AN OBJECTIVE METHOD FOR THE INTERCOMPARISON OF TERRAIN

STABILITY MODELS AND INCORPORATION OF

PARAMETER UNCERTAINTY

by

Kiran J. Chinnayakanahalli

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Civil and Environmental Engineering
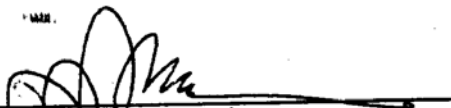
Approved:

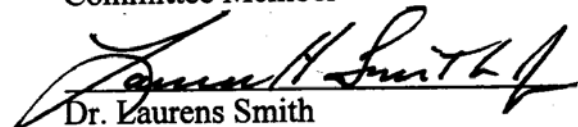_____
Dr. David G. Tarboton
Major Professor

_____
Dr. John C. Schmidt
Committee Member

_____
Dr. Robert T. Pack
Committee Member

_____
Dr. Laurens Smith
Interim Dean of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2004

ABSTRACT

An Objective Method for the Intercomparison of Terrain Stability Models and

Incorporation of Parameter Uncertainty

by

Kiran Chinnayakanahalli, Master of Science

Utah State University, 2004

Major Professor: Dr. David G. Tarboton
Department: Civil and Environmental Engineering

A terrain stability map quantifies the propensity for landslide initiation at a point in space. There are a number of process-based models used for terrain stability mapping. SHALSTAB and SINMAP are two such models that combine steady-state hydrology assumptions with the infinite-slope stability model to quantify slope stability. SHALSTAB quantifies terrain instability in terms of the critical effective rainfall required to trigger pore pressure-induced instability. SINMAP quantifies terrain stability in terms of the probability that the infinite-slope stability model factor of safety is greater than one over uniform probability distributions quantifying the uncertainty in model parameters. These are two models with similar physical basis but they use different indices to quantify instability. Because of this, the relative performance of these models, when compared to observed landslide locations, is difficult to assess.

A new statistic based on the cumulative fraction of the stability index at locations of observed landslides, relative to the cumulative fraction of the stability index over the entire terrain, has been developed. This statistic quantifies the discriminatory capability of a stability map in terms of the degree to which a terrain stability map is successful in "capturing" observed landslides in areas mapped as unstable while minimizing the extent of these areas. This statistic

provides a way to compare models that use different stability indices and is a quantity that can be used for the calibration or optimization of model parameters using observed landslide initiation locations. This is useful when physical data necessary to estimate model parameters is limited. Instead of searching for single optimal parameter sets, a modified Generalized Likelihood Uncertainty Estimation (GLUE) methodology is used to perform a random search over the model parameter space and establish behavioral parameter sets for each model. This approach provides an objective way for the selection of parameters to compare the discriminatory capability of different models and also provides a basis for quantifying the uncertainty associated with model predictions.

We show results where the performance of these models that use different indices is compared using data from the Chetwynd area in east central British Columbia where 696 landslide initiation locations were mapped. We found that in this particular region, slope was the dominant variable that discriminated terrain instability. The optimally calibrated SHALSTAB and SINMAP models that also use drainage area as a predictor offered improvement over slope alone. The additional flexibility of the probability framework used by SINMAP gave it a slight advantage over SHALSTAB in discriminating unstable terrain.

(114 pages)

ACKNOWLEDGMENTS

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Shallow landslides comprise mass wasting characterized by translational slides of rocks and debris at a relatively small depth. The sudden failure and high speed of shallow landslides result in the destruction of downstream resources, property, and lives (Montgomery and Dietrich, 1994). This necessitates the modeling of shallow landslide occurrence and mapping of their probability of occurrence in large spatial areas.

A model predicting the stability of a hillslope must capture the topographic and hydrologic influence on landslide susceptibility. The availability of continuously improving topographic data has helped landslide hazard mapping models to distinguish unstable terrain on a more refined scale. However this introduces the extra difficulty of estimating properties like soil strength, soil thickness, and root strength that have a strong influence on the potential for landslides but are spatially heterogeneous and difficult to measure at a refined topographic scale. These factors can vary significantly even at a scale of few meters (Dietrich, Bellugi, and de Asua, 2001). This makes landslide hazard mapping based on these criteria an arduous task.

There are many terrain stability models in use. The differences between models introduce the need to objectively compare models and evaluate their capability to discriminate unstable areas. The purpose of this study is to answer some of the problems of terrain stability mapping relevant to model comparison, uncertainty estimation, and, automatic calibration.

Problem Statement

Terrain stability mapping involves assessment of the propensity for landsliding based upon observations of landslide occurrence in similar topographic locations. Approaches range from manual mapping that relies on experience and expert knowledge to multivariate analysis and mechanistic based theory (e.g. Dietrich, Bellugi, and de Asua, 2001) utilizing computer programs and geographic information systems.

Manual mapping of landslide hazards (e.g., WFPB, 1997) is performed with aerial photography and field investigation. Based on knowledge of geology and topography, areas are classified into landslide hazard categories using expert judgment. This method relies heavily on an individual's expertise and experience, to differentiate between stable and unstable areas.

Multivariate analysis attempts to overcome the qualitative weaknesses of judgement based manual mapping by correlating terrain attributes with landslide occurrence to estimate landslide propensity in locations with similar terrain (Carrera, 1983; Carrera et al., 1991; Carrera et al., 1995). This approach, though more quantitative than manual mapping, is empirical in nature and hence there are limitations to extrapolation beyond the study region (Dietrich, Bellugi, and de Asua, 2001).

This study focuses on two physically based models, SHALSTAB (Montgomery and Dietrich, 1994) and SINMAP (Pack, Tarboton, and Goodwin, 1998). Both of these models couple digital representation of the topography, an infinite-slope stability model that balances the destabilizing components of friction and cohesion on a failure plane parallel to the ground surface, and, a shallow subsurface flow model to predict the spatial distribution of relative slope stability. The models differ in that SHALSTAB uses the

ratio of effective rainfall (or recharge) to transmissivity (R/T) that is required to induce instability as a stability measure while SINMAP uses a stability index based on the probability of stability recognizing uncertainty in model parameters. Parameter uncertainty is modeled through the use of uniform probability distributions between lower and upper bounds for uncertain parameters. Detailed descriptions of these models are given later, together with a review of other terrain stability modeling approaches available in the literature.

Given the choice of models or variables predicting terrain stability, it becomes an important question to ask, which one of them is performing better? The fact that different models use different indices to categorize stability makes the comparison task that much more difficult. It should be kept in mind that the performance may be dependent on the study area. Therefore, whenever it is decided to implement stability models to address issues of landslide stability, it is necessary to do model intercomparison studies to come up with a model that conforms better to the geographical region considered. Model intercomparison is an important field of study in hydrology, but although many studies have been carried out to validate terrain stability models, we are not aware of any studies that have compared different models objectively. This kind of study is important as it can help the end user with limited knowledge to choose the model with better predictive capability.

Process based distributed models such as SHALSTAB and SINMAP are parameter intensive in nature. In order for these models to be used, one needs to know the values of quantities like soil thickness, strength, and transmissivity in the study area. It is often difficult to quantify these variables or their spatial variability. Studies applying

process-based models typically use some kind of average or representative values, available from previous investigations in the area, that inevitably have some uncertainty associated with them. When a new site is selected for study, it may be a daunting task to collect the parameter data required to run the model. This necessitates the development of a methodology that can estimate the values for such parameters and may further help in quantifying the predictive and parameter uncertainty.

The input data set for SHALSTAB consists of topographic slope and specific catchment area and material properties, specifically friction angle and dimensionless cohesion. The input data set for SINMAP also consists of topographic slope and specific catchment area and parameters quantifying climate, hydrologic and material properties (Pack, Tarboton, and Goodwin, 1998). SINMAP, being a probabilistic model requires an input range for many of its parameters. To run SINMAP, one has to supply as inputs the upper and lower bounds for dimensionless cohesion, C, the ratio of transmissivity to recharge, T/R, and soil friction angle, ϕ. The ranges so specified must be those pertaining to the study area. Considering the spatial variability of these quantities, it can be difficult to obtain parameter values and assess parameter ranges.

SINMAP provides a graphical procedure for the manual, or interactive, calibration of model parameters while referring to a plot of slope versus specific catchment area that displays both observed landslides and threshold stability index lines. The user adjusts parameters within what he or she feels to be physically reasonable ranges trying to position the model threshold lines that are controlled by parameter inputs so that they discriminate the domain in slope area space that holds the densest clustering of observed landslides. This relies on the experience and skill of the user. An automatic

calibration procedure that can quantify the parameters with minimum knowledge requirement can be a very useful tool.

Automatic model calibration has been widely studied, often in conjunction with efforts to quantify parameter and model uncertainty (Kuczera, 1983; Madsen, 2000). Beven and Binley (1992) introduced the generalized likelihood uncertainty estimation method (GLUE) for parameter estimation and quantification of uncertainty associated with rainfall runoff models. This method has since been applied in other areas (Brazier et al., 2000; Christianes and Feyen, 2002). In the GLUE methodology, parameters are sampled at random from broad feasible parameter ranges and a likelihood (goodness of fit) measure used to identify those parameter sets that acceptably fit the observations. These are called behavioral parameter sets. Beven and Binley (1992) oppose the idea that there is a single optimum parameter set. Rather, they advocate equifinality, or non-uniqueness, recognizing that multiple parameter sets may adequately represent a modeled data set. Uncertainty in model predictions is quantified by sampling over the set of behavioral parameters. Kuczera and Parent (1998) have shown that there are more efficient ways (e.g., Metropolis algorithm) of sampling the parameter space than the general random approach of GLUE. This study utilizes the GLUE methodology to perform automatic calibration and to quantify parameter and predictive uncertainty.

Objectives

The goal of this work was to develop an objective method to compare and evaluate terrain stability models that differ in their approximation the physics, input parameters and output measure of terrain stability. The specific objectives were:

a) To develop a general objective measure to quantify the performance in terms of discriminating capability of a terrain stability model relative to observed landslide locations.

b) Implement the GLUE method with the general objective function to estimate behavioral parameter sets that acceptably fit the observations and quantify prediction uncertainty associated with parameter uncertainty, and

c) Apply the general objective function to compare SINMAP and SHALSTAB and quantify the space of behavioral parameters for each model.

The premise of this approach is that the propensity for landsliding can be quantified based upon topographic variables. A discriminating model is one that produces an output stability measure that, when used to classify the domain, captures a large number of landslides in a small fraction of the area. By using a terrain stability measure based on topographic variables the assumption is that future landslides are more probable in topographic settings similar to where past landslides have been observed.

CHAPTER 2

LITERATURE REVIEW

Hydrologic Models and Terrain Stability

A rainstorm or rapid snowmelt can lead to sufficient infiltration such that pore water pressure builds up at the contact between the soil mantle and the underlying impermeable or less permeable layer such as bedrock. The increased pore pressure reduces the effective normal stress, which through the friction angle, is related to the shear strength. The condition of reduced shear strength at the failure plane can trigger shallow landslides. Dietrich, Wilson, and Reneau (1986) suggest that topographically driven convergence of subsurface flow plays a role in increasing pore pressure resulting in the initiation of shallow landslides in fine-scale topographic hollows. The two main components of topography that are related to shallow landslides are slope and topographic control on convergence of shallow groundwater that leads to increased pore water pressure. The groundwater response following a rainfall event can be modeled by many hydrologic models like TOPOG (O'Loughlin, 1986) and TOPMODEL (Beven and Kirkby, 1979; Beven, 1997), which predict the spatial distribution of soil saturation. The slope stability component can use this saturation distribution to map unstable regions in the landform.

The topographic property used to quantify topographic convergence is specific catchment area (Beven and Kirkby, 1979). Specific catchment area, a, is defined as upslope area per unit contour width [m^2/m] (Figure 2-1). Given the tendency for water to flow downhill, specific catchment area serves as a useful surrogate for the subsurface

lateral water flux. The steady state hydrology parameterizations used by SINMAP and SHALSTAB slope stability models have their roots in hydrologic models like TOPOG (O'Loughlin, 1986) and TOPMODEL (Beven and Kirkby, 1979; Beven, 1997). These models assume that shallow subsurface flow is downslope following the topographic gradient, and that downslope subsurface discharge at each point is in equilibrium with a steady state recharge (R [m/hr]). It follows from these assumptions that depth integrated downslope discharge per unit contour length, q [m^2/hr] can be expressed as

$$q = R\,a \tag{2.1}$$

If one assumes that the soil profile has a transmissivity (depth integrated hydraulic conductivity) T [m^2/hr], the capacity for downslope flux (maximum subsurface discharge for a given soil thickness) at each point is obtained from Darcy's law as Tsinθ, where θ is
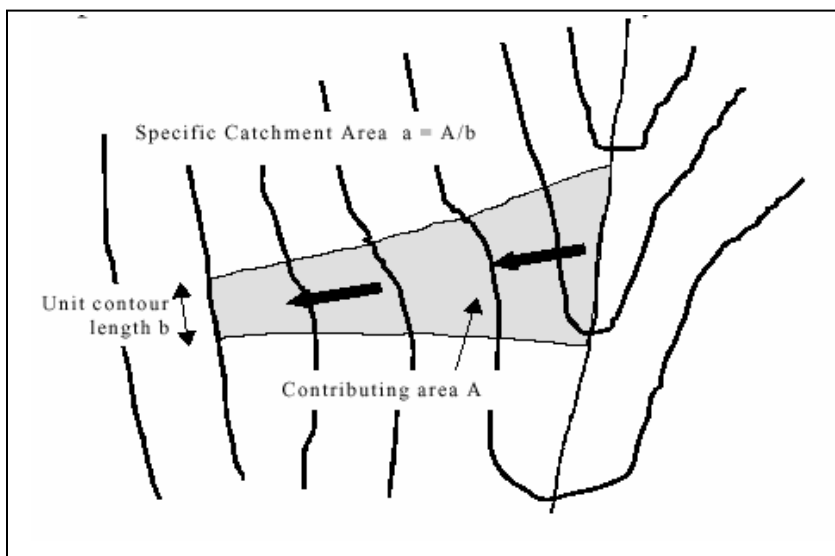
Figure 2-1. Definition of specific catchment area.

the topographic slope angle and the maximum hydraulic gradient is $\sin\theta = dh/dL$. The relative wetness of a soil profile is related to the ratio of the lateral moisture flux q to the lateral subsurface flow capacity $T\sin\theta$. This is used to define a wetness measure.

$$w = \min\left(\frac{q}{T\sin\theta},1\right) = \min\left(\frac{Ra}{T\sin\theta},1\right) \qquad (2.2)$$

The wetness measure has an upper bound of one because lateral flux q in excess of the lateral flow capacity is assumed to form overland flow.

TOPMODEL (Beven and Kirkby, 1979) further assumes that hydraulic conductivity decreases exponentially with depth. Lateral flow is assumed to occur in the saturated zone below the water table. The depth to the water table adjusts so that the lateral flow capacity below the water table is equal to the lateral flow. With this assumption the depth to the water table is proportional to $\ln\left(\frac{Ra}{T\sin\theta}\right)$. This subdivides into $\ln\left(\frac{R}{T}\right) + \ln\left(\frac{a}{\sin\theta}\right)$. The second term is a topographic variable and is defined as the TOPMODEL topographic wetness index $\ln(a/\sin\theta)$. TOPMODEL actually uses $\tan\theta$ in this expression, a difference that is inconsequential for small slopes. For steep slopes the approximation $\sin\theta = \tan\theta$ is not valid and it is correct to use $\sin\theta$ in the equation because flow is along the length of the slope.

SHALSTAB (Montgomery and Dietrich, 1994) and SINMAP (Pack, Tarboton, and Goodwin, 1998) use the simpler assumption that hydraulic conductivity is constant over a soil depth D. Transmissivity is then K.D and w in equation 2.2 represents the fraction of soil depth D that is required to be saturated to conduct the lateral flow q. The ratio R/T $[m^{-1}]$ is a time varying quantity that gives area averaged relative wetness in terms of assumed steady state recharge relative to the soil's capacity for lateral drainage of water. Although the term 'steady state' is used with lateral flux approximated using Equation (2.1), the quantity R is not a long term (e.g. annual) average of recharge. Rather, it is the effective recharge for a critical period of wet weather likely to trigger landslides. The ratio R/T combines both climate and hydrogeological factors. The quantity $(T/R)\sin\theta$ [m] may be thought of as the length of hillslope (planar, not convergent) required to develop saturation in the critical wet period being considered.

Once soil wetness is computed, it becomes the role of the infinite-slope stability model to utilize these data to categorize unstable regions in the landform through the relationship between relative wetness and pore pressure that affects shear strength. SHALSTAB uses w from Equation 2.2 to derive pore pressure in the infinite-slope stability model and then to solve for the critical R/T required to trigger instability at each location. SINMAP uses R/T as an input model parameter with a uniform distribution. The w from equation 2.2 is then used to derive pore-pressure used with the infinite-slope stability model to calculate factor of safety and the probability of stability.

<u>Slope Stability Models</u>

The infinite-slope stability equation used by many process-based models resembles the one used by Hammond et al. (1992) in Level I Stability Analysis (LISA) model which is based on the Mohr-Coulomb failure law. The infinite-slope stability model used by LISA calculates a factor of safety that is the ratio of the restoring components of friction and cohesion to the destabilizing components of gravity on failure plane. LISA has the following assumptions:

a) Downslope discharge flowlines and the groundwater level are parallel to the topographic surface.

b) The failure plane is of infinite length parallel to the topographic surface.

c) Only one soil layer is considered.

The failure plane of infinite length assumption applies when the length of the failure plane is large compared to soil thickness. Iverson and Major (1986, 1987) and
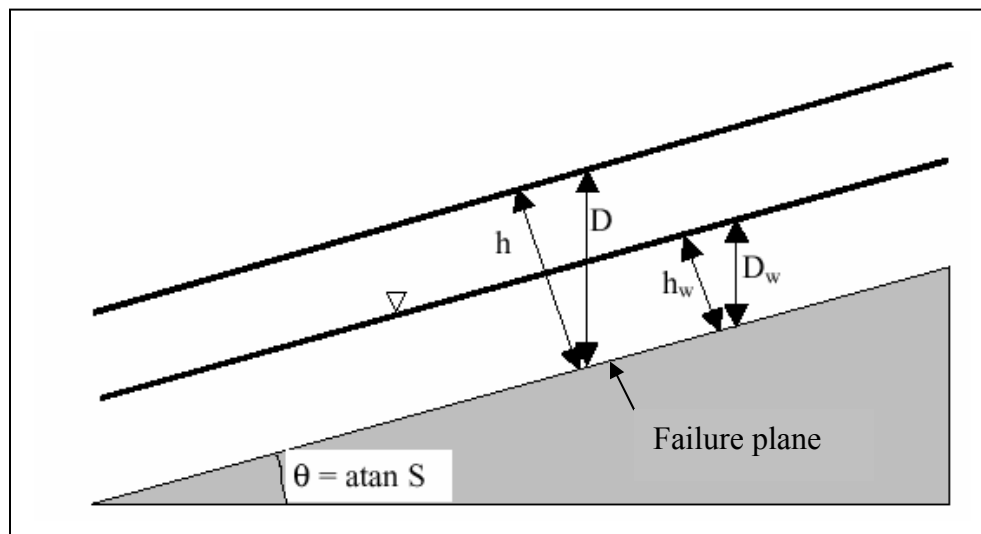
Figure 2-2. Infinite-slope stability model schematic.

Iverson (2000) discuss the limitations associated with the assumption of flow parallel to the topographic surface/failure plane. They report that the factor of safety may be significantly overestimated or underestimated, depending on the actual direction of seepage. At the level of the simplified models being evaluated here the information necessary to calculate seepage directions is rarely available so this work uses the parallel flow assumption. The infinite-slope stability model is illustrated in Figure 2-2. The infinite-slope stability model factor of safety (FS) is given by Hammond et al. (1992):

$$FS = \frac{C_r + C_s + \cos^2\theta\left[q_0 + \rho g(D - D_w) + (\rho_s g - \rho_w g)D_w\right]\tan\phi}{\sin\theta\cos\theta\left[q_0 + \rho g(D - D_w) + \rho_s g D_w\right]} \qquad (2.3)$$

where

C_r = apparent root cohesion [N/m^2]

C_s = soil cohesion [N/m^2]

θ = slope angle, the arc tangent of slope, S, which is the change in elevation per unit horizontal distance.

q_0 = tree load [N/m^2]

ρ_s = bulk density of saturated soil below the water table [kg/m^3]

ρ = bulk density of soil above the water table [kg/m^3]

ρ_w = density of water [kg/m^3]

D = vertical soil depth

D_w = vertical height of the water table within the soil layer [m]

ϕ = internal frictional angle.

In Figure 2-2,

$$h = D \cos\theta \tag{2.4}$$

where h[m] is the soil thickness perpendicular to the slope. This equation predicts the effect of pore pressure due to a water table thickness h_w and vertical height D_w on the shear strength at the failure plane. With the assumption of uniform hydraulic conductivity these water table variables are related to relative wetness from Equation 2.2 using

$$w = \frac{D_w}{D} = \frac{h_w}{h} \tag{2.5}$$

Pack, Tarboton, and Goodwin (1994) combined C_r and C_s with soil density and thickness to obtain a dimensionless cohesion C expressed as

$$C = \frac{(C_r + C_s)}{h\rho_s g} \tag{2.6}$$

C quantifies the contribution from cohesive forces to the slope stability. Dietrich et al. (2001) in developing SHALSTAB with cohesion, used a similar non-dimensional cohesion term, with the difference that C_s in SHALSTAB is taken as zero and SHALSTAB uses D instead of h Equation 2.6.

Pack, Tarboton, and Goodwin (1998) assumed q_0 in Equation 2.3 to be equal to zero and took $\rho = \rho_s$ in their slope stability equation.

Further, denoting the water to soil density ratio $\dfrac{\rho_w}{\rho_s} = r$ \hfill (2.7)

the factor of safety, FS, can be expressed as

$$FS = \frac{C + \cos\theta[1 - wr]\tan\phi}{\sin\theta}$$ (2.8)

In the above equation, the relative wetness measure w is the key quantity that connects the hydrologic model to the infinite-slope stability model. This equation is the basis for SINMAP (Pack, Tarboton and Goodwin, 1998) that uses slope (derived from the digital elevation model) and w (obtained from the hydrologic model) to calculate the relative stability at each grid point.

Pack, Tarboton and Goodwin (1998), account for uncertainty by assuming parameters to be uniformly distributed between the lower and upper bounds. The bounds represent the feasible ranges of the parameter expected in the study area. The model outputs a stability index (SI) at each digital elevation model (DEM) grid cell. SI is defined as the probability that a location is stable (FS>1) assuming uniform probability distributions for the parameters. The values of SI vary between 0 (most unstable) and 1 (least unstable). In the domain where the probability that FS>1 equals one, the model is not discriminating. This occurs when FS > 1 for the most unstable limits on the parameters. In these cases, SI is given the value of the factor of safety with the parameter values at their most unstable bound limits.

