeh-Mehrjardi [9, 10], Hamid Reza Pourghasemi [11], Sandeep Kumar [12], Dieu Tien Bui [13,14*]

[1] Geographic Information Science Research Group, Ton Duc Thang University, Ho Chi Minh City, Vietnam

[2] Faculty of Environment and Labour Safety, Ton Duc Thang University, Ho Chi Minh City, Vietnam

[3] Young Researchers and Elite Club, Gorgan Branch, Islamic Azad University, Gorgan, Iran

[4] Faculty of Natural Resources, University of Tehran, Karaj, Iran

[5] School of Agricultural, Computational and Environmental Sciences, Centre for Sustainable Agricultural Systems & Centre for Applied Climate Sciences, University of Southern Queensland, Springfield, QLD 4300, Australia

[6] Department of Earth and Marine Sciences (DISTEM), University of Palermo, Via Archirafi 22, 90123 Palermo, Italy

[7] Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede, Netherlands

[8] CSIRO Agriculture and Food, 15 College Road, Sandy Bay, TAS 7005, Australia

[9] Department of Geosciences, Soil Science and Geomorphology, University of Tübingen, Tübingen, Germany

[10] Faculty of Agriculture and Natural Resources, Ardakan University, Ardakan, Iran

[11] Department of Natural Resources and Environmental Engineering, College of Agriculture, Shiraz University, Shiraz, Iran

[12] Department of Agronomy, Horticulture and Plant Science, South Dakota State University, Brookings, USA

[13] Institute of Research and Development, Duy Tan University, Da Nang 550000, Vietnam

[14] Geographic Information System Group, Department of Business and IT, University of South-Eastern Norway, N-3800 Bø i Telemark, Norway

* Corresponding authors' Email addresses: Orahmati68@gmail.com; Dieu.T.Bui@usn.no

PMT: new analytical framework for automated evaluation of geo-environmental modelling approaches

Abstract

Geospatial computation, data transformation to a relevant statistical software, and step-wise quantitative performance assessment can be cumbersome, especially when considering that the entire modelling procedure is repeatedly interrupted by several input/output steps, and the self-consistency and self-adaptive response to the modelled data and the features therein are lost while handling the data from different kinds of working environments. To date, an automated and a comprehensive validation system, which includes both the cutoff-dependent and –independent evaluation criteria for spatial modelling approaches, has not yet been developed for GIS based methodologies. This study, for the first time, aims to fill this gap by designing and evaluating a user-friendly model validation approach, denoted as Performance Measure Tool (PMT), and developed using freely available Python programming platform. The considered cutoff-dependent criteria include receiver operating characteristic (ROC) curve, success-rate curve (SRC) and prediction-rate curve (PRC), whereas cutoff-independent consist of twenty-one performance metrics such as efficiency, misclassification rate, false omission rate, F-score, threat score, odds ratio, etc. To test the robustness of the developed tool, we applied it to a wide variety of geo-environmental modelling approaches, especially in different countries, data, and spatial contexts around the world including, the USA (soil digital modelling), Australia (drought risk evaluation), Vietnam (landslide studies), Iran (flood studies), and Italy (gully erosion studies). The newly proposed PMT is demonstrated to be capable of analyzing a wide range of

environmental modelling results, and provides inclusive performance evaluation metrics in a relatively short time and user-convenient framework whilst each of the metrics is used to address a particular aspect of the predictive model. Drawing on the inferences, a scenario-based protocol for model performance evaluation is suggested.

Keywords**:** PMT, Spatial modelling, Goodness-of-fit, Validation, Performance analysis; predictive model evaluation framework

Software and data availability

Name of tool:	PMT (Performance Measure Tool)
Developers:	Samadi M., Kornejady A., and Rahmati O.
Hardware required:	General-purpose computer (2 Gb RAM)
Software required:	ArcGIS 10.2
Programming languages:	Python© 2.7
Program size:	120 KB
Availability and cost:	Freely available in GitHub
Web link:	https://github.com/mahmoodsamadi/PMT
Year first available:	2018

1. Introduction

Spatially-applicable predictive models must include a mandatory step where different aspects of the model performance can be quantitatively benchmarked. Without considering the performance of such geo-environmental models, the users would not be confident about the veracity of the modelling results, and is unlikely to utilize them for practical decision making

(Pullar and Springer, 2000; Glade, 2005; Beguería, 2006). The accuracy of predictive models, which is a pertinent factor demonstrating the usefulness of the relevant models, can significantly result in the misclassification costs of the approach depending on the error magnitudes and types (Frattini et al., 2010). For example, in the modelling of natural hazards, the Error Type I (*i.e*., false positive) is likely to indicate that a stable part of a spatial region is classified as being unstable, and therefore, it can lead to unnecessary control and risk mitigation measures that are implemented. The Error Type II (*i.e*., false negative) can imply that a given terrain unit is susceptible to the hazard, and it can be incorrectly classified as being stable, and consequently, this terrain region can be allowed to be occupied by people or infrastructure without a responsible and actionable risk mitigation activity. These errors, if not assessed properly, can consequently incur social and economic costs, depending on the vulnerability and economic value of the elements at risk (*e.g*., infrastructures, lives, etc.). In light of this need, a robust investigation of such predictive errors in spatially-applicable models is highly warranted, to make the modelling approaches and model results more viable for real-life usage, risk mitigation and implementation.

Over the past couple of decades, a number of susceptibility assessment models have been built, each striving to portray the current and future spatial patterns of a specific phenomenon. Many studies have included a "model comparison" or a "performance assessment" step that was aimed to evaluate the spatial modelling result, and to select the most optimal spatially-relevant model. These sorts of models, largely promulgated as an operational tool, have largely been reported in different fields and applications, such as landslide susceptibility studies (*e.g.,* Kornejady et al., 2017; Kavzoglu et al., 2019; Yan et al., 2019), flood susceptibility studies (*e.g.,* Rahmati and Pourghasemi, 2017; Siahkamari et al., 2018; Choubin et al., 2019), forest fire modelling purposes (*e.g.,* Arpaci et al., 2014;Tien Bui et al., 2017), groundwater potential modelling studies (*e.g.,* Naghibi et al., 2017; Miraki et al., 2019), species distribution modelling tasks (*e.g.,* Bucklin et al., 2015; Shabani et al., 2016; Quillfeldt et al., 2017), land subsidence modelling (*e.g.,* Abdollahi et al., 2018; Ghorbanzadeh et al., 2018), soil digital mapping (*e.g.,* Minasny and McBratney, 2007; Wiesmeier et al., 2011; Malone et al., 2017), gully-erosion susceptibility (*e.g.,* Akgün and Türk, 2011; Conoscenti et al., 2014; Garosi et al., 2018). The evaluation of predictive models with different statistical metrics and their implemented approaches, especially in such a diverse range of studies, clearly warrant automated and coherent scientific strategies where performance evaluations are implemented by means of a universally acceptable and statistically robust tool.

A review of published literature in this respect reveals significant advancements in predictive model performance evaluations where the context of application and the respective model type were seen to play a pivotal role in how these evaluation tools were implemented. Recently, the study of Rahmati and Pourghasemi (2018) compared the performance of ten different advanced machine learning models for the modelling of landslide susceptibility, while the study of Fukuda et al. (2013) applied and compared seven different data-driven models for developing species distribution maps. These authors considered the receiver operating characteristic (ROC) curves and a number of cutoff-dependent methods for judging the capability of their model, and consequently, in preparing and transporting the results to their statistical software, although this was a relatively time-consuming task. Particularly, one must note that when susceptibility maps are supposed to be directly incorporated into land-use planning, the best performing model are likely to be highly favored for practical decision-making tasks (Youssef et al., 2016; Siahkamari et al., 2018). This is primarily because the model performance assessments provide immensely useful insights into the optimal structure of such models, and the possibility of their practical implementation for perceived risk mitigation (Van Westen, 2006).

Most performance evaluation metrics that are designed to evaluate the overall learning skill of the predictive model, and the validity of the generated results from them are based on comparing the predicted patterns in spatial models with the actual observation datasets (Chung and Fabbri, 2008). In a somewhat different approach to the traditional model evaluation approaches (*e.g.,* graphical check of the model's susceptibility maps in respect to the ground-truth datasets), the new generation of model performance metrics is mainly applicable for quantifying the traditional terms and the models' functionality. According to a general consensus, the performance indices in a predictive model can be classified into two different categories: cutoff-dependent metrics (*e.g*., Cohen's Kappa, sensitivity, and specificity) and the -independent metrics (*e.g*. receiver operating characteristic, ROC method) (Frattini et al., 2010). These approaches have been used in a number of spatial modelling sub-fields.

Meanwhile, there is little doubt that the ArcGIS software, by virtue of its wide flexibility, portability and the relevance in spatial modelling approaches (e.g. geostatistics, mapping tools, variogram, kriging, and local/global scale metrics), has been unceasingly used by many researchers to implement the most basic as well as the more complex spatial functions and statistical criterion that are available. In spite of this widespread usage of ArcGIS software as a spatial modeling platform , the absence of a dedicated GIS-based tool and its non-availability to aspiring researchers and practitioners who are outside of the major subscribed users and institutions, is still very challenging (Scott and Janikas, 2010). Furthermore, the GIS users need to employ cumbersome step-by-step procedures in order to calculate each of their performance indices, and occasionally, they need to reach out for additional commercial and/or freely available software platforms (*e.g*., Microsoft Excel, SPSS, and R packages). These types of external model evaluation frameworks and largely the expensive software that need to be used to analyze these data outside of GIS platforms, represent a challenging task when aiming at optimizing any modeling workflow.

In respect to these arguments for more robust evaluation of spatially-relevant predictive models, some of the freely available software, such as the R package in the form of “cvTools” (Alfons, 2012) or “CrossValidate” (Coombes, 2018), and the relevant modelling platforms in the R software have partially satisfied the need to compute these metrics. However, these add-in tools also seem to be relatively deficient in terms of their inclusiveness in the respective modelling approach, and also sometimes, they may require additional external coding skills, which in some cases may not available to the users. Furthermore, each of these add-in software are likely to include only some of the cutoff-dependent and/or –independent evaluation criteria, and not include the others (as necessary) within a universally desirable manner, and therefore, the external software may be less flexible and attractive to the novice modeler and other non-scientific stakeholders, practitioner and decision-makers.

To address inherent limitations posed by existing approaches adopted in the evaluation of spatial models, this research study aims to propose and construct a new, robust and comprehensive GIS-based package, denoted as the Performance Measure Tool (PMT), to scrutinize in a statistically sound manner the performance of spatially-relevant predictive models. The merits of the proposed PMT, augmented by its extensive validation in diverse regions, contextual applications and global studies, are likely to enable modelers and risk mitigation practitioners to calculate practically useful performance metrics (both cutoff-dependent and the –independent category). The PMT is designed in such way that it has the ability to provide information in a tabular and graphical format with a relatively simple platform and self-explanatory user interface. This proposed tool is likely to be useful for any spatially-relevant model, various types of end-users—from the beginner who are not familiar with advanced coding, to those who are comfortable with a 'click-based procedure' and also practitioners in any scientific sub-field who need to implement decisions about the model's versatility. To further ensure credibility and generalizability of the software, the proposed PMT has been benchmarked rigorously to evaluate its relative performance in different geo-environmental modelling contexts and in different parts of the world including studies in Australia, Asia, Europe, and America.

2. Basic Design Framework of the Performance Measure Tool

Implementing the notion that performance evaluation of a spatially-relevant predictive model must be an important cornerstone of any spatial modelling attempt, in this study different cutoff-dependent and –independent evaluation criteria, elaborated later in greater detail, have been proposed. A brief review of recent literature shows that most of these analyses are underpinned by a matrix-wise calculation, termed as the confusion matrix (and also, sometimes known as the table of confusion, error matrix, or the matching matrix) and the contingency table (also known as the cross tabulation or crosstab). Some researchers have interchangeably used these two names in their studies and considered the confusion matrix largely as a special derivative of the contingency matrix. Other researchers, however, pointed out a delicate, and logical difference, in that the former is more suitable for evaluating the performance of different classifiers (*i.e*., more common in data mining models), while the latter is used to evaluate the rules of association and interrelations between any two variables (Powers, 2011). However, the name "matching matrix" is well-adapted in unsupervised machine learning algorithms, whereas the confusion matrix is used in supervised learning (*i.e*., input data fed by the training instances).

In this research, the confusion matrix has been considered as a way to describe the primary basis for constructing the proposed spatially-relevant model evaluation tool. Consequently, a 2 × 2 confusion matrix is created where the rows are the instances in an actual class (*i.e.*, the observations) and the columns are the instances in the predicted class, as illustrated in Table 1. As the name "confusion" implies, the matrix is able to examine the degree of mislabeling one state (as another) by means of directly comparing the predictions and the observations. The statistics derived from the matrix are therefore all presented as either the row-wise (*e.g.*, positive and negative predictive values) or the column-wise (*e.g.*, sensitivity and specificity) in the implemented PMT tool.

Table 1 HERE

It should be emphasized that the process and various stages of model performance assessments can be rather a time-consuming and a complex task for the performance measures in a traditional approach must be calculated separately using the geo-statistical techniques. This is particularly the case for novice end-users (*e.g.*, risk mitigation practitioners who may be unfamiliar with various mathematical and statistical knowledge). More importantly, to the best of the authors' knowledge, there is hardly any reliable, comprehensive and end-user-friendly tool currently available that can be used to consider the most relevant performance metrics, particularly in the widely adopted ArcGIS environment. Considering this deficit, this paper aims to develop an efficient and automated approach that operates in a quick, reliable and organized manner, and also presents a relatively effective framework providing a user-friendly interface. The PMT has deliberately been written in the freeware, the Python programming environment using a portability feature that enables it to be installed easily within a geo-processing framework found in the ArcToolbox of the ArcGIS 10.2 software.

