# UNCERTAINTY IN DIGITAL ELEVATION DATA USED FOR GEOPHYSICAL FLOW SIMULATION

Laércio M. Namikawa[1,3] and Chris S. Renschler[2,3]
*[1]INPE - Inst. Nac. Pesquisas Espaciais, C.P. 515, S.J.Campos, SP, 12201, Brazil; [2]National Center for Geographic Information and Analysis (NCGIA), University at Buffalo—The State University of New York, 301 Wilkeson Quad, Buffalo, NY 14261-0023, USA; [3]Department of Geography, University at Buffalo—The State University of New York, 105 Wilkeson Quad, Buffalo, NY 14261-0023, USA*

Abstract: Elevation information plays a crucial role in simulations of geophysical flows providing the slope and curvature information required to solve a series of differential equations. Digital Elevation Models (DEMs) used as simulation input diverge in resolution, acquisition time and generation methodology. When more than one elevation data set is available, they can be analyzed to provide an estimate of the uncertainty. This work present a method to estimate elevation uncertainty based on at least two elevation data sets. Cluster analysis of uncertainty is also carried out and a strong correlation between uncertainty and terrain curvature is found. The results of this study can be used to modify the available DEMs along ridge and valley lines, to create a quality of DEM information based on the curvature, and to provide a valuable quality measure for the input and the effect on the geophysical flow model results.

Key words: Digital Elevation Model; Uncertainty; Cluster Analysis

## 1. INTRODUCTION

Environmental process models using elevation data derive products to be used in supporting decision makers in major problems, ranging from flooding risk areas, volcanic hazards, man-made structures positioning to ecological studies.

In hazard mapping models, such as simulation models for geophysical mass flow, the use of a certain elevation data set may result in mapping areas with none or low risk for a given hazard. However, these areas may turn into high risk areas with a small difference in the elevation data. In a similar way, a model that predicts values at a particular point to have a magnitude that is below a safe threshold may exceed the threshold to an unsafe value. This phenomenon is very common in cone-shaped volcano landscapes where data resolution and the initiation point for the debris flow determine the flow path down the slopes. In contrast to the high density of ridgelines of a volcano landscape, the surface flow path in watersheds in a non-cone shaped landscape have a low ridge density and the flow ends up almost automatically at a particular watershed outlet along the major drainage pattern. To overcome these limitations, a process-based mathematical model should be able to perform, in this case, hazard maps taking into account the uncertainties that exist in elevation data.

Uncertainty is known to be present in digital elevation models (DEMs) (Hunter and Goodchild 1997; Canters et al. 2002). The sources for elevation data are usually either measurements taken in-situ by surveys or by remote sensing. Any of these measurement methods has an uncertainty attached to it. The measurements are transformed in a common standardized digital data structure and the additional processing in the processing algorithms will add even more uncertainty (Renschler In Press). Elevation data from United States Geological Survey (USGS) DEM (USGS 2003), laser altimetry, radar interferometry (Hodgson et al. 2003), and Global Positioning System contain an uncertainty component.

Unfortunately, there is no spatially distributed information about the elevation uncertainty. If the uncertainty is available at all, it is specified as a global measure, and not a distributed quantity for every location covered by the DEM. For example, 10 and 30 meters resolution DEM from USGS specify the error in terms of a Root Mean Squared (RMS) for few locations (USGS 2003). This uncertainty could and should be estimated when the acquisition instrument, pre-processing, and interpolation methods used to create a particular DEM are available. This information about the DEM production is usually not attached to the data set released.

Given the constraints, a method that defines uncertainty in DEM by taking advantage of the existence of more than one data set for the same region is required. Since more than one observation about a phenomenon in a particular location should lead to a better understanding of the measured value, two or more DEMs from different sources can be used to estimate their uncertainties.

When an uncertainty value is attached to each location inside the coverage region of the DEM, the correlation between uncertainty and morphological feature can be defined. If such correlation exists, uncertainty

for DEMs where only one data is available can be estimated from DEM morphology. Such estimation will only be possible if the differences between the DEMs are not randomly distributed. Therefore, a method to divide regions with random distribution of high differences in elevation from areas with clustered high differences values is essential.

In this paper, a method to estimate the uncertainty in a DEM is presented, and to comprehend the uncertainty, clusters of high values of differences between two DEMs are defined and correlated with DEM morphology.

## 2. UNCERTAINTY ESTIMATION AND CORRELATION ANALYSIS

The proposed uncertainty estimation procedure is divided in three phases: descriptive statistics based on two or more DEMs; definition of a clustering map; and correlation with terrain morphology.