Montgomery and Dietrich (1994) built on the work on distributed landslide modeling carried out by Okimura and Ichikawa (1985) to develop a process-based model to predict land instability. The model combined digital terrain data with the infinite-slope stability model and the hydrologic model TOPOG (O'Loughlin, 1986), to map the spatial

patterns of potentially unstable zones in the terrain. Montgomery and Dietrich (1994) in their initial form of the model ignored cohesion and assumed uniform soil thickness. They solved for the critical steady state rainfall for each location that increases the wetness to the point where the factor of safety, FS, is equal to one and the slope becomes unstable. The steady-state-critical-rainfall, R, required to cause instability provides a measure of the relative potential for landslide instability because the occurrence of instability inducing wetness is more likely at locations with smaller critical R. The steady state critical rainfall, R, is given by the following equation:

$$R_{cr} = \frac{T \sin \theta}{a} \left(\frac{\rho_s}{\rho_w} \right) \left[1 - \frac{\tan \theta}{\tan \phi} \right].$$
(2.9)

Montgomery and Dietrich (1994) estimated both soil and hydrologic parameters required by the model based on their field experience in their study areas in Marin County, California, and Mettman Ridge, Oregon. They treated the parameters as spatially constant, though they recognized that spatial variability could be incorporated. The spatially constant assumption was merely because data were not available. While acknowledging the uncertainty in the estimation of hydrologic parameters, soil conductivity, thickness, bulk density, and friction angle, they did not adopt the Monte Carlo method for predicting the spatial distribution of the parameters because they felt that Monte Carlo method would obscure the topographic influence on slope stability. Later Dietrich et al. (1998) included the cohesion term as follows:

$$\frac{R}{T} = \frac{\sin\theta}{a}\left[\frac{\rho_s}{\rho_w}\left(1-\left(1-\frac{C^*}{\sin\theta\cos\theta}\right)\right)\frac{\tan\theta}{\tan\phi}\right]$$

(2.10)

Here $C^* = \dfrac{C_r}{\rho_s gD}$, with C_r the apparent cohesion due to roots, g the acceleration

due to gravity and D the vertical soil thickness. This is different from SINMAP (Pack,

Tarboton, and Goodwin (1998), which considers depths perpendicular to the soil surface.

Instead of critical rainfall, this equation represents instability using the ratio of rainfall to

transmissivity (R/T) as the index of instability.

Borga, Fontana, and Cazorzi (2002) used a quasi-dynamic wetness index model

developed by Barling, Moore, and Grayson (1994) to model the spatial distribution of the

soil saturation in response to a rainfall of specified duration. This concept is different

from the steady state wetness index model used by SHALSTAB and SINMAP. The

steady state hydrologic model used in SINMAP and SHALSTAB assumes that the

specific catchment area is predictive of the subsurface flow at any given point. However,

this is valid only if the recharge to a perched water table occurs "for the length of time

required for every point along the hillslope to reach subsurface drainage equilibrium and

experience drainage from its entire upslope contributing area" (Barling, Moore, and

Grayson, 1994). The quasi-dynamic wetness index model considers the time for the

subsurface water to redistribute across the catchment. Barling, Moore, and Grayson

(1994) assumed a uniform recharge and routed the subsurface flow using a spatially

variable velocity which depends on the local slope. Borga, Fontana, and Cazorzi (2002)

combined this subsurface flow model with the infinite-slope stability model for a

cohesionless soil of constant thickness to identify unstable topographic elements so as to account for the duration associated with the critical rainfall. Borga's approach replaces Equation (2.9) by

$$R(d)_{cr} = \frac{T \sin \theta}{a(d)} \frac{\rho_s}{\rho_w} \left[1 - \frac{\tan \theta}{\tan \phi} \right] \qquad (2.11)$$

where d is the duration of rainfall, a(d) the dynamic specific catchment area that contributes given rainfall of duration d and $R(d)_{cr}$ the consequent critical rainfall. The hydrologic relationship between duration and frequency of occurrence of the rainfall causing instability in the topographic element is evaluated for a range of durations and terrain stability is quantified using the shortest return period (highest frequency of occurrence) of the critical rainfall at each point over the set of rainfall durations tested. Thus, terrain stability is quantified based on the probability of critical destabilizing rainfall accounting for partial contributing area effects and rainfall intensity, duration, and frequency properties.

Wu and Sidle (1995) presented a dynamic model that integrated the infinite-slope stability equation, a kinematic wave groundwater model, and a continuous change root strength model. Wu and Sidle (1995) followed the work of Moore and Grayson (1991), to combine surface and subsurface kinematic models for the dynamic simulation of soil wetness in response to an input time series of precipitation. This approach allowed the investigation of FS both in the temporal and spatial dimension as well as examination of the impact of timber harvesting and forest management. Wu and Sidle used averages

from the study of Schroeder and Alto (1983) to obtain their soil parameters. They recognized, following the work of Gray and Megahan (1981), that the factor of safety given by the infinite-slope stability model is most sensitive to soil cohesion, root cohesion, soil depth, and slope angle while moderately sensitive to the groundwater height and frictional angle.

In Wu and Sidle's (1995) model landslides are assumed to occur when FS becomes less than 1. Landslide volumes were calculated and Wu and Sidle report the changes in landslide volumes modeled when the soil characteristics were varied. Uncertainties involved with the spatial distribution of the parameters like soil depth complicated the problem of comparing the simulated locations of failure with the survey data (Wu and Sidle, 1995). The study recognized the difficulties in accurately determining some of the sensitive parameters while still maintaining the advantages of spatially distributed models.

The discussion on parameters from different approaches so far has pointed to the fact that models are only as good as their parameters. Hence, it becomes an important task to estimate parameters. The following discussion of parameter uncertainty follows the work of Beven and Binley (1992).

Generalized Likelihood Uncertainty
Estimation (GLUE) Methodology

The common approach to calibration in hydrology is to search for a global optimum set of parameters. Many hydrologic studies (for example Ibbitt and O'Donnell, 1971; Kuczera, 1983; Kuczera, 1990) have illustrated the difficulties in finding such an optimum set of parameters. Difficulties arise from the presence of threshold parameters,

from the intercorrelation among parameters, from autocorrelation and heteroscedascity in the residuals and from insensitive parameters (Kuczera and Parent, 1998). Such effects result in local minima, and, valleys and plateaus in the parameter response surface (Beven and Binley, 1992). Difficulties increase with the number of parameters. Physically based distributed models generally have many parameters and hence are ridden with parameter estimation problems.

Beven and Wood (1983) addressed the problem related to the parameterization of physically based distributed hydrologic models by arguing that in order to realistically represent a physical process one has to acknowledge the uncertainty involved. Beven and Binley (1992) suggested a need for realistic estimation of predictive uncertainty. In this context, they introduced the Generalized Likelihood Uncertainty Estimation (GLUE) procedure for quantifying the predictive uncertainty of a model and parameter set. The GLUE procedure can also be used for parameter estimation.

GLUE assumes the equivalence or near equivalence of different parameter sets as simulators of systems. This property of different parameter sets that behave as equally capable simulators producing similarly acceptable behavior of the physical system, is termed equifinality (Beven and Binley, 1992). Finding behavioral parameter sets requires a large number of model runs for different sets of parameters that are chosen from probability distributions (often uniform) over their feasible range. An important requirement of the GLUE protocol is the generalized likelihood measure that ascribes each parameter set used to run the model a value, which is a measure of its predictive capability or goodness of fit. Such assignment may be based on direct comparison with the observed values or on *a priori* knowledge about the system. The likelihood measure

quantifies how well the model and parameter set conforms to the observed behavior of the system. Any interaction between the parameters will be implicitly reflected in the likelihood values. The likelihood measure should increase or decrease monotonically as the goodness of fit increases.

The requirements for application of the GLUE procedure are as follows (Beven and Binley, 1992):

a) Formal definition of a likelihood measure which quantifies the predictive capability of the parameter set.

b) Definition of the initial range or distribution of parameter values to be considered for a particular model structure.

c) A procedure for using likelihood measures in uncertainty estimation. This procedure results in the behavioral parameter sets whose distribution quantifies the parameter uncertainty and leads to the estimation of predictive uncertainty in step 5.

d) A procedure for updating likelihood measures recursively as new data become available. This refers to a Bayesian approach of updating the likelihood function conditioned on the new data.

e) A procedure for evaluating uncertainty such that the value of additional data can be assessed. This requires assessing how uncertainty quantified across the behavioral parameter sets changes as new data become available.

Although the concept of likelihood has formal meaning in probability theory, in the GLUE methodology it is extended to encompass any measure of goodness of fit of a model, not limited to the probabilistic definition.

In this study, we develop an objective measure, denoted Q, that quantifies the effectiveness of a terrain stability map in a general non-parametric way. We suggest this be used as a likelihood measure in the GLUE procedure so that the GLUE methodology can be applied to the terrain stability-mapping problem. Only the first three requirements of the GLUE method are pertinent to our study and we did not implement the last two steps. We therefore refer to the GLUE procedure used within our parameter optimization as the modified-GLUE method.

Terrain Stability Model Performance Evaluation

Model performance evaluation is a general process following model development which informs potential users of the effectiveness of the model. Performance evaluation of a terrain stability model is done by comparing the observed landslide locations with the model predictions (Dietrich et al., 1998; Dietrich, Bellugi, and de Asua, 2001; Montgomery et al., 2000; Montgomery, Sullivan, and Greenberg, 1998; Borga, Fontana, and Cazorzi, 2002; Zaitchik, van Es, and Sullivan, 2003). For the model to perform well, most of the observed landslides, should be contained in the unstable area of the terrain as predicted by the model. A better model would contain more landslides in its unstable region. Performance evaluation can be used to compare the performance of one model with different parameters, as in parameter calibration, or to compare the performance of different models.

Montogomery, Sullivan, and Greenberg (1998) used an inventory of 3224 mapped landslides in Oregon and Washington covering an area of 2993 km^2 over 14 watersheds to evaluate the SHALSTAB terrain stability model. Montgomery, Sullivan,

and Greenberg (1998) used a threshold of critical rainfall, R_{cr} (the minimum rainfall required to cause instability in a grid cell) as an index to categorize stability zones. They used seven R_{cr} categories (varying between least to stable zones) to create the terrain stability map. The distribution of landslides in R_{cr} categories was statistically tested to evaluate the model. The null hypothesis used by Montgomery, Sullivan, and Greenberg (1998) was that if the model is not discriminating the relative landslide hazard the distribution of landslides in each R_{cr} category occurs in proportion to the area of that R_{cr} category. That is, if the model is discriminating then it would have higher density of observed landslides in unstable categories and hence the null hypothesis would be rejected. The density of landslides is the number of landslides in each category divided by the terrain area in each category.

Montgomery, Sullivan, and Greenberg (1998) used Pearson's chi-square statistic for goodness of fit for categorized data to test their null hypothesis. They found that the test statistic was statistically significant for each watershed leading them to reject the null hypothesis.

Other studies by Dietrich et al. (1998) and Dietrich, Bellugi, and de Asua (2001), used the same data and the model (SHALSTAB) to study the effects of forest clearing on landsliding. Here, they use $\log(R_{cr}/T)$ as the stability measure instead of R_{cr}. Dietrich et al. (1998) and Dietrich, Bellugi, and de Asua (2001) used polygons to represent the observed landslide locations which are compared with model predictions. The polygons can cover more than one grid cell and in such cases the lowest value of $\log(R_{cr}/T)$ lying within the polygon is assigned to the landslide. This causes a bias towards lower values of $\log(R_{cr}/T)$, and hence an evaluation method like the one

followed by Montgomery, Sullivan, and Greenberg (1998) would be biased toward showing better performance of the model.

To see if there is a significant bias toward low values of $\log(R_{cr}/T)$, Dietrich et al. (1998) and Dietrich, Bellugi, and de Asua (2001) use a similarly biased random model as a point of reference to test the performance of SHALSTAB. The biased random-model creates as many landslides as were actually observed of approximately the same size as the median size of the observed landslides. It then places them randomly in the terrain. As with the observed landslides, Dietrich et al. (1998) and Dietrich, Bellugi, and de Asua (2001) assign the lowest value of $\log(R_{cr}/T)$ to each of the randomly generated landslides. Dietrich et al. (2001) hypothesize that if the bias is significant, then the distribution of landslides in various $\log(R_{cr}/T)$ categories must be similar to the distribution obtained by the random model. If not, the model is said to be discriminating the relative landslide hazard.

Dietrich et al. (1998) and Dietrich, Bellugi, and de Asua (2001) use plots of landslide density versus the $\log(R_{cr}/T)$ to evaluate the model performance relative to the random model (Figure 2-3). If there is no bias because of allocation of the lowest $\log(R_{cr}/T)$ to the landslide, then the landslide density curve from the random model will be flat as shown by the horizontal line in Figure 2-3. The bias caused the random model to have slightly higher landslide incidence at lower values of $\log(R_{cr}/T)$. For the model to perform better, it should have higher incidence of landslides than the random model at lower $\log(R_{cr}/T)$ categories. Dietrich et al. (1998) showed that SHALSTAB performed better than the random model in all the seven North California-watersheds

used in their study. They, however, do not report the statistical significance of differences between SHALSTAB and the random model performances.

Zaitchik, van Es, and Sullivan (2003) present a method that is different from the methods mentioned above to evaluate terrain stability models. Zaitchik, van Es and Sullivan (2003) acknowledge that when landslides are triggered by local factors not accounted for by the physically based models, the spatial distribution of the landslides will be different from the spatial structure of the model predictions. In general, landslides are clustered because of the local interaction or mutual correlation with a single destabilizing process which might be active locally and not accounted for by the model. Such clustering, Zaitchik, van Es, and Sullivan (2003) suggest would introduce a spatial bias in the model evaluation.

Zaitchik, van Es, and Sullivan (2003) applied SINMAP to an agricultural area in central Honduras to study the affect of hurricane Mitch. Zaitchik, van Es, and Sullivan (2003) noted that the observed landslide locations were more closely clustered than those predicted by SINMAP which indicates that fine scale physical phenomena not captured by the model may play a role in initiating landslides. Zaitchik, van Es, and Sullivan (2003) improved the model performance by aggregating the observed landslides on a scale of 150m. This method would imply that two different models can be made to perform well at different aggregation scales.
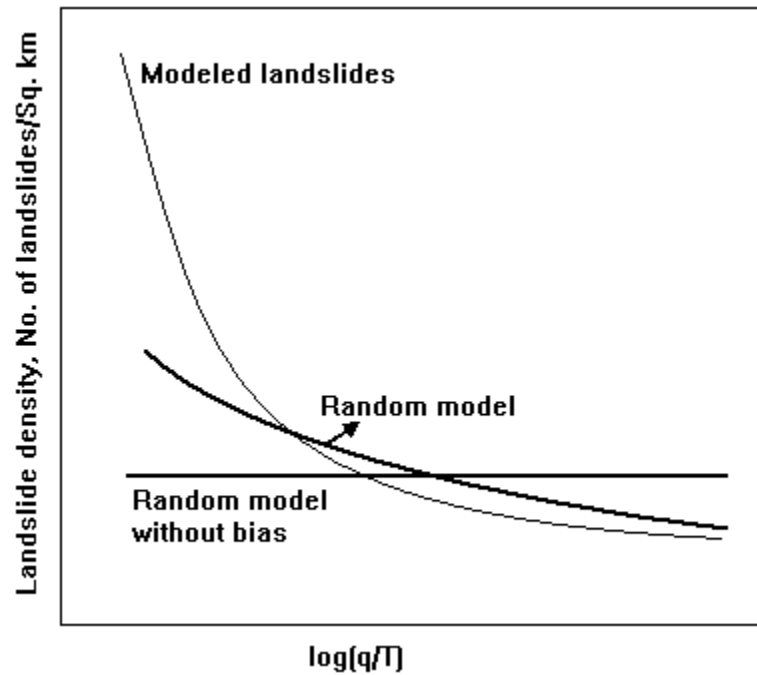
Figure 2-3. Example of landslide density plots used in the evaluation study by Dietrich et al. (1998).

Dhakal, Amada, and Aniya (2000) like others recognized that a large number of observed landslides in the unstable category of the hazard map provides an indicator of the performance of the model, but did not mention that such categorization must have smaller associated terrain area to imply a better model performance. They used a multivariate statistical approach to create a hazard map for the Kulekhani watershed located in the Lesser Himalayan region of the Himalayan belt in the central region of Nepal. A split sampling method was used to evaluate the model. The total number of observed landslides was divided into two sets, and one of them was used in the multivariate statistical method to create a landslide hazard map that had categories: Unstable, Moderately Unstable, Less Unstable and Least Unstable. The second set (test landslides) of observed landslides was mapped on the hazard map and the incidence of

the test landslides in the unstable category was compared with the previous set of landslides used in the creation of the hazard map. This is not the most stringent type of test to evaluate the model but Dhakal, Amada, and Aniya (2000) used it to study the general effect of sampling on the resulting hazard maps.

Borga, Fontana, and Cazorzi (2002) compare their quasi-dynamic terrain stability model with a steady state model (similar to SHALSTAB) by plotting the cumulative distribution of the model stability measure at observed landslides against the cumulative distribution of the model stability measure over the terrain. The objective measure, Q, that we develop below (see Chapter 4) for quantifying the discriminating capability of terrain stability maps, results from an integration of plots like this. Using this graphical approach, Borga, Fontana, and Cazorzi (2002) demonstrated that quasi-dynamic model performed better than the steady-state model in discriminating landslides. However, this finding is difficult to generalize because the comparison was done without optimizing the model parameters.

CHAPTER 3

STUDY AREA

This work was performed using data from the Chetwynd area in British Columbia, Canada (Figure 3-1). This area was chosen because the data were readily available from a previous study by Pack (personal communication) who mapped and checked the quality the landslide initiation locations.

The study area lies in the foothills of the Rocky Mountains to the north of Williston Lake (Figure 3-1), just before it flows into the Peace River, at 56° 15' north and 122° 23' west. The Canadian Cordilleran region is divided into five belts, distinguished by their different histories of uplift and subsidence allied with different geological structures (Canada Soil Survey Committee, 1977). The Rocky Mountains lie along the Foreland belt, which comprises the eastern mountain ranges and foothills of the Canadian Cordillera and are composed of imbricated and folded miogeoclinical and clastic wedge assemblages deposited on and adjacent to the stable craton of Ancestral North America (Gabrielse, 1991).

The area under investigation is 497 km^2, with a topographical relief of 1230 m (675 to 1909 m). Much of the bedrock is made up of layered Jurassic and Cretaceous sedimentary rock. The eastern side of the study area is composed of mudstone, siltstone and shale. The geology of the west and the center parts have not been mapped in detail are classified as undivided sedimentary rock and coarse clastic sedimentary rock. A major thrust fault, north-westward in direction, traverses the eastern side of the study area.
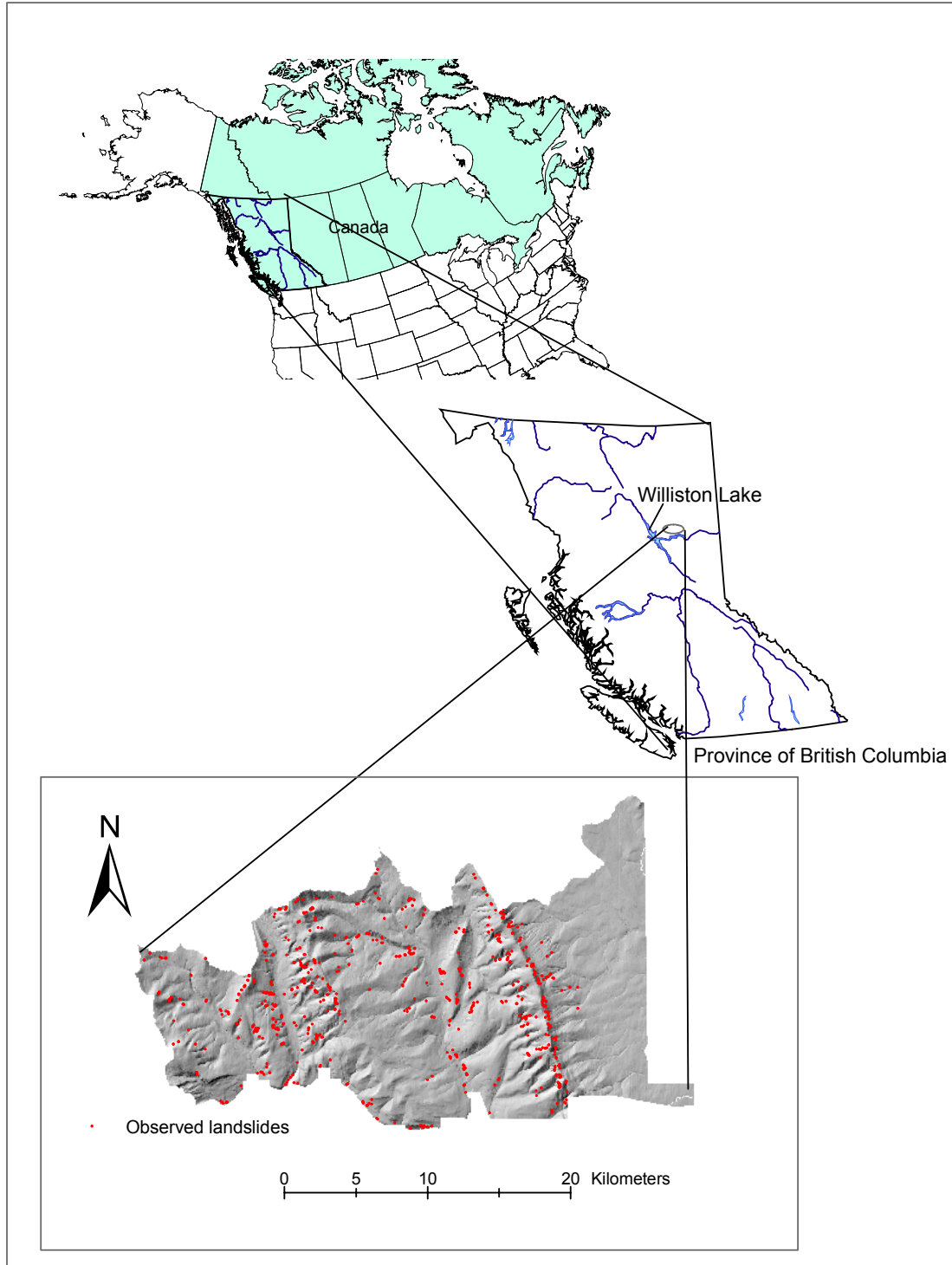
Figure 3-1. Map of Chetwynd study area.

The following information is obtained from Jukes and Andrew (2001). The study area is administered by the Forest District of Dawson Creek in the Prince George Forest Region. The nearest community is Hudson's Hope, a District Municipality, located approximately 10 miles, south-east of the study area. This area is one of the five blocks leased under Tree Farm License (TFL 48, also known as Chetwynd TFL) to Canadian Forest Products Ltd (Canfor) for tree farming. The Boreal White and Boreal Spruce communities are the dominant biogeo-climate in the region and the principal commercial species are White Spruce and Aspen.

The daily mean temperature of the region varies between $-9^{o}C$ in December to $15.3^{o}C$ in July (Environment Canada, 2004). The mean annual rainfall is 318 mm and the mean annual snowfall is about 170 cm in the region (Environment Canada, 2004). Landslides in the region are believed to be triggered by the increase in soil moisture following snowmelt or high intensity rainfall that results in soil saturation.

TRIM (Terrain Resource Information Management) DEM data for the project area were obtained from the Chetwynd Division of Canfor. These data are produced and sold by Base Mapping and Geomatic Services of MSRM (Ministry of Sustainable Resource Management), B.C. Data were digitally compiled from 1:60,000 scale photographs at an accuracy appropriate for a 20 m contour interval map (i.e spot elevation accuracies of plus or minus 5 m). The data were interpolated to a 15 m grid DEM from the raw irregularly spaced elevation points using a triangulated network interpolation method. The DEM created from the contour interval map with an accuracy of ±5 m is not ideally suited for modeling shallow landslides whose average depth is approximately 2m, because they may be triggered in terrain features to small to be

resolved by the DEM. Nevertheless, this was the best data available, and we assume that it is sufficient for our model comparison purposes.