Figure S1 (refer to supplementary information) illustrates the graphical user interface and the execution process of the proposed PMT.

Fig. S1 HERE

To illustrate the operational mechanism of the proposed PMT, one part of the Python code used for calculating the evaluation criteria is displayed in Figure S2. The required inputs used to execute the tool and the relevant outputs files are given in Tables 2 and 3, respectively. It is important to note that the PMT extension allows the end-users to evaluate the accuracy of the predictive model in both steps, composed of training/calibration and the validation phase. End-users can also adopt both parts of the training and validation process to check the accuracy of their predictive models, although investigating the accuracy of the model in the training step can also be left unchecked in this particular tool. This option is added because most of the interest is usually focused on the validation component, as it guarantees the viability of the model to be used for the prediction and decision-making process. Conversely, calibration is a component uniquely voted to build the reference model, and to evaluate the covariate effects, although these can be subjected to some degree of overfitting (Lombardo et al., 2018). These stages make the model easy-to-use with no special skills required to run the proposed tool.

Fig. S2 HERE

Table 2 HERE

Table 3 HERE

3. Statistical background of the performance metrics

3.1. Confusion matrix

In what follows next, the authors outline the kinds of information these metrics are able to convey regarding the model performance. In order to construct a confusion matrix from a spatial model, the users should define a cutoff (in percentile units) to split the spatial map into two distinct classes in which the PMT can calculate the cutoff-dependent performance metrics. This is the analogous operation to splitting a probability distribution into two distinct classes, although in our case, this is performed directly within ArcGIS into map form. In this process, the first class (*i.e.*, the lower percentage of susceptibility/ suitability map) is considered as the absence areas (*e.g.,* the landslide-free areas) and the upper part as the presence locations (*e.g.*, the landslide affected areas). For instance, let us assume a 50% cutoff for a landslide susceptibility map of particular interest with 20 landslides located within the lower 50% (*i.e.*, low to moderate susceptible areas). In this case, those 20 samples will be considered as error sources (denoted as the 'false negative error', that has been discussed later) by the proposed tool and consequently, it can reduce the performance of the predictive model since the landslides that have already occurred are supposed to be located within the areas with the highest susceptibility values. The 50% cutoff value is also quite common in existing literature, especially for the equally balanced presence/absence datasets (*e.g.,* Lombardo and Mai, 2018). However, the prevalence can be considered as the best alternative since it is able to represent the inherent predominance of a phenomenon and it is not controlled by the experimenter. Additionally, quantifying the prevalence of a natural phenomenon is somewhat problematic (discussed in Section *5.3*). Most of the data mining models can circumvent this issue by calculating the prevalence by means of estimating the best possible distribution of an event using generalized algorithms which is common in the presence-only models (*e.g.* Maximum entropy model).

3.2. Cutoff-dependent Approach

Cutoff-dependent metrics include True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), False Negative Rate (FNR), Misclassification rate, Accuracy, Positive Predictive Value (PPV), False Discovery Rate (FDR), Negative Predictive Value (NPV), False Omission Rate (FOR), F-score, Matthews Correlation Coefficient (MCC), Informedness (Bookmaker informedness; BM), Markedness (MK), Threat Score, Equitable threat score, True skill statistic, Heidke's skill score, Odds ratio, Odd ratio skill score, and Cohen's kappa. Table 4 details the equations for all of the cutoff-dependent metrics.

Table 4 HERE

The TPR, also termed as the sensitivity, recall, or hit rate, represents the probability of correctly predicting the positives as observed in reality (given as True positives (TP) / total number of positives (P)). The TNR, termed as the specificity, aims to quantify the probability of correctly predicting the negatives as observed in reality (given as true negatives (TN)/ total number of negatives (N)). The FPR, also known as the "1–specificity" or fall-out, aims to indicate the probability of incorrectly predicting a non-event location as an event (given as false positives (FP)/ total number of negatives (N)). Furthermore, the FNR, also denoted as the miss rate, indicates the probability of incorrectly predicting an event location as a non-event (given as false negatives (FN)/ total number of positives (P)). This quantity is used to express how often the model wrongly predicts absences. Misclassification rate undertakes both the false negative and false positive values and therefore reflects an overall error rate ((FP+FN)/total). The accuracy (or the model efficiency) is the opposite metric compared to the misclassification rate, since it is able to highlights the overall success of the predictive model ((*i.e.*, TP+TN)/total). Overall, this metric shows how often the predictive model is correct. The PPV, also denoted as the confidence or the precision in data mining approaches, or as Powers (2011) analogously calls it as the accuracy of predicted positives, is used to measure the proportion of predicted presences that correctly represent the real presence. As a complement component of the PPV, a false discovery rate is applied to conceptualize the Type I errors (*i.e.*, rejection of a true null hypothesis) (Benjamini and Hochberg, 1995). In accordance with the PPV, the NPV is used to measure the precision of the predictive model in predicting the absence (or non-event) locations. However, this metric largely ignores how well the model is able to handle the presence locations and that the FOR simply is the complement of the NPV. The F-score is also called the harmonic mean of the precision and the recall (i.e., sensitivity) where it reaches its best values at 1 (*i.e.*, best precision and recall) and the worst at 0. In essence, MCC is a correlation coefficient metric computed between the observed and the predicted binary classifications, and it is able to undertake a true and a false positive and negative value. The terms *informedness* and *markedness*, implemented in the PMT, were introduced initially by Powers (2011). Informedness, however, is likely to be the only unbiased indicator in the confusion matrix and it measures the probability that an informed decision that is being made rather than guessing, either the correct or the incorrect decision (due to overtraining, atypical data, or even deliberately) (Powers, 2011). Markedness, also referred to as *deltaP* in psychology, is the complementary pair of informedness indicating the probability that an outcome is marked by the predictor (marker). Threat Score also penalizes the rare events since some success of correct predictions of a less frequent event might be resulted out of random chance. Although Threat Score uses different statistics in conjunction, the actual sources of misclassification error are not discernible. Equitable Threat Score also known as the Gilbert's skill score (Gilbert, 1884; Schaefer, 1990), the equitable threat score functions as per above based on critical success score, but it is also used to eliminate the hit rates (*i.e.*, true positive rates) originated by random chance. True skill statistic (TSS) (also called the Hanssen and

Kuipers discriminant or Pierces skill score), is applied to measure the ability of a predicted value to discriminate between the events and the non-events, using all of the elements in the confusion matrix. The Heidke's Skill Score operates according to the accuracy level but it is also used to improve its meaning by eliminating the true positive rates that would be expected to occur by chance (Heidke, 1926). Odd Ratio is used to measure the odds that an event (or an outcome) will occur given a particular exposure, compared to the odds of the event occurring in the absence of that exposure (Pepe et al., 2004). Odd Ratio Skill Score (also known as the Yule's Q) rescales the values of the odds ratio into the -1 and the +1 range. In addition, Kappa is essentially a measure of how well the model has performed as compared to how well it would have performed purely by chance, and this would enable the modeler to better understand the true outcome of the model in respect to the random occurrence of that value.

3.3. Cutoff-independent approach

This approach, included in the PMT, includes two different methods that can be categorized as: (1) receiver operating characteristic (ROC) curve, and (2) success-rate curve (SRC) and prediction-rate curve (PRC).

3.3.1. ROC curve

The ROC curve, used typically in risk assessment through predictive model results, simply plots the sensitivity (*i.e.*, true positive rates) on the *Y*-axis against the 1–specificity (*i.e.*, false positive rate) on the *X*-axis (Gorsevski et al., 2006). The area under the ROC curve (denoted as AUROC, bounded by [0, 1]), is the actual measure of the model evaluation since it generates a quantitative value of the performance (Pontius and Schneider, 2001; Mas et al., 2013; Swets, 2014). The closer the AUROC is to unity, the better is the performance. The ROC curve can be interpreted differently depending on the dataset; it can address the learning capability (or the so-called goodness-of-fit) of the model if the training set is used for plotting; it can also infer the predictive skill of the model if the validation set is used (Fawcett, 2006; Lombardo and Mai 2018).

In this regard, the proportion between training and validation samples is highly relevant. A 70:30% split is quite common among the researchers (Pradhan and Lee, 2010). Although different partitions have also been used, such as 80:20% (*e.g.* Lipovetsky, 2009), 70:30% (*e.g.* Choubin et al., 2019) or even 50:50% (*e.g.* Deo et al., 2016; Deo et al., 2017), there is no empirical consensus on the best partition since this is more of an expert-user based decision. Irrespective of this, having a large amount of inventory data (*i.e.*, number of events), one can assign a greater percentage of such data to train the predictive model and a lesser percentage for validation. Opting for a suitable approach to partition the training and validation sets is yet another crucial matter that has been the subject of many studies, e.g. Kornejady et al. (2017). In this regard, the random sampling, self-organizing maps for input selection, Mahalanobis distance, excerpting separate training/validation areas, and temporal partitioning are all some of the common sample partitioning approaches. For more details, readers can refer to the references therein.

3.3.2. Success-Rate Curve (SRC) and Prediction-Rate Curve (PRC)

The SRC is a measure of the learning capability of the model, while the PRC is able to examine the predictive power. Although the SRC and the PRC may share some common features with the ROC, the ROC in particular uses almost all the elements of the confusion matrix. This includes positive (TPR and TNR) and negative (FPR and FNR) aspects of the model, while the SRC and the PRC are calculated independently from the confusion matrix. In fact, the SRC represents the cumulative areal percentage of the susceptibility classes (*i.e*, from the highest values to the lowest) on the *X*-axis against the areal cumulative percentage of the training set located within those susceptibility classes on the Y-axis (Chung, 2006; Blahut et al., 2010). In terms of its physical interpretation, a steeper SRC curve is used to indicate that more events fall within the highly susceptible classes; *i.e.*, a good learning skill. The PRC curve, however, follows the same plotting process as the SRC, but the training data are replaced by the validation set.

4. Testing the Efficacy of PMT: Selected Case Studies

In this section, the proposed PMT is applied to 5 distinct, real geo-environmental modelling tasks and case studies in order to robustly investigate its credibility and generalizability, and also to demonstrate the potential benefits in considering different evaluation criteria promulgated by the PMT. It is imperative to note that the selected case studies exhibited various noticeable characteristics in terms of the issue under investigation, the modelling strategies, the overall frameworks and the predictive model type, spatial or temporal scales considered and the geographical and climatic conditions that influence the results and implementation of the model.

To provide a robust evaluation of the proposed PMT, the most relevant and a relatively diverse range of data sets were obtained from most recently conducted research studies and also some newly implemented models based on: (1) gully erosion prediction mapping in two small catchments of central-western Sicily, Italy (Conoscenti et al., 2018) (2) flood hazard modelling in the Galikesh region, Iran (Rahmati and Pourghasemi, 2017) (3) drought risk modelling in south-east Queensland, Australia (Dayal, 2018; Dayal et al. 2018) (4) landslide susceptibility modelling in the Kon Tum province, Vietnam (5) soil digital modelling in South Dakota, USA (Fig. 1). Each of these studies employed a range of geo-spatial models where the PMT is used to provide a consolidated assessment of its efficacy in providing greater insights into the practicality of the modelling various frameworks.

An overall description of the study areas and the applied models are provided as follows whereas further details of the modelling approaches are provided in the references therein.

A detailed flowchart of the various studies is shown in Fig. 2.

Fig. 1 HERE

Fig. 2 HERE

4.1. Gully Erosion Modelling (Italy)

Intense farming activities in two small catchments of central-western Sicily, Italy, have expedited many erosion processes. In particular, the gully erosion has led to the landscape dissection and massive soil loss (Conoscenti et al., 2018). The gullies in the study area have developed as a result of the interrelation of several geo-environmental factors and human activities such as access roads, parcel borders, wheel tracks, and plow furrows. In addition to the multivariate adaptive regression splines (MARS) model already utilized by Conoscenti et al. (2018) for gully erosion prediction mapping, in this paper we used the generalized linear model (GLM) to conduct a fair comparison of their approach (Fig. 3).

Fig. 3 HERE

4.2. Flood Hazard Modelling (Iran)

Over the last few decades, the Galikesh region, located in the Golestan province, in the north-east of Iran, has witnessed severe flood events due to the particular climatic and topo-hydrological conditions that resulted in many economic losses and causalities attributable to environmental mismanagement (*e.g.*, deforestation, overgrazing, and over-exploitation). Since flood-inundation has been one of the major issues of the urban areas in Golestan province for decades, Rahmati and

Pourghasemi (2017) used evidential belief function (EBF) to investigate the flood-prone hotspots (Fig. 4). In this paper, we have implemented the proposed PMT as a statistical and decision-support tool to provide an inclusive performance evaluation of their model.

Fig. 4 HERE

4.3. Drought Risk Modelling (Australia)

An area located in the south-east of Queensland, Australia, encompasses intensive agricultural activities, such as grazing, horticulture, and animal production, other than the densely populated localities, which require a reliable water supply. As the study area is affected by severe and frequent drought events, Dayal (2018) and Dayal et al. (2018) attempted to develop a spatial drought risk map by employing the Bayes' theorem (*i.e.*, classifying spatial indicators), fuzzy logic (*i.e.*, standardizing spatial indicators), and fuzzy GAMMA overlay (*i.e.*, aggregating drought vulnerability, exposure, and hazard indices) technique (Fig. 5). Employing the findings of that study, in this paper we utilized their final drought risk map as a potential input to the proposed PMT, enabling us to examine the different aspects of its performance over the geospatial scale. In order to investigate the influence of the cutoff values on the performance analysis, three different cutoffs, *i.e.*, 50%, 70%, and 90% were taken into account and the results were compared, as illustrated in Fig. 6.

