DEFINING PROBABILITY OF SATURATION WITH
INDICATOR KRIGING ON HARD AND SOFT DATA

Steve W. Lyon
Graduate Researcher
Biological and Environmental Engineering
Cornell University
Ithaca, New York 14853
Email: sl336@cornell.edu

Arthur J. Lembo, Jr.
Senior Research Associate
Crop and Soil Sciences
Cornell University
Ithaca, New York 14853
Email: ajl53@cornell.edu

M. Todd Walter
Senior Research Associate
Biological and Environmental Engineering
Cornell University
Ithaca, New York 14853
Email: mtw5@cornell.edu

Tammo S. Steenhuis
Professor
Biological and Environmental Engineering
Cornell University
Ithaca, New York 14853
Email: tss1@cornell.edu

1 Corresponding Author
ABSTRACT

In humid, well-vegetated areas, such as in the northeastern U.S., runoff is most commonly generated from relatively small portions of the landscape becoming completely saturated. However, little is known about the spatial and temporal behavior of these saturated regions. Indicator kriging provides a way to use traditionally collected water table data to quantify probability of saturation to evaluate predicted spatial distributions of runoff generation risk, especially for the new generation of water quality models incorporating saturation excess runoff theory. When spatial measurements of a variable are transformed to binary indicators (i.e., 1 if above a given threshold value and 0 if below) and the resulting indicator semivariogram is modeled, indicator kriging produces the probability of the measured variable to exceed the threshold value. Indicator kriging gives quantified probability of saturation or, consistent with saturation excess runoff theory, runoff generation risk with depth to water table as the variable and the threshold set near the soil surface. The probability of saturation for a 120m x 180m hillslope based upon 43 measurements of depth to water table is investigated with indicator semivariograms for six storm events. The indicator semivariograms show high spatial structure in saturated regions with large antecedent rainfall conditions. The temporal structure of the data is used to generate interpolated (soft) data to supplement measured (hard) data. This improved the spatial structure of the indicator semivariograms for lower antecedent rainfall conditions. Probability of saturation was evaluated through indicator kriging incorporating soft data showing, based on this preliminary study, highly connected regions of saturation as expected for the wet season (April through May) in the Catskill Mountain region of New York State. Supplementation of hard data with soft data incorporates physical hydrology of the hillslope to capture significant patterns not available when using hard data alone for indicator kriging. With the need for water quality models incorporating appropriate runoff generation risk estimates on the rise, this manner of data interpolation will lay the groundwork for future model evaluation and development.

KEYWORDS: Indicator Kriging, Indicator Semivariogram, Geostatistics, Binary Logistic Regression, Risk Assessment
INTRODUCTION

To limit the pollution in waterways associated with agricultural land use (i.e., fertilizer, pesticide, and other chemicals coming from field applications), watershed managers need objective methods to identify high-risk, contaminant-contributing areas from agricultural fields. Pollution transport models are often used to rank vulnerability of contaminant sources with field-scale properties ignoring landscape position and current theories in hydrology on how and where runoff is generated [18]. Many studies demonstrate that this approach is not appropriate when considering chemical transport [9,11,17,23,42,31,38]. Quantifying probability of saturation to represent runoff production and nutrient transport is imperative to the development of tools and models based on physical hydrological mechanisms. For example, in humid, well-vegetated areas, such as in the northeastern U.S. and many other locations, most runoff is generated by saturation excess near stream channels. Since these hydrologically active areas vary in size seasonally and during individual storm events, they are often referred to as variable source areas (VSA) [12,16]. Simulation of VSA provides the groundwork for risk assessment tools usable by watershed planners to limit pollution that incorporate current theories in hydrology. Risk assessment tools and models predicting VSAs based on topographic index distributions similar to that in TOPMODEL [3] are beginning to emerge [13,45,18,34,25,26,1]. The extremely high spatial and temporal variability of soil moisture and groundwater significantly impacts many hydrological processes making development of such tools a challenge for researchers. Without methods to objectively quantify regions of runoff production, it is impossible to develop risk assessment procedures incorporating the appropriate hydrological mechanisms.

As new risk-assessment tools appear, the question is raised, “How can a high-risk region in a field be objectively quantified?” This question requires new methods of collecting and interpreting hydrologic data to evaluate these management tools. In recent years, progress has been made in collecting snap shots of soil moisture using various remote sensing techniques [6,4,14,39,40] and field measurements [46,29,28,41,52]. These methods are powerful, but can be limited by satellite coverage, accessibility to watershed, and watershed size. Also, while spatially robust, they are often too temporally sparse.
(i.e., low frequency of sampling) to encompass concepts of VSAs integrated into new risk-assessment tools. High temporal resolution measurements of depth to water table are becoming readily available due to in-expensive, self-contained, water level data loggers (e.g., TruTrack, Inc). Since water table height is crucial in runoff generation, especially for saturation excess theory and during the wet season, these measurement techniques provide new information for assessing risk at the field scale. The point nature of these datasets leads naturally to geostatistical techniques to characterize spatial patterns for objective measures saturation during a rain event. With appropriate spatial interpolation techniques, temporally rich, discrete point systems generate spatial measures for hillslopes capable of capturing the dominant characteristics of VSA hydrology.