Landslides were inventoried by Pack from 1994, 1:15,000 scale black and white photographs. Landslides associated with talus were disregarded in this inventory. This left a reminder of 696 landslides within the Tree Farm License boundary of Canfor. Of these 611 were categorized as "debris slides." In this data "debris slides" that comprise the majority of the landslides are shallow translatory landslides composed of mixture of coarse and fine grained soils.

The term "shallow" in shallow-landslide, is relative and in a general sense refers to translational landslides with a maximum depth up to 2 m. Landslides in the Chetwynd study region were reported (Pack, personal communication, 2004) to be less than 2 m in depth and can thus be classified as shallow landslides.

The landslides were also divided into three size classes:

a) Small: less than 15 m wide by less than 75 m long.

b) Medium: between the small and large size.

c) Large: wider than 30 m or longer than 75 m.

According to this classification, 397 landslides are "small," 230 are "medium," and 69 are "large."

Manual methods were employed to digitize the landslides located on the aerial photographs as orthophotos were not available for this area at the time of study. The, landslide locations were carefully plotted using TRIM contour form from the 1:20000 TRIM map sheets as a guide. At the time of transfer, it was noted that the TRIM data

frequently failed to detect small but critical gully walls or terrace faces where landslides commonly originate.

CHAPTER 4

METHODOLOGY

This chapter develops the discriminatory measure Q that quantifies the effectiveness of a terrain stability map to efficiently discriminate between stable and unstable terrain when compared to locations where landslides were observed. Where a terrain stability map is based on a model with adjustable parameters, the measure serves as a goodness of fit measure that may be used to optimize the calibration of parameters. Recognizing equifinality (Beven and Binley, 1992) and the non-uniqueness of parameters that result in equal model performance, this measure is also used to define behavioral parameter sets that are equally likely simulators of the physical system.

The Q measure is developed both to compare parameter sets for a single model and to compare differences between models that may use different indices to quantify stability.

The assumption that underlies this methodology is that the spatial propensity for landsliding is related to variables that quantify the topographic setting relevant to landslide potential and that future landslides are more probable in topographic settings similar to where past landslides have been observed. Terrain stability mapping involves assessment of the propensity for landsliding based upon observations of landslide occurrence in similar topographic locations.

<u>A Measure to Quantify the Discriminating</u>
<u>Capability of a Terrain Stability Map</u>

Some notation is needed to develop the measure of discriminating capability. Let x be an index that is predictive of slope stability. At its simplest, x could be the slope (the higher the slope, the lower the stability). In other cases x could be the stability index SI from SINMAP, or the critical destabilizing rainfall obtained from SHALSTAB, or the return period of critical rainfall with dynamic wetness, effectively accounting for the partial contributing area as developed by Borga, Fontana, and Cazorzi (2002). x is calculated at each point to provide a map of terrain stability The goal is to quantify the predictive capability of the complete map of x, given observations of landslide initiation locations. In general, the map of index x will also depend on the specific parameters of the model and can be represented as $x(\tau, \beta)$, where τ is a vector of terrain variables, specifically slope, θ, and specific catchment area, a, and β is a vector of model parameters. Model parameters consist of cohesion, C, friction angle, ϕ and also, in the case of SINMAP, the ratio of transmissivity to steady state recharge T/R. SINMAP takes as input upper and lower (max and min) bounds for these variables (Table 4-1).

Table 4-1. Terrain variables and model parameters for the models compared

Model	Terrain variables, τ	Model parameters, β
SINMAP	θ, a	C_{min}, C_{max}, T/R_{min}, T/R_{max}, ϕ_{min}, ϕ_{max}
SHALSTAB	θ, a	C, ϕ
Slope	θ	-

The goal is to define a measure that quantifies the capability of a map of x to discriminate unstable regions with respect to the observations of landslides. We require that x be monotonic, that is, greater or lesser x is predictive of greater or lesser likelihood of landslide initiation. Without loss of generality, we assume that decreasing x is predictive of more landslides (in the opposite increasing case, a measure –x or 1/x can be used to transform to the decreasing case). Both SHALSTAB and SINMAP quantify terrain instability with measures that decrease as terrain becomes unstable. In the case of SHALSTAB, x is R/T, and in the case of SINMAP, x is the stability index SI based on the probability of stability and factor of safety.

Conceptually, we want x values to be low at locations where landslides are observed to initiate, and high at other locations. If this is the case, then x has the capability to discriminate landslide locations. We will define a measure Q that quantifies this concept. The data used in this analysis include:

a) A grid map where the index x has been computed for each grid cell in the given domain.

b) A set of points within the domain where landslide initiations have been observed. Each landslide point falls within one particular grid cell in the domain, and is associated with the stability index for that grid cell.

If x is a good measure for discriminating terrain instability, we might have all the landslides occur in a small fraction, f, of the terrain for which $x < x_T$, where, x_T is some threshold value in x. Another measure that has all landslides occur in even a smaller fraction (< f) of the terrain is, in a discriminatory sense, better. We do not, however, expect our models to ever be perfect because they are descriptors of topographic settings

that quantify a probability for landslide occurrence and in any one survey only a fraction of high landslide probability locations will have actual occurrences. The fraction of the landslides F_O that occur within a fraction of the terrain F_R for a specific threshold value of x denoted x_T, quantify the discriminating capability of the index x at threshold x_T.

For example, lets consider two models A and B, which have for some threshold value x_T, 50% of the landslides occurring in 20% and 50% of the terrain, respectively (Figure 4-1a). Comparatively, Model A, when compared with the observed landslide fraction, has smaller fraction of the terrain categorized as unstable and hence, is a better discriminator. Another way of looking at the same trend is, for a model to have a higher fraction of the landslides for the same fraction of the terrain (Figure 4-1b). In both cases, Model A is performing better than Model B. It should be noted that in such a comparison, the stability index x and the threshold value, x_T can be different for these two models.
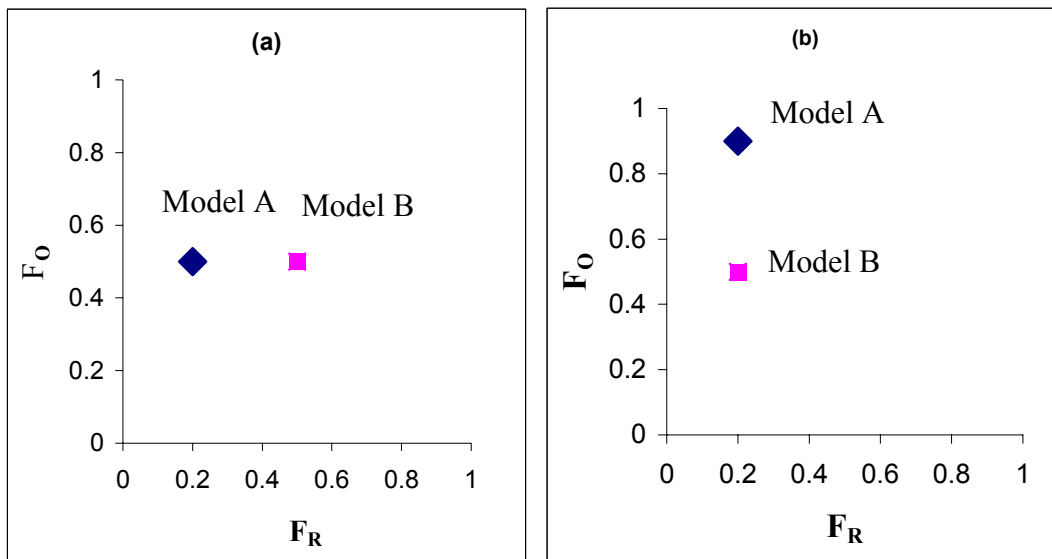
Figure 4-1. Example of F_O versus F_R . a) Model A and B discriminate the same fraction of landslides F_O, but in model A this occurs in a smaller fraction of terrain F_R. b) Model A and B discriminate the same fraction of terrain F_R, but the terrain discriminated by A contain a higher fraction of landslides F_O.

To demonstrate the nature of Figure 4-1, let us consider an example of a simple index of stability that has a value of either 1 or 0 with 0 indicating instability. In a more general case the stability measure can assume continuous values. Figure 4-2 shows two such predictions, (a) and (b), where red squares shows the observed landslide area, zeros on red squares depicts that the model has correctly predicted the landslide and ones on red region the incorrect predictions. Zero on a green area depicts a location where the model predicted instability but a landslide was not observed and one on green are the areas predicted by the model as stable and there were no landslides observed.

The two sets R and O are defined as

O=$\{x_i\}$, a set of x values for the observed landslides. This contains all the values in the red region. i.e. O_A=$\{0,0\}$ and O_B=$\{0,1\}$

R=$\{x_j\}$, a set of x values for the points representative of the terrain. For this example R includes all the points in the region shown in the Figure 4-2, i.e. R_A= R_B= $\{0,0,1,1,1,1,1,1,1,1\}$.
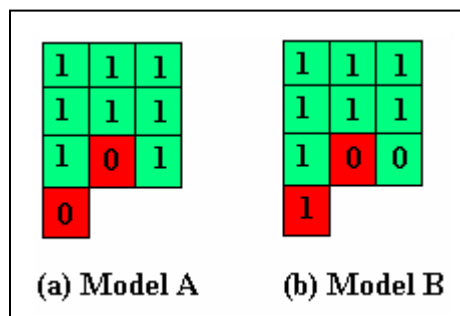
Figure 4-2. Example demonstrating the fractions F_O and F_R.

It can be seen that the prediction shown by model A (Figure 4-2(a)) is better as it is correctly predicting all the landslides. The basis for defining the measure Q is to represent this effect in a general way independent of specific x_T threshold values.

Define the ratios F_O and F_R for this set. For model A,

$$F_R = \frac{\text{Number of zeros in the set}}{\text{Number of zeros and ones in the set}} = \frac{2}{10} = 0.2 \tag{4.1}$$

$$F_O = \frac{\text{Number of zeros in the set O}}{\text{Number of zeros and ones in the set}} = \frac{2}{2} = 1 \tag{4.2}$$

Similarly, for model B, $F_R = 0.2$ and $F_O = 0.5$. Figure 4-1 b was actually the graphical representation of this example and shows the better performance of Model A at discriminating landslide locations.

For a continuous stability index x, F_O and F_R depend upon a threshold value of x. The functions, $F_O(x)$ and $F_R(x)$ are cumulative probability distribution functions for points selected at random from the observed landslide locations and anywhere on the terrain respectively. For a particular threshold x_T, if $F_O(x_T)$ is large for $F_R(x_T)$ small, the model is a good discriminator. This cumulative distribution comparison is a non-parametric approach that is independent of x, because one can compare different indices at the same value of F_O or F_R notwithstanding that underlying x values may be different between indices.

Another way to quantify the discriminating capability of index x is in terms of landslide density. If there are a total of N landslides in an area A and fraction F_O of them

is contained in a fraction of the area F_R demarkated by a threshold x_T, then the landslides density associated with that threshold is

$$\rho_{LS}(x_T) = \frac{F_O(x_T)N}{F_R(x_T)A}$$
(4.3)

In general, as x_T increases and more area is discriminated we should expect landslides density to decrease. A model that has a higher landslide density for the same fraction of the terrain demarcated (i.e. the demarcated terrain contains more of the observed landslides) is better. Equivalently a model that has higher landslide density for the same fraction of landslides contained in the terrain demarcated (i.e. the demarcated terrain containing these observed landslides is smaller) is better because it has been more efficient in discriminating unstable terrain.

Landslide density can also be used to estimate spatial probability in a time period consistent with the time period for which landslides were observed. This requires an estimate of the size (spatial extent) of landslide initiation area. The spatial probability of landslide occurrence associated with the threshold x_T is ($\rho_{LS}(x_T).y$), where y is the area of the landslide initiation zone. This spatial probability represents the probability that a location with index value less than or equal to x_T will be a landslide origination point in time period consistent with the period over which landslides were observed under the assumption that the probability of landslides is equal for points with equal x. There is a danger that if landslides are mapped as points, the analysis could be done at a scale smaller than y and the analysis could result in densities large enough so that spatial

probability can be greater than 1. This would be symptomatic of the scale of the analysis being reduced beyond what is physically reasonable.

The basis for defining the measure Q is to represent the integral discriminatory ability over a range of x_T threshold values so that a measure independent of specific x_T threshold values is obtained. In the example given earlier, the (F_O, F_R) pairs corresponded to a specific threshold value of each model. If more such pairs are plotted for different threshold values, the plot may look like Figure 4-3. In this example, because for each particular $F_R(x)$ the fraction of observed landslides $F_O(x)$ from model A is greater than model B, model A is a better discriminator.

In general, models may behave differently at different thresholds, i.e. Model B may perform better than A for some threshold values (Figure 4-4). In such cases, we
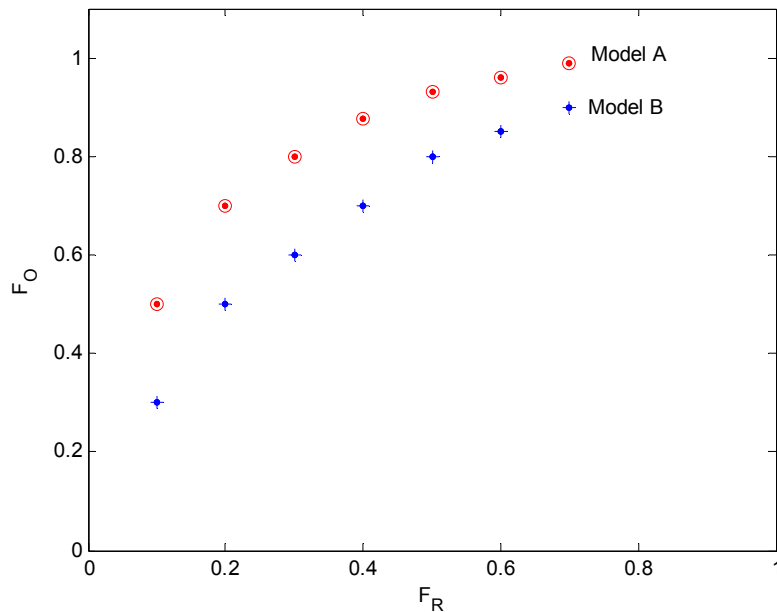
Figure 4-3. Comparison over different thresholds.

cannot readily say which model is performing better by merely looking at the plots. We are interested in which model performs best in an aggregate sense for range of x values. We therefore suggest that the integral under this graph be used as a goodness of fit measure to quantify the discriminatory power of a map of x. Our measure is defined as

$$Q = \int_0^1 F_O \, .dF_R \qquad (4.4)$$
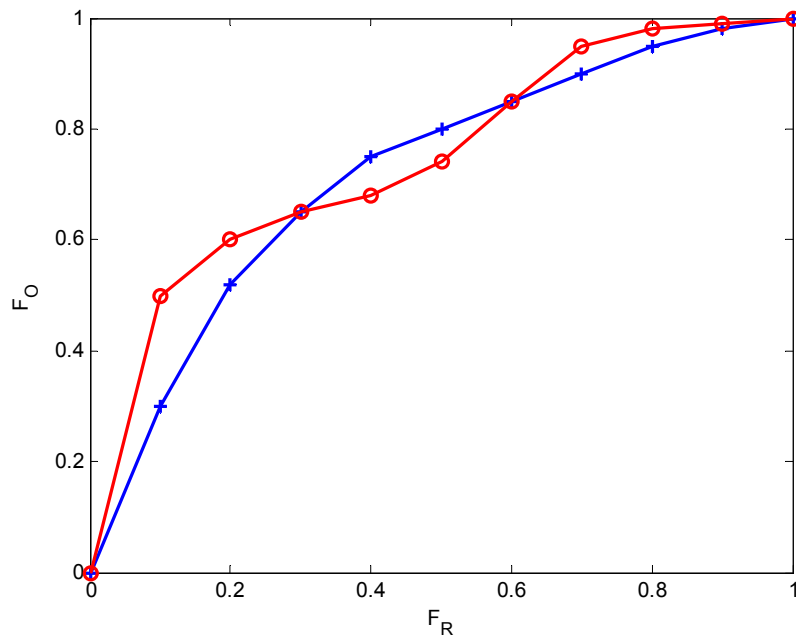
Figure 4-4. The graph of $F_O(x)$ and $F_R(x)$ for two models A and B.

The Numerical Evaluation of Q

The numerical evaluation of Q will be performed using two data sets:

a) Observed landslide points. The set of stability indices x_j evaluated at each observed landslide location point with $j = 1 \ldots n$, where n is the number of observed landslide initiation points

b) Representative terrain points. The set of stability indices x_i with $i = 1 \ldots m$, where m is the number of representative points from the domain, either sampled at random, or the entire domain (if computationally feasible, i.e. for small domain).

The subscripts i and j will be used to distinguish representative terrain points and observed landslide points. It is important to note that x_i and x_j are evaluated for one specific parameter set or combination of model parameters. The Q value so computed reflects the predictive capability of the model and specific parameter set. Q thus becomes a goodness of fit measure for the model and may be used for parameter estimation. More than one set may have the same Q value, implying non-uniqueness or equifinality in the model.

$F_O(x)$ and $F_R(x)$ are evaluated numerically from the sets $\{x_i\}$ and $\{x_j\}$, as follows,

a) Rank the data x_i and x_j from smallest to largest, such that we have the representative terrain points, x_i: $x_1 < x_2 < \ldots < x_m$ and the observed land slide points set: x_j: $x_1 < x_2 < \ldots < x_n$.

b) Pick the smallest number of m or n. Most commonly this will be n. If the smallest number is n, then the empirical cumulative probability is estimated using

a plotting position, $F_O(x_j)=j/(n+1)$. Each value of $F_O(x_j)$ gives a fraction of observed landslides for a threshold value of x_j.

c) Once $F_O(x)$ is defined, the next step is to compute the probability distribution function $F_R(x)$. This is done for each observed landslide point x_j, by interpolating between the plotting position estimates of $F_R(x_i)$ as follows:

(a) For each observed landslide point x_j, find the subscript value i in the ranked list of representative terrain points x_i such that, $x_i < x_j \leq x_{i+1}$.

(b) Interpolate $f = (x_j-x_i)/(x_{i+1}-x_i)$. In other words, f gives the position of x_j between x_i and x_{i+1}.

(c) Define, $ir=i+f$ with the following exceptions:

If $x_j \leq$ all x_i, set $ir=1$; i.e. the x_j being the lowest of all x_i will occupy the first rank.

If $x_j >$ all x_i, set $ir=m$; i.e. the x_j being the highest of all x_i will be ranked last.

Here, ir gives the interpolated position of x_j in the x_i set.

(d) Estimate $F_R(x_j)$ is as,

$$F_R(x_j) = \frac{ir}{(m+1)} \tag{4.5}$$

$F_R(x_j)$ gives the fraction of terrain points that are predicted to be unstable by the model for a threshold value of x_j.

Once $F_O(x_j)$ and $F_R(x_j)$ are determined, the area under the plot of $F_O(x_j)$ versus $F_R(x_j)$, Q is calculated by recognizing that y axis values $F_O(x_j)$ are evenly distributed with

spacing $\left(\dfrac{1}{n+1}\right)$ between points. The trapezoidal rule approximation of the integral in (5) is therefore

$$Q = 1 - \left(\frac{1}{n+1} \sum_{j=1}^{n} F_R(x_j) + \frac{1}{2(n+1)} \right) \qquad (4.6)$$

This sum has extended the integral to the end points (0,0) and (1,1). In the rare case in the above procedure when the smallest number is m, rather than n, F_R and F_O switch places along with their corresponding variables and the value of Q is calculated from.

$$Q = \frac{1}{m+1} \sum_{i=1}^{m} F_O(x_i) + \frac{1}{2(m+1)} \qquad (4.7)$$

Q as estimated above provides a measure of the predictive capability of the model.

GLUE Methodology Adaptation

Figure 4-5 gives a flowchart of the procedure that was followed. Ranges for the model parameters that define a feasible parameter space are input (Figure 4-5) to the Monte-Carlo simulation. The Monte-Carlo simulation then generates a large number of parameter sets, β_k from the feasible parameter space. The number of sets, k, to be generated is also input.

The parameters of SHALSTAB are the dimensionless cohesion, C and the friction angle, ϕ, i.e $\beta_k = (C, \phi)$. The inputs to SINMAP are specified as ranges and hence the upper and lower bounds of user input variables T/R, ϕ and C serve as the parameter set for the model, i.e $\beta_k = (C_{min}, C_{max}, \phi_{min}, \phi_{max}, T/R_{min}, T/R_{max})$. Minimum and maximum values T/R, ϕ and C will be generated, with the condition that $C_{max} > C_{min}$, $\phi_{max} > \phi_{min}$ and $T/R_{max} > T/R_{min}$. Recognizing that SINMAP parameters already quantify uncertainty in terms of uniform distributions, this procedure to quantify the uncertainty in distribution bounds is a second order quantification of the uncertainty.

The terrain stability model (SHALSTAB or SINMAP) is evaluated for each of the above parameter sets (i.e k times) at each observed landslide initiation point, $(\theta, a)_i$ and at each terrain point quantified by slope and specific catchment area, $(\theta, a)_j$. i ranges from 1 to m, the number of observed landslides, and, j ranges from 1 to n, the number of terrain points sampled. Terrain is sampled to reduce the computational burden. Wherever it is feasible the entire terrain should be considered. In this case, m is equal to the total number of terrain points.

Following model evaluation, stability index sets x_i and x_j for each of the parameter sets β_k are obtained. The objective function, Q is computed for each parameter set and parameter sets are ranked according to the Q value. Here Q is used as a goodness of fit, or likelihood measure in the GLUE procedure to determine the parameter sets for which the model is a good discriminator, or in a relative sense a better discriminator than other models or parameter sets.

On completion of the procedure, lists of Q, one for each model are obtained. Those parameter sets with Q close to the maximum Q obtained (say all parameter sets where Q is within 5% of its maximum value) are termed behavioral and can be used to quantify the uncertainty associated with model predictions. The differences (in terms of discriminatory power) between different models are assessed in terms of the maximum Q value from each model.

In following this procedure the assumption is that the sampling has been sufficient to cover the parameter space, recognizing the curse of dimensionality as the number of parameters increase. We assess sufficiency of the number of parameter sets generated by checking stability of the maximum Q and uncertainty bounds to repeats of the procedure.

The behavioral and the non-behavioral parameter sets from the GLUE methodology can also be used that assess the relative sensitivity of the parameter on the stability index. If a parameter is sensitive, then its probability distribution in the behavioral set should be different from that of the non-behavioral set. This Generalized Sensitivity Analysis (GSA; Spear and Hornberger, 1980) is useful for evaluating the sensitivity of multi parameter models and can easily be combined with GLUE methodology.