### 2.1 Statistical Analysis of Two or More DEMs

This study suggests that the uncertainty in a DEM can be estimated by taking advantage of the existence of more than one data set for the same region. In this context, elevation at one location is not the true value, but a random variable. From a set of observations at that particular location, a set of data values are recorded, and represent the variable. If only one observation is available, the variable description is poor. However, if more than one observation is available, descriptive statistics will better characterize the random variable and inferential statistics have the potential to estimate the true value within a confidence interval.

The range, standard deviation, and coefficient of variation measures of dispersion and variability can be used to assess the uncertainty. However a comparison of spatial variability requires the use of a measure that does not take into account the magnitude of the elevation, given that the variation in magnitude may be high. Therefore, coefficient of variation is selected to provide the relative measure. Coefficient of variation for every location in a DEM is defined by:

$$CV = \frac{S}{\overline{X}} 100 \qquad 1$$

Where $\boldsymbol{S}$ and $\overline{X}$ are respectively the standard deviation and the mean of the independent elevation measures at the same location. Coefficient of

variation is usually multiplied by 100 to represent a percentage of standard deviation in relation to the mean.

## 2.2 Clustering Analysis

If the uncertainty in the DEM is randomly distributed, measures from the descriptive statistics of elevation values in the region are expected to be similar. To test this hypothesis, the coefficient of variation measure of variability is calculated and analyzed for spatial clustering. If there are clusters of high coefficient of variation, a correlation analysis with the DEM morphology is executed on the clusters.

The cluster detection method is based on finding the significant peaks of a surface that results from the application of a smoothing filter based on Gaussian kernel (Rogerson 2001). The Gaussian kernel is applied on a standardized measure; therefore a z-score statistics is calculated on the coefficient of variation data. The clusters of significant high values of coefficient of variation are found where the smoothed grid value is greater than a critical value.

### 2.2.1 Z-Score Calculation

The z-score is calculated for coefficient of variation data using the mean and standard deviation of the whole area. Then, at every point of the DEM grid, the z-score is calculated by:

$$z_{rc} = \frac{CV_{rc} - \overline{CV}}{S_{CV}} \qquad 2$$

Where $z_{rc}$ is the z-score at row $r$ and column $c$, $CV_{rc}$ is the coefficient of variation in percentage at row $r$ and column $c$, $\overline{CV}$ and $S_{CV}$ are respectively the mean and the standard deviation of coefficient of variation in percentage for the whole region.

### 2.2.2 Gaussian Kernel

The search for significant clusters of high values of coefficient of variation requires the use of Gaussian kernel for smoothing. The standard deviation of the Gaussian kernel is selected based on the desired smoothing, with low values yielding less smoothing and high values generating more smoothing.

The Gaussian kernel is built by applying the weights calculated accordingly to the distance of each neighbor cell of the coefficient of variation grid to the center cell. The weights $w_{ij}$ are calculated using:

$$w_{ij} = \frac{e^{\frac{-d_{ij}^2}{2\sigma^2}}}{\sqrt{\pi\sigma}} \quad 3$$

Where $\sigma$ is the standard deviation of the Gaussian kernel and $d_{ij}$ is the distance from the center cell $i$ to neighbor cell $j$. The weights are used to smooth values at $i$ using:

$$y_i = \frac{\sum_j w_{ij} z_j}{\sqrt{\sum_j w_{ij}^2}} \quad 4$$

Where $y_i$ is the smoothed value at cell $i$, $w_{ij}$ is the weight for the cell at the distance from center cell $i$ to neighbor cell $j$, and $z_j$ is the z-score of cell $j$. Note that since z-score is being used, the distance used here is in the cell units, i.e. if cells are neighbors in the same grid row or column, the distance between them is one.

The selection of the standard deviation parameter of the Gaussian kernel is based on the ability to overcome the random differences, to enhance the clusters, and to obtain an adequate smoothing. To avoid using all the grid cells for each cell under calculation, the maximum distance at which a neighbor cell influences the center cell must also be defined.

### 2.2.3 Selection of Gaussian Kernel Standard Deviation

Gaussian kernel used for smoothing requires the selection of the kernel standard deviation. The selection criteria consider the result of the kernel application, with the goals for the kernel being to remove the random variations and to enhance the clusters of extreme values. The following criteria are used:

- Few clusters with areas smaller than five grid cells. This criterion contemplates the insignificance of small clusters to the overall objective of applying correlation analysis.
- Proportion of the area of clusters in relation to total area of all cells sufficiently small. When Gaussian kernel standard deviation is too high, the clusters will tend to occupy the whole region.

### 2.2.4 Maximum Distance

Considering the case when all the neighbor cells have the same value, the maximum distance to be considered in the Gauss kernel can be defined when a given percentage of the maximum value for a smoothed grid cell is reached. The reached percentage should be high enough to permit the use of small number of cell as the neighborhood instead of the whole region. This approach is possible given that the weights decrease exponentially with distance to the center cell.