Water table elevations have been estimated through various forms of kriging and cokriging [32,10,2,7]. At the heart of any kriging methodology is the semivariogram. The semivariogram model provides a quantitative estimate of the structure required to characterize a spatial pattern from a series of point data [47]. The major structural parameters of a traditional semivariogram model include the range, sill, and nugget. The range provides a measure of the maximum distance over which spatial correlation affects the variable of interest. The sill, if it exists, represents the spatial variance of two distant measurements. The nugget represents the variance between two relatively close measurements. The nugget gives the variance in the measurement due to the inherent variability of the sampling device and the occurrence of spatial patterns smaller than the shortest sampling interval. Within the realm of semivariogram and kriging techniques, indicator kriging provides a manner to capture extreme values [24]. Traditionally, indicator semivariograms and the resulting kriging assess risk of contamination in various constituents such as heavy metals [44,37,19] and assess uncertainty in soil properties [33,27,21]. In the most basic form, indicator semivariograms treat data as a binary indicator with respect to a threshold value (i.e., 1 if threshold is exceeded; 0 if threshold is not exceed). For a more complete discussion of indicator semivariograms and indicator kriging along with many possible derivatives in algorithms and methodology, see Goovaerts [20], Deutsch and Journel [8], and Chilès and Delfiner [5].
Western et al. [48] examined soil moisture patterns through indicator semivariograms and showed good spatial structure for high soil moisture conditions. They point out that there are shortcomings in the technique for distinguishing between connected and unconnected patterns in soil moisture. That is the semivariogram does not fully characterize the spatial structure of the pattern. Nevertheless, indicator semivariograms still have promise when applied directly to “hard data”, or the actual measurements of the variable of interest, in the form of depth to water table for distributing runoff source areas. Areas prone to saturation either have a high ground water table or hard pan (fragipan) at shallow depth indicating landscape factors such as soil depth (i.e., available water storage capacity), upland-watershed area, and local topography as important factors determining whether or not a particular area saturate. For wet periods, interflow will be higher and often expand the extent of saturation around saturation-prone areas; conversely, dry periods will decrease interflow and extent of saturation. Thus, if the threshold value is near the soil surface, indicator semivariograms and kriging quantify the probability of saturation for an area. Another option available with indicator kriging is the incorporation of “soft data”, or local information that is a proxy to the variable of interest and need not relate directly [20], to supplement hard data. Soft data can help compensate for a lack of spatially exhaustive observations by providing information about predictions where no hard data are available using prior probabilities [50]. With saturation excess theory and indicator kriging incorporating soft data, regions at high-risk in terms of chemical transport and runoff production can be quantified to aide in management tool development.

This study looks at the ability of indicator semivariograms to capture probability of saturation for a hillslope in Town Brook watershed in the Catskill Mountains of New York, USA. From this initial investigation, a manner to produce soft data to supplement the hard data observations is presented providing a method to compensate for spatially sparse data without compromising the significant hydrological features on the landscape. The spatial structure of the observations of saturated regions is improved with the incorporation of soft data. Probability of saturation maps are generated from indicator semivariograms based on the hard data, soft data, and the combination of the two. By
using soft data, realistic representations of high probability of saturation regions for the rain events that capture the physical hydrology of the hillslope can be made. This gives objective quantification to measure the risk of saturation at the field scale for future model development and watershed management. As tools for evaluating nutrient management become the mode in research, this physical representation of risk assessment is needed to validate predictions of new techniques.

SITE DESCRIPTION AND DATA

The 2.44 ha hillslope in this study located on New York State Department of Environmental Protection (DEP) owned lands is part of a 2 km² sub-watershed located in the southwest corner of the 37 km² Townbrook watershed in the Catskill Mountain region of New York State (Figure 1). The hillslope ranged in elevation ranged from 585 m to 600 m above mean sea level with slopes ranged from 0° to 8°. The soils are gravelly silt loams over fractured bedrock. These shallow soils are typified as higher conductivity (5 cm/hr) surface material (< 40 cm deep) overlaying less conductive material (0.5 cm/hr) base material (> 40 cm deep) with large fractures. The landuse on the hillslope is uniformly grass/shrub with forested regions above the study area. A 200+ point survey of the region was conducted to supplement the 10m digital elevation model (DEM) of the hillslope to derive 1-m interval contours for identifying small-scale topographic features. This supplemental elevation data was used to generate explanatory variables for the binary logistic regression analysis. Based on a multidirectional algorithm similar to that of Quinn et al. [35], upslope area was determined for the study site. Local flow paths were estimated as cells accumulating 500 m² of upslope area. Distance to these flow paths and the total distance to stream were evaluated along steepest surface gradient for each of the sampling locations. Topographic index, defined as ln(a/tanβ), and local slope at each sampling location were also evaluated.