To assess the statistical significance of the models with different Q, we use a Chi-square test. This test evaluates the difference between the fractions of the observed landslides captured at a stability index threshold corresponding to a specific fraction of the terrain mapped as unstable by each model. For a fraction of terrain $F_R(x_T)$ mapped as unstable by each model, terrain stability models may have a different proportion of
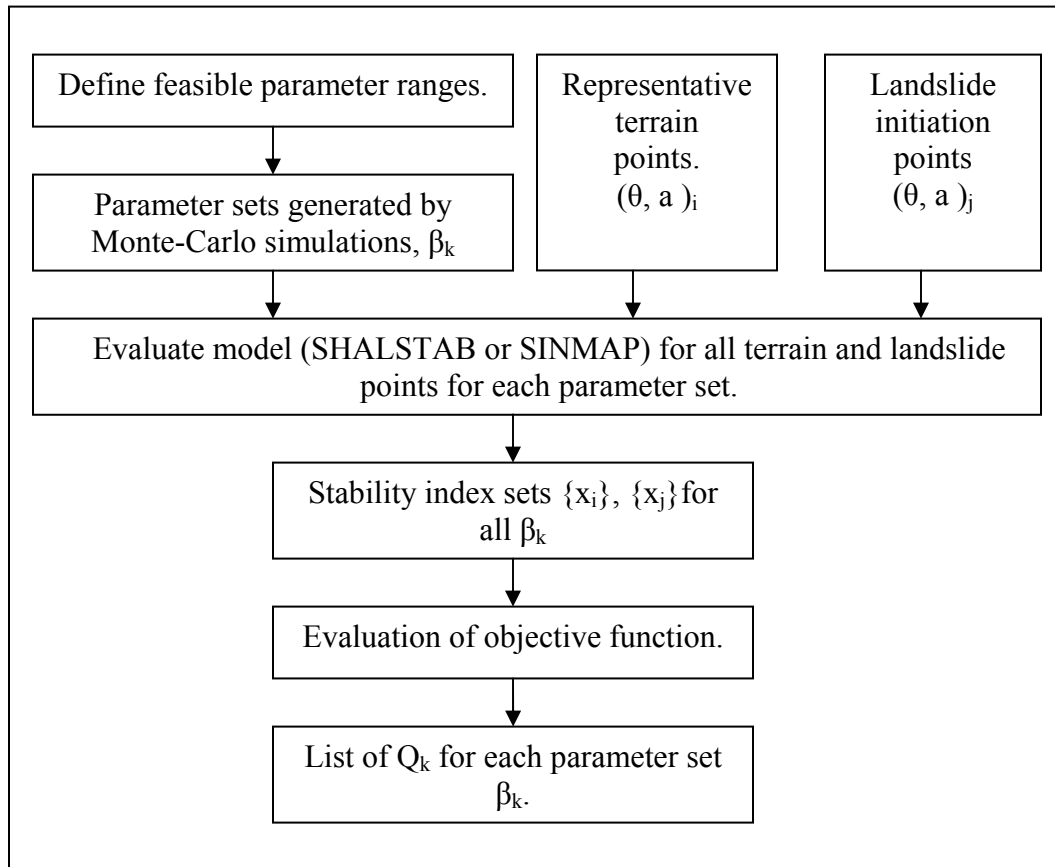
Figure 4-5. Flow structure of the GLUE methodology.

landslides contained within the unstable terrain, $F_R(x_T)$. x_T will generally be different for different models.

The chi-square statistic measures the discrepancy/difference between the observed and the expected frequencies of a random variable being tested. In our case, we are trying to check for the differences between the numbers of landslides observed in a region mapped as unstable by two different models. We assume as a null hypothesis the number provided by one of the models and evaluate whether the number obtained from the other model is significantly different, or could be obtained by chance.

We compare two models and select a threshold for each such that the fraction of terrain F_R mapped as unstable by each model is the same. We then have

e_1 = number of observed landslides in the reference model's unstable area

e_2 = number of observed landslides in the reference model's stable area

o_1 = number of observed landslides in the other model's unstable area

o_2 = number of observed landslides in the other model's stable area,

Then the chi-square statistic is given by

$$\chi^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2}$$ (4.8)

such that

$$\sum o_j = \sum e_j = N$$ (4.9)

where

N = Total number of landslides

χ^2 can be checked for the level of significance using the standard χ^2 distribution.

A split sample test was used to evaluate the effectiveness of optimized model parameters from the GLUE methodology in comparison to data not used in the parameter estimation. The study area was divided into two regions and parameters were optimized using the data set from one region through the GLUE procedure. The optimized parameters were then used to develop stability index maps for the other region. The performance of these stability maps in the other region was quantified using the Q measure.

CHAPTER 5

DATA ANALYSIS AND RESULTS

Terrain stability maps (Figure 5-1) that predict landslide occurrence were created for the study area using SINMAP and SHALSTAB, with their default parameter values (Table 5-1). Though most of the steeper region (see hillshading in Figure 3-1 and contours on the expanded portions of Figure 5-1) is predicted to be unstable by both the models, the models differ in how they categorize the areas of moderate to gentle slope. To illustrate this difference between SHALSTAB and SINMAP in differentiating stable zones from unstable ones, Figure 5-1 shows expanded detail for a small portion of the study area where the red regions are more unstable than the green. Both of these models use a continuous stability index scale to characterize the degree of instability. Thresholds within this continuous scale may be used to classify the domain into zones with different stability categories (Dietrich et al., 1998; Pack, Tarboton, and Goodwin, 1998) as has effectively occurred in Figure 5-1 with the selection of color schemes for depicting the continuous stability index surface.

SHALSTAB uses the ratio R/T as the index to classify stability classes. Following Dietrich et al. (1998) and Dietrich, Bellugi, and de Asua (2001), a log (base 10) transformation of R/T was used to facilitate interpretation. This results in negative log(R/T) values instead of the very small actual R/T values. As an example of the categorization of the stability indices mapped in Figure 5-1, a threshold of log(R/T) equal to −2.8 is used for SHALSTAB (red and its darker shades) to map unstable terrain. The
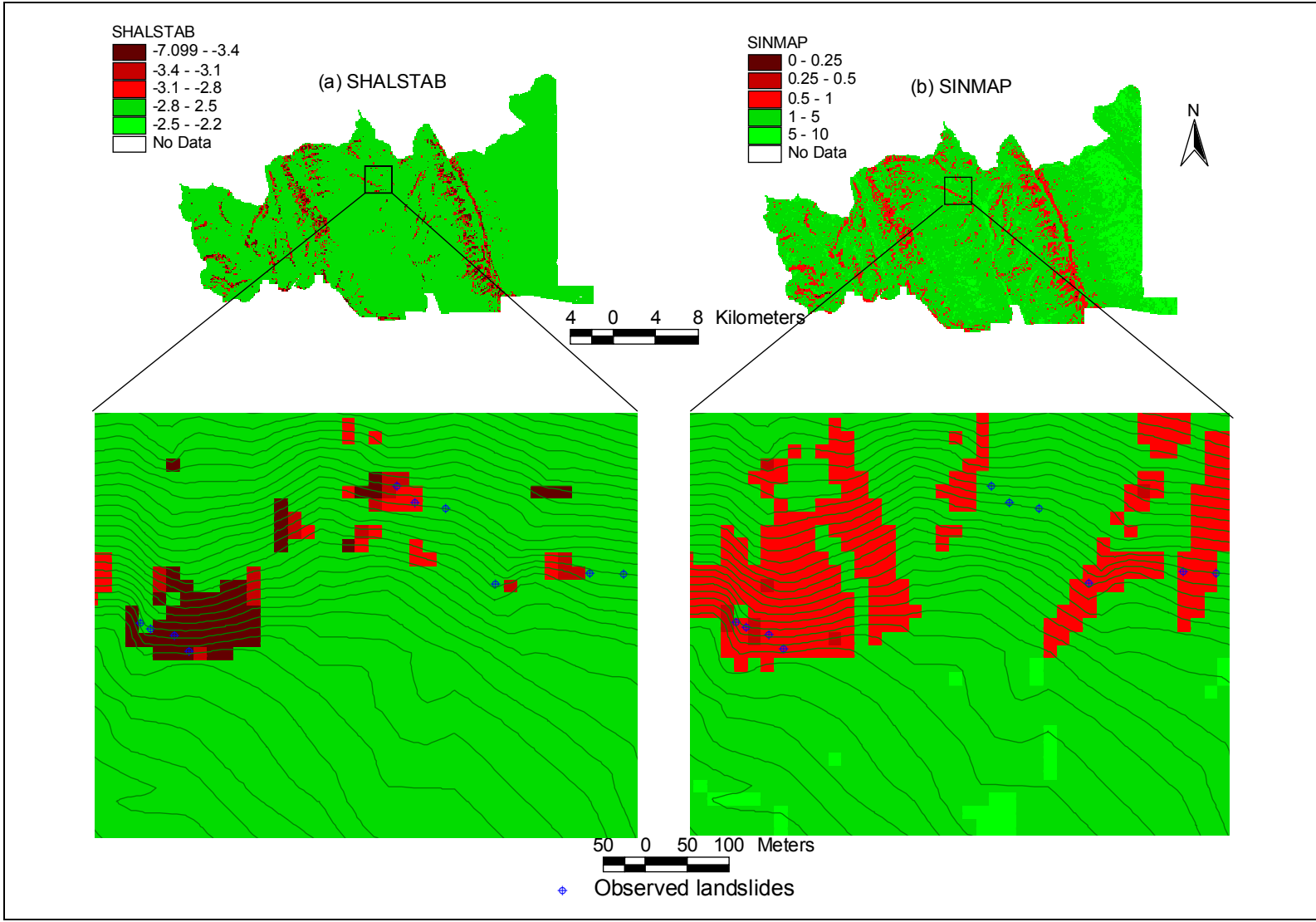
Figure 5-1. Terrain stability maps using SHALSTAB and SINMAP.

threshold of -2.8 was chosen for illustration because it has been used by Dietrich, Bellugi, and de Asua (2001) as a terrain stability threshold. SINMAP uses a stability index of one as a threshold (red and its darker shades) to map unstable terrain. The threshold of one in SINMAP delineates terrain where there is a finite probability of it being unstable according to the model.

Figure 5-1 also depicts the observed landslide locations. Within the expanded area SHALSTAB captures four of the ten observed landslides under its unstable category (log(R/T) less than -2.8) while SINMAP captures six of the observed landslides within its unstable category (SI less than 1) but categorizes more area as unstable. In an ideal case, each of the models should have identified all the landslides under their respective unstable categories which should only cover a small fraction of the domain. Note that the models cannot be easily compared because they have different thresholds categorizing stability zones and different unstable area size.

Table 5-1. Default parameter values for SHALSTAB[1] and SINMAP[2]

Model	Parameters	Default values
SHALSTAB	C	0.0
	Φ (degree)	45^0
SINMAP	$T/R_{max}[m]$	3000
	$T/R_{min}[m]$	2000
	C_{max}	0.25
	C_{min}	0.0
	Φ_{max} (degree)	45^0
	Φ_{min} (degree)	30^0

1. From Montgomery and Dietrich, 1994
2. Pack et al., 1998

For the whole region, SHALSTAB maps about 8% of the area as unstable with a threshold of –2.8. This area contains 395 observed landslides (57% of the landslides). A threshold of –2.5 contains 433 observed landslides (62% of the landslides) and 9% of the terrain. SINMAP, with a threshold stability index value equal to 1 maps a larger region, 11.5%, of the total area as unstable. This area contains 507 (73%) observed landslides.

If an inexperienced person has to choose between SHALSTAB and SINMAP, given the above knowledge and the region, he or she may choose SINMAP because the mapped unstable area contains more observed landslides. However, this does not consider the fact that SINMAP in identifying more landslides has categorized a larger area as unstable. In other words, he or she may still have not chosen the best model, because the information given is not sufficient to differentiate models based on their performance.

There is some subjectivity involved in choosing thresholds. Dietrich et al. (1998), use a threshold of -3.1 to categorize areas as high hazard zones in their studies conducted in California Coastal Ranges of coastal Mendocino and Humboldt counties. A threshold of -2.5 was not supported by Dietrich et al. (1998) as it was only as good as the random model used as a point of reference (see section 2.4). In SINMAP, a SI threshold of 1 is consistent with the idea of mapped locations having a finite (greater than 0) probability of instability. A threshold of 0.5 is consistent with mapped locations having a higher than 50% probability of instability. These probabilities are taken over the uniform distributions for the parameters. The point is that these models have different discriminating powers at different thresholds and hence the above example does not have sufficient information to select a better model.

Landslide Density

One way to interpret terrain stability maps is through estimates of landslide density. Landslide density is a common quantity used in landslide hazard mapping, expressed as number of landslides per unit area. We can associate a landslide density given by Equation 4.3 with each value of x for each model.

The landslide density, $\rho_{LS(x)}$, quantifies in a non-parametric way the number of landslides per unit area in the area associated with stability index values less than x. Computing the landslide densities for SHALSTAB and SINMAP, using the thresholds mentioned in the above paragraph, the $\rho_{LS}(-2.8)$ is equal to 9.9 LS/km^{2*} for SHALSTAB and $\rho_{LS}(1)$ is 9.3 LS/km^2 for SINMAP. This might be interpreted to show SHALSTAB to be performing better than SINMAP, because it achieves a higher landslide density, so is more discriminating. This comparison however is only good when both the models are capturing the same number of landslides or apply to the same fraction of the terrain. As mentioned earlier, SHALSTAB captures relatively fewer landslides (395) compared to SINMAP (507) and hence we cannot really say that SHALSTAB is performing better than SINMAP from this information.

Dietrich et al. (1998) use the plot of landslide density versus the stability index. log(R/T), to demonstrate the better performance of SHALSTAB over a random model (see section 2.4 for the definition of random model). Such a comparison was possible because the random model used by Dietrich et al., (1998) has the same stability index,

---

[*] Landslides per square kilometer.

log(R/T). A plot of landslide density versus x is informative for a particular model, but not amenable to comparing different models.

Figure 5-2 shows the landslide density plots for a) SHALSTAB and b) SINMAP. In a landslide density versus threshold plot, undulations occur where, due to sampling effects, a small change in threshold captures a large increment of landslides, for only a small increase in area. The curve reflects the behavior of the Equation 5.1(equation is introduced below), which can increase or decrease depending on the fraction of landslides identified, and the fraction of the total area mapped as unstable.

The cumulative distribution curves $F_O(x)$ and $F_R(x)$ for default parameter values are plotted on the landslide density graph to illustrate the role of thresholds. From this plot, one can determine the fraction of landslides and the fraction of the entire terrain associated with any threshold. Then the threshold value that best discriminates the unstable region is the one that gives the maximum difference between the landslide and the area cumulative curves. In other words, if one is using a specific parameter set to map stable and unstable zones using just one threshold, then the best discrimination is obtained by using the threshold that gives the maximum difference between the cumulative distributions of $F_O(x)$ and $F_R(x)$. The knowledge of the threshold with maximum discrimination is helpful in setting up bounds for the unstable category in the terrain stability map.
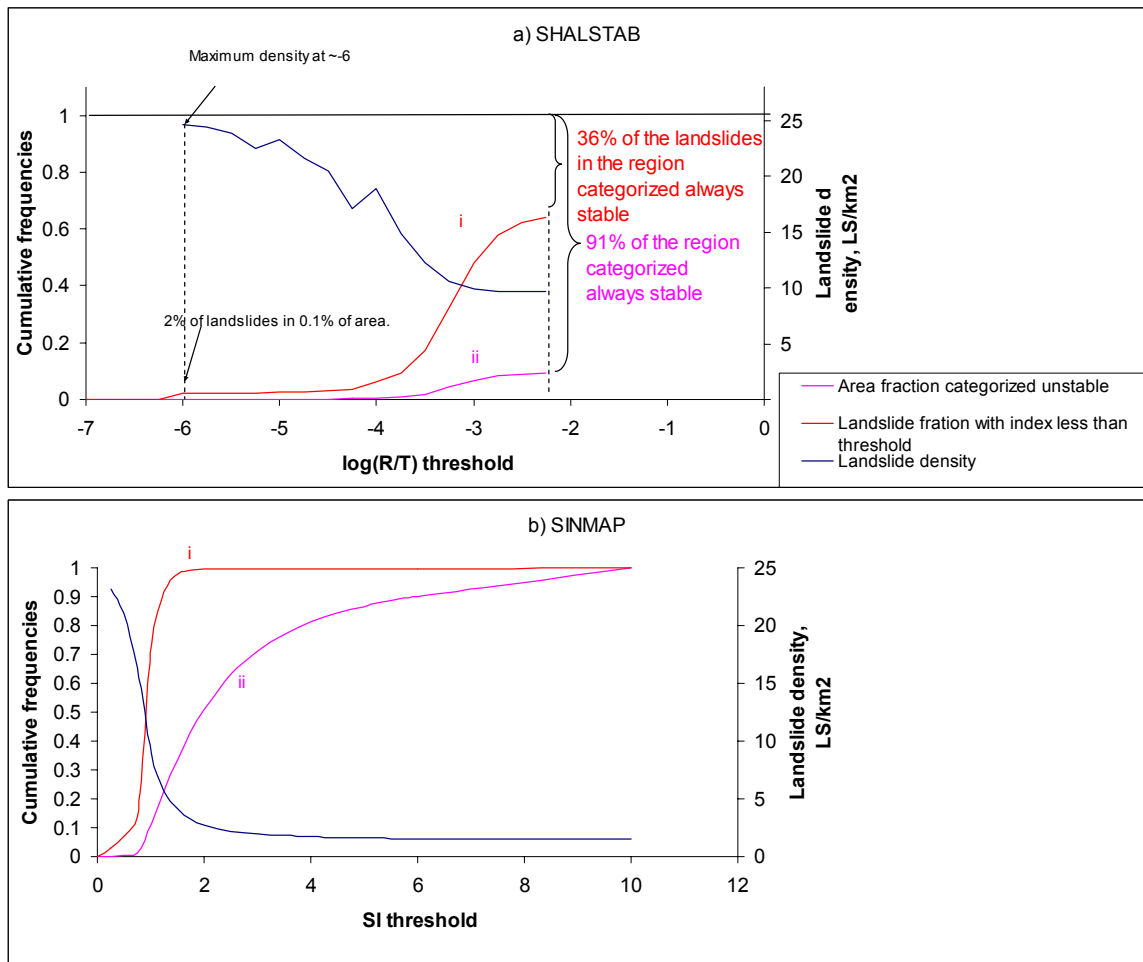
Figure 5-2. Landslide density plots for (a) SHALSTAB and (b) SINMAP.

SHALSTAB with its default parameter set can for this region identify at most 64% of the landslides in 9% of the area corresponding to a threshold of -2.2. Larger values of this threshold categorize more area as unstable but never capture any more observed landslides because 36% of the landslides are at unconditionally stable locations according to SHALSTAB with its default parameters (C=0, Φ=45°). Similarly, SINMAP can contain approximately 92% of the observed landslides at a threshold of 1.25, with

20% of the total area mapped as unstable. This suggests a better performance of SINMAP in identifying landslides (when default parameter sets are used).

For cohesionless soil, SHALSTAB simplifies the infinite-slope stability model (Equation 2.8) and expresses it using the following equation.

$$FS = \frac{\left[1 - w\left(\frac{\rho_w}{\rho_s}\right)\right]\tan\phi}{\tan\theta} \tag{5.1}$$

SHALSTAB equates the FS to one, and evaluates the critical rainfall (R_{cr}) from Equation 2.9. Setting FS equal to one, one can write Equation 5.1 as

$$\tan\theta = [1 - wr]\tan\phi \tag{5.2}$$

The above equation suggests that, if $\tan\theta \le [1 - wr]\tan\phi$, the slope is stable. Further w can not exceed saturation represented by w=1, so the condition for the unconditional stability is $\tan\theta \le [1 - r]\tan\phi$. For default parameter values of SHALSTAB (C=0, Φ=45°), slopes less than 26.6° are unconditionally stable. There are 245(~36%) observed landslides in the dataset that occur on slopes less than this value and hence SHALSTAB is not able to contain them within the unstable region with the default parameter set. If cohesion is included, without altering the friction angle, this number will be higher, because cohesion increases stability and the slope angle required to cause instability will be increased.

<u>Spatial Landslide Probability</u>

One goal of terrain stability mapping is to produce a spatial probability map of landslide hazard in the study area. The landslide density discussed above can be related to a spatial probability of landsliding within a time period associated with the occurrences of the observed landslides as was discussed in Chapter 4.

For the Chetwynd study area, the size of each landslide is taken as a single DEM grid cell, equal to $225m^2$(15m x 15m pixel), and we have 696 observed landslides in $497km^2$. From this, the spatial probability for the whole region would be equal to 0.00032. The size of the entire landslide scar can be bigger than $225m^2$, but we are considering only the landslide initiation which is here assumed to correspond to one grid cell. Landslide density can be assigned to each grid cell and hence we can compute the spatial probability for each grid cell. Such a procedure gives the spatial probability maps that could assist in land management. Figure 5-3 is the same as Figure 5-2, but one of the y-axes has been changed to the spatial probability, to show the conversion between landslide density and spatial probability.

Figure 5-4 depicts the spatial probability maps for a) SHALSTAB and b) SINMAP, created using the conversion from the Figure 5-3. These maps were created for default values of the parameters (Table 5-2). The expanded region shown is the same as the one shown in Figure 5-1. The color shading on a spatial probability map is essentially the ratio of the number of landslides that occurred in terrain with the same or more unstable stability index, relative to the number of grid cells in the area with the same or more unstable stability index. As such, a value, for example, of 0.005 might

represent that 20 landslides occurred in an area comprising 40000 grid cells (9 km^2). The

probability of a landslide occurring in any one of these grid cells during the time interval

when the observed landslides were recorded is 20/40000 = 0.005. The majority of areas
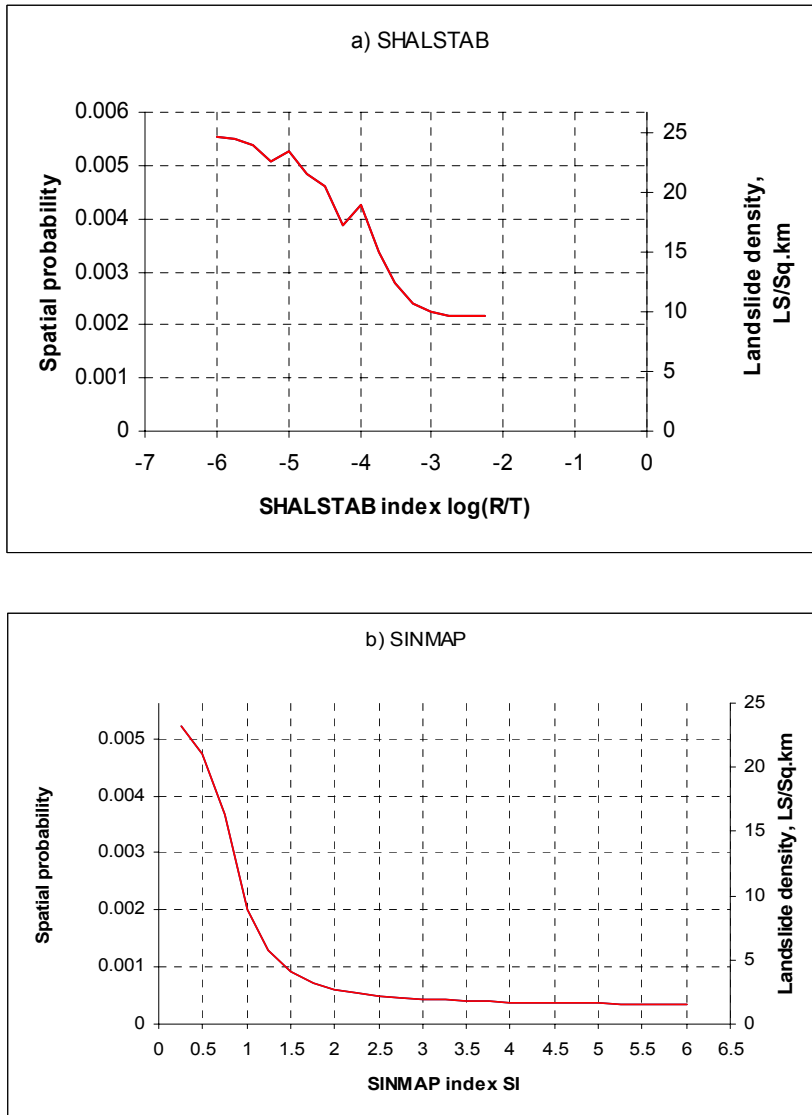

Figure 5-3. Spatial probability plots as a function of stability index threshold for (a)
SHALSTAB and (b) SINMAP with default parameter sets.

marked with high probability of landslides are in the region of steeper slopes showing the topographic influence on landslides.

