JOURNAL OF THE AMERICAN WATER RESOURCES ASSOCIATION
VOL. 38, NO. 6 AMERICAN WATER RESOURCES ASSOCIATION DECEMBER 2002

SUBWATERSHED SPATIAL ANALYSIS TOOL: DISCRETIZATION OF A DISTRIBUTED HYDROLOGIC MODEL BY STATISTICAL CRITERIA[1]

S. Haverkamp, R. Srinivasan, H. G. Frede, and C. Santhi[2]

ABSTRACT: Complex hydrologic models, designed for simulating larger watersheds, require a huge amount of input data. Most of these models use spatially distributed data as inputs. Spatial data can be aggregated or disaggregated for use as input to a model, which can impact model outputs. Although, it is efficient to minimize data redundancy by aggregating the spatial data, upscaling reduces the detail/resolution of input information and increases model uncertainty. On the other hand, a large number of model inputs with high degrees of disaggregation take more computer time and space to process. Hence, a balance between striving for a maximum level of aggregation and a minimum level of information loss has to be achieved. This study presents a definition of an appropriate level of discretization, derived by establishing a relationship between a model's efficiency and the number of subwatersheds modeled. An entropy based statistical approach/tool called SUbwatershed Spatial Analysis Tool (SUSAT) was developed to find an objective choice of an appropriate level of discretization. The new approach/tool was applied to three watersheds, each representing different hydrologic conditions, using a hydrologic model. Coefficients of efficiency and entropy estimated at different levels of discretization were used to validate the success of the new approach.
(KEY TERMS: watershed modeling; geographic information system; data preprocessing; spatial discretization; heterogeneity; entropy.)

INTRODUCTION AND OBJECTIVES

Geographic Information System (GIS) based spatial data are widely used for hydrologic modeling. Many hydrologic models, such as Agricultural Non Point Source (AGNPS) (Young *et al.*, 1994) and Soil and Water Assessment Tool (SWAT) (Arnold *et al.*, 1993), work on equal sized, grid based input data. Homogeneous areas of watersheds, however, result in substantial input data redundancy. Decreasing redundancy does not influence model efficiency but does save calculation time and disk space requirements. Saved calculation time enables researchers to implement more complex and highly parameterized approaches to modeling, which, in turn, provides a chance to reduce model uncertainty (Grunwald, 1997). On the other hand, it should be noted that reducing redundancy of input data by using a high degree of aggregation of those inputs that are spatially distributed can also affect model efficiency (Krysanova *et al.*, 1998). For example, upscaling reduces the detail/resolution of input information and increases model uncertainty. Hence, a balance must be achieved between maximum level of aggregation and minimum level of information loss. It is essential to capture enough of the heterogeneity of the input data in order to obtain optimal results.

Each hydrologic model is developed based on certain watershed delineation concepts. A common technique of delineation is to build clusters of cells, called Hydrologic Response Units (HRUs) (Leavesley *et al.*, 1983; Maidment, 1991; Krysanova *et al.*, 1998). In general, HRUs are defined by combining spatial attributes relevant to the model into discrete spatial features. The exact definition of HRUs varies depending on the model's conceptualization. In the case of SWAT, HRUs are defined by unique land use and soil combination. All grids grouped within one HRU should be as homogeneous as possible in terms of their hydrologic properties. This kind of distributed modeling approach for HRU implies a detailed transfer of soil and land use input data but does not consider the topography of a watershed. However, the direction and amount of water yield from one

[1]Paper No. 00036 of the *Journal of the American Water Resources Association.* **Discussions are open until June 1, 2003.**
[2](Haverkamp and Frede) Department of Agricultural Ecology and Natural Resources Management, University of Giessen, Heinrich-Buff-Ring 26-32, 3, D-35390 Giessen, Germany; (Srinivasan and Santhi) Blackland Research Center, Agricultural Experiment Station, 720 East Blackland Road, Temple, Texas 76502 (E-Mail/Srinivasan: Srin@brc.tamus.edu).

modeling unit to another depend on the topographic parameters such as hillslope and aspect. Since these parameters are not considered in the HRU approach, it is not possible to establish an appropriate routing structure through the investigated watershed.

To overcome this disadvantage, a strategy of subwatershed delineation is developed. Topography serves as the criterion for subwatershed extraction and also for slope and aspect (Martz and Garbrecht, 1998). Subwatershed delineation is entirely independent of land use and soil data. In contrast to the HRU approach, in the case of subwatershed delineation, lateral flow paths between subwatersheds are considered, which is essential for water yield calculations depending on the routing.

To minimize model uncertainty induced by input data aggregation, an "appropriate" number of subwatersheds for any watershed has to be identified. It is common to describe the watershed by different numbers of subwatersheds, keeping all other model input parameter values unchanged. Although there is no clear definition of an appropriate number of subwatersheds, the aim of this labor-intensive practice is to find the smallest number required for obtaining good and stable simulation results. This less than satisfactory procedure should be replaced by a statistically based method, which would be applicable to the input data before the first model run. The new approach should be able to:

- calculate a level of spatial aggregation, which will not lead to a significant loss of information;
- identify those configurations of modeling units that satisfy the demanded level of spatial aggregation; and
- consider both subwatersheds and HRUs approaches in delineation wherever appropriate.

The objective of this study is to develop a statistically based data preprocessing tool that helps to identify the level of spatial discretization required that optimizes for both accuracy of modeling results and the reduction of redundancy in a model's spatial input data.

STUDY AREA AND INPUT DATA

The Weiherbach and Dietzhoelze watersheds in the Federal Republic of Germany (FRG) and the Bosque River watershed in Texas, USA, were selected for this study. These watersheds differ widely in climatic and hydrologic conditions and sizes (Table 1). They were simulated for a period of five years at a daily time step. Measured stream flow data available at the outlets of these watersheds were used for validation.