Fig. 5 HERE

Fig. 6 HERE

4.4. Landslide Susceptibility Modelling (Vietnam)

Landslides are the dominant geo-hazardous elements in the Kon Tum province of Vietnam. Hence, this study has used two novel data mining models including maximum entropy (MaxEnt) and a recently developed model named as BayGmmKda (Bayesian-based ensemble of Gaussian mixture model and radial-basis-function Fisher discriminant analysis) (Tien Bui and Hoang, 2017) (Fig. 7). This study also uses the proposed PMT to highlight the potential asymmetries among the performance metrics.

Fig. 7 HERE

4.5. Soil Digital Modelling (USA)

Soil digital modelling has received significant attention amongst scientists in recent years, where computer-assisted pedometric-predictive mapping of soil properties has led to the creation of an inclusive geographically-referenced soil database. To this end, an attempt is carried out to map the soil bulk density (BD) predictive distribution in South Dakota, USA, by obtaining soil bulk density samples of the study area and using two data mining models, namely the artificial neural network (ANN) and decision tree (DAT) (Fig. 8). We have delineated the need for rendering quantitative suitability maps into probability values to be able to use the proposed PMT for further assessing the models' performance. In general, there is a few differences between models' requirements. For example, DAT model does not require a separate dataset to optimize parameters and just uses the training dataset for model building (i.e., learning and predicting), whereas ANN model uses both the training and validation datasets for model building, validation, and reevaluation and tuning parameters. Therefore, in ANN model, soil inventory dataset was divided into three subsets: training (50% of input data) and 25% each for validation and testing. For comparison sake, the same 25% testing dataset was kept in a vault and used for assessing the generalization power of both the ANN and DAT models.

Fig. 8 HERE

5. Results and Discussion

The following results and the subsequent discussions are based on Table 5, containing all the previously-described performance metrics that have been calculated by means of the newly proposed GIS-based PMT extension system. After a preliminary diagnosis of the models in each of the aforementioned case studies, a detailed comparison of the performance metrics is provided.[1]

Table 5 HERE

5.1. Gully Erosion Modelling, Italy

According to the AUROC values, both the GLM and the MARS model show excellent performance where the differences in the AUROC values were almost negligible. According to Conoscenti et al. (2018), the excellent performance of these two models is indebted to a well-investigation of the gullies in the study area and opting the main controlling factors that best defined the occurrence mechanism. This process has been carried out by building a base model comprised of the slope gradient and the contribution area and is then fed by nine other geo-environmental factors one at a time. Moreover, the exemplary features of the chosen model have also led to a significantly good performance, defined by measures such as the handling of all types of factors (*i.e.*, both categorical and continuous) and well detecting the interactions among the factors and also between the factors and the response event. Notably, Gómez-Gutiérrez et al.

[1] Note: the discussion provided here follows a particular way as the inferences derived from each case study is modified or reemphasized perpetually on the basis of the collective information obtained from different case studies and modelling scenarios. It is tried to be err on the side of caution to avoid raising any misleading points and engaging in dogmatic defense of one approach to the detriment of another.

(2015) also applied the MARS model to predict the gully occurrence in a relatively close (ca. 85 km) catchment with similar characteristics; however, the AUROC values stood at the range of about 0.75-0.85, which was lower than that of Conoscenti et al. (2018). This highlights the importance of making a well-structured input data and the calibration/ validation techniques. To this point, both models seem to have rather similar performances.

However, a greater discrimination between models become apparent, as present in the results, after breaking down these overall precision metrics into smaller components (*i.e.*, considering simpler indices) that explain the efficacy of the approach more elaborately. Considering the misclassification rate of both models, it is evident that the GLM approach has most likely misclassified the presence and the absence more than the MARS model. Also, accuracy, as understood to be the opposite concept of misclassification rate, attested the same pattern, where the MARS model exhibited a higher accuracy in the classification of the presence and the absence localities generated by the spatially-relevant model.

Further exploring the confusion matrix, it becomes evident that the higher value of the misclassification rate in the GLM approach is directly rooted in the false negative rate. That is, the GLM approach appears to have misclassified a number of ‘presence locations’ as the ‘absence locations’ (in fact, this happened almost 13 folds greater than the MARS model). This indicates that the GLM approach has somewhat failed to locate the gullies in notable study areas, and therefore, may require further careful consideration prior to its application for real-life decisions. In fact, the present analysis shows that this error appears to have also spread out to the other metrics such as the sensitivity, F-score, NPV, and the FOR. The reason for the high AUROC value for the GLM approach is plausibly due to that the latter is a cutoff independent metric, while the confusion matrix elements have been calculated based on a 50% cutoff value. However, this does not justify the GLM's underperformance at misclassifying the absence locations, since both predictive models are compared under the same situation.

As explained in the *Theory* section, in such situations, the MCC may be the best representative of the model's performance regarding the agreement between the observations and predictions. One reason for this is because, as opposed to AUROC, AUSRC, and AUPRC, the MCC values the cost of error and attempts to avoid to circumvent or truncate any error sources. Expectedly, the MCC has well differentiated the performance of both MARS and GLM approaches, where the MARS model with a value close to 1 almost represents a perfect model, while the GLM approach with a value below 0.5 has shown a lesser degree of agreement between the observations and predictions. This notion raises the possibility of some randomness (*i.e.*, being closer to zero). The underperformance of the GLM approach highlights the disadvantages of using a predictive model that is built on linear functions. Such a model is largely incapable of considering the nonlinear interactions between the causal factors and the response event, may be sensitive to the number of predictors, and more importantly, it could be sensitive to the outliers which are robustly handled by non-linear basis functions in the MARS model. Given that the asymmetries of the cutoff-dependent and –independent metrics are now more evident, a greater degree of scrutinization is perhaps required, as provided by a more extensive discussion in the following real-life case studies.

5.2. Flood hazard modelling, Iran

Recently, Evidential Belief Function (EBF), as a bivariate statistical model underpinned by the Dempster-Shafer theory (Shafer 1976), has been adopted for flood inundation and susceptibility mapping in Iran (Rahmati and Pourghasemi, 2017). Starting with the AUROC values, the overall performance is acceptable, with respectively, 0.86 as the learning capability (obtained from the training set) and 0.78 as a predictive skill (obtained from the validation set). Higher learning skill compared to the predictive capability is common, and generally expected since the model's parameters have been calibrated on a much larger data sample compared to the validation set. However, this might question the possibility of overfitting, where a statistical model begins to describe the random error in the data rather than the relationships between variables; that is, the model becomes accustomed to the pre-used set of data. In this regard, simple statistical assumptions have been identified as one of the main sources of overfitting issues, especially in bivariate statistical models. This can negatively influence the generalization power and the transferability of the model's results to the validation set/ areas/ time periods.

Considering the results presented here, all of the favorable qualities of the model (*i.e.*, all the performance metrics highlighting the success of the model) have deteriorated to some extent in the model validation stage. Although according to the AUROC classifications provided by Hosmer and Lemeshow (2000), the values greater than 0.7 and 0.8, respectively, indicate an acceptable and excellent performances, which in turn somewhat addresses the possibility of overfitting. This is also evident in the AUSRC and AUPRC values, indicating that the predictive model is respectively well-performing in terms of both the learning capability and the predictive skill. As for the AUSRC and AUPRC values, the differences are discernable when compared to the training- and validation-derived AUROC values. These differences are conceivable, given that the AUSRC and AUPRC are calculated merely based on the presence localities. Therefore, by using the AUSRC and AUPRC, the potential error sources (*i.e.*, polluting the presence population to some absences which are incorrectly classified as positives) are left unclear and some degree of success (*i.e.*, correctly detecting the absence locations) are also not acknowledged and not included in the final calculation. This makes using the AUSRC and AUPRC less favorable to use due to their erroneous behavior (Frattini et al., 2010).

A closer scrutinization appears to shed more light on the randomly-driven performances and consequently, the weakness of the model or the input data. Considering the MCC—so far suggested as an all-inclusive metric in this study—the values greater than zero (*i.e.*, random agreement) reveals a promising level of precision; however, the values may not be high enough (*i.e.,* far from a perfect precision to be certain of a non-random performance. In particular, the level of disagreement between the observed and the predicted values appears to increase in the validation stage. Other comprehensive measures, such as the true skill statistic, informedness, and markedness are also in concurrence with the MCC value.

The Heidke's Skill Score, well-known for providing a robust accuracy value by diminishing the TPR values generated by random chance, shows how the preliminary accuracy values (i.e., efficiency) is likely to decay. Similarly, the Cohen's Kappa aims to address the random aspect of the model performance and provides new values in agreement with the latter. However, as stated in our recent discussion, one should be cautious when using the cutoff-dependent metrics. Drawing relevance from a report given by Frattini et al. (2010), the score-based metrics, despite providing valuable insights, highly relies on certain cutoff values. That is, different cutoff values might result in different performance values. However, this assumption still does not contradict using the score-based metrics for a comparison purpose, since, as stated above, all the predictive models were supposed to be compared under the same cutoff value(s) (*e.g.*, the Italian case study). To test this concern, we have applied three different cutoffs for assessing the performance of a drought risk map developed in the south-east region of Queensland, Australia.

To elaborate further, we provide two assumptions regarding the reduction in the accuracy of the EBF metric. The first assumption pertains to the model's structure. Bivariate statistical models have long been criticized for ignoring the interactions among the predictors, which can have direct (and largely negative) influence on both the learning and the predictive skills. Moreover, as stated by Ruspini et al. (1992), and more recently Reineking (2014), a need for categorizing factors with continuous nature and also presenting a generalized probabilistic reasoning limit the application of the EBF metric only to some specific problems (*e.g.*, detecting the uncertainty sources) rather than a general use. However, a review of the previous work of Rahmati and Pourghasemi (2017) reveals that the two other well-known data mining models (*i.e.*, boosted regression trees and the random forest) have been used in addition to the EBF and surprisingly, we noted that the EBF outperformed both of the data mining models, although the differences were negligible (*i.e.*, AUROCs= 0.73-0.78), which leaves us with the second hypothesis.

Regarding the latter, the input data can be responsible for such limited performances of all three models. Reviewing the model input data shows that only 63 flooding points were used as an input for the modelling process in the period of 2001-2009, let alone that they were categorized into two sets of 47 (training) and 16 (validation) locations which seems to be rather small to build a proper predictive model. Complementing the inventory map by collecting more data from a broader time period would provide a larger information matrix for the models to rely on. This highlights a note given by Ruspini et al. (1992); "*the alleged lack of decision-support and counterintuitive nature of evidential belief models, in fact, indicates the lack of basic informational shortcomings*".

5.3. Drought Risk Spatial Attribution and Modelling, Australia

For a drought risk map produced in the south-east of Queensland, Australia, the following inferences can be derived from the validation stage only in order to focus on the alteration of the performance metric values. The question mentioned above regarding the liability of the cutoff-dependent metrics is answered by means of producing three cutoffs thresholds, *i.e*., 50%, 70%, and 90%, and then comparing these results.

It was evident that the AUROC and AUSRC expectedly yielded intact performance values through all of the three cutoffs (Table 5). Based on this, the predictive skill of the fuzzy model appears to be well performing. However, the values of all the cutoff-dependent metrics drastically change at each cutoff. It is evident that by a transition from 50% to 90% cutoff, the area of danger zone appears to shrink (as illustrated in Fig. 8). Moreover, at each cutoff threshold, a different population of the negatives and the positives appears to fall within the safe and danger zones.

The direct impact of these transitions on the results is transparent in Table 5. As appears, moving from 50% to 70% cutoff, the *FN* error decreases to a certain level and adds to the *TN*, serving as an advantage point for the model, while the false positives and true positives have remained intact. Moreover, a vivid increase is also discernible in the values of the cutoff-dependent metrics. However, another step towards the 90% cutoff backfires, where—similar to the previous transition—although the *FN* value decreases and adds to *TN*, most of the *TP* population migrates to *FP* category. This expectedly decreases the values of some cutoff-dependent metrics such as F-score and PPV. Although 70% cutoff performed better than 50% and 90% cutoffs. Such a choice would not be advisable for the other study areas and certainly not for the other predictive models, because it is only in favor of this particular predictive model and the specific distribution of the positive/negative points throughout the study area.

As previously mentioned in the *Theory* section, the only suitable substitute for the cutoff value is the prevalence of the phenomenon, which again is difficult to ascertain, unless one constructs an inclusive archive of the 'presence-absence locations' by visiting numerous sites. This type of data compilation is more common in species distribution assessment, whereas, in natural hazard-related studies, extracting absence locations are executed as an additional stage after inventory mapping, based on random selection or other analytical strategies. Drawing on these inferences, it is reasonable to ascertain that using cutoff-dependent performance metrics may not be practical for individual model assessment, unless it is accompanied by mentioning the cutoff value from which the metrics' values are extracted (*i.e*., 50% for Iran, Italy, and all the following case studies), or it is carried out by setting the prevalence as the cutoff value.