Spatial probability maps are results of integration over all thresholds and could be a possible way of comparing terrain stability models. A spatial probability map also overcomes the difficulty of comparing models with different indices. Another way to integrate over all thresholds is to use the F_O versus F_R plot for a visual depiction and the integrated measure representing the area under the plot (Equation 4.4) to quantify the discriminatory power of the stability indices. Figure (5-5) shows such a plot for SHALSTAB and SINMAP for their default parameters. The Q value for SINMAP equal to 0.909 is greater than that of SHALSTAB which is equal to 0.775, suggesting that for their respective default values, SINMAP performs better than SHALSTAB. A random model where landslides occur at locations independent of the stability index would result in a Q value of 0.5 corresponding to a 1:1 line.

To understand if slope and specific catchment area could be used as an index, their F_O versus F_R was also plotted. The Q value of slope and specific catchment area are 0.905 and 0.48 respectively. This illustrates the strong influence of slope on landslides in this study area. In fact, slope by itself is almost as good as the other two models used in this study. A Q statistic of specific catchment area close to 0.5 indicates the poor performance of the specific catchment area by itself to identify landslides locations. The statistical significance of the difference between values of 0.909, 0.775 and 0.905 for SINMAP, SHALSTAB and slope will be evaluated later. The Q statistic is capturing the inability of SHALSTAB to identify all the landslides. As discussed earlier, SHALSTAB
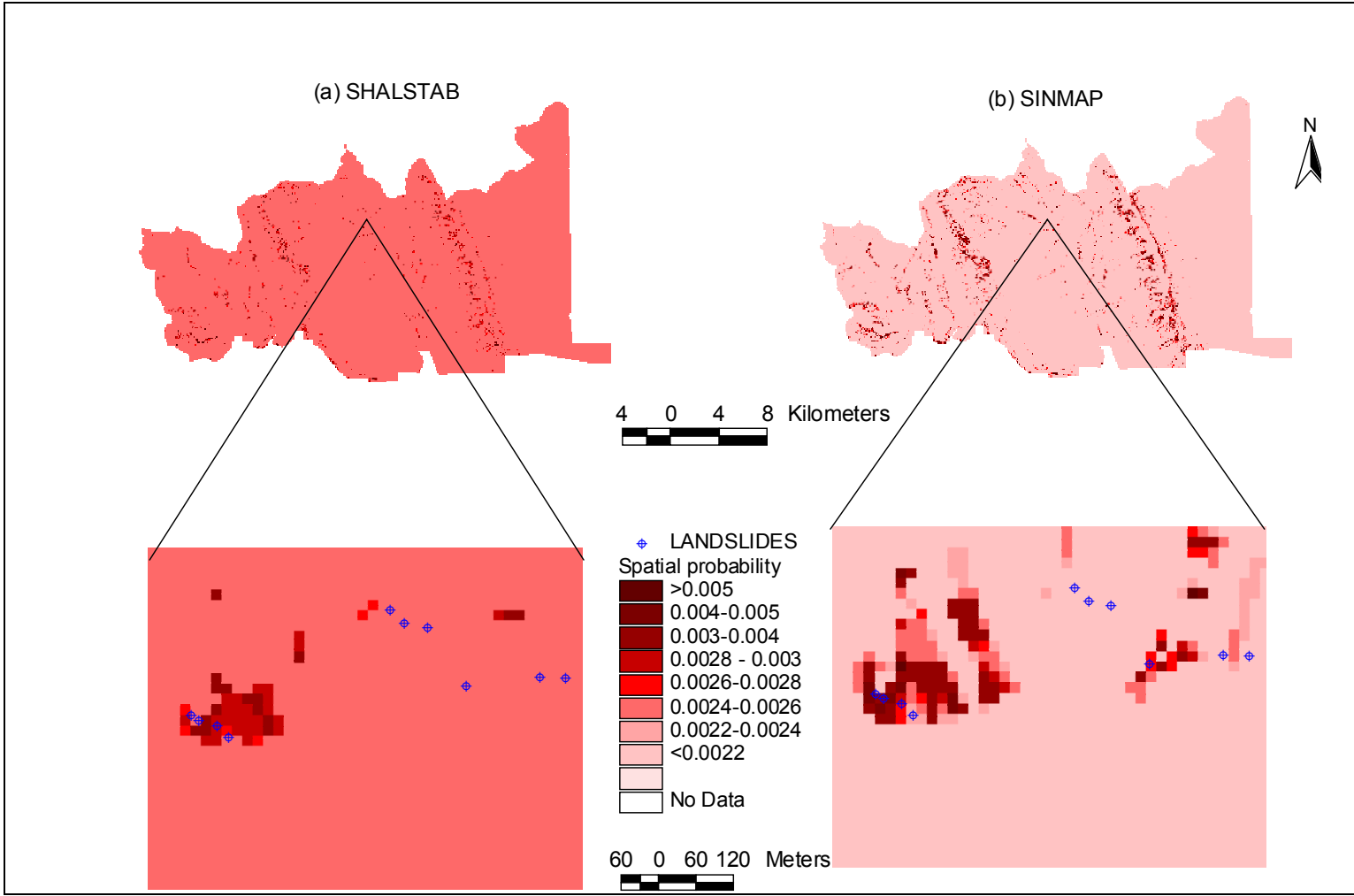
Figure 5-4. Spatial probability maps a) SHALSTAB b) SINMAP for their default parameters.

with the default parameter values was able to identify only 64% of the landslides. The low value for SHALSTAB is because the model is non-discriminating over the 91% of the terrain that is mapped as unconditionally stable, but still contains 36% of the landslides. Close examination of the curves (Figure 5-5) at the left indicate that SHALSTAB is less discriminating than slope initially, but as $F_R = 0.09$ is approached, is more discriminating than slope.

The actual numerical difference between SINMAP (Q=0.9099) and slope (0.9054) is very small (0.0045) but there is a visually distinguishable difference between the plots of SINMAP and slope in the steeper region of the terrain (Figure 5-5).
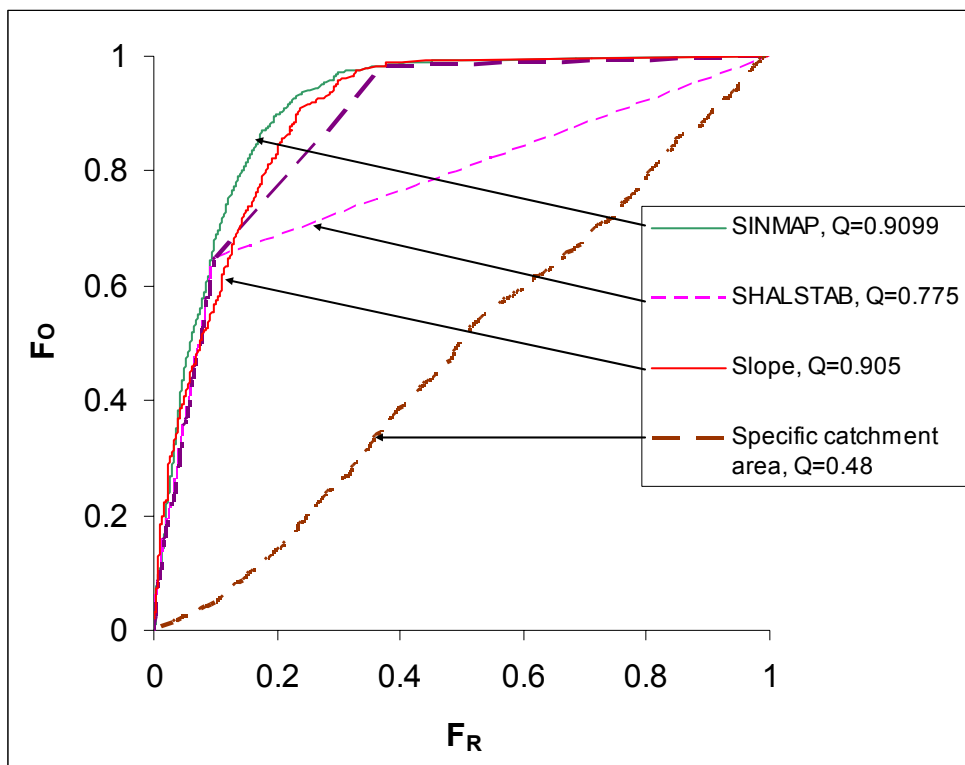

Figure 5-5. $F_O(x)$ versus $F_R(x)$ plot for different stability indicies.

<u>Implementation of GLUE Methodology</u>

F_O versus F_R plots; spatial probability maps and the Q statistic that integrate under the F_O versus F_R curve provide ways to objectively assess the discriminating capability of terrain stability indices. However, the terrain stability index maps produced by models, such as SHALSTAB and SINMAP depend on the input model parameters. Objective comparison therefore requires that the parameters be optimized. The modified GLUE random search methodology described in the previous chapter was used to find parameter sets that give the best discriminatory performance as measured by the Q statistic. These optimized parameter sets were used to compare the performance of different models.

According to the basic premises of GLUE, there could be more than one parameter set for which models may behave equally well in simulating the physical system (Equifinality). Such sets are termed as behavioral sets of the model. Here, we use modified GLUE methodology to generate behavioral sets of parameters to be used for the comparison of models.

Ten thousand sets of parameters were generated for each model using Monte Carlo simulations. Monte Carlo simulations take parameter ranges as input. Parameters are assumed to be uniformly distributed within their range which represents the initial uncertainty associated with each parameter. The GLUE procedure is intended to reduce parameter uncertainty; therefore we set the initial upper and lower parameter bounds fairly wide. Table 5-2 gives initial upper and lower limits of the parameters.

For computational purposes, representative terrain points were used instead of using the entire terrain. Representative terrain points are required, because we are

comparing the distribution of stability indices at the observed landslide locations with the terrain points to quantify the discriminatory power of the model. 5000 terrain points were sampled to obtain the set of representative terrain points. We performed repeat tests to verify that results were not sensitive to this terrain point sampling. For a single parameter set, when the terrain points used were increased to comprise the entire domain (all the terrain points), for SINMAP, the value of Q increased from 0.9143 (with 5000 terrain points) to 0.9158. This results in classifying 0.75 km^2 more area as stable when using a threshold that captures 95% of the landslides. A similar change of 0.36 km^2 was obtained with SHALSTAB. These differences are insignificant, relative to the 497 km^2 study area, so we concluded that the simplification involved in using 5000 sampled terrain points was justified. This expedited the analysis considerably because evaluating 10,000 simulations for the entire terrain is computationally burdensome.

For each parameter set from the Monte Carlo simulation a Q value was computed and parameter sets ranked according to their Q statistic (maximum to minimum). The

Table 5-2. Input range for Monte Carlo simulation for SHALSTAB

Model		Lower limit	Upper limit
SHALSTAB	C	0.0	0.5
	Φ[degree]	20^0	60^0
SINMAP	T/R$_{max}$ [m]	200	5000
	T/R$_{min}$ [m]	200	5000
	C$_{max}$	0.0	0.5
	C$_{min}$	0.0	0.5
	Φ_{max}[degree]	20^0	60^0
	Φ_{min}[degree]	20^0	60^0

first 5% of the sets are treated as the behavioral set of the parameters (the first 500 sets out of 10000 simulated sets). This results in two sets of behavioral parameters, one for each model. The Q-values from these behavioral sets can be used to compare models. The choice of 5% is somewhat subjective. Freer and Beven (1996) choose a series of these thresholds to demonstrate the sensitivity of parameters of TOPMODEL.

Figure 5-6 and Figure 5-7 give dot plots of the Q value for each of the parameter sets simulated for SHALSTAB and SINMAP, respectively. Each point in these figures is a plot of the value of one parameter from one of the 10000 sets versus the Q statistic evaluated for that particular parameter set. These plots are used to interpret the ranges of parameter values for which the model is behavioral in the sense of achieving a high Q statistic. Note that the Q statistic is for the combination of the parameters and not for just one value of it.

For SHALSTAB, we see that behavioral parameter sets are obtained for C values less than about 0.15 and for Φ values less than about 35º. Within these ranges there are combinations of C and Φ that can produce a high value of Q and good performance of the model. The result for SINMAP shows that model behaves consistently well no matter what the values of the parameters are. That is, all the dots tend to accumulate at Q value of 0.9. This is due to SINMAP still having discriminating capability in the conditionally stable domain due to the use of FS as SI in this part of the domain. This nature of SINMAP to consistently perform well is explored further in the Appendix. The sensitive range of Q is compressed due to a large part of the domain being flat.

To understand the parameter space, histograms of the behavioral parameter set were created. Figures 5-8 and 5-9 show the histograms for SHALSTAB and SINMAP

respectively. The histograms reveal the ranges of the model parameters that perform well and the absence of bars across the full range represents the reduction in the initial parameter uncertainty. Ideally if the initial ranges are fairly large, the resulting histogram of the behavioral set must be somewhere in the middle reducing the uncertainty at both ends. A good example of this is seen in Figure 5-9a. These histograms also indicate the potential for identifying the model parameters. A histogram with approximately the same height of bins across a large range of the parameter implies that the parameter cannot be identified/optimized within that range.

Figure 5-8 indicates that SHALSTAB performs better for lower values of C and Φ. There is a significant reduction in initial parameter uncertainty indicated by the histograms in Figure 5-8 relative to the initial ranges assumed for these parameters. Densities at the lower end of Φ suggest that the lower limit of Φ could be reduced for the analysis. Physically the lower limit of C is zero and hence the range of C cannot be further extended at the lower end.

The histogram plots for SINMAP show that the Φ_{min} has the highest density at its lower limit. So it may be possible that the parameter space for Φ might stretch beyond the lower limit. The same argument can be extended to other parameters of SINMAP, whose edges have relatively higher densities. When extending the parameter ranges, one should also consider whether it makes physical sense. For example, the Φ_{max}'s upper limit may also be extended, but an internal friction angle greater than 60 would rarely occur in the real world. T/R_{max} and C_{max} are other parameters that have some potential for stretching their limits. T/R_{max} showed the characteristic of parameter unidentifiability
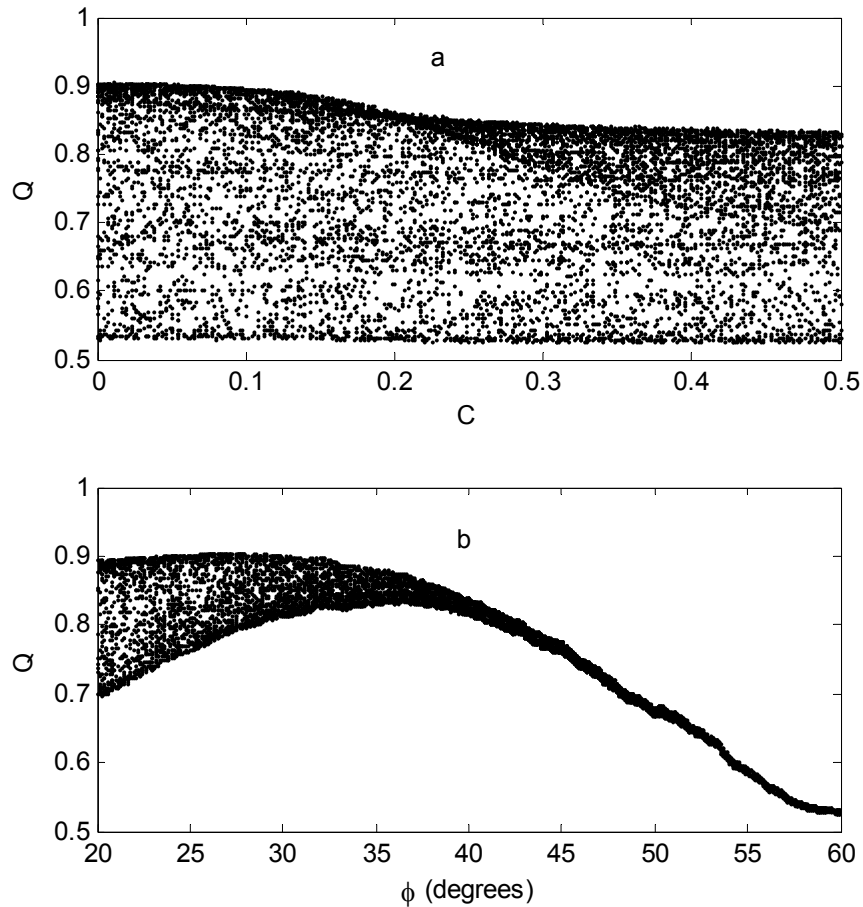
Figure 5-6. Scatter plots of Q goodness of fit measure versus individual SHALSTAB model parameters.

when upper limit was increased by having an approximately uniform distribution across the range. Nevertheless, from the available literature (Sidle, Pearce, and O'Loughlin, 1985; Sidle and Wu, 2001; Dietrich, Bellugi, and de Asua 2001; Dietrich et al., 1998), we believe the parameters ranges used are fairly large and hence they should not reasonably be extended any further.

Even though there is little basis for Φ values less than 20°, for illustrative purposes we performed tests extending the ranges for Φ_{min} and Φ_{max} to be 10-60 degree.

The GLUE procedure was repeated with these new ranges for Φ while the ranges of the remaining parameters are kept unchanged. 75000 parameter sets were generated from Monte Carlo simulation instead of 10000. SHALSTAB did not show any major changes while there was noticeable change in the results of SINMAP. The scatter plots for SINMAP from this run are illustrated in Figure 5-10. With the extension in Φ parameter range, we start to observe the poor performance of SINMAP for some parameter sets.

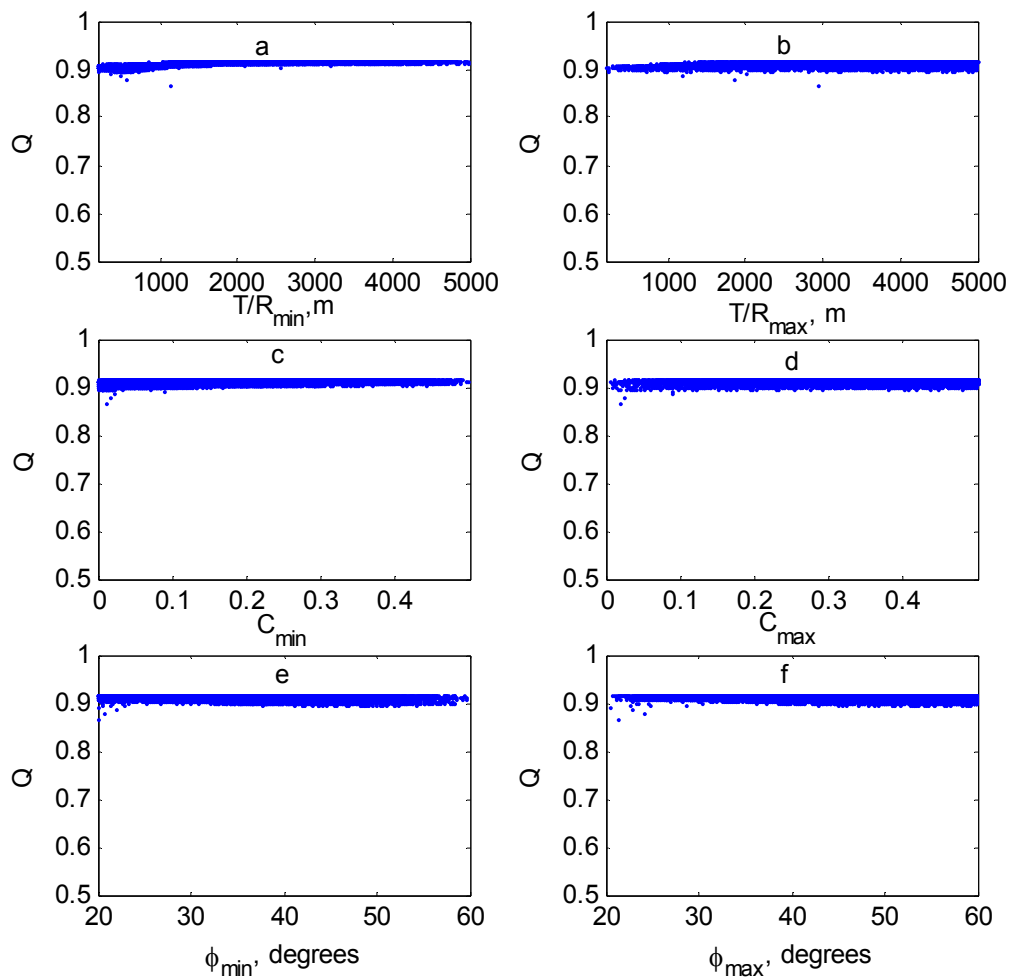Figure 5-7. Scatter plots of Q goodness of fit measure versus individual SINMAP model parameters.
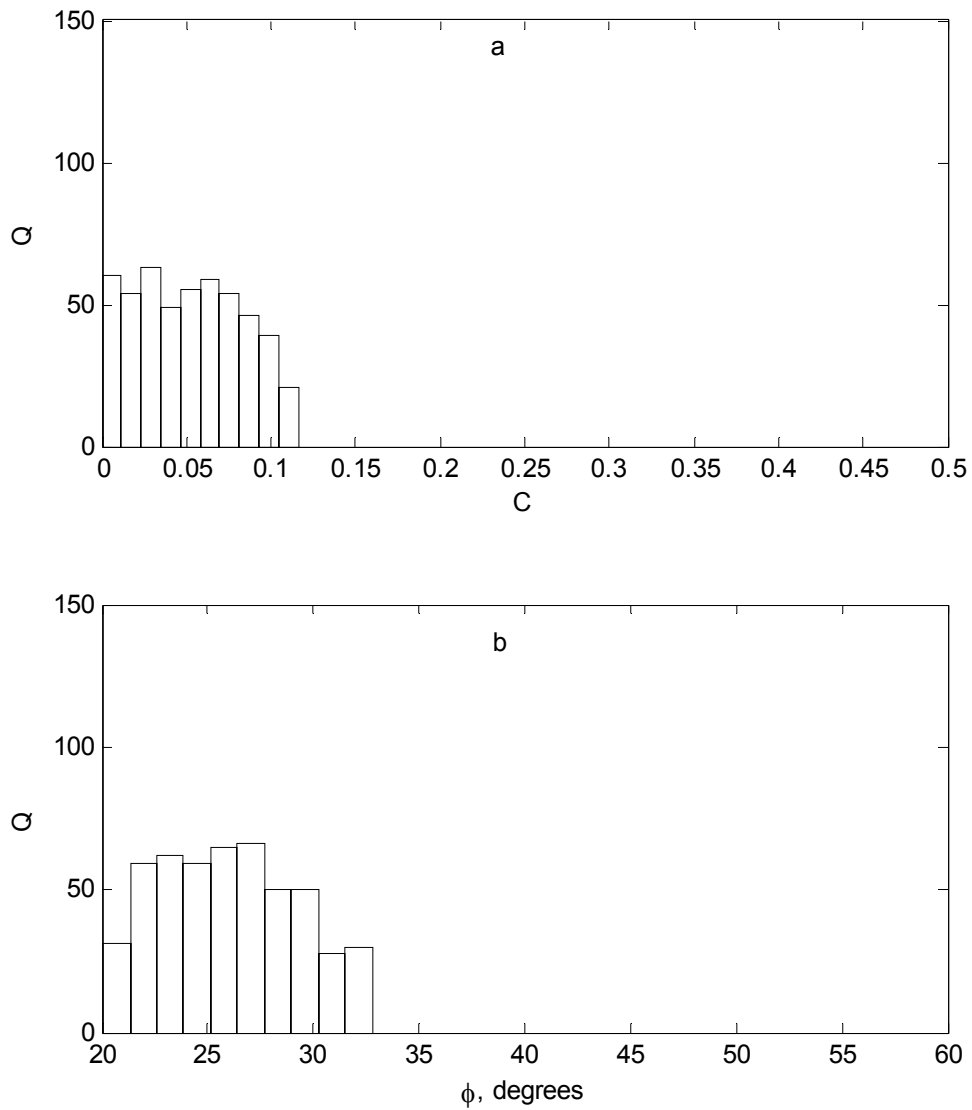
Figure 5-8. Histograms of behavioral parameter set for SHALSTAB.