The land use data for the German watersheds were obtained from the classified Landsat thematic map images (Noehles and Frede, 1998). Measured soil data of the German watersheds were made available by the Hessian Department of Soil Research (Frede and Bach, 1999). Elevation data were provided by the Hessian Land Survey Office. Land use, soil, and elevation data for the Bosque Watershed were obtained from the U.S. Department of Agriculture-National Resources Conservation Service (USDA-NRCS).

THEORY

In general, spatial modeling units such as subwatersheds or HRUs are parameterized by one representative value for each input parameter such as land use or soil. Like a stencil, a map of the spatial modeling units is laid over each map of the input parameter to find the representative value by the dominant category. The input parameter can be nominal data (continuous across grids such as land use and soil) or interval data (discrete for each grid like topography).

TABLE 1. Statistical Information of the Study Areas.

	Weiherbach (FRG)	**Dietzhoelze (FRG)**	**Bosque (USA)**
Watershed Area (km^2)	6.3	81.7	4,297.0
Grid Area (m^2)	625.0	625.0	40,000.0
Soil Category Frequency (> 10 percent)*	3	3	1
Land Use Categories	10	7	5
Land Use Category Frequency (> 5 Percent)*	5	5	2
C**	0.244	0.593	0.847

*User defined threshold frequency: 10 percent for soil categories and 5 percent for land use categories.
**Pearson's corrected contingency coefficient.

In the case of nominal data, input parameters can be easily derived from GIS grids by calculating average values for each spatial modeling unit. However, defining the representative values for nominal scaled parameters such as land use or soil categories are more difficult. Usually, the dominant category of the nominal scaled parameter is regarded as representative and thus chosen for subwatershed or HRU parameterization. However, it is not necessary that an increasing number of subwatersheds will increase the heterogeneity of the nominal scaled input parameter values for the entire watershed. The required number of subwatersheds for an appropriate level of spatial discretization depends on two important characteristics of the nominal data, namely, the number of categories and their distribution within the watershed. This indicates that sensitivity of the hydrologic model to spatial variation of the nominal data is important. However, all types of nominal data are not equally important to the model.

For illustration, Figure 1 shows the characteristic behavior of the coefficient of efficiency (Nash and Sutcliffe, 1970) obtained from simulation of the Dietzhoelze watershed using the SWAT model. A line fitted to the distribution of calculated efficiency values reveals that model efficiency becomes more constant and approaches asymptotically a particular value as the number of subwatersheds increases. Thereafter, the rate of improvement in efficiency declines with an increasing number of subwatersheds. A final, almost constant value of efficiency is reached only with a threshold number of subwatersheds. This is in agreement with the conclusions of S. Mamillapalli (1997, Ph.D. Dissertation, Purdue University, unpublished) which shows that a threshold level of disaggregation existed for each watershed beyond which the model accuracy could not be improved.

The deviation of calculated efficiency values from the fitted line of average efficiency values decreases as the number of subwatersheds increases. Average model efficiency rises to a limit at about 75 subwatersheds and remains constant beyond this threshold. The reason for the bias of efficiency around the fitted line of average efficiency can also be interpreted from Figure 1. In the case of interval data, average input parameter values are almost constant for all levels of discretization. However, in the case of nominal data, input parameter values vary at different levels of discretization. At a lower level of discretization, there are fewer subwatersheds and therefore, there is a small range of input categories. As a result, model efficiency is often poor. As the level of spatial discretization increases, the range of input categories also increases, resulting in better model efficiency. When the pattern of spatial modeling units is able to represent the combination of input parameter categories, the maximum efficiency is obtained and any further increase in number of modeling units will not remarkably change modeling results. Therefore, the limb of the fitted line of coefficient of efficiency (Figure 1) becomes constant indicating the spatial modeling units representing the constant portions (averages) of input parameter categories. Thus, an "appropriate" level of discretization can be defined by the maximum value of average model efficiency as is shown by the constant limb of the balanced line (Figure 1). The absolute value of the maximum is of no importance to this definition. The amount of bias around the fitted line of model efficiency is assumed to be induced by the variation (heterogeneity) in parameterizing the modeling units by their nominal scaled properties. If the modeling units were parameterized only by interval scaled properties, an aggregation would lead to a continuous (nonlinear) rising or falling function of model efficiency against spatial discretization.

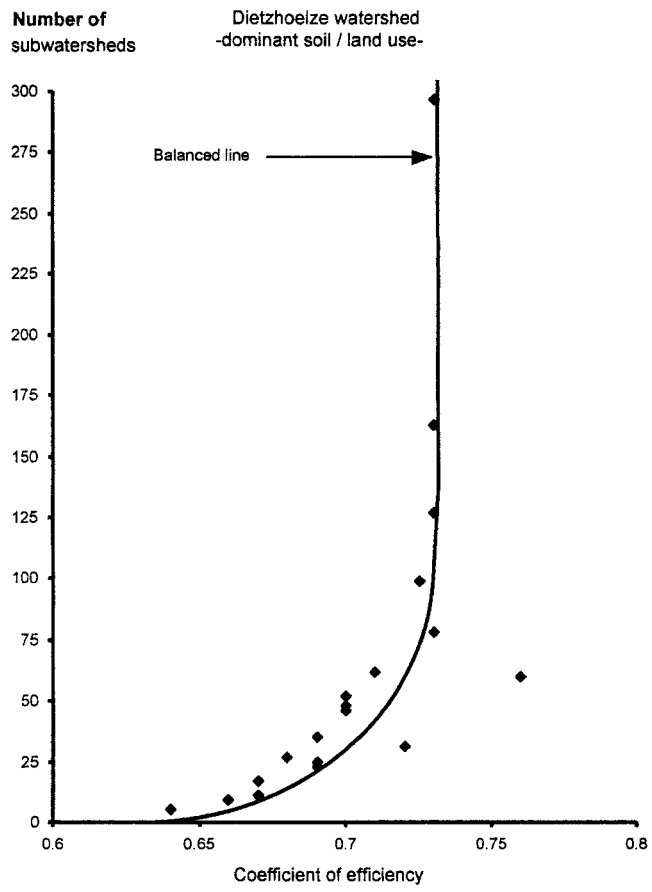

Figure 1. Coefficient of Efficiency Estimated Using SWAT at Different Levels of Subwatershed Discretizations With Dominant Land Use and Soil HRU Approach for the Dietzhoelze Watershed.