As with the case of Iran, the AUROC yielded the most accurate performance value that a spatial modeler can rely on. Thus, based on current arguments, we confirm the second assumption in which the incapability of the models (*i.e*., EBF, BRT, and RF) to progress is due to the unsatisfactory input data (*i.e*., either scarce inventory, inadequate spatial indicators or spatial resolution) rather than the models' structure. Analogously, the AUROC and AUPRC values are more representative for the fuzzy model's performance in Queensland, Australia. Also, they are comparatively in accordance with the validation method of Dayal (2018) and Dayal et al. (2018), based on which the correlation of the drought risk map and the soil moisture/ rainfall departure maps confirmed plausible predictive skills.

Comparing the different predictive models (*i.e*., choosing the premier model among the many alternatives) or different scenarios of a specific model (*i.e*., opting the best scenario from different sample partitioning techniques, different spatial resolution, and so forth), is still feasible by using the cutoff-dependent metrics as they do provide valuable information that can lead to a more transparent distinction between the choices. In particular, the cutoff-dependent indices can assist us with distinguishing the features of the GLM and the MARS models for the case study in Italy. Hence, in the following case studies, the cutoff-metrics are used only for a comparison and selection of the better-performing predictive model.

5.4. Landslide Susceptibility Modelling, Vietnam

In accordance with the analytical evidence from the results of previous case studies, this study avers that the use of the cutoff-dependent metrics can be informative for a predictive model comparison. The inferences of this case study are interesting in several ways, showing that how one should interpret the latter with some degree of caution. According to the AUROC and AUPRC values of MaxEnt and BayGmmKda models tested in Vietnam (Table 5), the MaxEnt appears to slightly excel in predictive skill, although both models show an excellent performance (AUROC> 0.8). On the other hand, asymmetries are evident in the values of the cutoff-dependent metrics, as we have categorized them as the ROC-accordant and -discordant metrics (see Table 6).

Table 6 HERE

According to Table 6 and the relevant equations provided in Table 4, both categories support high *TP* and *TN* values, while there is a subtle difference that makes them oppose. In fact, a model's success in *FP* stage is highly favored in the ROC-accordant metrics, while the discordant group leans towards penalizing a model's downfall in the *FN* stage. This is evident in the confusion matrix of the MaxEnt and BayGmmKda, in which the MaxEnt shows an outstanding performance with a zero *FP* value, while the *FN* population is drastically increased in such a way that it even surpasses the *FN+FP* population in BayGmmKda model. In this case, the BayGmmKda has well balanced the FP and FN population that accords to Table 7. As previously mentioned in the *Theory* section, although a zero FP (Type I error) in MaxEnt results cause no infrastructural and study costs, a drastic increase in FN (Type II error) values can cause massive casualties via misrepresenting an area as a safe location.

Table 7 HERE

Considering the structure of these predictive models, as opposed to the presence-absence nature of the BayGmmKda, MaxEnt is considered as a presence-only model where some randomly chosen pseudo-absence locations (*i.e.*, background samples) help the model differentiate the presence locations and eventually predict an occurrence pattern. Therefore, presence-absence-based validation metrics (*i.e.*, all the metrics provided in this study) may not be a good fit for the performance assessment of MaxEnt. This being the case, AUPRC might be the best fit for MaxEnt and in fact, it has clearly distinguished the performance of both models. However, according to Phillips et al. (2006), at least, those background locations should be considered as 'pure absences' to be able to graph a ROC curve, and also to calculate the metrics derived from confusion matrix. This is an inevitable process for the MaxEnt. Another critical inference of this case study underlines that although cutoff-dependent metrics are valuable metrics for comparing different models, they are not necessarily supposed to be in line with cutoff-independent metrics. This is the reason why MaxEnt and BayGmmKda both excel, but in different areas. Therefore, relying on what we have conceived so far, each cutoff-dependent or -independent metric has a unique indication of a model's performance.

There is a consensus that selecting the best predictive model can be a matter of the user preference and study area's goals, which has been previously well-delineated in Goetz et al. (2015). This can be carried out by relying on a pros and cons list for all the metrics and assessing whether they work in agreement with the objective(s) of the project. Taking aside the disadvantages of cutoff-dependent metrics, some critics have also been moved towards AUROC (Lobo et al., 2008). The main complains pertain to ignoring the PPV (addressed earlier in *Theory* section) and equally weighting omission (not recording some instances) and commission (miss-recording some instances) errors. However, this directly stems from predefining a series of thresholds and the presence-absence fabric of AUROC which is not only specific to AUROC but rather all the performance metrics. Furthermore, these limitations do not question the metric itself, but rather the application of them. For instance, ROC curves were first employed in the study of "*discriminator systems for the detection of radio signals in the presence of noise in the 1940s"*, following the attack on Pearl Harbor, USA (Garrett et al., 2008). Even the use of AUROC in clinical biochemistry is carried out under a presence-absence condition (Obuchowski et al., 2004). Therefore, in order to employ AUROC and other cutoff/prevalence- independent metrics in a probabilistic environmental modelling context, their limitation should be accepted in favor of their valuable outcomes regarding the performance evaluation.

Under these premises, we aver that the project study goal can assist the decision maker with opting the well-performing model. For instance, if the number of opposing metrics matters the most, the BayGmmKda would be the well-performing one. In particular, many municipal authorities may decide in favor of public safety, which in turn can end in an immediate rejection of the MaxEnt due to having considerable Type II error that can also cause notable fatalities. Comparatively, if the uncertain nature of the cutoff value is in question, one can choose the decisive judgment of the AUROC.

5.5. Soil Digital Modelling, USA

As previously mentioned, this case study represents a unique application of the proposed PMT for performance assessment of the Bulk Density (BD) lateral distribution in South Dakota, USA. In contrast to the previous applications of data mining methods that deal with predicting the probability of an occurrence, in this study we employed the ANN and DT approaches for predicting an actual quantity of BD whose actual amounts can be measured in the field. Measuring the BD samples from different location of the study area, root mean square error (RMSE) can be a good metric to test the accuracy of the results (*i.e.*, an approximated standard deviation of data) if the data are Gaussian (*i.e.*, rich data) and devoid of any outliers (Chai and Draxler, 2014). However, RMSE or accuracy, in general, can be biased and may not reflect the total precision of a predictive model, warranting the need for a consolidated list of model evaluation metrics that provide more extensive insights into the predictive performance.

In respect to the above discussion, the proposed PMT approach can be a good alternative, but the nature of the prediction map should be rendered into its probability terms or at least as an indication of the probability. That is, the higher values of the prediction map can indicate the greater probability of having higher BD values, and vice versa. By doing so, the cutoff-dependent and -independent metrics have been calculated based on which, almost all the indices congruently introduce ANN as a better-performing model compared to DAT; the rest of opposing metrics (e.g. specificity and PPV) show negligible differences. This is in agreement with those reported by Taghizadeh-Mehrjardi et al. (2017) where the ANN was seen to outperform the support vector machine (SVM) model in the mapping of soil organic matter distribution.

6. Synthesis and Conclusion

This paper provides a novel scientific contribution towards the design and implementation of an adaptive, largely automated and user-friendly GIS-based spatial model assessment system, denoted as the Performance Measure Tool (PMT). PMT can be used to address existing challenges in pragmatic evaluation of predictive models in diverse contexts, and generally, for any scientific branch where information has a spatial connotation. The PMT encloses the relevant mathematical formulations to make it an easy-to-use software; it has the added capability to evaluate the accuracy of the spatial modelling approach based on the different cutoff-dependent and -independent evaluation criteria. The PMT is considerably flexible, and hence, it can be widely applicable in multiple scientific and engineering applications where spatially-relevant predictive models are tested. The approach has the potential to be applied in diverse contexts, as verified in this research study, to extend its usage from geo-environmental spatial models to fields such as medical geography and epidemiology where data-driven approaches are adopted to generate predictive models and such models require robust comparison with several benchmark models and real-life (observed) datasets.

In context of proposing an additional GIS-based predictive model assessment tool, the consolidated metrics that are generated and evaluated by the proposed PMT, certainly provides a new practical pathway for real-life decision-makers who are seeking a better performing predictive model (relative to any other comparative model). Based on contested reasons, and evaluations of PMT with several studies collated in this research paper, real-life decision-makers can deduce the grounds on which their predictive models performs better than the others prior to implementing them for practical use. By accommodating multiple types of real-time geo-environmental modelling instances in this study, the take-home messages are as follows. The use of a merely row-wise or a column-wise calculated index from the confusion matrix is not a robust approach for model selection as this can ignore the more practical concepts considered by their counterpart tools.

In contrast, some of the model evaluation indices (*i.e.* cutoff-dependent and –independent ones) generally use a collective information of the matrix in such a way that a set of multiple statistics are used in conjunction with each other. Notwithstanding this, some cutoff-dependent metrics may infer the same connotation which they can be used interchangeably (*e.g*., threat score and equitable threat score, or the odds ratio and the odds ratio skill score). Moreover, the choice of using the cutoff-dependent metrics over each other without a prior knowledge can also constitute an unjust approach since each metric is able to tackle a different aspect of the model performance. However, all metrics can be highly sensitive to the cutoff values so, they should be suggested only for the model comparison.

As demonstrated in the theory of PMT and relevant case studies, it becomes unambiguous that the measurement of the prevalence of the studied phenomenon is highly advisable in order to ascertain reliable cutoff-dependent values. Doing so, they are likely to be applicable even for the performance assessment of an individual model, and also, they could be comparable with cutoff-independent metrics.

On the other hand, the cutoff-independent metrics (*i.e*., AUROC, AUSRC, and AUPRC) can decisively screen the premier model regardless of the changes in their cutoff values. However, the AUROC is also underpinned by some specific assumptions so that using it would require accepting its mathematical fabric. Furthermore, AUSRC and AUPRC only support presence locations, they show an erroneous behavior and in particular may result in an underestimation of performance compared to AUROC. Moreover, all cutoff-dependent and -independent metrics can occasionally mislead by providing different results and consequently different model ranks. In such case, selecting the reference model is strictly tied to the aim of the research and specific aspect(s) of interest. We also concluded that compartmentalizing models in different performance categories is not feasible since the matter of performance itself is quite relative.

We also propose the following scenario-based decision-making inferences:

I. Italy and USA case studies: having more than one model→ if AUROC values converge and the changes are negligible→ using other cutoff-dependent metrics to derive the better-performing model.

II. Iran and Australia case studies: having one model→ no access to prevalence value change→ cutoff-dependent metrics change drastically by altering cutoff values→ use AUROC as the decisive metric.

III. Vietnam case study: more than one model→ metrics are opposing and taking different parts (i.e. each selecting a different model) → decision should be made based on the project goal by making pros and cons list for all the metrics.

As our final upshot, ROC and AUC are metrics that tend to lump together the prediction as a whole; however, studying confusion matrices, accuracy and precision of a model ensure a better insight on a model hit and misses. This is something that can be rarely found in the literature, despite its great importance. The PMT quickly provides a full suite of performance metrics allowing the users to better evaluate their spatial model and supporting a more critical judgment, which in turn can promote better decision-making procedures.

Acknowledgments

This study supported by the United States Department of Agriculture–NIFA (Award Number 2014-51130-22593), University of Southern Queensland Office of Research and Graduate Studies Postgraduate Research Scholarship, and project FLUMEN (project number: 318969) at University of Palermo, funded by the EU (call identifier: FP7-PEOPLE-2012-IRSES). R C Deo is thankful to AQ Queensland-Smithsonian Fellowship for provision of research time in writing phase of the paper. Meteorological data of Australian case study were obtained from Australian Terrestrial Ecosystem Research Network Data Discovery Portal and Australia Water Availability Project (AWAP). Flood data was acquired from Iranian Department of Water Resources Management (IDWRM).

References

Abdollahi, S., Pourghasemi, H. R., Ghanbarian, G. A., Safaeian, R. 2018. Prioritization of effective factors in the occurrence of land subsidence and its susceptibility mapping using an SVM model and their different kernel functions. B. Eng. Geol. Environ.. doi: 10.1007/s10064-018-1403-6.

Akgün, A., Türk, N., 2011. Mapping erosion susceptibility by a multivariate statistical method: a case study from the Ayvalık region, NW Turkey. Comput. Geosci. 37(9), 1515-1524.

Alfons, A. 2012. Package "cvTools": Cross-validation tools for regression models. https://cran.r-project.org/web/packages/cvTools/index.html

Allouche, O., Tsoar, A., Kadmon, R., 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). J. Appl. Ecol. 43(6), 1223-1232.

Arpaci, A., Malowerschnig, B., Sass, O., Vacik, H., 2014. Using multi variate data mining techniques for estimating fire susceptibility of Tyrolean forests. Appl. Geogr. 53, 258-270.

Beguería, S., 2006. Validation and evaluation of predictive models in hazard assessment and risk management. Nat. Hazards 37 (3), 315–329

Benjamini, Y., Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society. Series B (Methodological), 289-300.

Blahut, J., van Westen, C. J., Sterlacchini, S. 2010. Analysis of landslide inventories for accurate prediction of debris-flow source areas. Geomorphology 119(1-2), 36-51.

Boughorbel, S., Jarray, F., El-Anbari, M. 2017. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PloS One 12(6), e0177678.

Bucklin, D.N., Basille, M., Benscoter, A.M., Brandt, L.A., Mazzotti, F.J., Romanach, S.S., Speroterra, C., Watling, J.I., 2015. Comparing species distribution models constructed with different subsets of environmental predictors. Divers. Distrib. 21(1), 23-35.

Chai, T., Draxler, R. R. 2014. Root mean square error (RMSE) or mean absolute error (MAE)?– Arguments against avoiding RMSE in the literature. Geosci. Model Dev. 7(3), 1247-1250.

Chapi, K., Singh, V.P., Shirzadi, A., Shahabi, H., Bui, D.T., Pham, B.T., Khosravi, K., 2017. A novel hybrid artificial intelligence approach for flood susceptibility assessment. Environ. Modell. Softw. 95, 229-245.