These results show that poor discriminatory performance of SINMAP occurs when the range of Φ is allowed to be less than 20° but that for the entire reasonable range of parameters SINMAP gives good discriminatory performance.
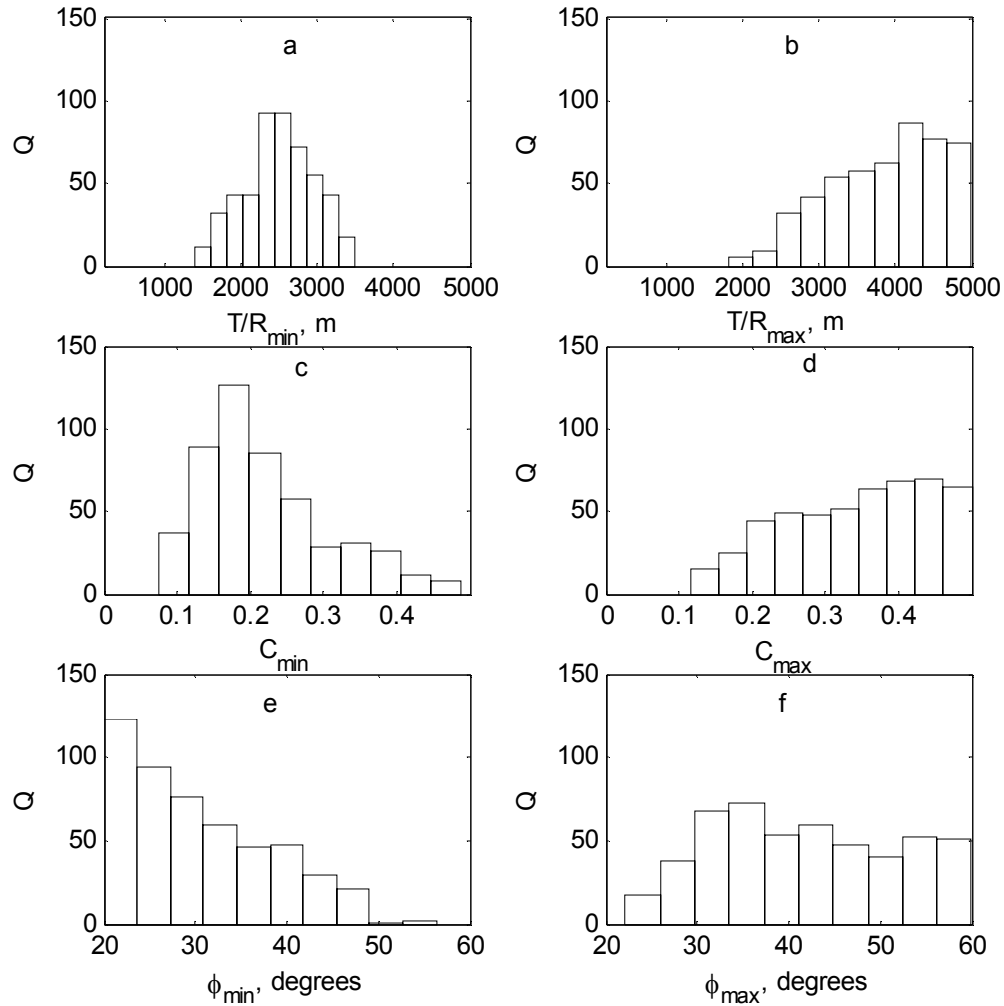
Figure 5-9. Histograms of behavioral parameter set for SINMAP.

Though the above analysis gives insight into the behavior of Q-statistic and ranges of behavioral parameters, it tells us very little about the sensitivity of parameters. The following section demonstrates the use of modified GLUE procedure to perform sensitivity analysis to assess the relative influence of each parameter on the model utilizing the cumulative distribution of parameters in behavioral and non-behavioral sets.

Figure 5-10. Scatter plots of Q goodness of fit measure versus SINMAP model parameters with extended range for Φ.

Sensitivity Analysis

Although the dot plots give some information of parameter sensitivity, further information is obtained by comparing the cumulative distribution of landslides of behavioral and non–behavioral parameter sets.
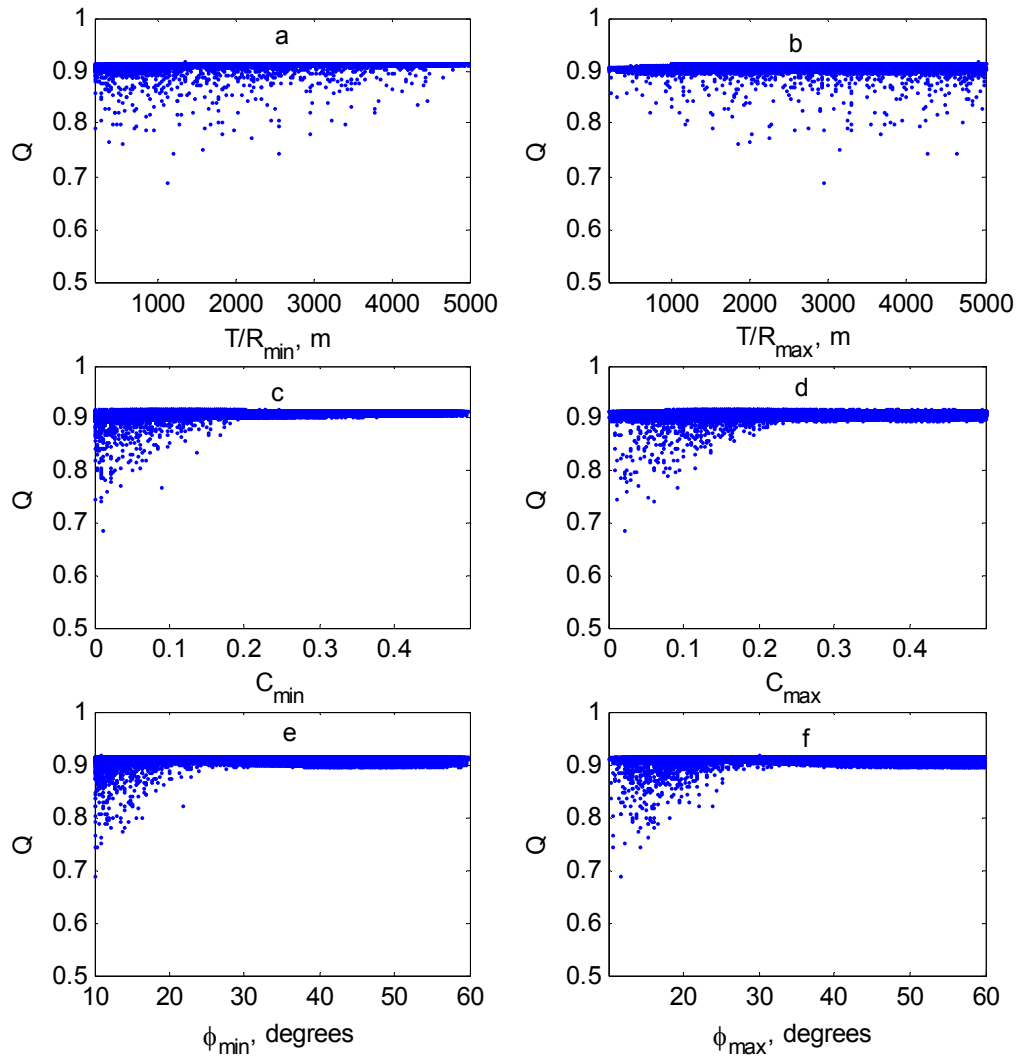
The behavioral and non-behavioral sets have 500 and 9500 parameter sets, respectively. The sensitivity of a parameter is quantified in terms of how different the parameter cumulative distributions are between behavioral and non-behavioral sets.

Figure 5-11 shows a marked difference between the behavioral and non-behavioral distributions (for SHALSTAB). Similar plots for SINMAP (Figure 5-12), suggests that the least sensitive parameters are C_{max} and Φ_{min}, and the most sensitive parameters are T/R_{min} and C_{min}.
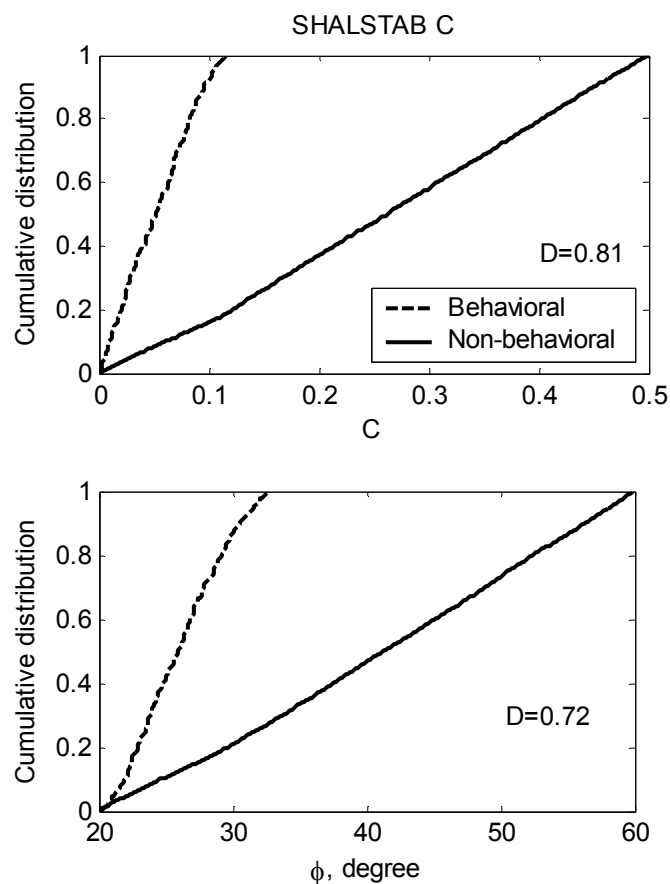
Figure 5-11. Comparison of behavioral and non-behavioral parameter distribution for SHALSTAB.
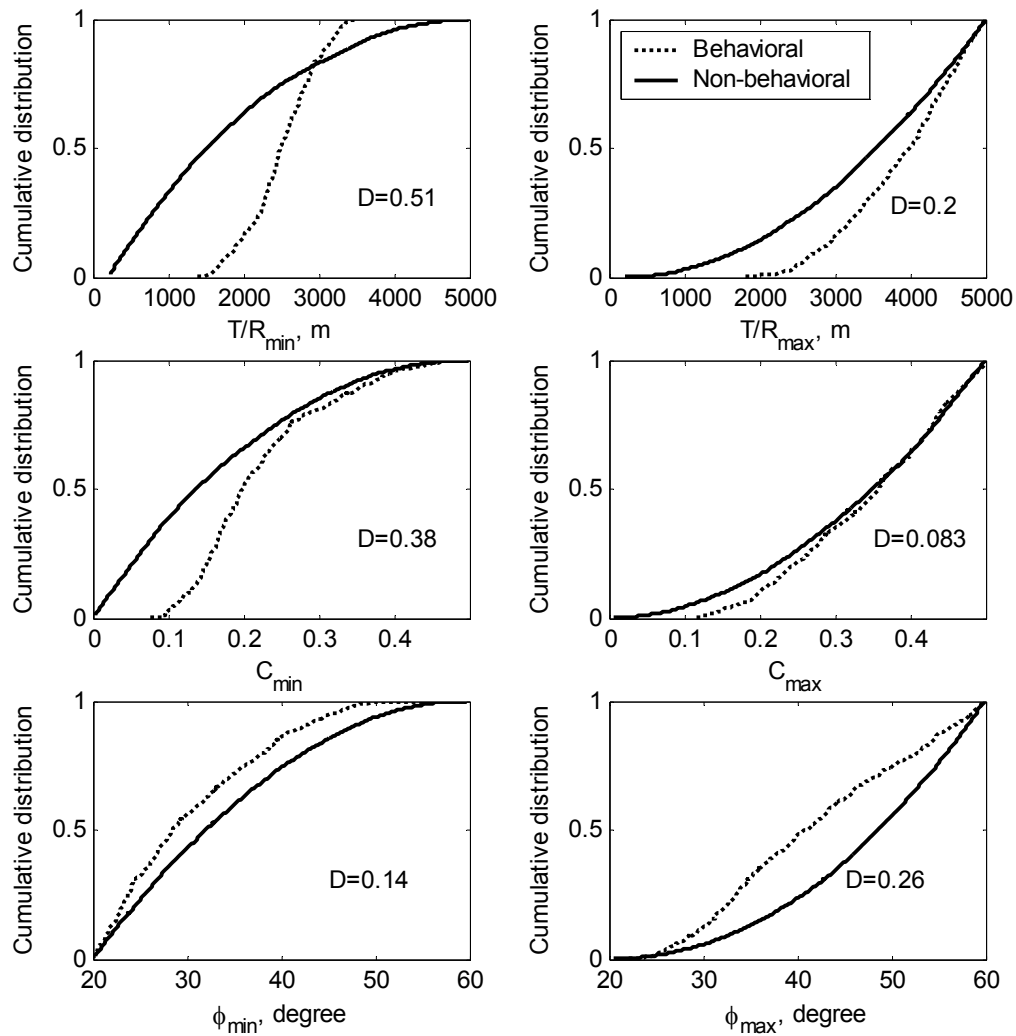
Figure 5-12. Comparison of behavioral and non-behavioral parameter distribution for SINMAP.

The Kolmogorov-Smirnov test (KS-test) was used to quantify the statistical significance of differences between two parameter set distributions. The KS-statistic (D) is the maximum vertical distance between the two distributions. D varies between 0 and 1, with higher values signifying greater differences between distributions.

The level of significance of D depends on the sample size and the confidence of the statistical test result increase with the number of samples. The sample size in this case is the number of parameter sets in the behavioral set, which is equal to 500. The non-behavioral set is considered as an underlying distribution. For different levels of statistical significance (α) the critical values of the test statistic D are given in Table 5-3. If the D statistic evaluated in KS-test is greater than the critical value, then the distributions are significantly different and hence the parameter is sensitive.

As is evident from the graphs (Figure 5-12), both the parameters of SHALSTAB have D values closer to one, suggesting a very significant difference between distributions. This means that the model is sensitive to both the parameters. The D values of SHALSTAB were significantly greater than the critical values for 99% confidence given in Table 5-3. In the case of SINMAP, the parameters are not as sensitive as SHALSTAB's parameters. The most sensitive parameter is T/R_{min} which has D=0.51 and the least sensitive is C_{max} which has D=0.08. Even though numerically smaller, these D values are greater than the critical values of the KS-test for 99% confidence level suggesting that parameters are sensitive. Note that C_{max} comes very close to being insensitive.

Figure 5-13 and 5-14 explore the interaction between parameters within the behavioral parameter sets. The axes in these figures represent the ranges used in the Monte Carlo simulation. The empty space depicts the reduction in parameter uncertainty from the GLUE random search. For example in Figure 5-13, if we plot all the 10000 points from the simulation, it would fill the entire box. After implementation of GLUE

methodology, the 500 behavioral parameter sets are seen to cluster in the lower left of the

domain with C<0.12 and Φ <32°.
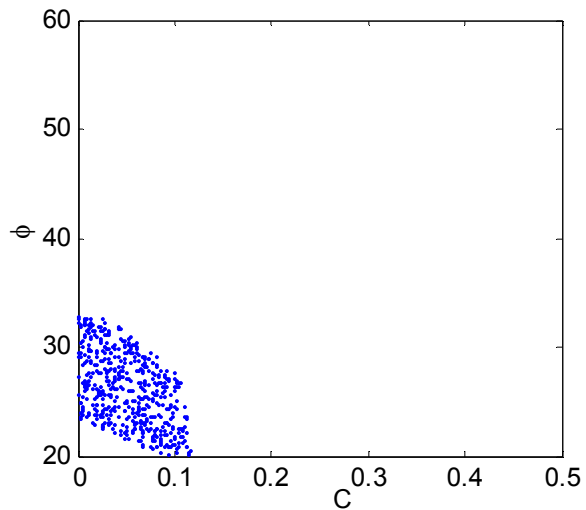
Figure 5-13. Interactions among behavioral paramters, SHALSTAB.

Table 5-3. Critical values of the Kolmogorov-Smirnov test

α	0.1 (90% confidence)	0.05 (95% confidence)	0.01 (99% confidence)
Approximations to critical values when sample size, n, is greater than 40	$\dfrac{1.22}{\sqrt{n}}$	$\dfrac{1.36}{\sqrt{n}}$	$\dfrac{1.63}{\sqrt{n}}$
When n=500,	0.0546	0.0608	0.0729

Adapted from (Davis, 1986)
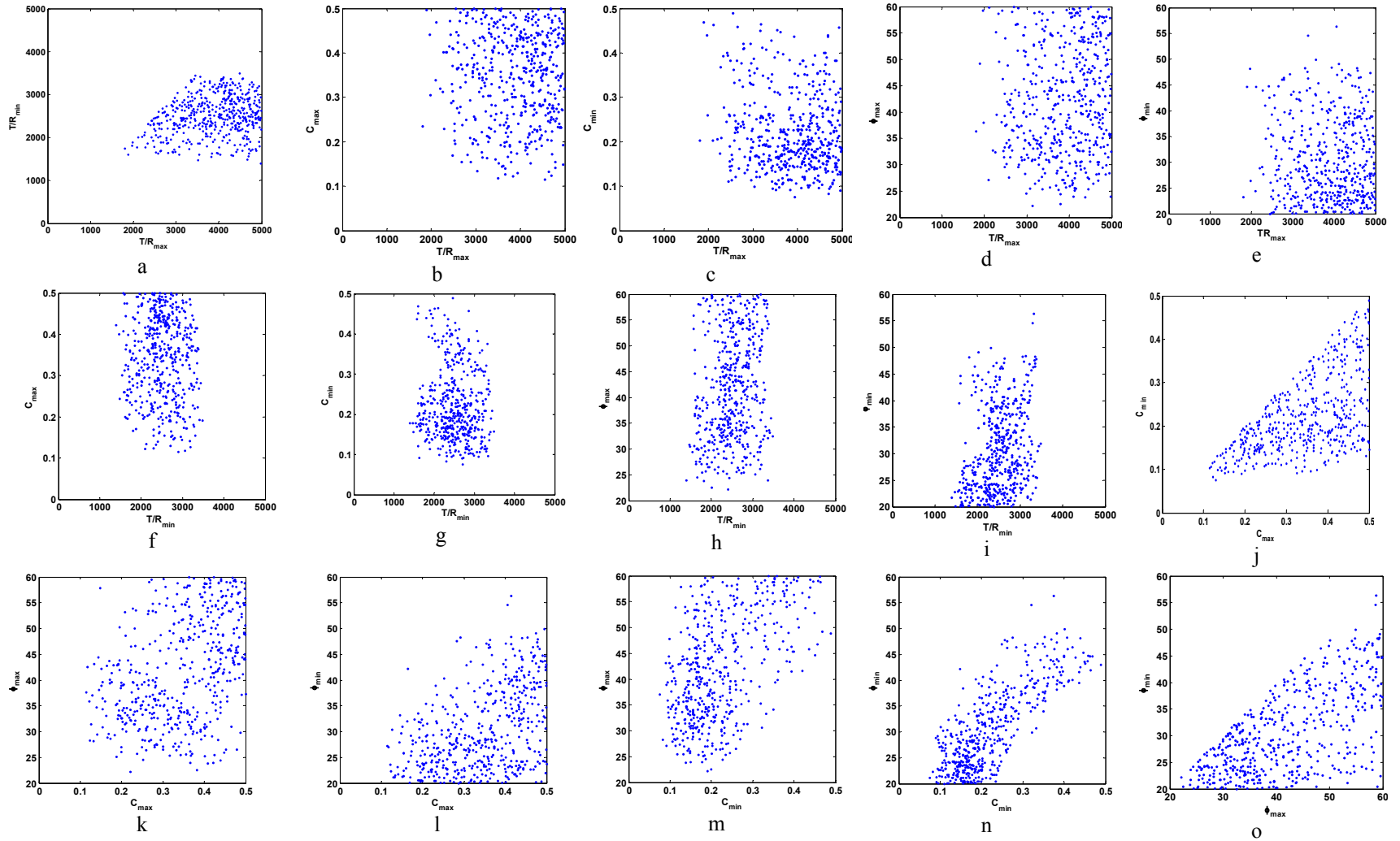
Figure 5-14. Interactions among behavioral parameters, SINMAP (Units: T/R (m), Φ(degree)).

The initial Monte-Carlo simulation ranges were large so we generally expect a reduction in the parameter range following calibration. This occurred with SHALSTAB, but was not the case with SINMAP. In Figure 5-14 Φ_{max} and Φ_{min} exhibit no reduction in their parameter space, while all other parameters show at least some reduction. The empty spaces in the upper diagonal triangle of the box in Figure 5-14 (a), (j), and, (o) should not be confused with the reduction in parameter uncertainty, because they are due to the conditions ($T/R_{max} > T/R_{min}$, $C_{max}>C_{min}$, and $\Phi_{max}> \Phi_{min}$) used while generating these parameters using Monte-Carlo simulation. According to these conditions, for a plot of Φ_{max} versus Φ_{min} (Figure 5-15 (o)), Monte–Carlo simulation filled up the entire lower diagonal triangle with its 10000 points. The behavioral set (500 points) still spreads throughout the lower triangle, suggesting no reduction in parameter space with respect to Φ_{max} and Φ_{min}. On observing carefully, this characteristic can be seen in any of the plots involving Φ_{max} or Φ_{min}.

Model Comparison

To have an objective model comparison we compare models using the parameter set with highest Q value from the modified GLUE random search. We also examine the improvement in model performance over the default parameter values. Results for SHALSTAB and SINMAP are given in Table 5-4. The optimized parameter sets from GLUE random search show an increase in the Q statistic over the default parameter values in both the models, but the improvement in SINMAP is small. This is the result of clustering of optimized parameters near 0.9 shown in Figure 5-7. The plot of $F_O(x)$ versus $F_R(x)$ for the best sets is given Figure 5-15. Presence of large proportion of flat

area in the terrain, where the models are similar in discriminating the stable area, results in very small differences in the numerical values of Q between SINMAP and SHALSTAB, though there is visually discernable difference in their graphs (Figure 5-15). At the end of this section, we evaluate the statistical significance of these differences using a chi-square statistical test at selected values of F_R.

The thresholds identifying 50%, 90%, and 95% of the landslides, and the corresponding fraction of area categorized as unstable were computed for each of the models for the default and the best parameter sets given in Table 5-4. The effectiveness of the GLUE methodology can be observed by comparing the performances of the default and the parameter set with highest Q value. The results are given in Table 5-5 (SHALSTAB) and 5-6 (SINMAP).