The use of the autocorrelation matrix calculated on the coefficient of variation grid allows the use of a distance smaller than defined considering all cells with the same value. The influence of each cell is weighted by the corresponding value in the correlation matrix. Therefore, the maximum distance is calculated by:

$$d_{\max} = d_j \ni \frac{\sum w_j a_j}{2\sqrt{\pi\sigma^3}} 100 > p_{\min} \qquad 5$$

Where $d_{max}$ is the maximum distance to be considered, $d_j$ is the distance of neighbor cell $j$ to center cell $w_j$, is the weight of the Gaussian kernel for cell $j$, $a_j$ is the autocorrelation value at cell $j$, $\sigma$ is the Gaussian kernel standard deviation, and $p_{min}$ is the selected percentage of the maximum value (given by $2\sqrt{\pi\sigma^3}$).

### 2.2.5 Critical Value

The clusters of significant values of coefficient of variation are found based on the critical value $M^*$, that makes the probability of finding a value greater than $M^*$ equal to a selected significance level $\alpha$. The critical value $M^*$ is calculated using the following equation (Rogerson 2001):

$$p\left(\max z_i > M^*\right) = \frac{AM^*\varphi\left(M^*\right)}{4\pi\sigma^2} + \frac{D\varphi\left(M^*\right)}{\sqrt{\pi}\sigma} + \left[1 - \Phi\left(M^*\right)\right] \qquad 6$$

Where $A$ is region area (in cell size units, i.e. one cell area is one), $D$ is the caliper diameter, $\sigma$ is the Gaussian kernel standard deviation, $\varphi$ is the probability density function of the normal distribution, and $\Phi$ is the cumulative distribution function of the normal distribution. For a rectangular grid, $D$ is half of the sum of the rectangle height and width.

The contribution of the third term in **(6)** to the result is sufficiently small to be discarded (Rogerson 2001), resulting in the simplified equation:

$$p\left(\max z_i > M^*\right) = \frac{AM^*\varphi\left(M^*\right)}{4\pi\sigma^2} + \frac{D\varphi\left(M^*\right)}{\sqrt{\pi}\sigma} \quad 7$$

Furthermore the following approximation can be used:

$$M^* = \sqrt{-\sqrt{\pi}\ln\left(\frac{4\alpha\left(1+.81\sigma^2\right)}{A}\right)} \quad 8$$

The approximation of $M^*$ given in **(8)** is valid only if $A$ is smaller than 10000 or if $\sigma$ is not smaller than one (Rogerson 2001). When $A$ is greater than 10000, the approximation can be used if:

$$\sigma_t \Big/ \sqrt{A} > 0.01 \quad 9$$

Where $\sigma_t$ is the total smoothing given by $\sigma_t = \sqrt{\sigma_0^2 + \sigma^2}$, with $\sigma_0$ equal to 10/9 for a square grid.

For most cases using elevation data, $A$ is greater than 10000, and the selected $\sigma$ is expected to be between 1 and 4. In these conditions, the restriction of **(9)** is not satisfied. However, the critical value obtained by using the approximation of **(8)** is only slightly smaller than the value calculated using **(7)**. Moreover, the difference in cluster area size is not significant due to the high incline of the smoothed grid in clustered regions.

## 2.3 Correlation with Terrain Morphology

Although the coefficient of variation indicates uncertainties in the DEM, additional knowledge about the structure of a DEM quality can be extracted if correlation between uncertainty and morphological features of the DEM is obtained. The correlation with uncertainty is searched for the surface geometry represented by slope and curvatures.

### 2.3.1 Correlation with Slope

Slope is defined by the angle of the plane tangent to the surface $z(x,y)$ at a given location $(x_0,y_0)$ and the plane $z=0$. Using the gradient of the surface, the slope at $(x_0,y_0)$ is:

$$Slope = \arctan\sqrt{\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2} \quad 10$$

The partial derivatives ***δz/δx*** and ***δz/δy*** are estimated using a third order finite difference method (Horn 1981), using 8-neighborhood. The method is consistently considered one of the best estimation methods for elevation data (Skidmore 1989; Hodgson 1998; Jones 1998).

Correlation of uncertainty in elevation of a DEM and slope is expected given that a difference on the horizontal location ***(Δx,Δy)*** will lead to:

$$z(x+\Delta x, y+\Delta y) = z(x,y) + \Delta z(\Delta x, \Delta y) \quad 11$$

Where:

$$\Delta z(\Delta x, \Delta y) = \frac{\partial z}{\partial x}\Delta x + \frac{\partial z}{\partial y}\Delta y \quad 12$$

Therefore, the uncertainty in elevation is expected to be correlated to the slope.