Choubin, B., Moradi, E., Golshan, M., Adamowski, J., Sajedi-Hosseini, F., Mosavi, A., 2019. An Ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines. Sci. Total Environ. 651, 2087-2096.

Chung, C. J. 2006. Using likelihood ratio functions for modelling the conditional probability of occurrence of future landslides for risk assessment. Comput. Geosci. 32(8), 1052-1068.

Chung, C. J., Fabbri, A. G. 2008. Predicting landslides for risk analysis—spatial models tested by a cross-validation technique. Geomorphology 94(3-4), 438-452.

Cohen, J. 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20(1), 37–46.

Conoscenti, C., Agnesi, V., Cama, M., Caraballo-Arias, N. A., Rotigliano, E. 2018. Assessment of gully erosion susceptibility using multivariate adaptive regression splines and accounting for terrain connectivity. Land Degrad. Dev. 29(3), 724-736.

Conoscenti, C., Angileri, S., Cappadonia, C., Rotigliano, E., Agnesi, V., Märker, M., 2014. Gully erosion susceptibility assessment by means of GIS-based logistic regression: a case of Sicily (Italy). Geomorphology 204, 399-411.

Coombes, K.R., 2018. Package 'CrossValidate': Classes and Methods for Cross Validation of "Class Prediction" Algorithms. https://cran.r-project.org/web/packages/CrossValidate/index.html

Dayal, K. S., Deo, R. C., Apan, A. A. 2018. Spatio-temporal drought risk mapping approach and its application in the drought-prone region of south-east Queensland, Australia. Nat. Hazards, 1-25.

Dayal, K.S., 2018. Development of Statistical and Geospatial-Based Framework For Drought-Risk Assessment. PhD Thesis. University of Southern Queensland, Australia. 248pp.

Deo, R. C., Samui, P., Kim, D. 2016. Estimation of monthly evaporative loss using relevance vector machine, extreme learning machine and multivariate adaptive regression spline models. Stoch. Env. Res. Risk. A. 30(6), 1769-1784.

Deo, R. C., Tiwari, M. K., Adamowski, J. F., Quilty, J. M. 2017. Forecasting effective drought index using a wavelet extreme learning machine (W-ELM) model. Stoch. Env. Res. Risk. A. 31(5), 1211-1240.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recogn. Lett. 27(8), 861-874.

Frattini, P., Crosta, G., Carrara, A. 2010. Techniques for evaluating the performance of landslide susceptibility models. Eng. Geol. 111(1-4), 62-72.

Garosi, Y., Sheklabadi, M., Pourghasemi, H.R., Besalatpour, A.A., Conoscenti, C., Van Oost, K., 2018. Comparison of differences in resolution and sources of controlling factors for gully erosion susceptibility mapping. Geoderma 330, 65-78.

Garrett, P. E., Lasky, F. D., Meier, K. L. 2008. User protocol for evaluation of qualitative test performance; approved guideline. CLSI.

Ghorbanzadeh, O., Blaschke, T., Aryal, J., Gholaminia, K., 2018. A new GIS-based technique using an adaptive neuro-fuzzy inference system for land subsidence susceptibility mapping. J. Spat. Sci. 1-17.

Gilbert, G. K. 1884. Finley's tornado predictions. American Meteorological Journal. A Monthly Review of Meteorology and Allied Branches of Study (1884-1896), 1(5), 166.

Glade, T. (Ed.), 2005. Landslide Hazard and Risk. Wiley, Chichester, pp. 41–74.

Goetz, J. N., Brenning, A., Petschko, H., Leopold, P. 2015. Evaluating machine learning and statistical prediction techniques for landslide susceptibility modelling. Comput. Geosci. 81, 1-11.

Gómez-Gutiérrez, Á., Conoscenti, C., Angileri, S. E., Rotigliano, E., Schnabel, S. 2015. Using topographical attributes to evaluate gully erosion proneness (susceptibility) in two mediterranean basins: advantages and limitations. Nat. Hazards 79(1), 291-314.

Gorsevski, P.V., Gessler, P.E., Foltz, R.B., Elliot, W.J., 2006. Spatial prediction of landslide hazard using logistic regression and ROC analysis. T. GIS 10(3), 395-415.

Hanssen, A. W., Kuipers, W. J. A. 1965. On the Relationship Between the Frequency of Rain and Various Meteorological Parameters:(with Reference to the Problem Ob Objective Forecasting). Staatsdrukkerij-en Uitgeverijbedrijf.

Heidke, P., 1926: Berechnung des Erfolges und der Gute der Windstarkevorhersagen im Sturmwarnungsdienst (Measures of success and goodness of wind force forecasts by the gale warning service) Geogr. Ann. 8, 301–349

Hosmer, D. W., Lemeshow, S. 2000. Applied Logistic Regression., 2nd edn.(Wiley: New York.). NY, USA.

Kavzoglu, T., Colkesen, I. and Sahin, E.K., 2019. Machine Learning Techniques in Landslide Susceptibility Mapping: A Survey and a Case Study. In Landslides: Theory, Practice and Modelling (pp. 283-301). Springer, Cham.

Kornejady, A., Ownegh, M., Bahremand, A. 2017. Landslide susceptibility assessment using maximum entropy model with two different data sampling methods. Catena 152, 144-162.

Lipovetsky, S., 2009. Pareto 80/20 law: derivation via random partitioning. Int. J. Math. Educ. Sci. Technol. 40(2), 271-277.

Lobo, J. M., Jiménez-Valverde, A., Real, R. 2008. AUC: a misleading measure of the performance of predictive distribution models. Global Ecol. Biogeogr. 17(2), 145-151.

Loiselle, B.A., Howell, C.A., Graham, C.H., Goerck, J.M., Brooks, T., Smith, K.G., Williams, P.H., 2003. Avoiding pitfalls of using species distribution models in conservation planning. Conserv. Biol. 17, 1591–1600.

Lombardo, L., Mai, P. M. 2018. Presenting logistic regression-based landslide susceptibility results. Eng. Geol. 244, 14-24.

Lombardo, L., Opitz, T., Huser, R. 2018. Point process-based modelling of multiple debris flow landslides using INLA: an application to the 2009 Messina disaster. Stoch. Env. Res. Risk. A. 32(7), 2179-2198.

Malone, B.P., Minasny, B., McBratney, A.B., 2017. Combining Continuous and Categorical Modelling: Digital Soil Mapping of Soil Horizons and Their Depths. In Using R for Digital Soil Mapping (pp. 231-244). Springer, Cham.

Mas, J.F., Soares Filho, B., Pontius, R.G., Farfán Gutiérrez, M., Rodrigues, H., 2013. A suite of tools for ROC analysis of spatial models. ISPRS Int. Geo-Inf. 2(3), 869-887.

Minasny, B., McBratney, A.B., 2007. Incorporating taxonomic distance into spatial prediction and digital mapping of soil classes. Geoderma 142(3-4), 285-293.

Miraki, S., Zanganeh, S.H., Chapi, K., Singh, V.P., Shirzadi, A., Shahabi, H., Pham, B.T., 2019. Mapping Groundwater Potential Using a Novel Hybrid Intelligence Approach. Water Resour. Manage. 33(1), 281-302.

Naghibi, S.A., Ahmadi, K., Daneshi, A., 2017. Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping. Water Resour. Manage. 31(9), 2761-2775.

Obuchowski, N. A., Lieber, M. L., Wians, F. H. 2004. ROC curves in clinical chemistry: uses, misuses, and possible solutions. Clin. Chem. 50(7), 1118-1125.

Peirce, C. S. 1884. The numerical measure of the success of predictions. Science 4(93), 453-454.

Pepe, M.S., Janes, H., Longton, G., Leisenring, W., Newcomb, P., 2004. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am. J. Epidemiol. 159(9), 882-890.

Perruchet, P., Peereman, R. 2004. The exploitation of distributional information in syllable processing. J. Neurolinguist. 17(2-3), 97-119.

Phillips, S. J., Anderson, R. P., Schapire, R. E. 2006. Maximum entropy modelling of species geographic distributions. Ecol. Modell. 190(3-4), 231-259.

Pontius Jr, R.G., Schneider, L.C., 2001. Land-cover change model validation by an ROC method for the Ipswich watershed, Massachusetts, USA. Agr. Ecosyst. Environ. 85(1-3), 239-248.

Pourghasemi, H.R., Rahmati, O., 2018. Prediction of the landslide susceptibility: Which algorithm, which precision?. Catena 162, 177-192.

Powers, D. M. 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. http://hdl.handle.net/2328/27165.

Pradhan, B., Lee, S., 2010. Landslide susceptibility assessment and factor effect analysis: backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modelling. Environmental Modelling & Software, 25(6), 747-759.

Pullar, D., Springer, D., 2000. Towards integrating GIS and catchment models. Environ. Modell. Softw. 15(5), pp.451-459.

Quillfeldt, P., Engler, J.O., Silk, J.R., Phillips, R.A., 2017. Influence of device accuracy and choice of algorithm for species distribution modelling of seabirds: a case study using black-browed albatrosses. J. Avian Biol. 48(12), 1549-1555.

Rahmati, O., Pourghasemi, H. R. 2017. Identification of critical flood prone areas in data-scarce and ungauged regions: A comparison of three data mining models. Water Resour. Manage. 31(5), 1473-1487.

Reineking, T. 2014. Belief functions: Theory and algorithms. https://pdfs.semanticscholar.org/eb77/3cd7c84617bfd9e3abbb7695e113e94c9524.pdf

Ruspini, E. H., Lowrance, J. D., Strat, T. M. 1992. Understanding evidential reasoning. Int. J. Approx. Reason. 6(3), 401-424.

Sahana, M., Ganaie, T.A., 2017. GIS-based landscape vulnerability assessment to forest fire susceptibility of Rudraprayag district, Uttarakhand, India. Environ. Earth Sci. 76(20), p.676.

Schaefer, J. T. 1990. The critical success index as an indicator of warning skill. Weather Forecast. 5(4), 570-575.

Scott, L. M., Janikas, M. V. 2010. Spatial statistics in ArcGIS. In Handbook of applied spatial analysis (pp. 27-41). Springer, Berlin, Heidelberg.

Shabani, F., Kumar, L., Ahmadi, M., 2016. A comparison of absolute performance of different correlative and mechanistic species distribution models in an independent area. Ecol. Evol. 6(16), 5973-5986.

Shafer, G. 1976. A mathematical theory of evidence (Vol. 42). Princeton university press.

Siahkamari, S., Haghizadeh, A., Zeinivand, H., Tahmasebipour, N., Rahmati, O., 2018. Spatial prediction of flood-susceptible areas using frequency ratio and maximum entropy models. Geocarto Int. 33(9), 927-941.

Swets, J. A. 2014. Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers. Psychology Press.

Tien Bui, D., Bui, Q.T., Nguyen, Q.P., Pradhan, B., Nampak, H., Trinh, P.T., 2017. A hybrid artificial intelligence approach using GIS-based neural-fuzzy inference system and particle swarm optimization for forest fire susceptibility modelling at a tropical area. Agr. Forest Meteorol. 233, 32-44.

Tien Bui, D., Hoang, N. D. 2017. A Bayesian framework based on a Gaussian mixture model and radial-basis-function Fisher discriminant analysis (BayGmmKda V1. 1) for spatial prediction of floods. Geosci. Model Dev. 10 (9), 3391-3409

Van Westen, C. J., Van Asch, T. W., Soeters, R. 2006. Landslide hazard and risk zonation—why is it still so difficult?. B. Eng. Geol. Environ. 65(2), 167-184.

Wiesmeier, M., Barthold, F., Blank, B., Kögel-Knabner, I., 2011. Digital mapping of soil organic matter stocks using Random Forest modelling in a semi-arid steppe ecosystem. PlantSoil 340(1-2), 7-24.

Yan, F., Zhang, Q., Ye, S., Ren, B., 2019. A novel hybrid approach for landslide susceptibility mapping integrating analytical hierarchy process and normalized frequency ratio methods with the cloud model. Geomorphology 327, 170-187.

Yule, G. U. 1900. VII. On the association of attributes in statistics: with illustrations from the material of the childhood society, &c. Phil. Trans. R. Soc. Lond. A, 194(252-261), 257-319.

Figure Captions

Fig. 1 Study sites on the world map

Fig. 2 Methodological flowchart adopted in this study

Fig. 3 Gully erosion prediction maps of the central-western Sicily (Italy) generated by using the GLM (a) and MARS (b) models

Fig. 4 Flood-inundation susceptibility map of the Galikesh region (Iran) obtained from the EBF model

Fig. 5 Drought risk map of the south-east of Queensland (Australia) produced by using fuzzy GAMMA overlay technique

Fig. 6 Effects of 50% (a), 70% (b) and 90% (c) cutoff values on the extent of safe/danger zones and classification of presence/absence samples in south-east of Queensland

Fig. 7 Landslide susceptibility maps of the Kon Tum province (Vietnam) obtained from BayGmmKda (a) and MaxEnt (b) models

Fig. 8 Bulk density predictive distribution maps of South Dakota (USA) generated from ANN (a) and DT (b) models

Table 1 Confusion matrix elements.