In Figure 1, model efficiency increases as the level of discretization increases. Model efficiency depends on the heterogeneity of spatial input parameters, the level of discretization at which the constant limb of the balanced line (Figure 1) is reached. When subwatersheds are parameterized by their dominant categories of input data, the spatial heterogeneity of input layers is the most important factor in determining an appropriate level of subwatershed discretization.

The heterogeneity of spatial data can be quantified by an entropy function (Shannon and Weaver, 1949; Ventsel, 1964; Krasovskaia, 1997).

$$H(p_1, p_2, ..., p_k) = -\sum_{i=1}^{k} p_i \log_2 p_i \quad (1)$$

where H is entropy; p_1, p_2, . . ., p_k are the probability; p_1, p_2, . . ., p_k are ≥ 0 for all i and $\sum_{i=1}^{k} p_i = 1$; k is the number of categories; and i is the index of category.

The entropy function describes the uniformity of a probability distribution. It is a continually rising function of k, if probabilities (p_i) are all equal. H reaches its maximum value ($\log_2$ k) if all p_i equal 1/k and H becomes zero, the minimum possible value, if p_i is equal to 1. Entropy becomes the maximum where the heterogeneity of spatial input data is high, that is, where the maximum model efficiency is expected. Entropy is calculated from probabilities of the categories in the map. The probability of each category is estimated by percentage of frequency (the number of occurrence of each category).

In recent years, many researchers have attempted to use the entropy concept for analyzing spatial variability (Grunwald, 1997; USDA-ARS, 1997). The USDA-ARS laboratory at Tektran, Arizona, conducted a study on evaluating the performance of the hydrologic model, KINmatic simulation of catchment runoff and EROSion processes (KINEROS) (Woolhiser *et al.*, 1990) applying the entropy concept as a function of precipitation network density in the Walnut Gulch Experimental Watershed in Tombstone, Arizona. The study concluded that the maximum efficiency of the model was obtained with seven rain gages and the efficiency declined when the raingages were increased from seven (USDA-ARS, 1997). Grunwald (1997) conducted a validation of the event based modified AGNPS model in four German watersheds with contrasting landscape characteristics. The spatial variability of various GIS data was analyzed using the entropy concept. The conclusions showed that the model output data were most sensitive to loss of topographic information when aggregated due to high heterogeneity.

The advantage of the entropy approach is that it can be used to study the effects of several spatial data, either individually or in combination, by a single measure. As spatially distributed model inputs usually consist of more than one type of input parameter information or GIS layer, entropy has to be combined. Equation (2) shows an example for two input layers (L_1, L_2). Combined entropy can be calculated by the sum of $H(L_1)$ and $H_{L1}(L_2)$.

$$H(L_1, L_2) = H(L_1) + H_{L1}(L_2) \quad (2)$$

where L_1, L_2 are layer1 and layer2; $H(L_1)$ is the entropy of L1; and $H_{L1}(L_2)$ is the entropy of L_2 if L_1 is known.

In the case of complete independence between two layers, combined entropy [$H(L_1, L_2)$] is the sum of both entropy values $H(L_1)$ and $H(L_2)$, because, when categories of layer L1 are independent of layer L2, then $H_{L1}(L_2)$ equals $H(L_2)$, the entropy of layer 2. If input layers are not completely independent, $H(L_1, L_2)$ is less than the sum of the single entropy values. $H_{L1}(L_2)$ is a measure of entropy about L_2 if L_1 is known (similar to conditional probability). Land use and soil data are typical grid-based model input data and serve as examples to illustrate the essence of Equation (2). If the land use layer is correlated with soil layer, the randomness of the land use associated with a location is reduced if the soil attribute is already known and hence, the chance of determining land use is higher. Therefore, the combined entropy of land use and soil layers is reduced.

Each modeling unit is represented in the model by only one value for each input parameter. The complete information of all input parameters can be transferred to the modeling units, if there is only one combination of input parameter values for each modeling unit. If areas of modeling units are homogeneous with respect to input parameter values, model efficiency will not be improved by further subdividing the watershed into more spatial modeling units. It is assumed that if the corresponding entropy of spatial modeling units (subwatersheds or subwatersheds with HRUs) equals the entropy of the input parameter values, model efficiency will be maximized (Figure 1). Therefore, the identification of an appropriate level of discretization leads to the postulation formulated by Equation (3).

$$H(INP) \leq H(SUB) \quad (3)$$

where H(INP) is the entropy of combination of input layers and H(SUB) is the entropy of modeling units (subwatershed or subwatershed with HRUs).

The concept of entropy is applied as follows to develop the new tool that helps in identifying an appropriate level of subwatershed discretization. All maps of input layers are combined or spatially intersected to create a new map. Each unique combination of input parameter values in the new map is assigned a category. For example, corn on sandy loam soil forms Category 1, and wheat on clay loam soil forms Category 2. The new category names do not quantify the difference between encoded soil and land use but indicate a set of other categories. Maximum entropy (H_{max}) of this new map is calculated from the probabilities of new categories for the entire watershed using Equation (1). This maximum entropy could be reduced by a user defined percentage of information loss (percentage reduction in entropy that is acceptable; e.g., 5 percent in this study). Let this maximum entropy be denoted as a user defined maximum entropy ($H_{u\text{-}max}$).