Figure 1 shows the layout of 43 sampling locations on the hillslope instrumented with water level loggers for monitoring depth to water table. The water level loggers were WT-HR 500 capacitance probes from TruTrack, Inc. These loggers measured depth to water table in the upper 50 cm of the soil at 5-minute intervals starting from March 4th,
2004 that were averaged over 15-minute intervals. The loggers were placed on two grid systems. The first consisted of 19 loggers on a 10 x 10 m grid near the stream at the bottom of the hillslope. In addition, 24 loggers were located on a large spacing 30 x 40 m grid to gauge water table reactions up the hillslope. Above and below the hillslope site, two loggers were placed in the stream to function, along with a rating curve established from velocity measurements, as stream gauges. Finally, a tipping bucket rain gauge with data logger was set on the site to sample rainfall amounts at an interval of 10 minutes. Due to rain gauge malfunction, the rainfall data from March 23rd, 2004 to April 16th, 2004 was taken from a National Oceanic and Atmospheric Administration (NOAA) weather station located in Stamford, NY located approximately 1 km north of the hillslope. With the on-site rain gauge and NOAA station data, measures of daily rainfall amounts were created for the sampling period. From this period, six rainfall events were chosen corresponding to the largest events for the study period. Table 1 gives a summary of the six rainfall events considered in this study. These events all consisted of rainfall amounts greater than 1.5 cm on a given day and were used to investigate the formation of saturated areas on the hillslope.

METHODS
Indicator values \( I_i \) were calculated by setting a threshold depth to water table \( Z_c \) and assigning a value of 1 to sampling locations where the measured depth to water table \( z_i \) was less than or equal to the threshold depth to water table and a value of 0 to locations where the depth to water table was greater than the threshold depth to water table for the rainfall event. Thus, indicator values of saturation for sampling locations during each rainfall event were defined as

\[
I_i(z_c) = \begin{cases} 
1 & \text{if } z_i \leq z_c \\
0 & \text{if } z_i > z_c 
\end{cases}
\]  

(1)

where, \( I_i(z_c) \) is the indicator value at sampling location \( i \), \( z_i \) is the measured depth to water table at sampling location \( i \) (in cm), and \( z_c \) is the threshold depth to water table (in cm). Sampling locations were considered in exceedence of \( z_c \) if they fit the criteria of
equation (1) for at least one 15-minute interval during a rainfall event. On inspection, sampling locations that fit this criterion remained in exceedence for the majority of the rainfall event. Indicator semivariograms were constructed with GSLIB software [8] for the six rainfall events using $z_c$ of 0 cm, 5 cm, and 10 cm based on the indicator values defined by equation (1) using the sample semivariance, $\gamma_s(h)$, at a lag, $h$, of

$$\gamma_s(h) = \frac{1}{2N(h)} \sum_{(i,j)} (I_i(z_c) - I_j(z_c))^2$$

where, $N$ is the number of pairs, $I_i(z_c)$ and $I_j(z_c)$ are the indicator values at $i$ and $j$, respectively, with summation over pairs $(i,j)$ for the lag bin. This varied from standard semivariance calculations by using indicator values at points $i$ and $j$ instead of measured values. In order to make comparisons across threshold values, semivariograms were normalized with variance of the indicator values at a given threshold to create a sill value of 1 for all indicator semivariograms [8]. This method of analysis was consistent with that used by Western et al. [48] for soil moisture.

To investigate landscape factors influencing saturation, binary logistic regression was performed using MINITAB release 12 (MINITAB, Inc., 1998) to identify the most significant predictor variables of saturated regions. From the indicator semivariogram analysis, the threshold depth to water table for saturation was set 5 cm for the binary response variables. At this threshold, the semivariograms for the various rainfall events exhibit well-defined sills and ranges and there are a proportionate number of saturated and unsaturated sampling locations. Binary logistic regression gives (maximum likelihood) parameter estimates that can be used to calculate the probability of saturation ($p_i$) at sampling location $i$ for each rainfall event as

$$\pi_i = \log\left(\frac{1}{1 - p_i}\right) = c + \beta_1X_1 + \beta_2X_2 + ... + \beta_iX_i$$
where $p$ is the probability of saturation, $X_i$ is the set of explanatory variables, $c$ as the intercept and $\beta_i$ as the set of parameter estimates. From this, the probability of saturation was calculated from the explanatory variables as

$$p_i = \frac{\exp(\pi_i)}{1 + \exp(\pi_i)}$$

(4)