Observed	Predicted	
	Class stable (−)	Class unstable (+)
Class stable (−)*	(−|−) True negative (TN)	(+|−) False positive (FP; Error Type I)
Class unstable (+)**	(−|+) False negative (FN; Error Type II)	(+|+) True positive (TP)

* Absence areas ** Presence areas

Table 2 The PMT input files

ID	Setting	Description	ID	Setting	Description
1	Input raster layers	The raster maps generated by any spatial model representing the susceptibility or suitability of a phenomenon over an area (you can add different maps for the same area as many as desired).	5	Validation positives	Import the shapefile of all the validation samples of the phenomenon of interest (discarded dataset in the training stage).
2	Cutoff	An a-priori cutoff percentage to split the input raster into two segments (50% is set as default).	6	Validation negatives	Import the shapefile of the non-event validation locations.
3	Training positives	Import the shapefile of all the training samples of the phenomenon of interest.	7	Output workspace	The pass to contain the outputs (a folder address).
4	Training negatives	Import the shapefile of the absence training locations (should be prepared beforehand by different methods mentioned in the text)	8	Number of classes (for SRC and PRC curves)	The number into which the spatial raster is to be reclassified (100 classes are set as default). The reclassification method is based on an equal interval. A higher number of classes will result in smoother SRC and PRC curves with more precise AUSRC and AUPRC values.

Table 3 The PMT output files

ID	Setting	Description
1	Html file	It explains the main results of the performance analyses including confusion matrix, cutoff-dependent metrics, and cutoff–independent metrics. ROC, SRC, and PRC curves are other parts of this html file. In addition, all results were classified into two groups of cutoff-dependent and cutoff-independent approaches with some useful explanations regarding these approaches.
2	Microsoft excel file	This file summarize all of quantitative results (without explanations)

Table 4 Equations of cutoff-dependent performance metrics

Performance metric	Equation	Performance metric	Equation
True positive rate (TPR; sensitivity)	$\frac{TP}{P} = \frac{TP}{TP + FN}$	Matthews correlation coefficient (MCC)	$\frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$
False positive rate (FPR; fall-out; 1–specificity)	$\frac{FP}{N} = \frac{FP}{FP + TN} = 1 - \frac{TN}{TN + FP}$	Informedness (Bookmaker informedness; BM)	$TPR + TNR - 1$
True negative rate (TNR; specificity)	$\frac{TN}{N} = \frac{TN}{TN + FP}$	Markedness (MK)	$PPV + NPV - 1$
False negative rate (miss rate)	$\frac{FN}{P} = \frac{FN}{FN + TP} = 1 - TPR$	Threat score	$\frac{TP}{TP + FN + FP}$
Efficiency (accuracy)	$\frac{TP + TN}{T}$	Equitable threat score	$\frac{TP - TP_{random}}{TP + FN + FP - TP_{random}}$ $where, \quad TP_{random} = \frac{(TP + FN)(TP + FP)}{T}$
Misclassification rate	$\frac{FP + FN}{T}$	True skill statistic (Pierce's skill score)	$\frac{TP}{TP + FN} - \frac{FP}{FP + TN} = \text{Sensitivity} + \text{Specificity} - 1$
Positive predictive value (PPV; precision)	$\frac{TP}{TP + FP}$	Heidke's skill score	$\frac{TP + TN - E}{T - E}$ $where\ E = \frac{1}{T}[(TP + FN)(TP + FP) + (TN + FN)(TN + FP)]$
False discovery rate (FDR)	$1 - PPV = \frac{FP}{FP + TP}$	Odds ratio	$\frac{TP \times TN}{FN \times FP}$
Negative predictive value (NPV)	$\frac{TN}{TN + FN}$	Odd ratio skill score (Yule's Q)	$\frac{(TP \times TN) - (FP \times FN)}{(TP \times TN) + (FP \times FN)}$
False omission rate (FOR)	$1 - NPV = \frac{FN}{FN + TN}$	Cohen's kappa	$\frac{(TP + TN) - [\{TP + FN)(TP + FP) + (FN + TN)(FP + TN)]/T}{T - [\{(TP + FN)(TP + FP) + (FN + TN)(FP + TN)\}/T\}}$
F-score	$2\frac{PPV.TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$	-	-

Table 5 Performance metrics calculated for each case study

Country	Subject	Model	Modelling step	Efficiency (accuracy)	True positive rate (TPR)	False positive rate (FPR)	Threat score	Equitable threat score	Hedke skill score	Odds ratio	Odd ratio skill score
Australia	Drought risk mapping	Fuzzy function: 50% cutoff	Validation	0.625	0.580	0.222	0.545	0.142	0.25	4.8462	0.657
		Fuzzy function: 70% cutoff		0.85	0.818	0.111	0.75	0.538	0.7	36	0.945
		Fuzzy function: 90% cutoff		0.625	1	0.428	0.25	0.142	0.25	0	1
Iran	Flood inundation mapping	EBF	Training	0.808	0.891	0.245	0.647	0.446	0.617	25.33	0.924
			Validation	0.718	0.769	0.315	0.526	0.28	0.437	7.22	0.756
USA	Distribution of soil organic matters	DAT	Validation	0.442	0.431	0	0.431	0.014	0.028	0	1
		ANN	Validation	0.730	0.625	0.1	0.588	0.315	0.48	15	0.875
Italy	Gully susceptibility mapping	MARS	Training	0.970	0.963	0.022	0.942	0.888	0.940	1151	0.998
			Validation	0.976	0.970	0.016	0.954	0.910	0.953	1885	0.998
		GLM	Training	0.656	0.592	0	0.592	0.185	0.312	0	1
			Validation	0.674	0.605	0	0.605	0.211	0.348	0	1
Vietnam	Landslide susceptibility mapping	MaxEnt	Validation	0.601	0.556	0	0.556	0.112	0.202	0	1
		BayGmmKda		0.739	0.731	0.2521	0.592	0.314	0.478	8.08	0.779

Table 5 (continued)

Country	Subject	Model	Modelling step	True skill statistic	Cohen's kappa	True negative rate (TNR)	False negative rate (miss rate)	Misclassification rate	Positive predictive value (PPV)	False discovery rate (FDR)	Negative predictive value (NPV)
Australia	Drought risk mapping	Fuzzy function: 50% cutoff	Validation	0.358	0.25	0.778	0.419	0.375	0.900	0.100	0.350
		Fuzzy function: 70% cutoff		0.707	0.7	0.889	0.182	0.150	0.900	0.100	0.800
		Fuzzy function: 90% cutoff		0.571	0.25	0.571	0.000	0.375	0.250	0.750	1.000
Iran	Flood inundation mapping	EBF	Training	0.646	0.617	0.754	0.108	0.192	0.702	0.298	0.915
			Validation	0.453	0.437	0.684	0.231	0.281	0.625	0.375	0.813
USA	Predictive distribution of soil bulk density	DAT	Validation	0.431	0.028	1.00	0.569	0.558	1.000	0.000	0.033
		ANN	Validation	0.525	0.48	0.90	0.375	0.269	0.909	0.091	0.600
Italy	Gully susceptibility mapping	MARS	Training	0.941	0.940	0.978	0.037	0.030	0.978	0.022	0.963
			Validation	0.953	0.953	0.983	0.030	0.024	0.983	0.017	0.970
		GLM	Training	0.592	0.312	1.000	0.407	0.344	1.000	0.000	0.313
			Validation	0.605	0.348	1.000	0.394	0.326	1.000	0.000	0.349
Vietnam	Landslide susceptibility mapping	MaxEnt	Validation	0.556	0.202	1.000	0.444	0.399	1.000	0.000	0.203
		BayGmmKda		0.479	0.478	0.748	0.269	0.261	0.757	0.243	0.722

Table 5 (continued)

Country	Subject	Model	Modelling step	False omission rate (FOR)	F-score	Matthews correlation coefficient (MCC)	Informedness (Bookmaker informedness; BM)	Markedness (MK)	AUROC	AUSRC	AUPRC
Australia	Drought risk mapping	Fuzzy function: 50% cutoff	Validation	0.650	0.706	0.299	0.358	0.250	0.873	-	74.400
		Fuzzy function: 70% cutoff		0.200	0.857	0.704	0.707	0.700	0.873		74.400
		Fuzzy function: 90% cutoff		0.000	0.400	0.378	0.571	0.250	0.873		74.400
Iran	Flood inundation mapping	EBF	Training	0.085	0.786	0.632	0.646	0.617	0.866	79.710	-
			Validation	0.188	0.690	0.445	0.453	0.438	0.787	-	75.209
USA	Predictive distribution of soil bulk density	DAT	Validation	0.967	0.603	0.120	0.431	0.033	0.839	-	77.620
		ANN	Validation	0.400	0.741	0.517	0.525	0.509	0.879	-	79.630
Italy	Gully susceptibility mapping	MARS	Training	0.037	0.971	0.941	0.941	0.941	0.992	99.141	-
			Validation	0.030	0.977	0.953	0.953	0.953	0.995	-	99.285
		GLM	Training	0.687	0.744	0.430	0.593	0.313	0.987	97.134	-
			Validation	0.651	0.754	0.460	0.606	0.349	0.992	-	97.542
Vietnam	Landslide susceptibility mapping	MaxEnt	Validation	0.797	0.715	0.336	0.556	0.203	0.889	-	0.855
		BayGmmKda		0.278	0.744	0.479	0.479	0.479	0.819	-	69.460

Table 6 Opposing performance metrics for Vietnam's case study

ROC-accordant	ROC-discordant
Informedness	Markedness
PPV	MCC
TNR	NPV
TSS	Misclassification rate
1-Specificity	FNR
FDR	Cohen's Kappa
	F-score
	Hedke skill score
	Equitable threat score
	Threat score
	Sensitivity
	Accuracy
	FOR

Table 7 Comparing confusion matrix variants of MaxEnt and BayGmmKda models as implemented in Vietnam