The entropy function can be computed for grid based input data by using a moving window technique (Alberto, 1994). The moving window technique is used on the new map to identify the reference window size as follows. The reference window size is used to compare the entropy values of the input layers to the entropy values of subwatersheds according to Equation (3) to find the appropriate level of subwatershed discretization. During this process, a window size of certain number of grids is chosen and moved focusing on each grid of the new map and the entropy for each window is estimated from those categories of the new map that are covered by the window. Entropy of all the windows is summed to get the average entropy of the entire map. This procedure is repeated by increasing the window size iteratively until the estimated average entropy of the categories of new map approaches the user defined maximum entropy ($H_{u\text{-}max}$). The reference moving window size is defined by the window size when the average entropy of the new map reaches the user defined maximum entropy.

The entropy of input layers H(INP) in Equation (3) is estimated from the categories of the new map (combined soil and land use map) by moving the reference window on the new map similar to the procedure of the moving window technique.

Subwatershed maps with various levels of discretization (different number of subwatersheds) are developed from topographical data. A detailed description of the subwatershed delineation procedure is explained in the methodology section below. Similar to the moving window technique, the reference window is moved, focusing on each grid of the subwatershed map, and the entropy for each window is estimated. This process is continued to find the average entropy of the entire subwatershed map (H(SUB)). This procedure is repeated for other subwatershed maps with different levels of discretization. The entropy estimated for each subwatershed map (H(SUB)) is compared to the entropy of input layer H(INP). According to Equation (3), the subwatershed map whose entropy approaches the entropy of input layers is chosen as having the appropriate level of discretization.

Application of the moving window technique and entropy procedure is illustrated with an example (Figure 2). The example demonstrates two, 4x4 grid spaced watersheds. Let us consider that the combined map of land use and soil data looked as shown in Watersheds A and B. These input maps have to be compared with various subwatershed configurations (Subwatershed 1 through Subwatershed 3). It is to be noted that the category name of the subwatershed is not important but the category distribution and their arrangement are important. Let us say that a moving window size of 4x4 grids is used to estimate the entropy using Equation (2) (Figure 2). Comparing the entropy of the two watersheds with entropy of the three subwatershed configurations (Equation 3), it is concluded that:

- Subwatershed Configuration 1 is appropriate for Watershed B but not for Watershed A,
- Subwatershed Configuration 2 is appropriate for Watershed B and for Watershed A, and
- Subwatershed Configuration 3 is appropriate for Watershed B but not for Watershed A.

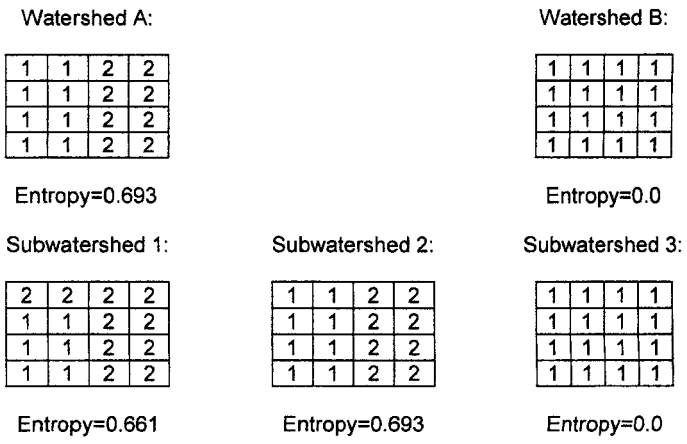

Figure 2. Application of the Entropy Procedure (i.e., watersheds and subwatersheds).

Certain hydrologic models, such as SWAT, are capable of treating HRUs within a subwatershed (Srinivasan and Arnold, 1994; Krysanova *et al.*, 1998). The level of discretization and heterogeneity in such

models tend to be greater than those models that use subwatersheds without HRUs. In SWAT, HRUs within each subwatershed are defined by first selecting land uses whose frequencies (frequency expressed in percent) are greater than the user defined land use threshold frequency and within those selected land uses, and by selecting the soils whose frequencies are greater than the user defined soil threshold frequency. The threshold frequency is an arbitrary number defined by the user to select certain category combinations from a large number of category combinations based on their frequency values. Spatial heterogeneity increases in the HRU approach. However, HRUs are not spatially allocated to certain areas within the subwatershed but are statistically distributed (Srinivasan and Arnold, 1994).

As HRUs are not spatially allocated within the subwatersheds, it is not possible to calculate the entropy induced by HRUs by combining the subwatershed map with HRUs to a new map as it was carried out with soil and land use maps. Consequently, the entropy of spatial modeling units, consisting of subwatersheds with HRUs, cannot be calculated by applying the moving window technique. Hence, the following approach is used to estimate the entropy of HRU with reference to Equation (2). The entropy of HRU is calculated by Equation (4).

$$H(HRU) = H(L_1) + (1\text{-}C) \times H(L_2) \tag{4}$$

where H(HRU) is the entropy induced by HRUs; $H(L_1)$ and $H(L_2)$ are the entropy of land use and soil layers; and C is Pearson's corrected contingency coefficient.

The soil and land use categories whose frequencies (frequency values) exceed the user defined soil and land use threshold frequencies are selected to form HRUs. Those categories whose frequencies are less than the user defined threshold frequencies are excluded. Hence, there is no information on the spatial allocation of the soil/land use categories/grids and it is not possible to create a map of HRUs and calculate $H(L_1, L_2)$ directly. As shown by Equation (2), however, the entropy of the two layers can be calculated from the sum of both single entropy values under the condition that the value of layer1 is known when calculating the entropy of layer2 $[HL_1(L_2)]$. To account for this uncertainty about L_2 if L_1 is known, the strength of correlation between soil and land use is calculated by Equation (5) (Koehler *et al.*, 1984). Pearson's contingency coefficient (C) calculated by Equation (5) provides the strength of correlation between soil and land use layers. C is estimated from the soil and land use maps for only those category combinations exceeding the user defined threshold frequencies using Equation (5). The number of different soil and land use categories are taken into consideration for the purpose of normalizing to 1 (Koehler *et al.*, 1984).

$$C = \sqrt{\frac{\sum \frac{(B_{ij} - E_{ij})^2}{E_{ij}} \times m}{\left(\sum \frac{(B_{ij} - E_{ij})^2}{E_{ij}} + \sum B_{ij}\right) \times (m-1)}} \tag{5}$$

where i is the number of soil categories; j is the number of land use categories; Bij is the observed frequency of combined categories of soil and land use; Eij is the expected frequency of combined categories and it is the sum of soil (row) frequency*sum of land use (column) frequency*1/N; m is the minimum of i (row) and j (column) of the contingency table; and N is the number of grids of the combined map (soil and land use maps).