### 2.3.2 Correlation with Curvature

A surface curvature is not unique and differs with direction, which is selected accordingly to the application. For this study, the selected directions are the maximum slope direction and the perpendicular to the maximum slope direction. The curvatures on these directions are respectively the profile curvature and tangential curvature (Zevenbergen and Thorne 1987; Mitasova and Hofierka 1993; Schmidt et al. 2003). Tangential curvature is preferred here instead of contour curvature, which has the direction of the contour line as the direction of interest, given that the contour curvature is more sensitive to errors in elevation (Schmidt et al. 2003) and converges to infinite at locations where the contour line radius tends to zero, such as at the top of hills or bottom of pits (Shary et al. 2002).

The profile curvature is given by:

$$k_p = \frac{\frac{\partial^2 z}{\partial x^2}\left(\frac{\partial z}{\partial x}\right)^2 - 2\frac{\partial^2 z}{\partial xy}\frac{\partial z}{\partial x}\frac{\partial z}{\partial y} + \frac{\partial^2 z}{\partial y^2}\left(\frac{\partial z}{\partial y}\right)^2}{\left(\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2\right)\left(\sqrt{\left(\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2 + 1\right)^3}\right)} \quad 13$$

And the tangential curvature is given by:

$$k_t = \frac{\frac{\partial^2 z}{\partial x^2}\left(\frac{\partial z}{\partial y}\right)^2 - 2\frac{\partial^2 z}{\partial xy}\frac{\partial z}{\partial x}\frac{\partial z}{\partial y} + \frac{\partial^2 z}{\partial y^2}\left(\frac{\partial z}{\partial x}\right)^2}{\left(\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2\right)\left(\sqrt{\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2 + 1}\right)} \qquad 14$$

Where $\delta z/\delta x$, $\delta z/\delta y$, $\delta^2 z/\delta x^2$, $\delta^2 z/\delta y^2$, $\delta^2 z/\delta xy$ are the partial derivatives of first and second order of the surface $z(x,y)$. The curvatures calculated by **(13)** and **(14)** have positive values for concavity and negative values for convexity.

Elevation differences between the DEMs are expected to be correlated to changes in slope at its maximum and orthogonal to maximum directions given that elevation values of DEM are obtained at a resolution that may not capture the rapid changes in the gradient of elevation. Therefore, in regions where curvatures are close to zero, elevations differences should be low, while where convexity is high, elevation values are expected to be in average higher for high resolution DEMs compared to low resolution ones. Conversely, areas with high concavity, elevation values are expected to be, in average, lower for high resolution DEMs compared to low resolution ones.

## 3. CASE STUDY – COLIMA VOLCANO

Models of mass flows related to volcanic activities require terrain elevation data. The model used in the TITAN simulation (Patra et al. In Press) treats flows as averaged granular flows governed by Coulomb type interactions where elevation information plays a crucial role. The elevation model used is considered to be a perfect representation of the topography. Estimates of the uncertainty in a DEM could be used in the simulation model if they were available. Volcan de Colima is a validation site for the TITAN mass flow model and its DEM is used in this test case, comprising an area of approximately 38 squared kilometers, between coordinates 635030, 2144510 and 665990, 2170250 (easting and northing in meters, respectively), in Universal Transverse of Mercator (UTM) map projection.

For the TITAN mass flow simulation at Volcan de Colima, a DEM available at Arizona Image Archive (www.aria.arizona.edu) was used. A second DEM was obtained from the Shuttle Radar Topography Mission (SRTM) data (available at ftp://edcsgs9.cr.usgs.gov/pub/data/srtm/North_ America/3arcsec/). The DEM from Arizona Image Archive will be named ARIADEM, and the DEM from SRTM, SRTMDEM, from here on.

To allow the analysis of differences between the DEMs, they must use the same mapping references obtained from a published map. For the Colima site, four maps (Comala (INEGI 1999b), San Gabriel (INEGI 1999d), Cuauhtemoc (INEGI 1999c), and Ciudad Guzman (INEGI 1999a)), in 1:50000 scale using UTM projection and ITRF92 datum, were used as base maps. ARIADEM is available in UTM projection with NAD27 datum and SRTMDEM is available in a 3-arc second grid using WGS84 datum. Thus, both DEMs were reprojected and resampled to a 90 meter resolution grid in UTM with ITRF92 datum from their original projections. SRTMDEM contains grid points with no values, where the interferometric process of generating elevation data failed due to low reflectance, shadowing, or layover effects.

## 3.1 Coefficient of Variation between ARIADEM and SRTMDEM

The coefficient of variation statistics is used to analyze the difference between the two DEMs since it provides a measure that does not take into account the magnitude of the elevation, which ranges from elevation in the 671 and 4212 meters in the study area. Coefficient of variation is calculated using **(1)** and requires standard deviation and mean, which are calculated on all grid points of the DEM, except where SRTMDEM has no values.