Observed	Models	
	MaxEnt	BayGmmKda
TN	330	1175
TP	1627	1231
FN	1297	452
FP	0	396

Figure
Click here to download Figure: Figures_Revised.doc

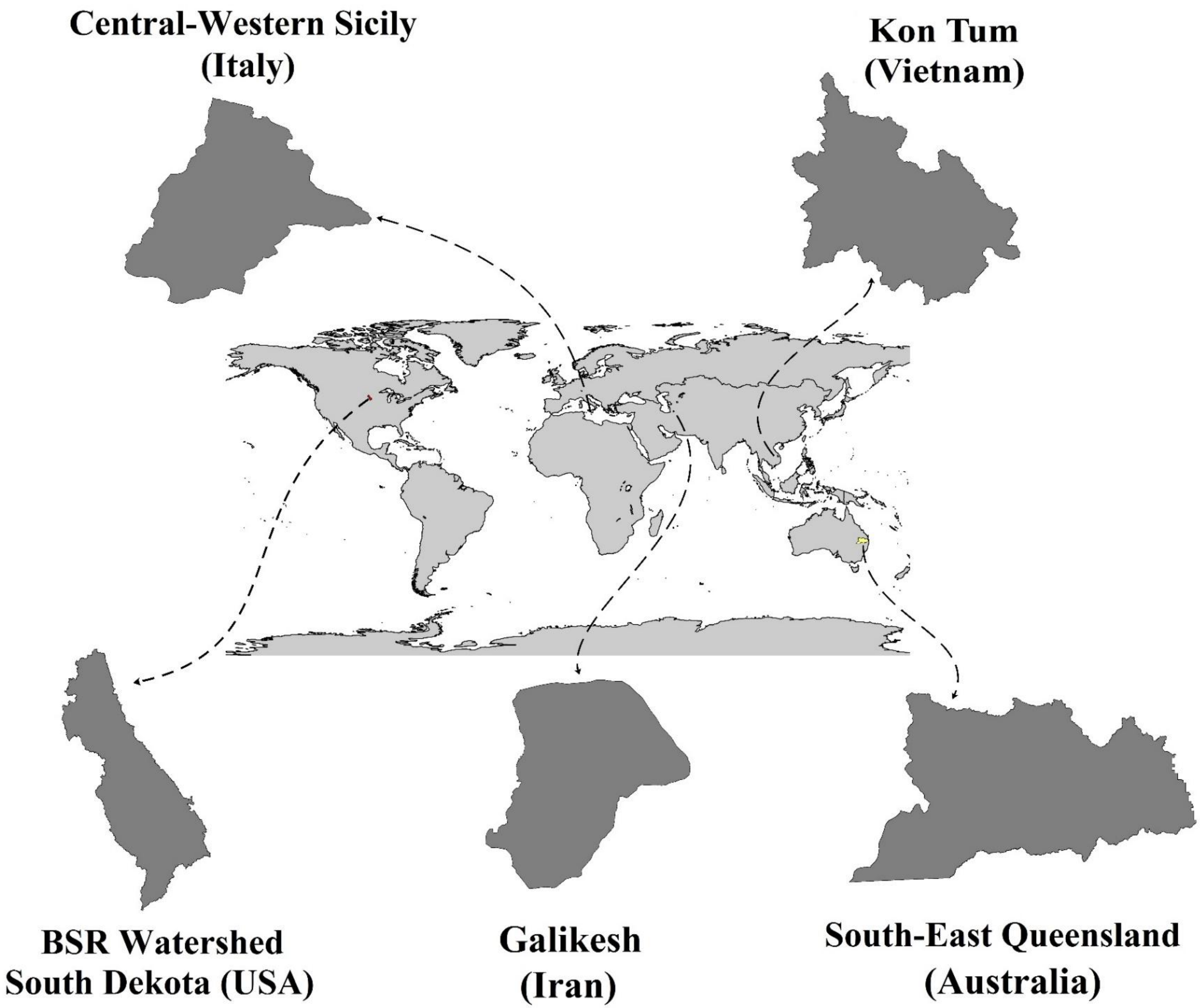
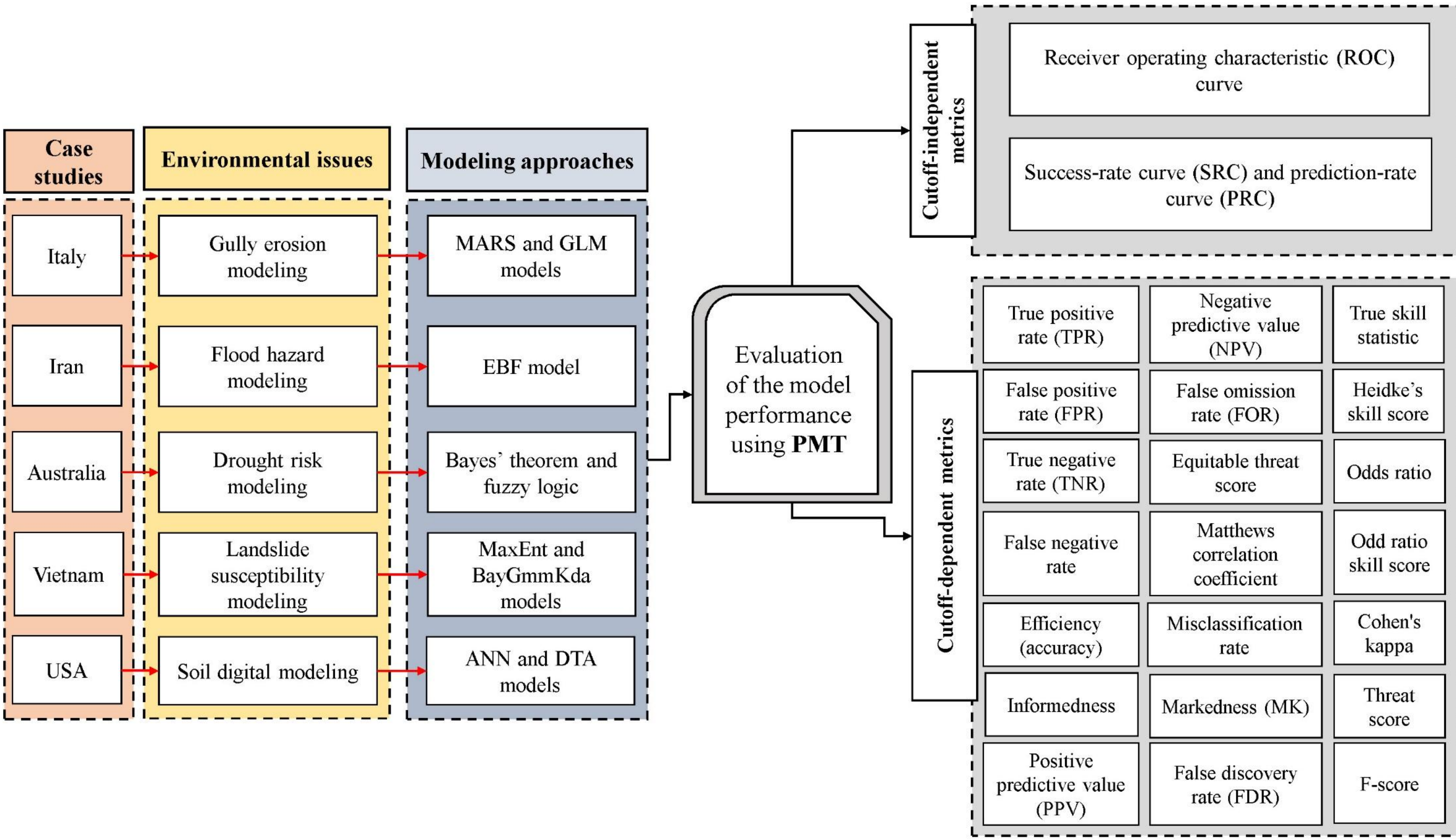
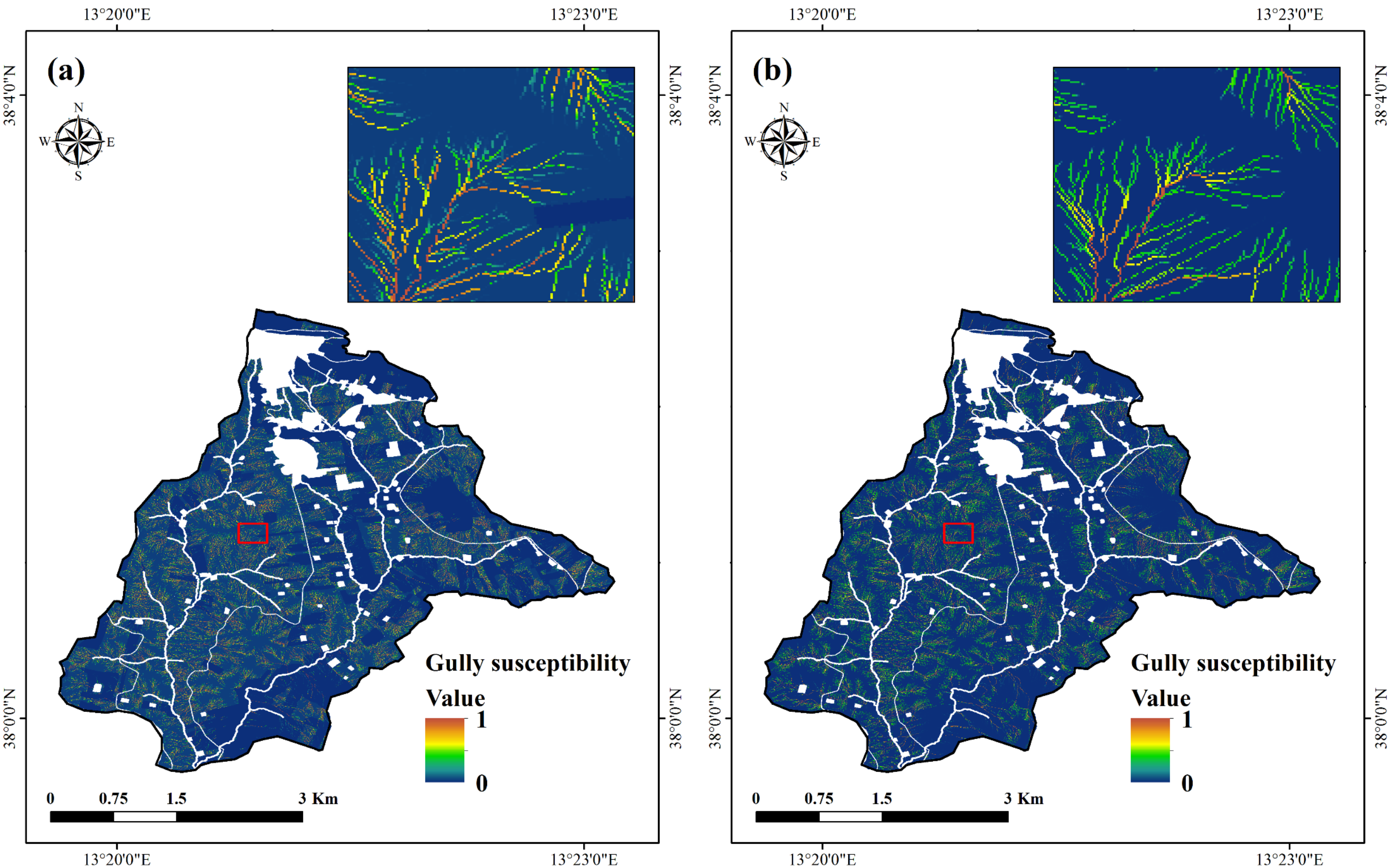


Fig. 1



Case studies
Environmental issues
Modeling approaches
Italy
Iran
Australia
Vietnam
USA
Gully erosion modeling
Flood hazard modeling
Drought risk modeling
Landslide susceptibility modeling
Soil digital modeling
MARS and GLM models
EBF model
Bayes' theorem and fuzzy logic
MaxEnt and BayGmmKda models
ANN and DTA models
Evaluation of the model performance using PMT
Cutoff-independent metrics
Receiver operating characteristic (ROC) curve
Success-rate curve (SRC) and prediction-rate curve (PRC)
Cutoff-dependent metrics
True positive rate (TPR)
Negative predictive value (NPV)
True skill statistic
False positive rate (FPR)
False omission rate (FOR)
Heidke's skill score
True negative rate (TNR)
Equitable threat score
Odds ratio
False negative rate
Matthews correlation coefficient
Odd ratio skill score
Efficiency (accuracy)
Misclassification rate
Cohen's kappa
Informedness
Markedness (MK)
Threat score
Positive predictive value (PPV)
False discovery rate (FDR)
F-score

Fig. 2



13°20'0"E
13°23'0"E
(a)
38°4'0"N
38°0'0"N
N
W
E
S
Gully susceptibility
Value
1
0
0
0.75
1.5
3 Km
(b)

Fig. 3

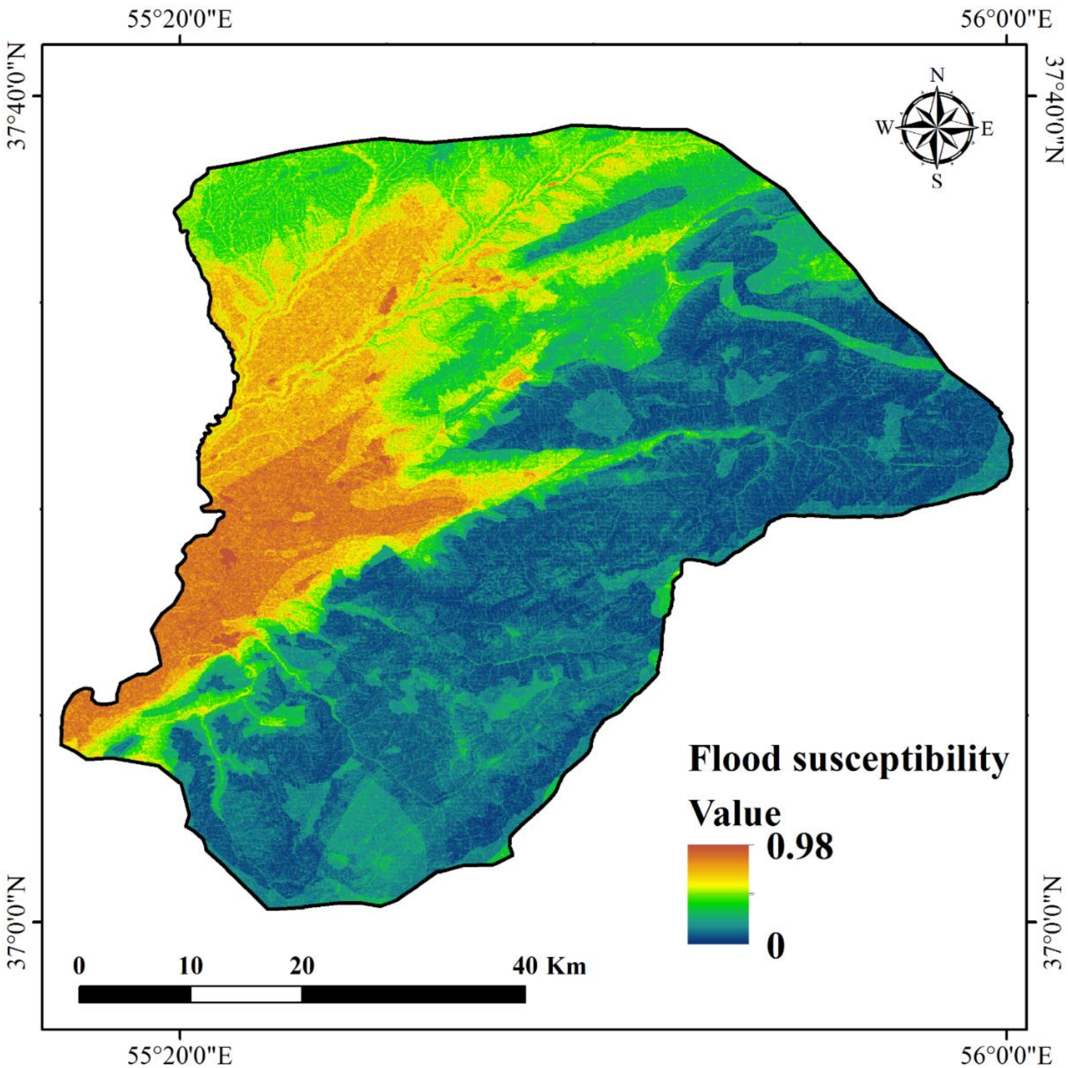

55°20'0"E
56°0'0"E
37°40'0"N
37°0'0"N
N
W
E
S
Flood susceptibility
Value
0.98
0
0
10
20
40 Km