Table 5-4. Best set of the parameter for SINMAP and SHALSTAB with their Q statistic

Model	Parameter	Default set	Q	Parameters with the highest Q value	Q
SHALSTAB	C	0	0.78	0.011	0.9
	Φ (degree)	45		27.6	
SINMAP	$T/R_{max}(1/m)$	3000		4926	
	$T/R_{min}(1/m)$	2000		2621	
	C_{max}	0.25	0.9	0.23	0.91
	C_{min}	0.0		0.16	
	Φ_{max}(degree)	45		31.5	
	Φ_{min}(degree)	30		22.1	

Figure 5-15. F_O versus F_R plot for the best set parameter set of SHALSTAB and SINMAP.

Table 5-5. Landslide percent – area percent relation for SHALSTAB

	Default parameter set		Parameters with the highest Q value		
Landslide percentage identified	50	64	50	90	95
Threshold $\log_{10}(R/T)$	-2.95	-2.12	-6.00	- 3.66	-3.33
Percentage of area marked unstable	7.16	9.04	7.39	20.7	29.24

Table 5-6. Landslide percent – area percent relation for SINMAP

	Default parameter set			Parameters with the highest Q value		
Landslide percentage identified	50	90	95	50	90	95
Threshold SI	0.97	1.2	1.35	0.99	1.34	1.47
Percentage of area marked unstable	5.95	20.56	27.34	5.95	19.59	24.68

From the comparison, the default values of SHALSTAB can at the most only identify 64% of the landslides, associating them with 9% of the area. The remaining 36% of the landslides fall in the region classified as always stable according to the model (explained earlier). Compared to this, the best set for SHALSTAB contained 95% percent of landslides within 29% of the area, showing an improvement in the model's discriminatory power.

With SINMAP the difference between the Q statistic for the default and the parameter set with the largest Q value (Table 5-6) was only 0.1. This difference of 0.1 is due to calibration that categorizes 2% less area as unstable when identifying 95% of the landslides. Two percent of the area is approximately equal to 10 km^2 which could be significant in land management as this applies to steeper terrain.

In comparing SINMAP and SHALSTAB the optimized Q values are very close. However the small difference in Q value still results in appreciable differences in the percentage of unstable area for identifying 50%, 90%, and 95% of the landslides. Furthermore SINMAPS's default parameter set is better than SHALSTAB with optimized parameters in classifying 50%, 90%, and 95% of the landslides.

One important question, this study tries to answer is, is it possible for a simple variable like slope to categorize unstable regions as efficiently as complex models? To answer this, an analysis using slope and specific catchment area as slope-stability thresholds by themselves was performed. The result of this analysis is provided in Tables 5-7 and 5-8 for slope and specific catchment area, respectively.

The specific catchment area by itself shows little discriminatory power and categorizes large area as unstable. This is because specific catchment area may be large where the slope is too small to initiate a landslide. The strong topographical influence is evident with slope, which performs better than the calibrated SHALSTAB, in identifying 95% of the landslides. Slope has one percent less area classified as unstable while identifying 95% of the landslides than SHALSTAB. Similarly, the calibrated SINMAP model is only marginally better than the slope.

The above explanation about the differences between model performances is helpful from a land management perspective but does not necessarily say if there is any

Table 5-7. Landslide percent – area percent relation for slope threshold

Landslide percentage identified	50	90	95
Slope threshold	$0.55(29^{o})$	$0.37(20^{o})$	$0.32(18^{o})$
Percentage of area marked unstable	6.47	20.39	27.19

Table 5-8. Landslide percent – area percent relation for specific catchment area threshold

Landslide percentage identified	50	90	95
Specific catchment area threshold (m)	182	56	39
Percentage of area marked unstable	50.54	84.88	90.4

statistically significant difference between the model performances. We used a Chi-square goodness of fit test to evaluate the level of statistical significance in the differences between the models.

We pick specific values of $F_R(x)$ where in Figure 5-15 there is a visually discernable difference between the models in capturing the number of landslides.

The null hypothesis (H_o) of the test is that there is no significant difference between the models in capturing the number of landslides within their unstable zones while mapping $F_R(x)$ fraction of the terrain as unstable. We evaluate the level of significance at which the null hypothesis must be accepted.

For the case being tested, the chi-square statistic is given by the Equation 4.8. $\chi^2 = 0$ means that models are identical in their fraction of landslides identified and the higher the value of χ^2 the higher the discrepancy between the models. Results from this test are given in Table 5-9, where α is the significance level at which the null hypothesis is accepted.

Lower values of α represent a more significant difference between the models. The null hypothesis is rejected with 95% confidence (i.e. there is significant difference between the models), when α is less than 0.05. The bold values show significantly different model performances.

Uncertainty Analysis

Uncertainty in the predictions from distributed models has long been recognized and any modeling exercise should make an attempt to quantify the uncertainties associated with the model prediction. A better model should have smaller predictive

Table 5-9. Chi-square statistic for the differences in the model performances

	At $F_R(x)$	χ^2	α
SINMAP versus SHALSTAB*	0.05	3.67	0.055
	0.1	4.56	**0.033**
	0.15	7.82	**0.005**
	0.2	19.50	**0.00001**
	0.25	16.44	**0.0001**
	0.3	14.26	**0.0002**
	0.35	1.42	0.233
SINMAP versus SLOPE*	0.05	6.46	**0.011**
	0.1	34.71	**4×10^{-9}**
	0.15	25.40	**5×10^{-7}**
	0.2	23.47	**1×10^{-6}**
	0.25	5.71	**0.017**
	0.3	3.01	0.083
	0.35	0.06	0.799
SHALSTAB versus SLOPE*	0.05	0.38	0.538
	0.1	14.63	**0.0001**
	0.15	5.72	**0.017**
	0.2	0.25	0.614
	0.25	4.48	**0.034**
	0.3	8.36	**0.004**
	0.35	1.20	0.272

***model that is assumed to have an expected frequency of landslide discrimination.

uncertainty. The predictive uncertainty at each grid cell in the case of terrain stability mapping can be defined as the range of the stability index that the model produces given the uncertainty in the model parameters.

In this case, the parameter uncertainty is given by the 500 vector behavioral parameter set. The range of each parameter within the behavioral set is the parameter uncertainty. At a grid cell, if we obtain the stability index value for each of the 500 parameter vectors in the behavioral parameter set, then the range of the stability index so obtained quantifies the predictive uncertainty of the model at that grid cell. Instead of

giving the entire range, we take the difference between 5th and 95th quantiles of the stability index at each grid cell to quantify the predictive uncertainty.

Figure 5-16 shows maps giving the differences between 5th and 95th quantiles of their respective stability indices for SHALSTAB and SINMAP. The following steps were carried out to create these maps

a) At each grid cell, stability index was evaluated for each of the best 500 parameter combinations.

b) The stability index at each grid cell was then sorted in the ascending order to obtain a vector, x_i, where i=1…500, and $x_i > x_{i-1}$.

c) x_{25} was assigned to the grid cell for the map of 5th quantile values. And x_{475} was assigned to the grid cell for the map of 95th quantile values.

d) Step 3 was carried out in each grid cell to obtain two maps, one with 5th quantile values and another with 95th quantile values.

e) A predictive uncertainty map is created by subtracting the 5th quantile map from the 95th quantile map.

Note that all 500 behavioral parameter sets are utilized to generate these maps. Therefore the values at grid cells will not represent stability index values from a single parameter set. For example two adjacent grid cells can have 5th quantile stability index values derived from different parameter sets. The range depicted in the map is a measure of uncertainty of the model in predicting landslides at each location.
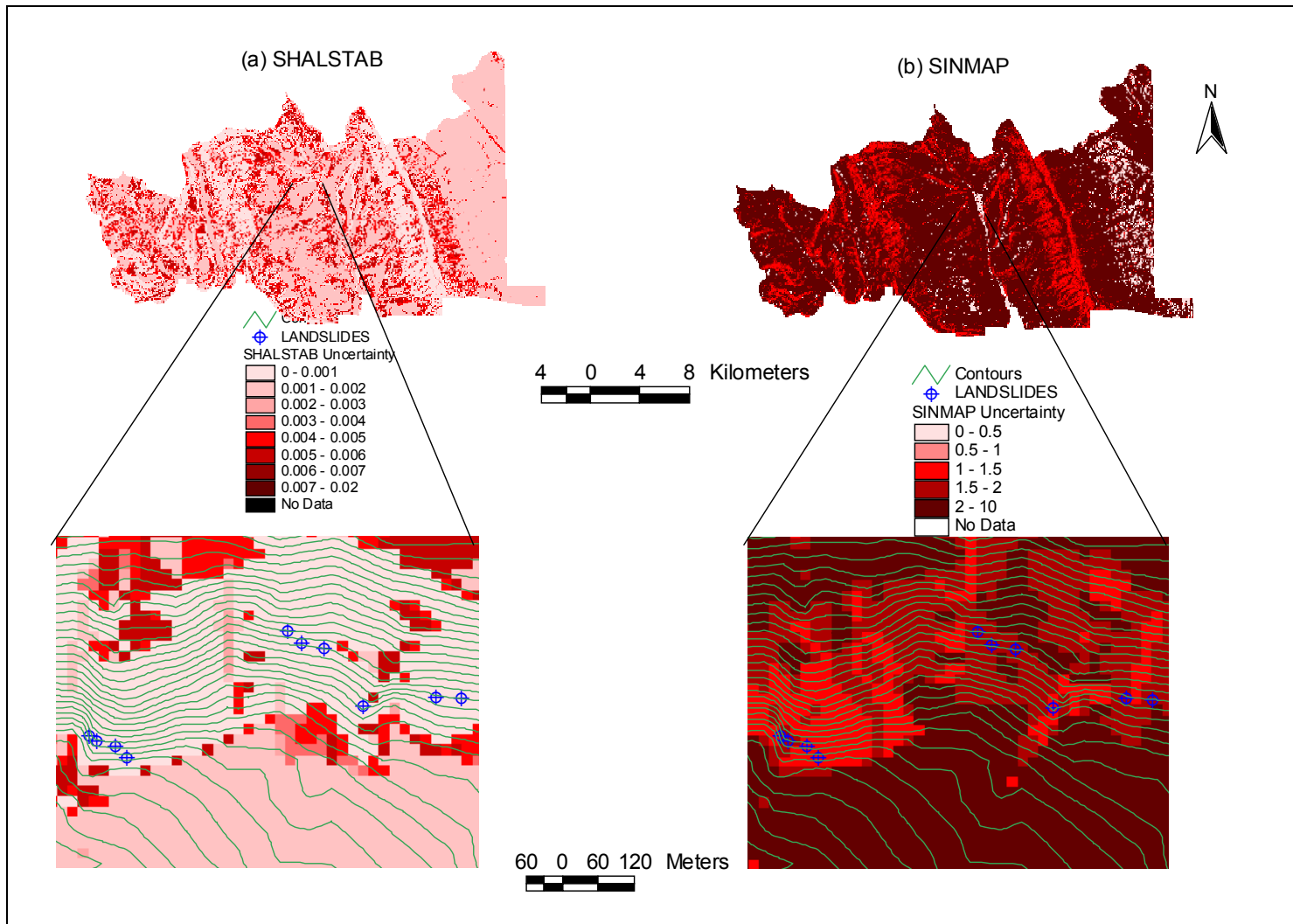
Fgure 5-16. Maps representing uncertainty in a) SHALSTAB's stability index log(R/T) and b) SINMAP's stability index SI.

SINMAP demonstrates smaller uncertainties in steeper regions of the terrain. That is, SINMAP predicts landslides better in steeper areas than in flat regions for this study area. SINMAP has relatively higher uncertainty in flatter region. The opposite is found for SHALSTAB. SHALSTAB has lower uncertainty in flatter region than the steeper region. Note that even when the model is highly certain (low certainty) about it stability index, one can still have a landslide observed in that area. This demonstrates the failure of the model to discriminate that landslide.

Split Sample Testing

Split sample testing is one of the preferred ways to evaluate the performance models in hydrology. To asses the efficiency of the calibration procedure, this test was used here. The study region was divided into two topographically similar parts, the GLUE random search methodology was used to estimate parameters in one part and these parameters were then used on the other part to evaluate the performance of the parameter sets. This is done for both the models to understand the effectiveness of the GLUE methodology in optimizing the parameters.

The study region was subdivided into two parts, north and south, that are topographically similar. A horizontal axis running approximately at the center of the study region was used to subdivide the region while minimizing topographical bias. This was done because the geological trends in the region run north-south so both divisions contain similar geological structure. The number of landslides in each of the divisions is approximately the same. The basic features of each of the subdivisions are given in Table 5-10. Figure 5-17 shows the split regions.
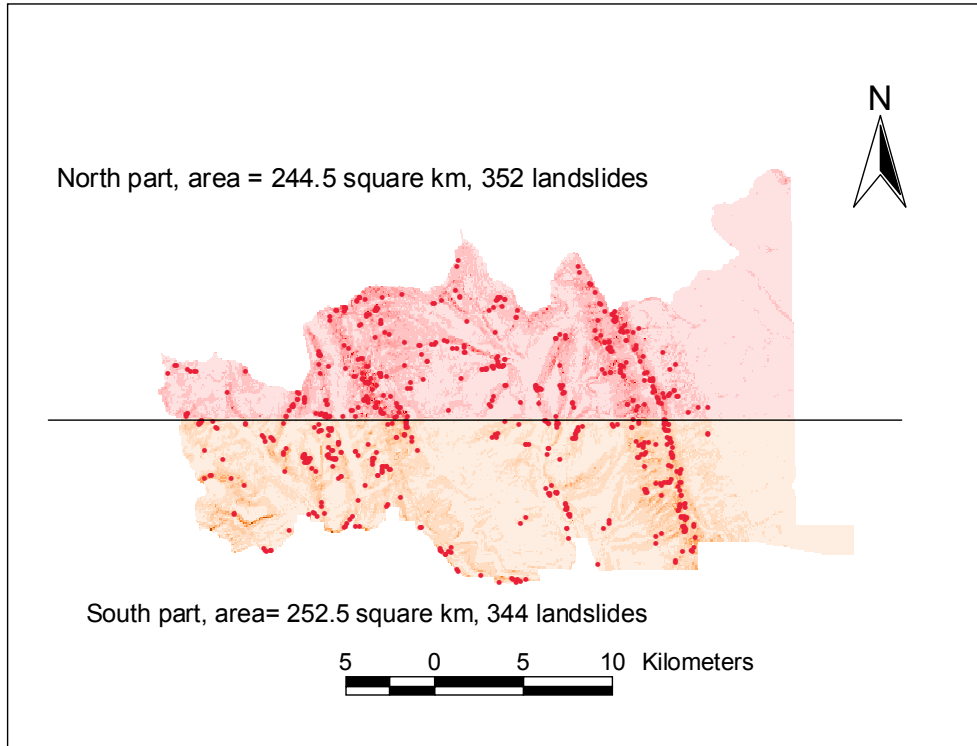
Figure 5-17. Study area divided into north and south part for the split sampling analysis.

Table 5-10. Basic characteristics of the divided regions

	Area (km^2)	Number of landslides
North	244.5	352
South	252.5	344

The GLUE random search methodology was applied to the northern region and the models evaluated on the southern part. The best parameter set was used to evaluate the models. As before the comparison was done using the Q statistic, and, the relation between the landslide percentage and area percentage.

Table 5-11 gives the result of split sample analysis for SHALSTAB. Parameters were optimized, for the north region using the parameter sets from the Monte-Carlo

simulation, landslide points from the north region and representative terrain points from the north region. The best parameter set so obtained, was then used along with the landslide points, and, the representative terrain points from the southern part to compute the value of Q for the south region. SHALSTAB gives similar result for both north and south regions, but note that the thresholds that discriminate similar percentage of landslides are different for these two regions. These results suggest that calibration could be extended to adjacent spatial regions. Similar results for SINMAP are presented in Table 5-12. In case of SINMAP, thresholds categorizing the same percentage of landslides in the two regions (north and south) do not change much.

These results suggest that both SHALSTAB and SINMAP are robust for prediction in regions not used in calibration. The better performance of SINMAP relative to SHALSTAB persists through this split sample test verifying that it is due to the structural differences between the models, rather than the calibration of parameters.

Table 5-11. Results from the split test analysis for SHALSTAB

	North-region			Q	South-region			Q
Landslide percentage identified	50	90	95	0.9	50	90	95	0.9
Threshold $\log_{10}(R/T)$	-6.00	-3.7	-3.4		-5.5	-3.6	-3.3	
Percentage of area marked unstable	7.2	19.5	27.5		6.9	23.2	30.0	

Table 5-12. Results from the split test analysis for SINMAP

	North-region			Q	South-region			Q
Landslide percentage identified	50	90	95		50	90	95	
Threshold SI	0.5	1.0	1.2	0.91	0.6	1.0	1.2	0.91
Percentage of area marked unstable	5.5	18.4	25.2		6.5	19	25	

CHAPTER 6

SUMMARY AND CONCLUSIONS

Terrain stability models provide an index of stability that may be used to map a geographic region into stable and unstable regions. A model's effectiveness at discriminating unstable terrain is evaluated by comparing the stability map it produces with the observed landslide points. A terrain stability model with good discriminating capability should contain most of the observed landslides within its unstable category while limiting the size of the unstable category. The major contribution of this study is the development of an objective measure, Q, to quantify the discriminating capability of a terrain stability model and provide the capability to objectively compare models that quantify terrain stability with incompatible stability indices. Two process-based terrain stability models, namely SHALSTAB (Montgomery and Dietrich, 1994) and SINMAP (Pack, Tarboton, and Goodwin, 1998), were used in this study to illustrate this model intercomparison.

The objective CDF vs CDF integral measure, Q, is the integral under the plot of Cumulative Distribution Function (CDF) of index at observed occurrence points versus the predictive index over the domain. This is a general measure that can quantify the capability of any predictive index field at discriminating the locations of point occurrence events. Here this has been applied in the context of landslides to compare terrain stability indices against point landslide initiation locations. The range of Q is between 0.5 and 1, with higher values quantifying better discrimination of unstable regions by the model.

We have addressed the following problems that arise when comparing process based terrain stability models: a) models use different indices to quantify terrain stability so they cannot be compared directly; b) models may use different thresholds of their stability index to categorize stability zones resulting in different fractions of the terrain being categorized as unstable; c) models require parameters to be estimated or calibrated. An intercomparison needs to guard against bias due to inferior parameter estimates. This motivates the need for a consistent parameter optimization.

Other tools that have been used with terrain stability models, namely a) Landslide density and b) Spatial probability, were analyzed to test their capability to compare terrain stability models. Landslide density was found to be informative for a particular model, but not amenable to comparing models with different stability indices. Spatial probability is presented as a possible method for overcoming the issue of different indices and thresholds.

Spatial probability provides an objective and non-parametric quantification of terrain stability. Because the measure mapped is probability, the selection of a threshold to delineate stability classes can appropriately accommodate commission and omission errors depending on the stability map application purpose. A lower spatial probability threshold should be used in cases where consequences due to landslide occurrence are more severe.

Spatial probability does, nevertheless have some deficiencies, namely a) modeling at a scale less than the size of the landslides that can result in the probability of landslide occurrence being greater than one; b) parameters still have to be calibrated; and c) the

distributed nature of spatial probability maps makes general interpretation of performance over an area difficult.

The measure Q identifies the model that has a greater fraction of the observed landslides within a smaller fraction of the unstable terrain as a better model. Q uses fractions instead of the actual stability index to quantify the predictive capability of the model and hence overcomes the difficulties of comparing models with different indices and thresholds. The measure Q still needs the models to be calibrated before they can be compared. The Q measure was used within the Generalized Likelihood Uncertainty Estimation (GLUE) random search methodology as a goodness of fit measure to find parameter sets that optimize the performance of terrain stability models so that terrain stability models with different stability indices could be compared objectively.

To the best of our knowledge, ours is the only study that has to date implemented the GLUE methodology in the field of terrain stability mapping. The procedure adopted in this study is slightly different from the original GLUE methodology. The aspects of GLUE that are common to our study are, the Monte-Carlo simulation of large number of parameter sets, and, the evaluation of each parameter set using a generalized likelihood (goodness of fit) measure. In this study the Q goodness of fit measure was developed to quantify the ability of the output stability index map associated with each parameter set to partition unstable and stable zones with respect to the observed landslide points. The output from the GLUE methodology is two sets of parameter vectors, a) behavioral and b) non-behavioral. The behavioral parameter sets for each model establish model calibration, i.e. for these sets, each model perform the best within its own modeling

framework. The Q measure from the behavioral set is used to compare model performance between different models.

The initial ranges provided to the Monte Carlo simulation for generating parameter sets represent the modeler's assessment of the initial parameter uncertainty and the behavioral set reduces the initial parameter uncertainty, provided the model is sensitive to the parameter. Sensitivity analysis revealed that SHALSTAB is sensitive to both of its parameters. SINMAP shows significant sensitivity to its parameters but it is not as sensitive to some of them as SHALSTAB is to its parameters. Parameter interaction plots revealed that the parameter uncertainty was considerably reduced in both parameters of SHALSTAB (C and Φ). SINMAP did not show significant reduction in parameter uncertainty in case of C_{max}, Φ_{min} and Φ_{max}. These were the parameters that had relatively smaller sensitivity among the SINMAP parameters. This demonstrates the use of GLUE methodology to explore parameter uncertainty.

The comparison between the default parameter set and the set with the highest Q value suggested that the parameter selection is more important for SHALSTAB. The Q value improved from 0.78 for the default parameter set to 0.9 for the best parameter set from the GLUE method, while there was no such significant increase with Q values of SINMAP. The distribution of Q values within the parameter sets from the GLUE method, suggested that SINMAP has good discriminatory capability for almost all realistic parameter sets. This is due to SINMAP using FS as its stability index SI in the domain where the infinite-slope model predicts unconditional stability, while SHALSTAB is non-discriminating in this domain. The accommodation of uncertainty through a uniform

distribution of C and Φ in SINMAP increase the range of slope that SINMAP is sensitive to, relative to SHALSTAB, resulting in better discriminatory capability by SINMAP.

Another important result from this work was the analysis of simple terrain variables like slope and specific catchment area as stability indices by themselves. Specific catchment area when used by itself as a terrain stability index shows very little discriminating power. This is because there are locations with large specific catchment area where slope is not sufficient to cause initiation of landslides. In this particular study area slope on its own showed considerable ability to identify landslides. Slope in fact performed slightly better than the optimized SHALSTAB and SINMAP was only marginally better than slope by itself. In the Chetwynd study area high topographic influence on landslides was expected and hence the success of slope alone as a stability index is not really surprising. This is though not universal, and in some other cases where slope is not so dominant, we believe process-based models can perform better than slope in discriminating unstable terrain.

The discriminating capability of terrain stability models are similar in flat regions because there are very few landslides in the flat region. The presence of a large proportion of flat terrain in Chetwynd study area compresses the scale of Q and leads to very small numerical differences between the values of Q from SHALSTAB, SINMAP and slope alone. Despite these small numerical differences there is a visually discernable difference in the model performance notable in the CDF vs CDF plots where the terrain is relatively steeper. We performed chi-square tests to evaluate the significance of the difference between model performances at fractions of terrain where the terrain is relatively steeper. The chi-square tests revealed that the difference in model

performances is statistically significant over the range of stability index where the terrain is steep.

One of the benefits of the GLUE methodology is the identification of behavioral sets that result in equivalent model performance (Beven's concept of equifinality, Beven and Binley, 1992). The distribution of model predictions over the behavioral set provides quantification of uncertainty associated with model predictions. The predictive uncertainty was quantified at each grid cell as the difference between the 5th percentile and 95th percentile values of the stability index from the behavioral parameter set. SINMAP and SHALTAB differ from each other in terms of the uncertainty they produce in steeper and flatter regions. SINMAP has lower uncertainty in steeper regions than in flatter areas whereas SHALSTAB demonstrated the opposite. Lower uncertainty at a grid cell does not does not necessarily identify the landslide. For example a model may be highly certain that a grid cell is stable because it is flat and has small specific catchment area, but it is possible to have a landslide at that point. This suggests the failure of the model in identifying the landslide at that point. Uncertainties at a grid cell should be viewed as the range of the stability index that can be expected because of the uncertainty in the model parameter and structure.