Binary logistic regression gave the probability of saturation based on a series of explanatory variables that were either continuous or discrete on the spatial domain. In the case of discrete variables, explanatory variables were made continuous by spatially interpolating with ordinary kriging. Thus, probability of saturation at a sampling location $i$ was estimated using equation (4) with the most significant explanatory variables from equation (3) for each of the rainfall events. The probability model resulting from equation (4) provided an estimate of prior probabilities to supplement the indicator values for each location from equation (1) based on the significant explanatory variables found in equation (3) for each rainfall event. The prior probabilities are based on variables that are not direct measures of saturation, but provide a proxy to the observation of saturated areas. Thus, for each rainfall event, hard data (indicator values) and soft data (prior probabilities) were defined. These data were merged in a method consistent with Goovaerts [20] with the soft data as local indicator means respecting hard data where available. The residuals between the hard and soft data were computed at each sampling location and the semivariogram modeled.

Finally, to visualize the effects of incorporating soft data and analyze the indicator semivariograms, kriging was performed using the Geostatistical Analyst extension in ArcMap 8.2 (ESRI, Inc., 2002) with a spherical model of the form

$$\gamma_s(h) = \begin{cases} 
\theta_s & \text{for } 0 \leq h \leq \theta_r \\
\theta_s \left[ \frac{3}{2} h - \frac{1}{2} \left( \frac{h}{\theta_r} \right)^3 \right] & \text{for } \theta_r < h
\end{cases}$$

(5)
where \( \gamma_s \) and \( \gamma_r \) as the partial sill and range parameters, respectively. The spherical model was fit by visual inspection using an interactive plotting program to avoid errors associated with blind fits by geostatistical software packages. This produced three estimates for probability of saturation derived from (1) hard data alone, (2) soft data alone, and (3) hard and soft data. The Geostatistical Analyst extension was also used to generate semivariogram surfaces for the largest and smallest rain events to investigate anisotropy in the observations on the hillslope.

RESULTS

**Indicator Semivariograms – Hard Data**

Figure 2 shows the indicator semivariogram surfaces normalized with variance of the indicator values for saturation thresholds of 0 cm, 5 cm, and 10 cm during rainfall event 1 and event 6 to investigate anisotropy of the sampling created with the Geostatistical Analyst extension in ArcMap 8.2 (ESRI, Inc., 2002). Darker areas on the semivariogram surface correspond to larger values on the indicator semivariogram and lighter areas correspond to smaller values on the indicator semivariogram. The near the center of the semivariogram surface corresponds to shorter distances on the semivariogram. Anisotropy appears as bands of dark regions occurring across the semivariogram surface showing directional trends in the semivariogram. There appears to be slight anisotropy occurring in both the northwest-southeast orientation and northeast-southwest orientation. This occurrence is not a contradiction, but corresponds to the shape of the sampling grid and topography of the hillslope (northwest-southeast orientation in the upslope and northeast-southwest orientation in the near stream region). The topography of the hillslope is too diffuse to justify incorporating anisotropy into the indicator semivariograms. In addition, the range of values of semivariance for both events is relatively low in terms of anisotropy. Thus, omnidirectional indicator semivariograms are appropriate and used for analysis of the study site.

Table 1 gives the total number of sampling locations for the six rainfall events at threshold depth to water tables of 0 cm, 5 cm, and 10 cm. These threshold depth to water
Tables correspond to the 19th, 49th, and 69th percentiles of observations saturating averaged over the six rainfall events. Figure 3 shows the indicator semivariograms for the three threshold depth to water tables created with ten lag bins with a 20-meter lag distance and a tolerance of 50% so that the number of lag bins multiplied by the lag distance is half the maximum distance of the study site and the lag distance captures the sampling grid spacing. Table 2 summarizes the normalized nugget and range components of the indicator semivariograms from fitting the spherical model in equation (5). The indicator semivariograms exhibit pure nugget effect for all events when the threshold depth to water table is set at 0 cm attributed to spatial scarcity of the data. The sampling locations considered saturated have no spatial correlation because they are sparse and not connected on a scale capable of being captured by the sampling grid. Sampling locations are sensitive to slight undulations in the localized topography and uncertainty associated with defining soil surface. As the threshold depth to water table is increased to 5 cm, the indicator semivariograms become smooth and exhibit better structure for events 2, 3, 5, and 6; however, events 1 and 4 still have virtually pure nugget structure. As the threshold depth to water table is raised to 10 cm, events 1 and 4 begin to show a range and structured sill. Events 2, 3, 5, and 6 begin to lose the well-defined sill and become slightly nontransitional. At a threshold depth to water table of 10 cm, the indicator semivariograms are beginning to model the water table spatial correlation and no longer capturing purely surface saturation.