Fig. 4

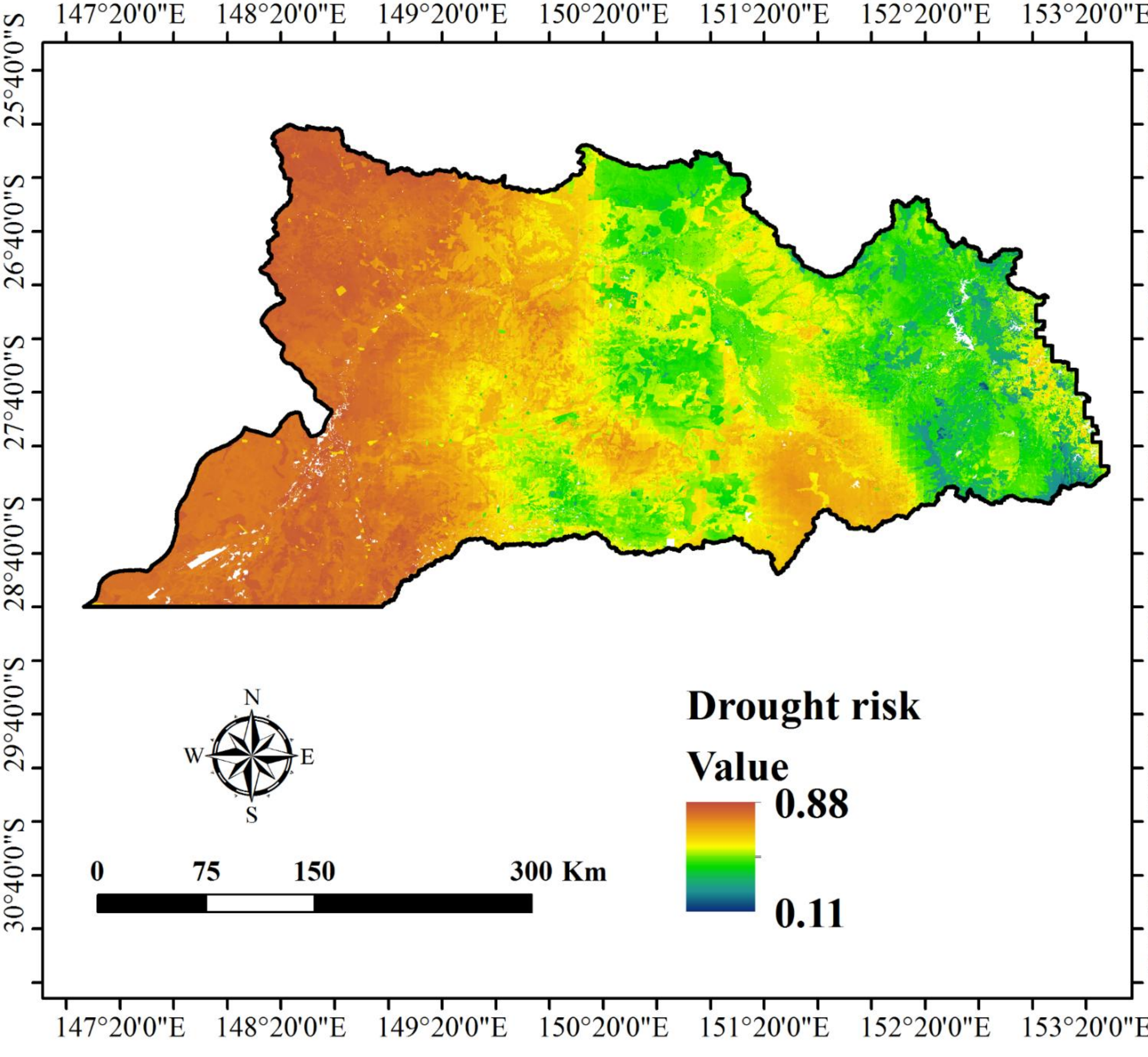

147°20'0"E
148°20'0"E
149°20'0"E
150°20'0"E
151°20'0"E
152°20'0"E
153°20'0"E
25°40'0"S
26°40'0"S
27°40'0"S
28°40'0"S
29°40'0"S
30°40'0"S
26°0'0"S
27°0'0"S
28°0'0"S
29°0'0"S
30°0'0"S
31°0'0"S
N
W
E
S
0
75
150
300 Km
Drought risk
Value
0.88
0.11

Fig. 5

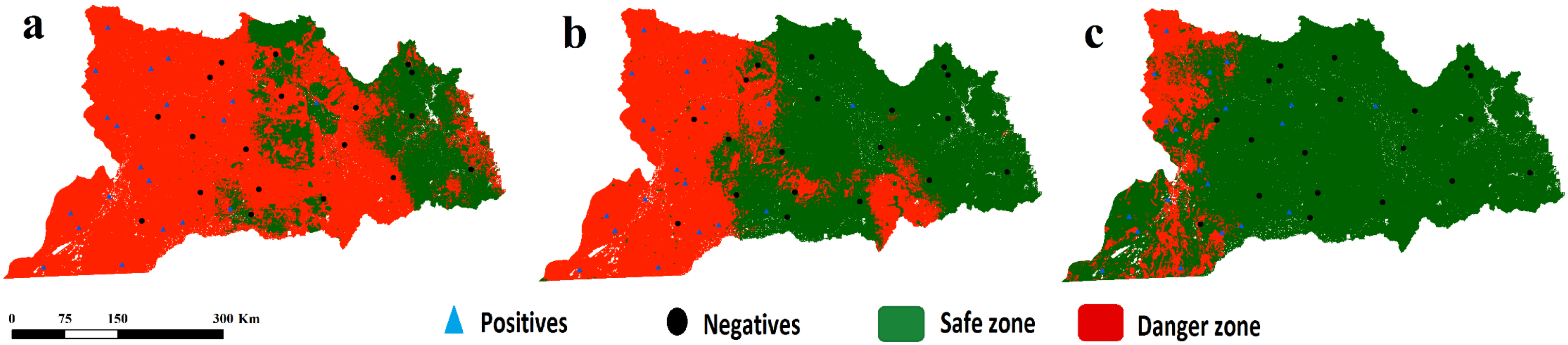
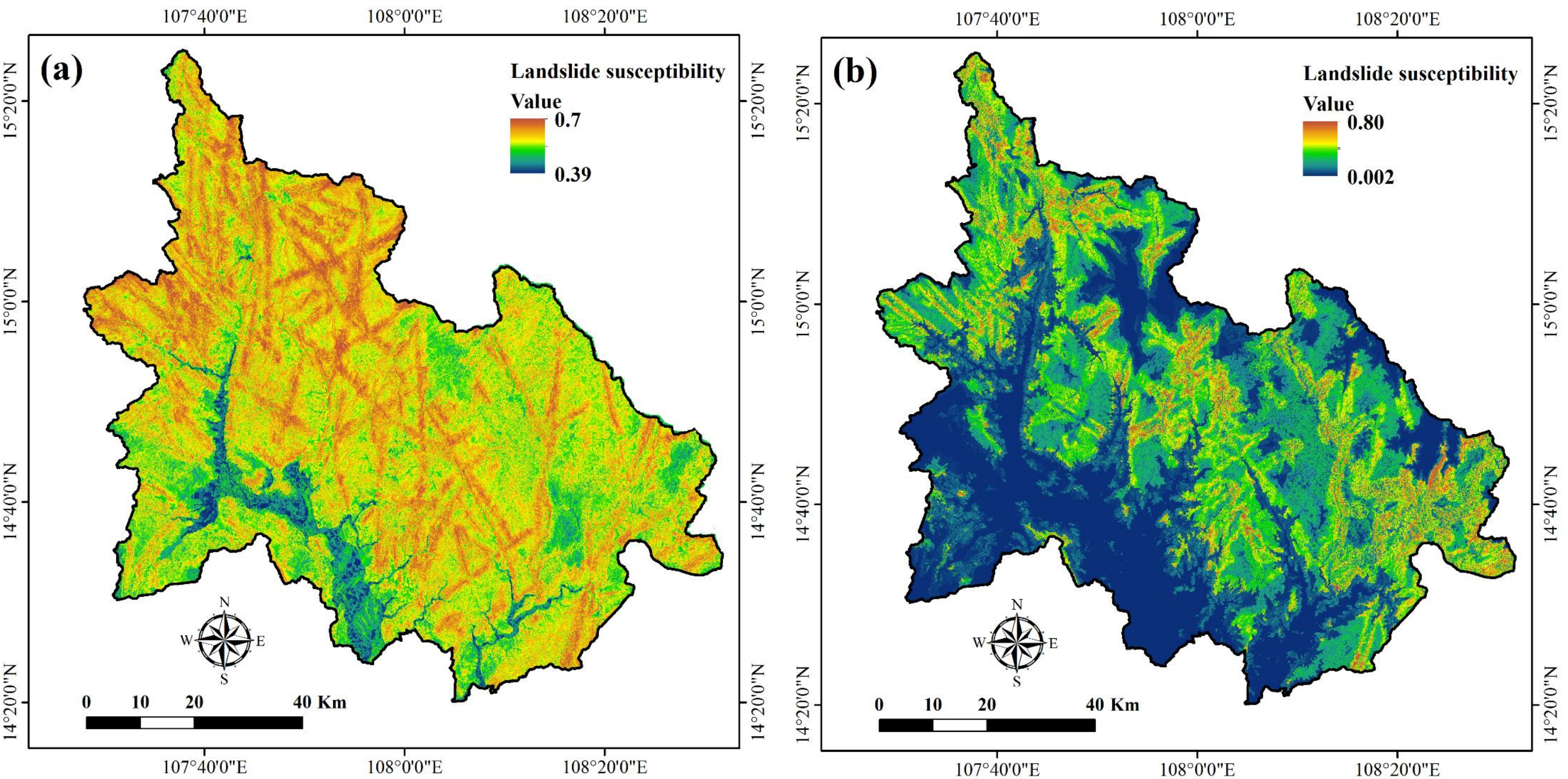
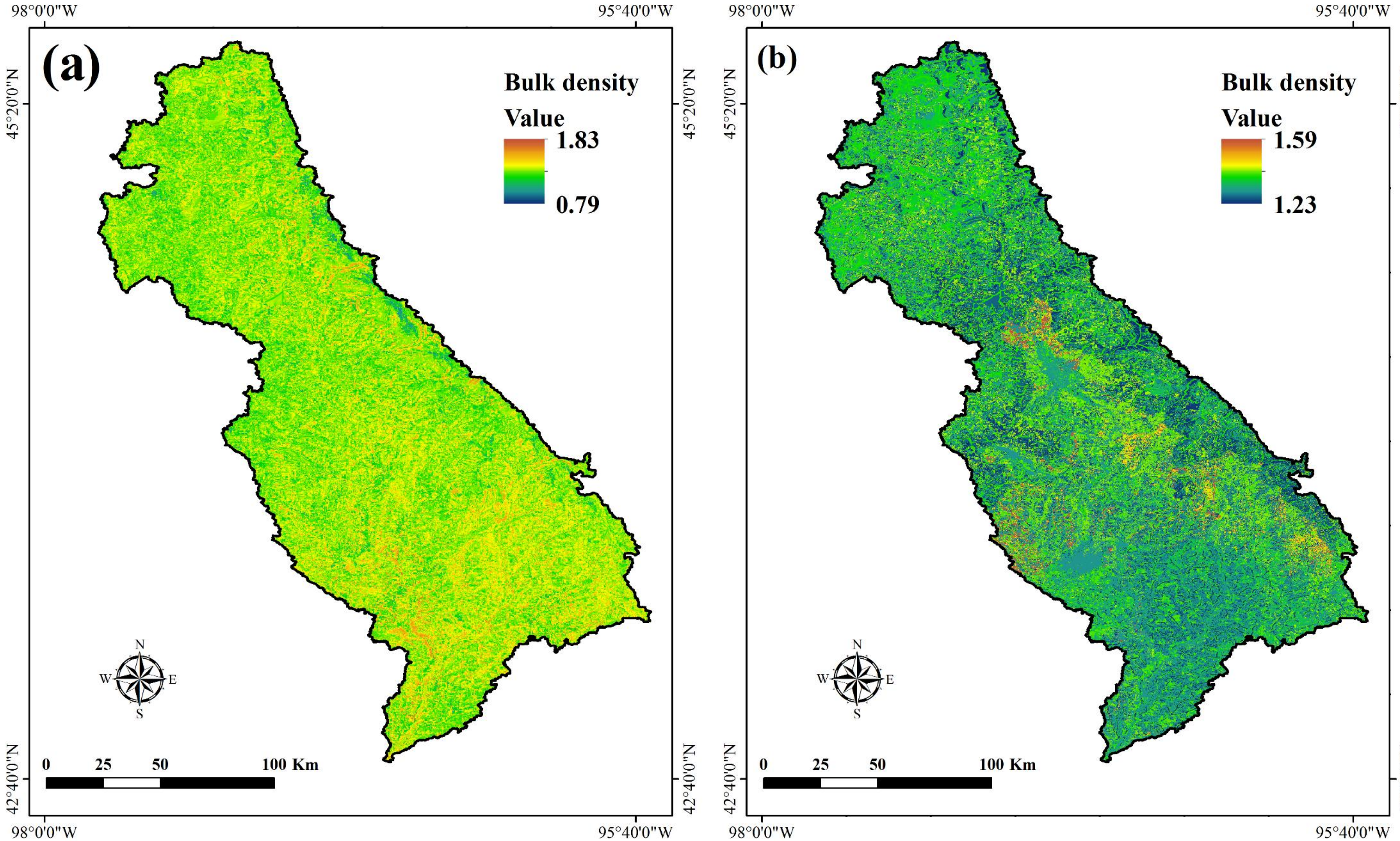

a
b
c
0 75 150 300 Km
Positives
Negatives
Safe zone
Danger zone

Fig. 6



(a)
Landslide susceptibility
Value
0.7
0.39
107°40'0"E
108°0'0"E
108°20'0"E
15°20'0"N
15°0'0"N
14°40'0"N
14°20'0"N
0
10
20
40 Km
(b)
Landslide susceptibility
Value
0.80
0.002

Fig. 7



(a)
Bulk density
Value
1.83
0.79
(b)
Bulk density
Value
1.59
1.23
98°0'0"W
95°40'0"W
45°20'0"N
42°40'0"N
N
S
E
W
0
25
50
100 Km

Fig. 8