## 3.2 Clustering Analysis

The cluster detection method requires the generation of the standardized measure z-score on the coefficient of variation grid. On the created z-score grid of values, the Gaussian kernel is applied using discrete convolution using a mask with values that corresponds to the kernel weights. The mask size is given by the maximum distance to be considered defined using the z-score grid autocorrelation matrix. The Gaussian kernel standard deviation and the critical value, used to define clustering, are selected based on the ability to undermine random variations and to enhance the clusters of high differences.

### 3.2.1 Z-Score

The standardized measure z-score for the coefficient of variation is calculated for every point of the grid using the mean (0.7680 meters) and standard deviation (0.7613), using **(2)**.

### 3.2.2 Maximum Distance

The autocorrelation matrix of the z-score grid is used to predict the maximum distance to be considered in the Gaussian kernel. Table 1 presents the distances for various Gaussian kernels and the distances at which 90%, 95% and 99% of the maximum is reached considering the autocorrelation matrix and without considering it.

*Table 1.* Distances to reach percentage of the maximum value for various Gaussian kernel σ, considering the autocorrelation matrix (a) and without considering it (b).

| σ | Distance (in grid cell unit) | | | | | |
|---|---|---|---|---|---|---|
| | 90% | | 95% | | 99% | |
| | (a) | (b) | (a) | (b) | (a) | (b) |
| 1 | 2.23 | 2.23 | 2.24 | 2.24 | 3.00 | 3.16 |
| 2 | 4.00 | 4.24 | 4.47 | 5.00 | 5.83 | 6.08 |
| 3 | 5.83 | 6.40 | 6.71 | 7.28 | 8.54 | 9.22 |
| 4 | 8.00 | 8.54 | 9.06 | 9.90 | 11.40 | 12.08 |

The selected maximum distance is the one when 95% of the maximum value for the Gaussian kernel is reached, given that the 95% threshold for the distance, considering the autocorrelation matrix, yields a convolution matrix that is capable of being handled by a neighborhood operation in map algebra, available in most Geographic Information Systems tools.

### 3.2.3 Selection of Gaussian Kernel Standard Deviation and Critical Value

The z-score grid of coefficient of variation was smoothed using Gaussian kernel with various standard deviation to select the most appropriate standard deviation for detecting clusters. Using the critical values $\boldsymbol{M}^*$, the smoothed z-score grid of coefficient of variation is divided in regions that exceed $\boldsymbol{M}^*$. Table 2 presents the critical values for Gaussian kernel with various standard deviation, the critical values $\boldsymbol{M}^*$ from **(7)** and **(8)**, the number of grid cells exceeding $\boldsymbol{M}^*$, the difference in number of cells using $\boldsymbol{M}^*$ from **(7)** and **(8)**, the percentage of regions smaller than 5 cells, and the percentage of clusters area to total area of the grid. The total number of cells in the study region is 98384.

Table 2 shows that the use of Gaussian kernel with standard deviation equal to one yields too many small regions. The proportion of cluster areas to total using values three and four for standard deviation of the Gaussian kernel is over twenty percent, which is considered too high. Therefore, the standard deviation selected here is two, with the critical value being selected by **(8)**, given that the difference when compared to using **(7)** is not

significantly different, and **(8)** does not require the use of a detailed *z*-table (Rogerson 2001).

*Table 2.* Clusters characteristics for various Gaussian kernel and critical values M*.

| σ | M* using | Clusters Above M* | | | |
|---|---|---|---|---|---|
| | | # of Cells | Underestimated % | Small Region (%) | Prop. Total (%) |
| 1 | (8): 5.032 | 5600 | 87.91 | 33.2 | 5.7 |
| | (7): 4.709 | 6370 | | 28.8 | 6.5 |
| 2 | (8): 4.739 | 13480 | 95.45 | 12.6 | 13.7 |
| | (7): 4.546 | 14122 | | 12 | 14.4 |
| 3 | (8): 4.562 | 19435 | 97.68 | 4.4 | 19.8 |
| | (7): 4.414 | 19896 | | 4.4 | 20.2 |
| 4 | (8): 4.436 | 23470 | 98.11 | 2.6 | 24.4 |
| | (7): 4.308 | 23921 | | 5.4 | 24.3 |

## 3.3 Correlation with Terrain Morphology

Tests of linear correlation between coefficient of variation grid and each of the terrain morphology measures indicates that the highest correlation is between slope and coefficient of variation, as indicated in Table 3, where Bonferroni adjustment for spatial autocorrelation was not used, although the rank of correlation would not have changed with the adjustment.