The first comparison of the indicator semivariograms is the normalized nugget. This represents the ratio of measurement error to the variance of the pattern. The normalized nugget decreases for all rainfall events when the threshold depth to water table is raised from 0 cm to 5 cm. At 0 cm, few sampling locations saturate per rainfall event resulting in decreased variance for saturation (i.e., more locations are not saturated). Additional sampling locations saturate as the threshold depth to water table is raised causing the variance to increase. This result is similar to that seen by Western et al. [48] for varying thresholds of soil moistures. Also, there is a higher normalized nugget for the two lowest antecedent rainfall conditions (event 1 and event 4) at any threshold compared with the other rainfall events. This can be attributed to saturated regions smaller than the
resolution captured by the sampling locations during low antecedent rainfall conditions. The next comparison of the indicator semivariograms structure is the range or the maximum distance over which pairs are considered spatially correlated. When applicable, there is an increase in range as the threshold depth to water table is raised from 5 cm to 10 cm. The indicator semivariograms begin to capture the large-scale spatial structure inherent to the water table. With threshold depth to water table at 5 cm, the indicator semivariograms are capture the saturated areas that are well defined for rainfall events with high antecedent rainfall conditions. For this threshold, lower antecedent rainfall conditions do not exhibit well-defined sills and ranges. Since this threshold depth to water table captures saturated areas reasonably well without incorporating the spatial structure of the water table itself, saturation is defined as a depth to water table less than 5 cm for the remainder of this article.

**Binary Logistic Regression – Soft Data**

Binary logistic regression was performed to identify the significant landscape features influencing saturation. Table 3 lists and describes the explanatory variable considered for binary logistic regression analysis. Also, Table 3 gives the corresponding significance through the p-value for each explanatory variable and measures of goodness of fit for each regression with the squared Pearson’s correlation coefficient, which is comparable to an $R^2$ for binary logistic regression. This list is by no means exhaustive, but contains explanatory variables selected based on their preeminence and recurrence in hydrological theories on runoff prediction. From the binary logistic regression, the only explanatory variable that significantly estimates saturation is the pre-event depth to water table (with $a = 0.05$). Event 5 and event 6 do not indicate pre-event depth to water table as a significant explanatory variable for predicting saturation. These rainfall events have the largest amounts of rainfall for all rainfall events (approximately double) with relatively high antecedent rainfall conditions. This leads to sampling locations that would typically not saturate in smaller rainfall events saturating due to the extremely large amount of the rainfall. Therefore, pre-event depth to water table is not a good predictor of saturation for these events. Using binary logistic regression, the pre-event
depth to water table for the remaining four events can be used to estimate probability of
saturation producing soft data to supplement hard data observations.

Soft data based on binary logistic regression was merged with the hard data observations
of saturation. Ordinary kriging was used to interpolate the observations of the pre-event
depth to water table for the hillslope and then probability of saturation determined from
binary logistic regression. The residuals (i.e., the difference between observed saturation
and probability predicted at each sampling location) were computed and modeled using a
simple kriging. Figure 4 shows the semivariograms of these residuals for events 1
through 4. Events 5 and 6 were not modeled because they did not show pre-event depth
to water table as a significant predictor of saturation. Table 4 summarizes the normalized
nugget and range components of the residual semivariograms fitting the spherical model
from equation (5). Events 2 and 3 have extremely well-defined sill and ranges of the
same magnitude of the indicator semivariograms using hard data alone. The sills of the
residual semivariograms for events 1 and 4 are more defined than those of the traditional
indicator semivariograms. Events 1 and 4 have higher normalized nuggets than events 2
and 3. This is consistent with the traditional indicator semivariogram analysis.