$H(L_1)$ and $H(L_2)$ are calculated from the selected soil/land use categories using Equation (1). The calculated $H(L_1)$, $H(L_2)$ and C are used in Equation (4) to calculate entropy of HRU, [H(HRU)]. H(HRU) is added to the entropy of the subwatershed layers [H(SUB)]. The resulting sum of H(SUB,HRU) is the total entropy value of subwatersheds including HRUs because subwatersheds (derived from topography) are not correlated to soil or land use input data of HRUs. According to Equation (3), the appropriate level of discretization is defined where the entropy of the combined input layers equals the entropy of subwatersheds. Hence, the entropy of subwatersheds including the HRUs, H(SUB,HRU) is substituted for H(SUB) in Equation (3) to find the appropriate level of discretization.

The entropy of the combined input layers for spatial modeling units calculated by the moving window technique shows a characteristic shape (Figure 3). Initially, as the moving window size increases, entropy also increases with a higher rate of change. Later, the rate of change in entropy is lesser, although the moving window size increases exponentially. The approach of average entropy reaching its maximum with the increasing moving window size is controlled by the spatial heterogeneity of input data category distribution within the watershed. The rate of change in entropy with increasing window size is lesser for homogeneous data than for heterogeneous data. Hence, in the case of heterogeneous watersheds, smaller moving window sizes are used to capture the heterogeneous distribution of input data and estimate the entropy.

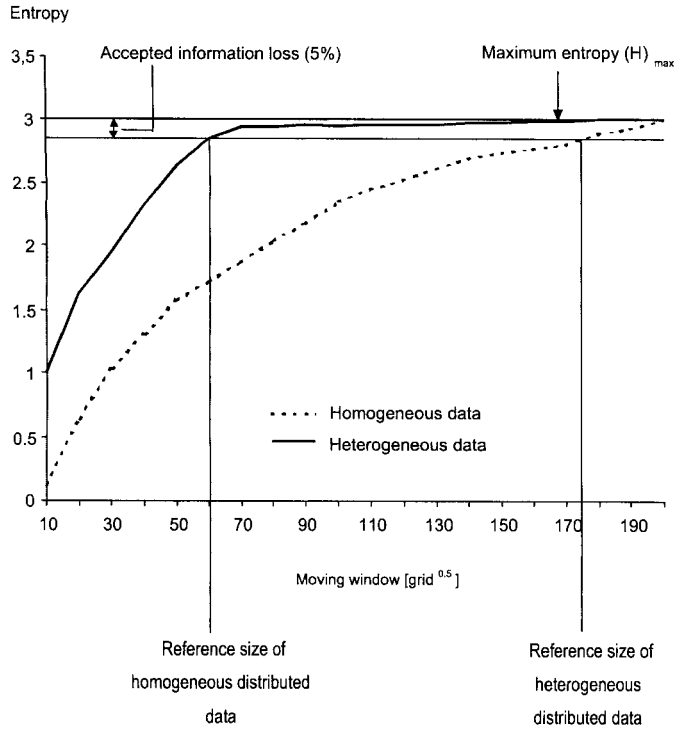

Figure 3. Iterative Determination of the Reference Moving Window Size.

METHODOLOGY

SWAT, a distributed hydrologic model (http://www.brc.tamus.edu/swat) developed by USDA-ARS, is used to validate the technique of choosing an appropriate number of subwatersheds on the basis of entropy (Arnold *et al.*, 1993). SWAT is used for a wide range of applications in water quantity and quality and agricultural management programs (Arnold *et al.*, 1999). It is integrated with the U.S. Environmental Protection Agency (USEPA) watershed modeling framework, Better Assessment Science Integrating point and Nonpoint Sources (BASINS). The Geographic Resources Analysis Support System (GRASS) GIS interface for SWAT (Srinivasan and Arnold, 1994) generates SWAT input files by collecting data from relational as well as distributed databases. The essential spatial input data used in this model are land use, soil, elevation and subwatershed maps.

In SWAT, the subwatersheds can be characterized either by the combination of dominant land use and soil categories or by HRUs within the subwatersheds. In the dominant land use and soil approach, the dominant type of soil within the dominant type of land use in each subwatershed is extracted. In the HRU approach, HRUs within the subwatersheds are delineated if the frequency of input data (land use and soil) categories exceed the user defined land use and soil threshold frequencies.

In this study, the subwatersheds were delineated by the Topographic PArameterZation tool (TOPAZ) (Martz and Garbrecht, 1998). TOPAZ requires a user defined minimum channel length and critical source area to determine sizes and shapes of the delineated subwatersheds. By varying these parameters, different numbers of subwatersheds and therefore different levels of spatial discretization were generated for each study area.

To identify an appropriate level of discretization for a watershed, a stand alone tool called SUbwatershed Spatial Analysis Tool (SUSAT) was developed (S. Haverkamp, 2000, Ph.D. Dissertation, Institut für Landeskultur Heinrich-Buff-Ring, unpublished). SUSAT calculates the entropy of the spatial modeling units and the entropy of the spatial input layers of the model for a given watershed. SUSAT functions as described below to identify the appropriate level of spatial discretization according to Equation (3).