*Table 3.* Correlation between coefficient of variation and terrain morphology measures, using all grid cells.

| Terrain Morphology Parameter | Correlation Coefficient | Correlation t-score, N=95866, Critical t = ±1.98 for 95% significance |
|---|---|---|
| Slope | 0.1739 | 54.70 |
| Profile Curvature | 0.0613 | 19.02 |
| Tangential Curvature | -0.0131 | -4.067 |

Correlations with curvatures are distorted given that slopes and coefficient of variation are all positive and curvatures can be positive and negative. By using difference between the two DEMs, a measure that also has positive and negative values, the bias against curvature is eliminated. Table 4 presents the linear correlation coefficients between differences in elevations and each of the terrain morphology parameters.

Although the correlation with slope is higher than the correlation with curvatures for grids of coefficient of variation and of difference in elevation, analysis of correlation in the clusters of high coefficient of variation provide information more significant on the uncertainty of the DEM.

*Table 4.* Correlation between differences in elevation and terrain morphology measures, using all grid cells.

| Terrain Morphology Parameter | Correlation Coefficient | Correlation *t*-score, N=95866, Critical t = ±1.98 for 95% significance |
|---|---|---|
| Slope | -0.1284 | -40.11 |
| Profile Curvature | -0.0486 | -15.08 |
| Tangential Curvature | 0.0738 | 22.93 |

The clustered areas are used to define correlation of the differences with the surface geometry represented by slope and curvatures to find which parameter of the DEM contributes most for high uncertainty. For each cluster of high coefficient of variation, the mean difference between ARIADEM and SRTMDEM, the mean slope, and mean curvatures from ARIADEM are obtained. Table 5 shows linear correlation coefficients between the mean differences and slope, profile curvature, and tangential curvature, with the respective *t*-score statistics.

*Table 5.* Correlation between mean differences in elevation and mean terrain morphology measures, using clusters.

| Terrain Morphology Parameter | Correlation Coefficient | Correlation t-score, N=117, Critical t = ±1.98 for 95% significance |
|---|---|---|
| Slope | -0.1175 | -1.269 |
| Profile Curvature | -0.1096 | -1.183 |
| Tangential Curvature | 0.3622 | 4.168 |

The results from Table 5 indicates that tangential curvature has a significant linear correlation with difference in elevation between ARIADEM and SRTMDEM, when comparing the mean for the cluster regions, while the other surface morphology parameters have no significant correlation with mean difference. Figure 1 presents the scatter plot of differences in function of tangential curvature.

The regression line of the correlation is given by:

$$Diff = 23020 * k_t - 33$$

Where ***Diff*** is the difference in elevation between ARIADEM and SRTMDEM and $\boldsymbol{k_t}$ is the tangential curvature.

The equation of difference in function of tangential curvature is used to define the expected difference for each cluster region. The map showing the regions with high difference between the observed and expected differences is presented in Figure 2.

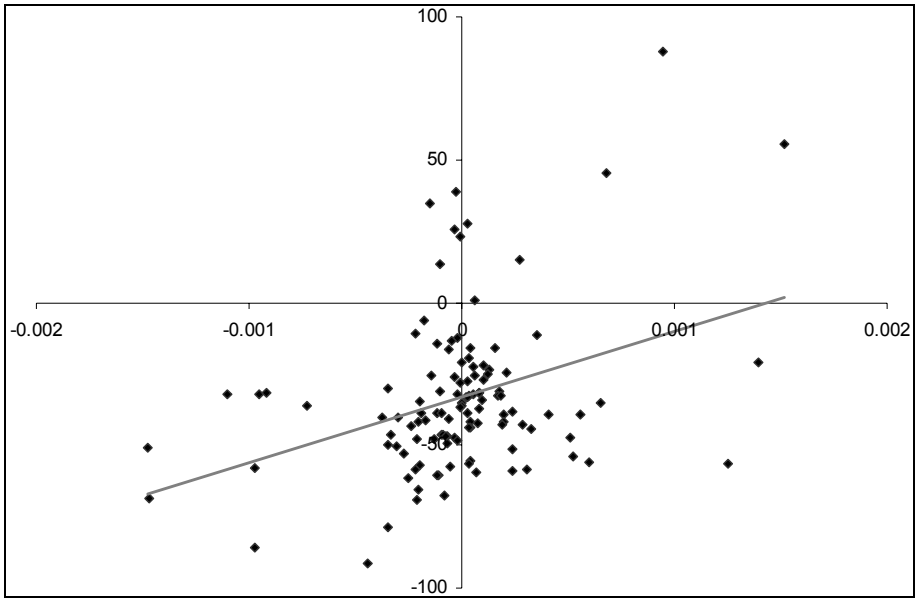


*Figure 1.* Scatterplot of differences in elevation between ARIADEM and SRTMDEM (in the vertical axis) in function of tangential curvature (in the horizontal axis).