Indicator Kriging – Hard and Soft Data

Figure 5 gives the indicator kriging distributions for probability of saturation from hard
data alone, soft data alone, and the hard and soft data indicator semivariograms for event
1 to visually demonstrate the effect of incorporating soft data on estimating saturated
areas. The kriging was performed with ArcMap Geostatistical analyst that uses filters to
smooth regions near actual observations producing smooth predictions maps as opposed
to true indicator kriging, which would honor observations of indicator values. Figure 5a
shows the probability of saturation measurement from the model developed through an
indicator semivariogram with hard data alone. Isolated observations of saturation located
high on the hillslope produce anomalies in the prediction of regions with high probability
of saturation occurring in topographically divergent areas. Given that saturation excess
runoff mechanisms are generally accepted to dominate in the region, it is unlikely that these high-probability saturation regions predicted with hard data alone are representative of physical hydrology of the hillslope. The kriging model developed from indicator semivariograms based on hard data alone does not define a spatial pattern consistent with hydrological theories on how runoff is generated. The probability of saturation estimated using the soft data alone is shown in Figure 5b to demonstrate predictions made through traditional binary logistic regression techniques on spatial information. Due to smoother spatial structure provided by kriging the pre-event depth to water table than the spatial structure in observed saturated areas, this method produces a more uniform region of high probability of saturation for the hillslope. However, by disregarding observations (i.e., the hard data) during the event, this method predicts high probability of saturation for sampling locations do not saturate in the rainfall event. These sampling locations are positioned with little upslope area contributing to accumulation of flow. The use of soft data alone (i.e., binary logistic regression) predicts with only pre-event depth to water table and cannot capture the dynamic nature of the saturated areas during the rainfall event. Figure 5c provides the kriging through incorporating hard and soft data. This realization provides a smoothed version of the indicator kriging using hard data alone created using the physical hydrology of the site by incorporating the structure of the pre-event depth to water table. Thus, the high probability of saturation region is more connected during this event with a low antecedent rainfall condition. Figure 5c gives a prediction of the region with high probability of saturation that agrees with the topography of the hillslope. Also, probability of saturation at locations that do not saturate during the event is better represented then through predications made from soft data alone.

This improvement is demonstrated in the cross-validation of methods presented in Table 5. Cross-validation was performed by removing one observation from the set of sample locations, performing the kriging, comparing the kriging to the left out observation, and then cycling through all the observations in all events. Table 5 shows the distribution of
observations (divided into saturated or unsaturated) among the probability of saturation. Thus, 13% of saturated observations for all the events where predicted as having a probability of saturation of 0 to 25% from the cross-validation using hard data alone and only 4% of observations were in this range for the cross-validation using hard and soft data. Conversely, 3% of all unsaturated observations where predicted in the 75 to 100% probability of saturation range using hard data alone while 0% of observations were in this range for soft data alone or hard and soft data. This method of cross-validation shows that the proportion of false-positives using hard data alone is higher then when using hard and soft data combined. That is, more saturated observations are predicted as having low probability of saturation and more unsaturated observations are predicted as having high probability of saturation using the hard data alone. These false-positives are especially important when data analysis of the type is used in developing nutrient management plans or validating chemical transport models. By honoring hard data, probability of saturation is measured and not merely estimated, as it is by using soft data alone. Physically-based measures for probability of saturation are created providing an improvement over predictions made with pre-event depth to water table alone or estimates developed from observed saturation alone (i.e., soft data alone or hard data alone).

**DISCUSSION**

By incorporating the pre-event depth to water table, we are able to supplement the hard data with soft data derived from physical hydrology of the hillslope. However, it is seen that pre-event depth to water table alone is not sufficient for predicting saturated areas. Since the pre-event water table is often considered as a steady state condition, this demonstrates the transient nature of saturated regions during rain events in natural systems. This result does not support the assumption linking groundwater levels and runoff as a succession of steady-states, but is similar to the results of Seibert et al. [36] where allowing transient groundwater storage produced more realistic process representation. Also, from binary logistic regression, it is demonstrated that traditional
variables used to describe hillslope hydrology (i.e., topographic index and its constituents) may not be able to capture spatial distribution of saturated areas. Many studies have identified other controls on hydrology [22,15,49,51]. For this study, topographic features not captured in the site survey or poor resolution in subsurface characteristics may have contributed to lack of significance topographical controls in the binary logistic regression. However, the resolution of the elevation data used is much better than that for most topographically driven models where the resolution is on the order of a 10x10 m DEM. It could also indicate that new methods need to be developed for considering runoff generation, especially in shallow soil systems. New consideration of the assumptions in calculating topographic index including soil characteristics [43] or solar radiation effects [30,49] or in the topographic index’s relation with groundwater [36] could be appropriate. This study demonstrates the importance in accurate prediction of pre-event ground water levels for the appropriate simulation of surface runoff.

Traditional indicator semivariograms for the 6 rainfall events (Figure 3) show that there is spatial structure with higher antecedent rainfall conditions (events 2, 3, 5, and 6). This structure was captured with 43 sampling locations monitoring depth to water table. Due to limitations in the ability to collect high spatial resolution hydrological data for most regions, this provides a method to identify areas on the hillslope with high probability of saturation without exhaustive spatial knowledge of the entire region. From a saturation excess point of view, these regions directly correspond to regions of runoff production. During rainfall, saturated regions follow topography of the hillslope and saturate laterally from flow paths. This is similar to the results seen by Western et al. [48] for soil moisture contents during higher antecedent conditions. Thus, when high antecedent rainfall conditions (i.e., wetter initial condition) coupled with a rainfall event, saturated regions expand and are identified more readily using the water level loggers. However, due to the sparse sampling grid, it is only for rain events with high antecedent rainfall conditions when this lateral expansion of saturated regions is large enough to be represented by the indicator semivariograms. For the smaller antecedent conditions, flow
paths remain small with respect to their lateral extents and there is little spatial structure observed in saturated regions. These saturated regions are still likely to be influenced by topography and connected to the stream, but on a spatial scale smaller than that being sampled due to micro topographical features on the hillslope.