Soil and land use maps are required as input for SWAT. These two maps are combined into a new map. Maximum entropy of the new map is calculated from the combined categories for the entire watershed (H_{max}) using Equation (1). Maximum entropy is reduced by an acceptable level of information loss to a user defined maximum entropy ($H_{u\text{-}max}$). A moving window technique is used on the new map to estimate the average entropy and to define the reference window size. The reference moving window size is defined by the actual moving window size when the average entropy equals the user defined maximum entropy.

The reference window is moved over the new map to estimate the entropy H(INP) of the combined soil and land use categories. The reference window also is moved over each of the subwatershed maps that are discretized at various levels to estimate the entropy H(SUB). SUSAT compares the entropy of subwatershed maps [H(SUB)] with the entropy of the soil and land use layers [H(INP)]. The subwatershed map for which its entropy approaches the entropy of input layers is chosen as the appropriate level of dicretization as per Equation (3).

SWAT uses either the dominant land use and soil approach or HRU approach within subwatersheds for simulating the hydrological process. In the dominant land use and soil approach, each subwatershed is expected to have one dominant land use and one dominant soil. Hence, when the distribution of land use and soil categories of the new map matches the distribution of (subwatershed) categories of a particular subwatershed map, then that subwatershed configuration is chosen as the appropriate level of discretization. At this level, entropy of both maps [H(INP) and H(SUB)] are equal and the soil and land use (input)

information is completely transferred to the subwatershed map. In the HRU approach, the entropy of HRU [H(HRU)] calculated in Equation (4) is added to the entropy of the subwatershed for identifying the appropriate level of discretization. The frequencies of soil and land use categories exceeding the user-defined threshold frequencies for the soil and land use maps are shown in Table 1. SUSAT was integrated with the data extraction procedure of the SWAT-GRASS Interface (Srinivasan and Arnold, 1994). Performance of SUSAT was quantitatively assessed using the coefficient of efficiency estimated for each watershed from the measured and model simulated flow data. SUSAT was validated for dominant land use and soil and HRU parameterization of subwatersheds.

RESULTS AND DISCUSSION

Although SUSAT was validated for all three watersheds in this study, the detailed results and discussion are presented for the Dietzhoelze watershed. Figure 4 shows entropy and efficiency values for dominant land use and soil approach at different levels of subwatershed discretization for the Dietzhoelze watershed. Calculated entropy for the subwatershed distribution ranges from 1.8 for 27 subwatersheds to 3.8 for 297 subwatersheds. The coefficient of efficiency rises from 0.68 for 27 subwatersheds to 0.73 for 78 subwatersheds and remains constant thereafter. The minimum number of required subwatersheds could be set to 78, as the coefficient of efficiency did not improve beyond this threshold. Comparison of the result (coefficient of efficiency of 0.73 for 78 subwatersheds) shown by the SUSAT approach (Figure 4) with the result (coefficient of efficiency of 0.76 for 60 subwatersheds) of the traditional trial and error approach (Figure 1), indicates that the SUSAT result seems to be within the acceptable range. The advantage is that SUSAT provides information on the appropriate level of discretization during data preprocessing itself without making several trial runs.

In this study, for all the study areas, a user defined land use threshold frequency of 5 percent and a user defined soil threshold frequency of 10 percent are used to form HRUs (Table 1). Additional discretization of HRUs within the subwatersheds increases chances of heterogeneity of the GIS input data. Therefore, entropy values for subwatersheds including HRUs (Equation 4) are much higher than the entropy values of subwatersheds that are directly characterized by the dominant land use and soil approach (Figure 4). Figure 5 shows the effect of the HRU approach on the Dietzhoelze watershed, where all other input parameters remain the same as that of Figure 4. The coefficient of efficiency increased from 0.77 for HRUs with five subwatersheds to the coefficient of efficiency of 0.78 for HRUs with nine subwatersheds and remained constant thereafter even for 60 subwatersheds. This result is in agreement with the entropy calculations of SUSAT as the entropy of all investigated levels of spatial discretization exceeds the entropy of combined spatial input parameters. Thus, they are characterized as appropriate levels of discretization by SUSAT (Equation 3).

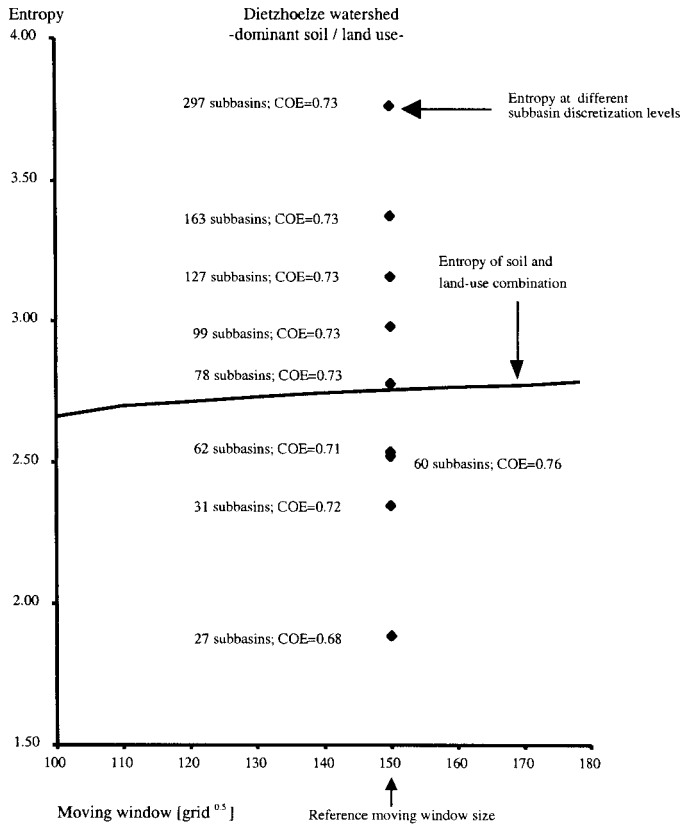

Figure 4. Validation Results (Entropy and Coefficient of Efficiency) From SUSAT With Dominant Land Use and Soil Approach for the Dietzhoelze Watershed.