A split sample test was conducted to assess the efficiency of the parameter optimization procedure. The results from the split sample test suggest that both SHALSTAB and SINMAP are robust for prediction in regions not used in calibration. The better performance of SINMAP relative to SHALSTAB persists through this split sample test verifying that it is due to the structural differences between the models, rather than the calibration of parameters.

This study presented an objective measure Q that was found to be effective in quantifying the discriminating capability of a terrain stability model. Our results from the application of this method in the Chetwynd area are encouraging, but additional evaluation in other areas is needed ascertain whether this approach is general and broadly effective.

REFERENCES

Barling, R. D., I. D. Moore, and R. B. Grayson. 1994. A quasi-dynamic wetness index for characterizing the spatial distribution of zones of surface saturation and soil water content. Water Resources Research 30:1029-1044.

Beven, K. 1997. Topmodel: A critique. Hydrological Processes 11:1069-1085.

Beven, K., and A. Binley. 1992. The future of distributed models: model calibration and uncertainty prediction. Hydrological Processes 6:279-298.

Beven, K. J., and M. J. Kirkby. 1979. A physically based variable contributing area model of basin hydrology. Hydrological Sciences Bulletin 24:43-69.

Beven, K., and E. F. Wood. 1983. Catchment geomorphology and the dynamics of runoff contributing areas. Journal of Hydrology 65:139-158.

Borga, M., G. D. Fontana, and F. Cazorzi. 2002. Analysis of topographic and climatic control on rainfall-triggered shallow landsliding using a quasi-dynamic wetness index. Journal of Hydrology 268:56-71.

Brazier, R. E., K. J. Bevan, J. Freer, and J. S. Rowan. 2000. Equifinality and uncertainty in physically based soil erosion models: application of the GLUE methodology to WEPP-The water erosion prediction project-for sites in the UK and USA. Earth Surface processes and Landforms 25:825-845.

Canada Soil Survey Committee. 1977. Soils of Canada: a cooperative project of the Canada soil survey committee and the soil research institute. Research Branch, Canada Dept. of Agriculture, Ottawa, Ontario.

Carrera, A. 1983. Multivariate models for landslide hazard evaluation. Mathematical Geology 15:445-468.

Carrera, A., M. Cardinali, R. Detti, F. Guzzetti, V. Pasqui, and P. Peichenbach. 1991. GIS techniques and statistical models in evaluating landslide hazard. Earth Surface processes and Landforms 16:427-445.

Carrera, A., M. Cardinali, F. Guzzetti, and P. Peichenbach. 1995. GIS technology in mapping landslide hazard. 135-175. In A.Carrera and F. Guzzetti (Eds.). Geographical information systems in assessing natural hazards. Kluwer Academin Publishers, Dordrecht, The Netherlands.

Christianes, K., and J. Feyen. 2002. Constraining soil hydraulic parameter and output uncertainty of the distributed hydrological MIKE SHE model using the GLUE framework. Hydrological Processes 16:373-391.

Davis, H. C. 1986. Statistics and data analysis in geology, Second edition. John Wiley & Sons, Inc., New York, 646p.

Dhakal, A. S., T. Amada, and M. Aniya. 2000. Landslide hazard mapping and its evaluation using GIS: An investigation of sampling schemes for a grid-cell based quantitaive method. Photogrammetric Engineering and Remote Sensing 66:981-989.

Dietrich, W. E., D. Bellugi, and R. R. de Asua. 2001. Validation of shallow landslide model, SHALSTAB, for forest management. Water Science and Application 2:195-227.

Dietrich, W. E., R. R. de Asua, J. Coyl, B. Orr, and M. Trso. 1998. A validation study of the shallow slope stability model, SHALSTAB, in forested lands of Northern California. Stillwater Ecosystem, Watershed & Riverine Sciences, Berkeley, CA.

Dietrich, W. E., C. J. Wilson, and S. L. Reneau. 1986. Hollows, colluvium, and landslides in soil-mantled landscapes. Pages p.361-388. In A. D. Abrahams, (ed.). chapter 17 in Hillslope Processes. Allen & Unwin, Boston.

Environment Canada. 2004. http://www.ec.gc.ca/envhome.html

Freer, J., and K. Beven. 1996. Bayesian estimation of uncertainty in runoff prediction and the value of data: an application of the GLUE approach. Water Resource Research 32:2161-2173.

Gabrielse, H. 1991. Late paleozioc and mesozoic terrane interactions in north-central Bristish Columbia. Canadian Journal of Earth Sciences 28:947-957.

Gray, D. H., and W. F. Megahan. 1981. Forest vegetation removal and slope stability in the Idaho Batholith. Forest Serv., U.S. Dep. of Agriculture, Res. Pap. INT-271,Ogden, UT.

Hammond, C., D. Hall, S. Miller, and P. Swetik. 1992. Level I stability analysis (LISA) documentation for version 2.0. USDA Forest Service Intermountain Research Station.

Ibbitt, R. P., and T. O'Donnell. 1971. Fitting methods for conceptual models. Journal of the Hydraulics Division, ASCE 97:1331.

Iverson, R. M. 2000. Landslide trigerring by rain infiltration. Water Resource Research 36:1897-1910.

Iverson, R. M., and J. J. Major. 1986. Groundwater seepage vectors and the potential for hillslope failure and debris flow mobilization. Water Resource Research 22:1543-1548.

Iverson, R. M., and J. J. Major. 1987. Rainfall, groundwater flow and seasonal movement at Minor Creek landslide, Northwestern California: physical meaning of the empirical relations. Geological Society of America Bulletin 99:579-594.

Jukes, W., and A. de Vries. 2001. Sustainable Forest Management Plan. Chetwynd, BC.

Kuczera, G. 1983. Improved parameter inference in catchment models. 2. Combining different kinds of hydrologic data and testing their compatibility. Water Resource Research 19:1163-1172.

Kuczera, G. 1990. Assessing hydrological model nonlinearity using response surface plots. Journal of Hydrology 118:143-162.

Kuczera, G., and E. Parent. 1998. Monte carlo assessment of parameter uncertainty in conceptual catchment models: the metropolis algorithm. Journal of Hydrology 211:69-85.

Madsen, H. 2000. Automatic calibration and uncertainty assessment in rainfall-runoff modeling. In Joint Conference on Water Resources Engineering and Water Resources Planning and Management, Hyatt Regency Minneaplois, USA.

Montgomery, D. R., and W. E. Dietrich. 1994. A physically based model for the topographic control on shallow landsliding. Water Resource Research 30:1153-1171.

Montgomery, D. R., K. M. Schmidt, H. M. Greenberg, and W. E. Dietrich. 2000. Forest clearing and regional landsliding. Geology 28:311-314.

Montgomery, D. R., K. Sullivan, and H. M. Greenberg. 1998. Regional test of a model for shallow landsliding. Hydrological Processes 12:943-955.

Moore, I. D., and R. B. Grayson. 1991. Terrain-based catchment partitioning and runoff prediction using vector elevation data. Water Resources Research 27:1177-1191.

Okimura, T., and R. Ichikawa. 1985. A prediction method for surface failures by movements of infiltrated water in a surface soil layer. Natural Disaster Science 7:41-51.

O'Loughlin, E. M. 1986. Prediction of surface saturation zones in natural catchments by topographic analysis. Water Resources Research 22:794-804.

Pack, R. T., D. G. Tarboton, and C. N. Goodwin. 1998. Terrain stability mapping with SINMAP and users guide for version 1.00. 41140-0, Terratech Consulting Ltd., Salmon Arm, B. C., Canada.

Schroeder, W. L., and J. V. Alto. 1983. Soil properties for slope stability analysis; Oregon and Washington coastal mountains. Forest Sci. 29:823-833.

Sidle, R. C., A. J. Pearce, and C. L. O'Loughlin. 1985. Hillslope stability and land use. American Geophysical Union, Water Resources Monograph series, 11:1-140.

Sidle, R. C., and W. Wu. 2001. Evaluation of the temporal and spatial impacts of timber harvesting on landslide occurrence. Water Science and Application 2.

Spear, R. C., and G. M. Hornberger. 1980. Eutrophication in peel inlet, II, identification of critical uncertainties via generalized sensitivity analysis. Water Research 14:43-49.

WFPB. 1997. Board manual: standard methodology for conducting watershed analysis. Version 4.0. Washington Forest Practices Board, Olympia, Washington.

Wu, W., and R. C. Sidle. 1995. A Distributed slope stability model for steep forested watersheds. Water Resources Research 31:2097-2110.

Zaitchik, B. F., H. M. van Es, and P. J. Sullivan. 2003. Modeling slope stability in Honduras: parameter sensitivity and scale of aggregation. Soil Science Society of America Journal 67:268-278.

APPENDIX

Explanation of Consistently High Values of Q-statistic

The Q statistic with SINMAP was consistently good over a wide range of input parameters. To understand why this is the case we need to understand how the combination of thresholds and parameter-sets work, in discriminating landslides, and, categorizing stable and unstable terrain. The simplest way to understand this concept is to use slope versus specific catchment area plots (SA-plots). Such a plot is shown in Figure A-1, where the red points are the observed landslide points and the blue ones are the representative terrain points. Slope and the specific catchment area are the two principal input quantities that are spatially variable in the terrain stability models evaluated.

The basis for maximizing the Q statistic is to include most of the landslides in the smallest possible terrain area. Fitting a model through estimating parameters and setting thresholds lines will divide the S-A space into two parts each containing some landslides and terrain area. The best threshold with the best parameter set would have in the unstable category, the highest number of landslides associated with least possible terrain area. In terms of the SA plot, such a threshold and parameter set would have in the unstable part, most of the red dots and very few blue dots.

The simplest threshold that could be used to divide the SA plot would be slope. The slope threshold used in the section 5.5 that identified 90% of the landslides was equal to 20°. Drawing a vertical line at 20° in Figure A-1, we can divide the domain into the left and right parts. The right hand part (unstable) contains the values shown in Table 5-7. i.e 90% of the landslides in 20.39% of the terrain. Similarly using the specific catchment

area threshold that identified 90% of the landslide (Table 5-8), we can draw a horizontal line that divides the SA-space into top and bottom parts. The top part (unstable) has 90% of the landslides associated with 85% of the area, which is much larger than the area associated with the slope threshold (21%). This shows why slope is better than the specific catchment area alone in discriminating unstable area.

Process based models like SHALSTAB and SINMAP use a line similar to the line C shown in Figure A-1 to discriminate stable and unstable terrain. Geometrically this line results in an improvement relative to the vertical line, because it does not include the points with smaller specific catchment area in intermediately steep slope. The line C in
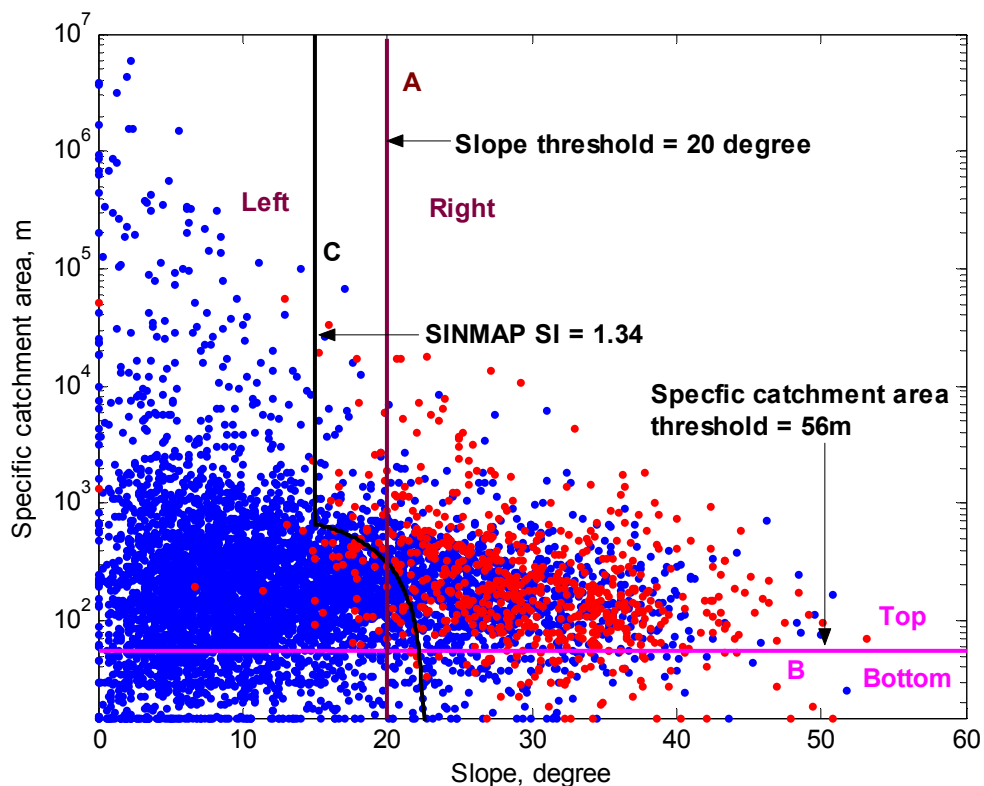
Figure A-1. SA-plot with slope and specific catchment area as thresholds.

Figure A-1 is drawn for SINMAP, using the parameter set with highest Q and a threshold of SI equal to 1.34 which identifies 90% of the landslides in 19.6% of the terrain area. In general the form of the threshold line used by the models and in actual cases; the form of the curve at the bottom of the vertical line and where it starts may change depending on the model, parameter set and the threshold used.

A threshold line lying to the right of the line C would identify smaller fraction of the observed landslides and a line lying to the left would categorize more terrain area as unstable. In order to maximize the discrimination power of a model; we need to optimize the position of the threshold line. This threshold line must be related to the Q values obtained from the behavioral parameter sets.

The statistic Q is constructed by integrating over the full range of stability indices, and hence a parameter set with high Q value should have at least one threshold that gives high discrimination of unstable terrain based on the observed landslide points. A close examination of the S-A domain shows that in order to have a high discrimination, one has to draw the threshold line within a small region hypothetically represented in Figure A-2. Any model and parameter set that contain a threshold line in this region will achieve a high Q value.

Figure A-3 shows different threshold lines for SHALSTAB using its parameter set with highest Q value (0.9). For a particular parameter set, the position of the vertical line shown in Figure A-3 is fixed and hence changing threshold would shift the threshold lines only in the vertical direction. As a result of this, SHALSTAB for a particular parameter set can never identify the landslides in the unconditionally stable domain by changing its threshold. Notice that for the parameter set used in the plot, the hypothetical

region in figure A-2 overlaps the domain of the threshold lines that can be drawn and hence this parameter set has a high power for discriminating unstable region.

SHALSTAB can have a parameter set whose threshold lines can fall to the right of the vertical lines in Figure A-3. The threshold line's domain in such a case would not overlap with the region in Figure A-2 and would also have more landslides in the unconditionally stable domain (which it cannot identify). Such parameter sets will have lower Q values. The point is that SHALSTAB because of its limited threshold line's
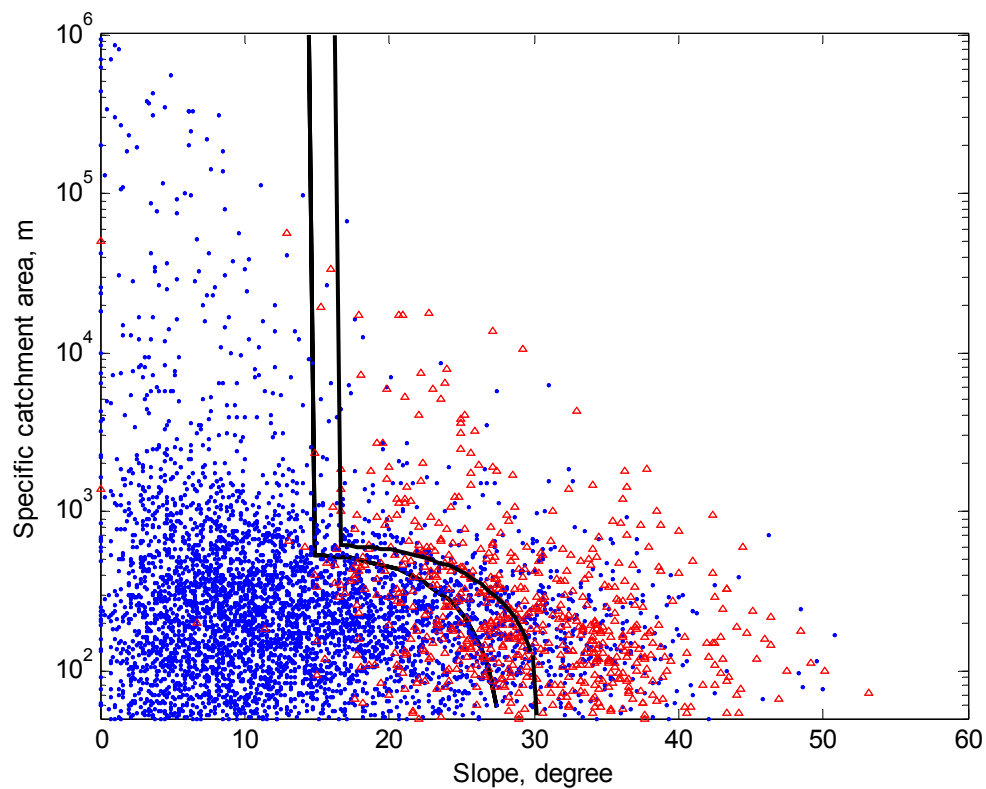

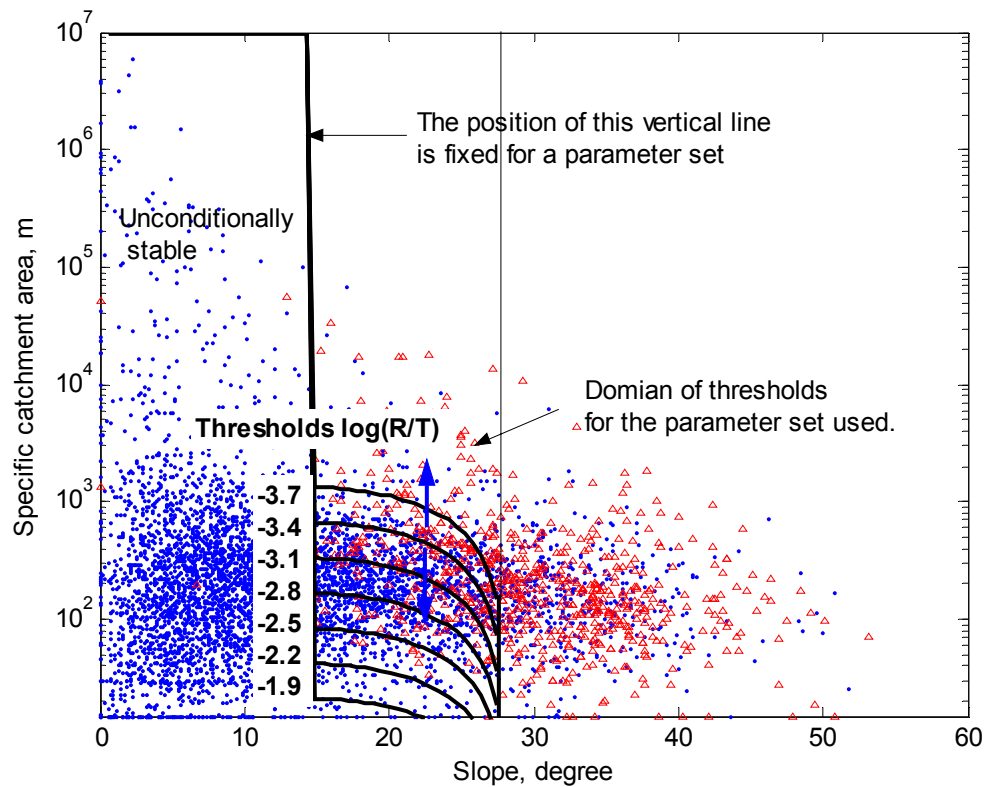Figure A-2. Hypothetical region where the threshold lines have high discriminating power.

Figure A-3. Different threshold lines for drawn for the best parameter set of SHALSTAB.

domain does not perform well all the time and optimization of the parameters becomes important.

SINMAP differs from SHALSTAB in how it draws the threshold lines in the SA-space. Figure A-4 shows SINMAP's different threshold lines for one particular parameter set (randomly selected). The domain of the threshold lines cover the entire S-A region left of the threshold line with SI=0. This makes it possible for SINMAP to discriminate even those landslides that lie in the unconditionally stable region, by just changing the threshold. As long as the zero threshold line of SINMAP is to the right of the hypothetical region in Figure A-2, there will always be some threshold that would

give high discrimination of unstable terrain. Therefore for most of the realistic parameter
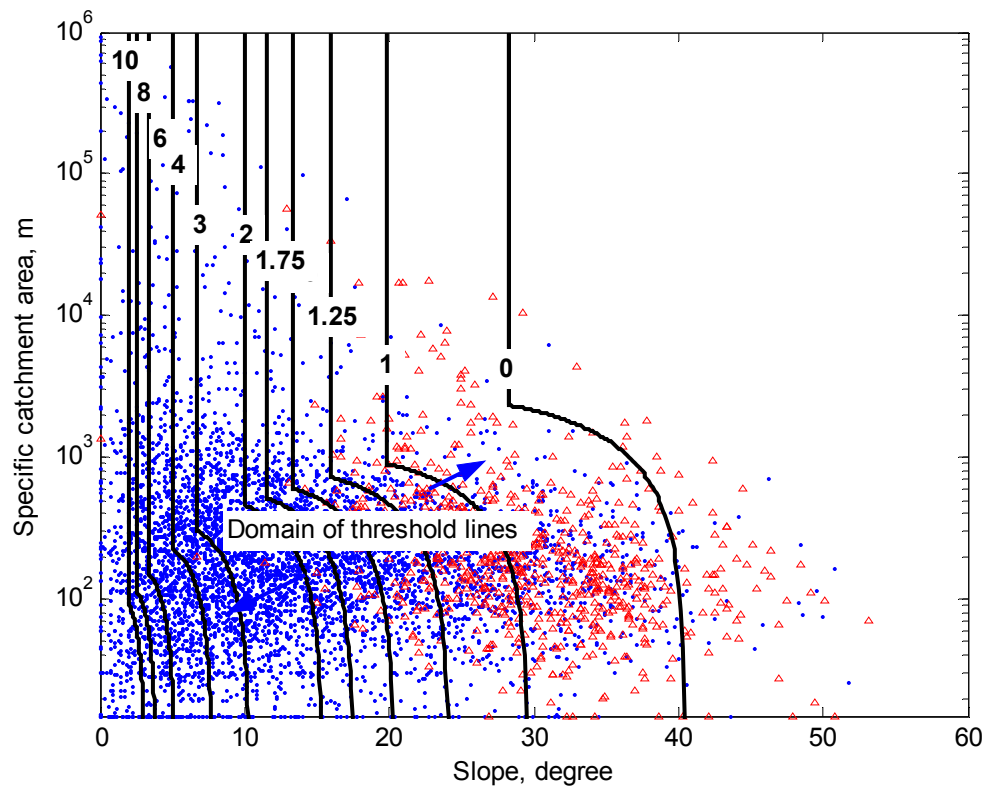
sets, SINMAP consistently gives higher Q values.

Figure A-4. Different threshold lines of SINMAP for some random parameter set.