The use of high-temporal, low-spatial resolution data acquisition techniques coupled with indicator semivariogram analysis represents saturated areas in locations where other techniques are not available. Indicator approaches can incorporate soft data and this study presents one method to estimate soft data from pre-event depth to water table improving identification of high probability of saturation areas. The method takes advantage of high temporal resolution to compensate for sparse spatial coverage. As data loggers of the type used in this study become less expensive, there will be increased datasets of this variety available to the research community. These datasets are easily obtained in locations with poor satellite coverage or not accessible to mobile TDR methods of Western and Grayson [46]. By taking advantage of the data’s temporal structure, fewer sampling sites are needed to obtain high quality data from a study site and provide more information about the formation of saturated areas in response to rain events of various size and intensity. We have presented only one such method to utilize the temporal resolution of the data, but others may be developed using time-series or geostatistical analysis. Results from this style of investigation lead to development of appropriate techniques to estimate pre-event depth to water table and better model saturated area formation during events. Figure 5 gives a practical example of using geostatistics to produce a spatial distribution usable in risk assessment. Incorporation of soft data leads to a more realistic representation of the reaction of the hillslope to the rainfall event by including processes involved in the formation of saturated areas.

CONCLUSION

Data to quantify probability of saturation using 43 point observations and indicator kriging methods for a hillslope in the Townbrook watershed in the Catskill Mountain
region of New York State has been presented. For the 6 events considered, traditional
indicator kriging on hard data alone captured spatial patterns of saturated regions for
large antecedent rainfall conditions. However, during events with low antecedent
conditions, the spatial structure provided by the indicator semivariograms were on a scale
smaller than that sampled and lead to poor representations. A method to produce soft
data was introduced using binary logistic regression to identify significant hydrological
features that improved the spatial structure of the indicator semivariogram for lower
antecedent rainfall conditions. This method to produce soft data also resulted in more
realistic probability of saturation maps when kriging was conducted using the indicator
semivariograms. Spatially sparse datasets supplemented with temporal data can provide
better estimates of runoff source areas. As nutrient management begins to focus on
timing and application of pollution sources, quantifiable identification of risk at the field
scale becomes increasing important. This provides a method to collect data that,
although not spatially exhaustive, still captures the physical hydrology of region in
question. This type of data provides the groundwork as the next generation of nutrient
transport models and risk-assessment tools develop.

ACKNOWLEDGMENTS

Research is made possible by support and collaboration of the Department of Interior,
U.S. Geological Survey and the Cornell University, New York Resources Institute under
Grant Agreement No 01HQGR0208. In addition, funding of the Integrative Graduate
Education and Research Traineeship (IGERT) Biogeochemistry and the Environmental
Biocomplexity Program at Cornell University is acknowledged. The authors would also
like to acknowledge the extremely helpful comments made by reviewers.
REFERENCES


Table 1. Summary of six rainfall events for the study site on the DEP lands in the Townbrook watershed

<table>
<thead>
<tr>
<th>Rainfall Event</th>
<th>Date</th>
<th>Antecedent Rainfall (10 days, mm)</th>
<th>Rainfall Amount (mm)</th>
<th>Number of Sampling Locations Saturating</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4/13/2004</td>
<td>1.0</td>
<td>16.3</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>29</td>
</tr>
<tr>
<td>2</td>
<td>4/26/2004</td>
<td>37.9</td>
<td>16.5</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>22</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>31</td>
</tr>
<tr>
<td>3</td>
<td>5/2/2004</td>
<td>39.1</td>
<td>18.3</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>22</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>5/18/2004</td>
<td>18.3</td>
<td>17.5</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>21</td>
</tr>
<tr>
<td>5</td>
<td>5/24/2004</td>
<td>37.1</td>
<td>31.0</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>25</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>33</td>
</tr>
<tr>
<td>6</td>
<td>5/26/2004</td>
<td>60.5</td>
<td>36.1</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>35</td>
</tr>
</tbody>
</table>
Table 2. Summary of normalized indicator semivariogram characteristics using hard data alone fit with spherical model with units of m for ranges with threshold depth to water tables at 0 cm, 5 cm, and 10 cm

<table>
<thead>
<tr>
<th>Rainfall Event</th>
<th>Normalized Nugget</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 cm</td>
<td>5 cm</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.35</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0.1</td>
</tr>
</tbody>
</table>
Table 3. Explanatory variables with corresponding p-values (significant at $a = 0.05$ in bold) for threshold depth to water table of 5 cm the squared Pearson’s correlation coefficient in parenthesis for the binary logistic regression of each rainfall event.