Table 2 shows the validation results of entropy for the three watersheds for the dominant land use and soil approach and the HRU approach at different levels of subwatershed discretization. The grey shaded areas indicate those levels of discretization that are identified by SUSAT to fulfill the minimum heterogeneity demand from input data and tolerated information loss. These levels of discretization indicate a maximum level of aggregation considering the user defined level of information loss. A further aggregation of spatial modeling units results in decreasing model efficiency (Table 2).

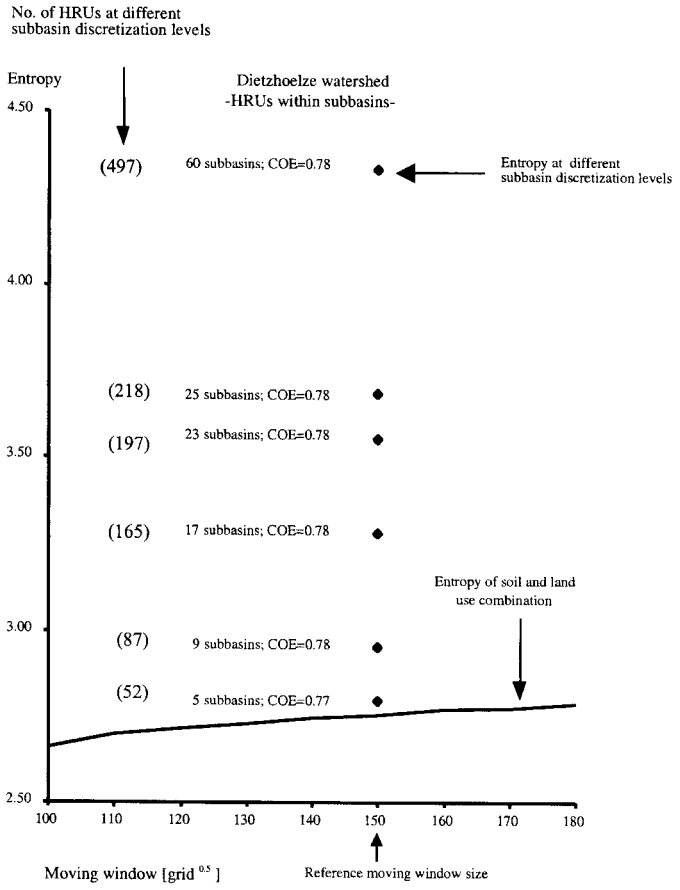

Figure 5. Validation Results (Entropy and Coefficient of Efficiency) From SUSAT for the HRU Approach for the Dietzhoelze Watershed.

TABLE 2. Validation Results (Entropy) of the Study Areas.

No. of Subwatersheds (TOPAZ)	Dietzhoelze / Entropy		Weiherbach / Entropy		Bosque / Entropy	
	Dominant	HRU	Dominant	HRU	Dominant	HRU
3	-	-	0.48	2.63	-	-
5	1.07	2.88	-	-	-	-
6	-	-	-	-	0.87	1.39
9	-	-	1.51	3.66	-	-
14	-	-	1.67	-	1.04	1.56
17	1.49	3.30	-	-	-	-
20	-	-	1.95	4.1	1.66	2.18
25	-	3.69	-	-	-	-
27	1.89	-	-	-	-	-
35	2.35	-	2.64	-	1.71	2.23
52	2.52	4.33	-	-	-	-
54	-	-	-	-	2.28	2.80
60	2.54	4.35	3.01	-	-	-
78	2.78	4.59	3.09	-	-	-
99	2.98	-	-	-	-	-
127	3.16	-	-	-	-	-
163	3.37	-	-	-	-	-
297	3.76	-			-	-

For all the investigated watersheds, the entropy of the HRU approach is much higher than the entropy of the dominant land use and soil approach. This could be due to the additional heterogeneity induced by the HRUs created within the subwatersheds. The number of HRUs created within a subwatershed depends on the threshold frequencies used for land use and soil. The required number of subwatersheds for an appropriate level of discretization varies (increases/decreases) depending on the number of HRUs created, so as to achieve at least the amount of entropy of the combined input layers as shown in Equation (3).

Similarly, Table 3 shows the validation results of the coefficient of efficiency for the three watersheds for the dominant land use and soil approach and the HRU approach at different levels of subwatershed delineation. The grey shaded areas indicate those levels of discretization that are identified by SUSAT. The coefficient of efficiency for the Weiherbach watershed varied between -4.54 (three subwatersheds) and 0.51 (78 subwatersheds) for the dominant land use and soil approach. The results of the HRU approach are much better than the dominant land use and soil approach. The coefficient of efficiency for the dominant land use and soil approach are lower because the number of land use and soil categories chosen for discretization are lesser and they are assigned to a small number of subwatersheds in this approach. The dominant land use and soil approach does not allow representation of all the land use and soil categories and causes differences between model representation and reality, and subsequently results in a poor coefficient of efficiency. For a realistic representation of land use and soil categories in the Weiherbach watershed more than three subwatersheds are required. The difference in the coefficient of efficiency between -4.54 and 0.51 for the Weiherbach watershed is a much more extended range than for the Bosque and the Dietzhoelze watersheds. Consequently, even for those levels of spatial discretization, which were identified as appropriate levels by SUSAT, the amplitude of variation in the coefficient of efficiency is higher for this watershed compared to the Bosque and the Dietzhoelze watersheds. The distribution of soil and land use categories for the Bosque watershed requires a very large reference moving window size. The Bosque watershed has a stretched shape and does not fit very well to the square shape of the moving window, and thus results are extrapolated from the maximum applicable reference moving window sizes.

CONCLUSIONS

A statistically based approach/tool (SUSAT) was developed to identify an appropriate level of discretization for any watershed. Three watersheds were studied for validating this approach. Before running the hydrologic model SWAT, SUSAT was able to identify those levels of detail that will provide a user-defined minimum level of input data heterogeneity. This approach could serve as an objective choice of spatial heterogeneity instead of a subjective guess.