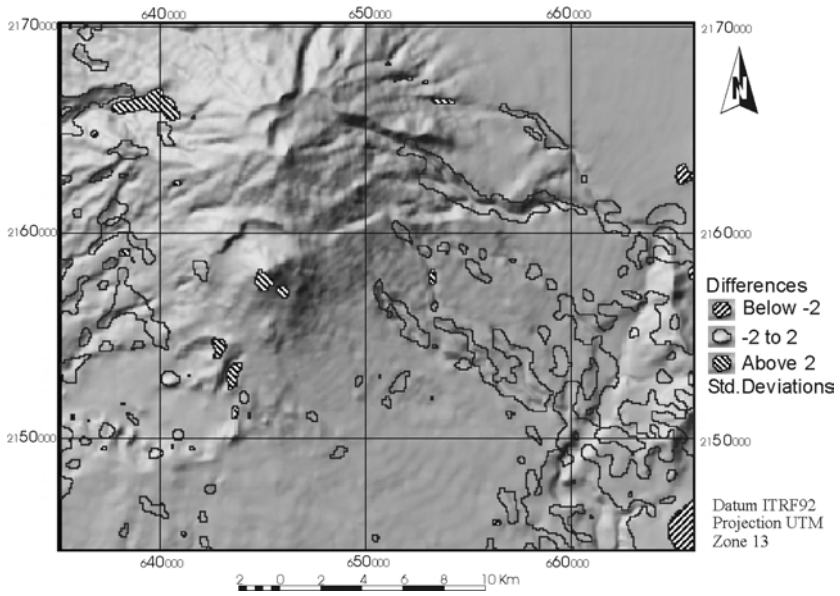


*Figure 2.* Clusters of high coefficient of variation, with differences between predicted and observed differences in elevation for each cluster. Background image is a pseudo-shaded image of the ARIADEM.

# 4. DISCUSSIONS

Correlation between coefficients of variation from DEM and slope is significant when the whole grid is considered. However, a significant source of uncertainty is not considered, as it can be asserted through the visualization of the observed coefficient of variation and the difference between observed and expected coefficient of variation, where they look similar. A further analysis using scatter plot of the differences on elevation in function of slope supports the previous finding. The correlation is statistically significant but there are other sources of uncertainty that are revealed only by the clusters analysis.

## 4.1 Discussions on Clusters

The correlation results on clusters show that the differences between ARIADEM and SRTMDEM are linearly correlated with tangential curvature inside the 117 clustered area units, when average of the area is considered. The statistical significance of the correlation using *t*-statistic is 99.99994% (the calculated t-statistic is 4.168).

To predict differences between ARIADEM and SRTMDEM at each cluster, the regression line is used, resulting in an average difference of 0.85 meters, with standard deviation of 25.51 meters. Of the 117 clusters, nine have differences between predicted and observed greater than two standard deviations and an analysis of the spatial context of the clusters is presented next.

### 4.1.1 Clusters with High Differences between Predicted and Observed Values

Three of the nine clusters with differences between predicted and observed higher than two standard deviations are located at the edges of the DEM, indicating that the cluster analysis is not efficient at these regions. This result is expected since the Gaussian kernel relies on the neighborhood that cannot be correctly defined at the limits of the DEM.

Other three of the nine clusters have areas smaller than 5 grid cells, indicating that clusters must have a minimum size to provide useful analysis, through a representative average values.

### 4.1.2 Other Clusters with High Differences between Predicted and Observed Values

There are seventeen clusters with differences between one and two standard deviations. From this set, four are located close to the Colima Volcano, in an unstable area with prevailing block-and-ash flow (Saucedo et al. 2002), and, since there is a time difference between acquisitions of data, the underperformance of the correlation may be related to erosion and deposition of material in the region of the cluster.

## 5. SUMMARY

The importance of elevation information as model input and the effects on simulating geophysical flows for hazard mapping demonstrates the need for knowledge and understanding of dealing with uncertainties. We present an estimation procedure to describe uncertainties in Digital Elevation Models by taking advantage of the availability of an additional elevation data set for the same area. Since only two samples of elevation are available, the coefficient of variation is used here as an uncertainty measure.

A cluster analysis of uncertainty is also carried out to determine the significant regions of high coefficient of variation using a cluster detection method. This approach is required given that the correlation of the uncertainty measure with slope considering individual cell of the DEM grid does not provide all the knowledge about the DEM parameters that affect its uncertainty.

Correlation analysis considering the average difference between elevations and average curvature within a cluster region is the method proposed here. Using this approach, the influence of the slope and random variations between DEMs are diminished and a strong correlation between averaged uncertainty and averaged terrain tangential curvature is found. Given that the extreme values of tangential curvature are found along ridge and valley lines, the available DEM can be modified in these regions by incorporating additional elevation information.