<table>
<thead>
<tr>
<th>Explanatory Variable</th>
<th>Variable Description</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Event 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.46)</td>
</tr>
<tr>
<td>$c$</td>
<td>Constant</td>
<td>0.158</td>
</tr>
<tr>
<td>$X_1$</td>
<td>Distance from location to stream along flowpath</td>
<td>0.111</td>
</tr>
<tr>
<td>$X_2$</td>
<td>Distance to local flowpath accumulating more than 500 m$^2$</td>
<td>0.312</td>
</tr>
<tr>
<td>$X_3$</td>
<td>Soil type</td>
<td>0.172</td>
</tr>
<tr>
<td>$X_4$</td>
<td>Topographic Index</td>
<td>0.728</td>
</tr>
<tr>
<td>$X_5$</td>
<td>Pre-event depth to water table</td>
<td>$\mathbf{0.010}$</td>
</tr>
<tr>
<td>$X_6$</td>
<td>Total upslope area</td>
<td>0.481</td>
</tr>
<tr>
<td>$X_7$</td>
<td>Upslope area with maximum value truncated at 1 ha</td>
<td>0.405</td>
</tr>
<tr>
<td>$X_8$</td>
<td>Local slope at location</td>
<td>0.052</td>
</tr>
</tbody>
</table>
Table 4. Summary of normalized residuals semivariogram characteristics using hard and soft data fit with spherical model with units of m for ranges.

<table>
<thead>
<tr>
<th>Rainfall Event</th>
<th>Normalized Nugget</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.60</td>
<td>65</td>
</tr>
<tr>
<td>2</td>
<td>0.25</td>
<td>55</td>
</tr>
<tr>
<td>3</td>
<td>0.05</td>
<td>50</td>
</tr>
<tr>
<td>4</td>
<td>0.81</td>
<td>35</td>
</tr>
</tbody>
</table>
Table 5. Occurrence in units of percentage of saturated and unsaturated sampling locations in estimated probability of saturations from the cross-validation analysis using all events

| Estimated Probability of Saturation | Hard Data | | | Soft Data | | | Hard and Soft Data | | |
|---|---|---|---|---|---|---|---|---|
| | Saturated (%) | Unsaturated (%) | Saturated (%) | Unsaturated (%) | Saturated (%) | Unsaturated (%) |
| 0 to 25% | 13 | 41 | 8 | 9 | 4 | 15 |
| 25 to 50% | 25 | 41 | 53 | 52 | 28 | 49 |
| 50 to 75% | 29 | 15 | 39 | 39 | 54 | 36 |
| 75 to 100% | 33 | 3 | 0 | 0 | 14 | 0 |
FIGURE CAPTIONS

Figure 1. Sampling locations instrumented with water level loggers on DEP lands located within Townbrook watershed in the Catskill Mountain region of New York State.

Figure 2. Normalized indicator semivariograms surfaces for event 1 and event 6 for threshold depth to water tables of 0 cm, 5 cm, and 10 cm.

Figure 3. Normalized indicator semivariograms for all six rainfall events with threshold depth to water tables of 0 cm (long dash), 5 cm (solid), and 10 cm (short dash) using hard data alone.

Figure 4. Normalized residual semivariograms for threshold depth to water table of 5 cm for four rainfall events (dot-dash – event 1, solid – event 2, short dash – event 3, and long dash – event 4) between hard and soft data.

Figure 5. Indicator kriging for threshold depth to water table of 5 cm performed using a) hard data alone, b) soft data alone, and c) hard and soft data for event 1 showing the probability of saturation for the event with indicator values marked.
New York State
(NTS)

Explanation

- DEP Lands Boundary
- Townbrook Watershed

Kilometers

0 0.5 1 2 3

42°23'52"N + 74°34'16"W

74°40'07"W
Figure 2

Event 1

0 cm 5 cm 10 cm

Event 6

0 cm 5 cm 10 cm

Semivariance

- 0.4 to 0.5
- 0.3 to 0.4
- 0.2 to 0.3
- 0.1 to 0.2
- 0.0 to 0.1
Figure 3

![Graphs showing semivariance for Event 1 to Event 6 with different threshold levels at 0 cm, 5 cm, and 10 cm.](image-url)

- Threshold at 0 cm
- Threshold at 5 cm
- Threshold at 10 cm
Figure 4
Legend

**Indicator Values**
- 0 (Unsaturated)
- 1 (Saturated)

**Probability of Saturation**

- 0 to 10%
- 10 to 20%
- 20 to 30%
- 30 to 40%
- 40 to 50%
- 50 to 60%
- 60 to 70%
- 70 to 80%
- 80 to 90%
- 90 to 100%

A) Indicator kriging with hard data alone

B) Indicator kriging with soft data alone

C) Indicator kriging with combination of hard and soft data