TABLE 2. Validation Results (Coefficient of Efficiency) of the Study Areas.

No. of Subwatersheds (TOPAZ)	Dietzhoelze / COE		Weiherbach / COE		Bosque / COE	
	Dominant	HRU	Dominant	HRU	Dominant	HRU
3	-	-	-4.54	0.30	-	-
5	0.64	0.77	-	-	-	-
6	-	-	-	-	0.04	0.43
9	-	-	-0.55	0.22	-	-
14	-	-	0.31	-	0.30	0.48
17	0.67	0.78	-	-	-	-
20	-	-	0.45	0.35	0.38	0.48
25	-	0.78	-	-	-	-
27	0.68	-	-	-	-	-
35	0.69	-	0.45	-	0.35	0.49
52	0.70	0.78	-	-	-	-
54	-	-	-	-	0.39	0.49
60	0.76	0.78	0.50	-	-	-
78	0.73	0.78	0.51	-	-	-
99	0.73	-	-	-	-	-
127	0.73	-	-	-	-	-
163	0.73	-	-	-	-	-
297	0.73	-			-	-

The performance of SUSAT was tested by comparing the SWAT simulated flow and measured flow. SUSAT provided valuable information on choosing the appropriate level of spatial heterogeneity.

The developed approach has several advantages. Two of them are as follows:

1. It is no longer necessary to do a time consuming "trial and error" based search for appropriate levels of discretization.
2. The proposed method is not limited to a certain number of layers. Therefore, it is flexible and independent of the architecture of the hydrologic model to be used. In principle, it is applicable for small and large scale watersheds.

Despite these advantages, SUSAT has some limitations. It is only applicable on nominal data. In the case of interval scaled values, prediction of an absolute range of efficiency is difficult due to the similarity between the frequency classes. SUSAT does not consider the influence of the routing structure through the subwatersheds to the watershed outlet. However, the effect of the routing on model results seems to be quite low for the three watersheds studied using SWAT. Also, there is no definite level of discretization that can be defined as optimum as shown by Figure 1. The optimum is a subjective definition. The choice still depends on a user defined limit on acceptable level of information loss. However, the proposed new approach helps to formulate different levels of discretization on an objective basis, and quantify their impacts on model results. The present study is limited to the application of a single hydrological model. However, the concept can be extended to other models.

LITERATURE CITED

Alberto, L. G., 1994. Probability and Random Processes for Electrical Engineering. Addison-Wesley Publishing Company, Bonn, Germany.

Arnold, J. G., P. M. Allen, and G. Bernhardt, 1993. A Comprehensive Surface-Groundwater Flow Model. J. Hydrology 142: 47-69.

Arnold, J. G., R. Srinivasan, R. S. Muttiah, P. M. Allen, and C. Walker, 1999. Continental Scale Simulation of the Hydrologic Balance. J. American Water Resources Association 35(5): 1037-1052.

Frede, H. G. and M. Bach, 1999. Perspektiven für Periphere Regionen. Perspectives for Peripheral Regions 40 (5/6):193-197.

Grunwald, S., 1997. GIS-Gestützte Modellierung des Landschaftswasserhaushaltes mit dem Modell AGNPSm, Boden und Landschaft.

Koehler, W., G. Schachtel, and A. Voleske, 1984. Biometrie, Heidelberger Taschenbuecher.

Krasovskaia, I., 1997. Entropy Based Grouping of River Flow Regimes. J. of Hydrology 202:173-191.

Krysanova, V., D. I. Mueller-Wohlfeil, and A. Becker, 1998. Development and Test of a Spatially Distributed Hydrologic/Water Quality Model for Mesoscale Watersheds. Ecological Modeling 106: 261-289.

Leavesley, G. H., R. W. Lichty, B. M. Troutman, and L. G. Saindon, 1983. Precipitation-Runoff Modeling System – User's Manual. U.S. Geological Survey. Water Resources Investigations Report 83-4238.

Maidment, D. R., 1991. GIS and Hydrologic Modeling. First International Symposium/Workshop on GIS and Environmental Modeling. Boulder, Colorado.

Martz, L. W. and J. Garbrecht, 1998. The Treatment of Flat Areas and Closed Depressions in Automated Drainage Analysis of Raster Digital Elevation Models. Hydrol. Processes 12: 843-855.

Nash, J. E. and J. V. Sutcliffe, 1970. River Flow Forecasting through Conceptual Models – Part I: A Discussion of Principles. J. Hydrology 10:282-290.

Noehles, I. and H. G. Frede, 1998. Landnutzungsklassifikation mit Multitemporalen Landsat-TM Szenen unter Besonderer Berücksichtigung von Sukzessionsbrachflächen. DFG. 1. Zwischenbericht.

Shannon, C. E. and W. Weaver, 1949. The Mathematical Theory of Communication. University of Illinois Press.

Srinivasan, R. and J. G. Arnold, 1994. Integration of a Basin-Scale Water Quality Model with GIS. Water Resources Bulletin 30(3): 453-462.

USDA-ARS (U.S. Department of Agriculture-Agricultural Research Service), 1997. Hydrologic Model Performance Evaluation Applying the Entropy Concept as a Function of Precipitation Network Density. USDA-ARS Laboratory, Tektran, Arizona.

Ventsel, E. S., 1964. Teorija verojatnostej (Probability Theory), Nauka, Moscow, Russia.

Young, R. A., C. A. Onstad, D. D. Bosch, and W. P. Anderson, 1994. Agricultural Non-Point Source Pollution Model, Version 4.03 – AGNPS Users Guide.

Woolhiser, D. A., R. E. Smith, and D. C. Goodrich, 1990. KINEROS, A Kinematic Runoff and Erosion Model: Documentation and User Manual. USDA-ARS, USDA-ARS Publ-77.