With the cluster analysis, the focus is placed on regions where estimated uncertainty is high and clustered allowing a study that goes beyond the expected influence of slope on elevation uncertainty. Using the results of the cluster analysis, one can evaluate either if additional information has to be collected when the cluster region coincides with the study area, or if there is no need to spend resources on additional data gathering.

The uncertainty of a DEM can be defined when an additional DEM is available. The importance of the existence of a global coverage DEM

provided by the SRTM dataset is highlighted in this paper, even with the limitations of having gaps in the data and the relatively lower resolution of three arc-second.

## ACKNOWLEDGEMENT

The National Science Foundation is acknowledged for providing funds through the research project "ITR/AP+IM: Information Processing for Integrated Observation and Simulation Based Risk Management of Geophysical Mass Flows", ITR-0121254.

## REFERENCES

Canters, F., W. D. Genst and H. Dufourmont (2002). "Assessing effects of input uncertainty in structural landscape classification." International Journal of Geographical Information Science **16**(2): 129-149.

Hodgson, M. E. (1998). "Comparison of Angles from Surface Slope/Aspect Algorithms." Cartography and Geographic Information Systems **25**(3): 173-185.

Hodgson, M. E., J. R. Jensen, L. Schmidt, S. Schill and B. Davis (2003). "An evaluation of LIDAR- and IFSAR-derived digital elevation models in leaf-on conditions with USGS Level 1 and Level 2 DEMs." Remote Sensing of Environment **84**: 295–308.

Horn, B. K. P. (1981). "Hill Shading and the Reflectance Map." Proceedings of the IEEE **69**(1): 14-47.

Hunter, G. J. and M. F. Goodchild (1997). "Modeling the uncertainty of slope and aspect estimates derived from spatial databases." Geographical Analysis **29**(1): 35-49.

INEGI, Ed. (1999a). Ciudad Guzman E13B25. Mexico City, Mexico, INEGI, Instituto Nacional de Estadistica, Geografia e Informatica.

INEGI, Ed. (1999b). Comala E13B34. Mexico City, Mexico, INEGI, Instituto Nacional de Estadistica, Geografia e Informatica.

INEGI, Ed. (1999c). Cuauhtemoc E13B35. Mexico City, Mexico, INEGI, Instituto Nacional de Estadistica, Geografia e Informatica.

INEGI, Ed. (1999d). San Gabriel E13B24. Mexico City, Mexico, INEGI, Instituto Nacional de Estadistica, Geografia e Informatica.

Jones, K. H. (1998). "A comparison of algorithms used to compute hill slope as a property of the DEM." Computers & Geosciences **24**(4): 315-323.

Mitasova, H. and J. Hofierka (1993). "Interpolation by Regularized Spline with Tension .2. Application to Terrain Modeling and Surface Geometry Analysis." Mathematical Geology **25**(6): 657-669.

Patra, A. K., A. C. Bauer, C. C. Nichita, E. B. Pitman, M. F. Sheridan, M. Bursik, B. Rupp, A. Webb, A. Stinton, L. M. Namikawa and C. Renschler (In Press). "Parallel Adaptative Numerical Simulation of Dry Avalanches over Natural Terrain." Journal for Volcanology and Geothermal Research **In Press, Corrected Proof**.

Renschler, C. S. (In Press). "Scales and uncertainties in using models and GIS for volcano hazard prediction." Journal of Volcanology and Geothermal Research **In Press, Corrected Proof**.

Rogerson, P. A. (2001). "A Statistical Method for the Detection of Geographic Clustering." Geographical Analysis **33**(2): 215-227.

Saucedo, R., J. L. Macias, M. I. Bursik, J. C. Mora, J. C. Gavilanes and A. Cortes (2002). "Emplacement of pyroclastic flows during the 1998-1999 eruption of Volcan de Colima, Mexico." Journal of Volcanology and Geothermal Research **117**(1-2): 129-153.

Schmidt, J., I. S. Evans and J. Brinkmann (2003). "Comparison of polynomial models for land surface curvature calculation." International Journal of Geographical Information Science **17**(8): 797-814.

Shary, P. A., L. S. Sharaya and A. V. Mitusov (2002). "Fundamental quantitative methods of land surface analysis." Geoderma **107**(1-2): 1-32.

Skidmore, A. K. (1989). "A comparison of techniques for calculating gradient and aspect from a gridded digital elevation model." International Journal of Geographical Information Systems **3**(4): 323 - 334.

USGS (2003). USGS Digital Elevation Model Data. **2003**.

Zevenbergen, L. W. and C. R. Thorne (1987). "Quantitative-Analysis of Land Surface-Topography." Earth Surface Processes and Landforms **12**(1): 47-56.