Soil erodibility and processes of water erosion on hillslope

Rorke B. Bryan

Soil Erosion Laboratory, Faculty of Forestry, University of Toronto, 33 Wilcocks Street, Scarborough, Ontario, Canada M5S 3B3

Received 1 June 1998; received in revised form 1 June 1999; accepted 17 August 1999

Abstract

The importance of the inherent resistance of soil to erosional processes, or soil erodibility, is generally recognized in hillslope and fluvial geomorphology, but the full implications of the dynamic soil properties that affect erodibility are seldom considered. In Canada, a wide spectrum of soils and erosional processes has stimulated much research related to soil erodibility. This paper aims to place this work in an international framework of research on water erosion processes, and to identify critical emerging research questions. It focuses particularly on experimental research on rill and interrill erosion using simulated rainfall and recently developed techniques that provide data at appropriate temporal and spatial scales, essential for event-based soil erosion prediction. Results show that many components of erosional response, such as partitioning between rill and interrill or surface and subsurface processes, threshold hydraulic conditions for rill incision, rill network configuration and hillslope sediment delivery, are strongly affected by spatially variable and temporally dynamic soil properties. This agrees with other recent studies, but contrasts markedly with long-held concepts of soil credibility as an essentially constant property for any soil type. Properties that determine erodibility, such as soil aggregation and shear strength, are strongly affected by climatic factors such as rainfall distribution and frost action, and show systematic seasonal variation. They can also change significantly over much shorter time scales with subtle variations in soil water conditions, organic composition, microbiological activity, age-hardening and the structural effect of applied stresses. Property changes between and during rainstorms can dramatically affect the incidence and intensity of rill and interrill erosion and, therefore, both short and long-term hillslope erosional response. Similar property changes, linked to climatic conditions, may also significantly influence the stability and resilience of plant species and vegetation systems. Full understanding of such changes is essential if current event-based soil erosion models such as WEPP and EUROSEM are to attain their full potential predictive precision. The complexity of the interacting processes involved may, however, ultimately make stochastic modelling more effective than physically based modelling in predicting hillslope response to erodibility dynamics. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: soil; erodibility; water erosion

1. Introduction

Transport of sediment from hillslopes to valleys where it is accessible to fluvial processes is of central importance in geomorphology. In soil-mantled parts of the world, the geomorphic and hydrologic processes involved in hillslope sediment transport are strongly influenced by soil properties. This influence has been recognized in principle (e.g., Chorley, 1959), but in practice, the effect of the dynamic complexity of surface soil properties on...
hillslope sediment delivery processes has been largely ignored by geomorphologists. More attention has been paid in mass wasting studies using soil engineering techniques, but these usually involve subsoils with significantly less complex properties. Many erosion studies by agricultural engineers have studied true, “pedogenic” surface soils, but have concentrated on disturbed, homogenized “agricultural” soils and gentle slopes where some properties and processes of importance in geomorphology and hydrology are suppressed.

Many erosional processes are active on hillslopes, but in this paper, we focus exclusively on those involving rainsplash and runoff. Bennett (1926) initiated research on soil properties related to erosion resistance in Cuba, but Middleton (1930) formalized the erodibility concept, proposing two indices of soil erodibility combining properties affecting runoff and particle detachability, linked to field behaviour of soils in Carolina. Over the next fifty years, many studies reviewed by Smith and Wischmeier, 1962; Bryan, 1968, 1991; De Ploey and Poesen, 1985; Romkens, 1985; Morgan, 1986; Lal, 1990 tested, refined, modified or replaced these indices in an elusive search for consistent, universally applicable indices of erodibility. These studies have been useful in agricultural land management, but have provided limited insight on geomorphologic or hydrologic problems. The object of the present paper is to review concepts of soil erodibility in the light of recent research, particularly in Canada, to examine why the available information has not been more useful in geomorphology and hydrology, and to identify important directions for future research.

Several obstacles restrict application of soil erodibility information to problems of sediment transport and slope development. The distribution of erosive forces and soil resistance differs for each of the sub-process active on hillslopes, and soil erodibility can be defined only in relation to specific processes. Few studies clearly define the processes active or their temporal and spatial variation. In many cases, study methodology precludes precise process examination. In the plot studies which led to the Universal Soil Loss Equation, for example, the soil erodibility (“K”) factor was inferred from soil loss measured at the end of a 22-in. long erosion plot, rather than by direct observation. The K factor is therefore a “grey box” variable, at best, which may integrate the effects of process: property interactions throughout the plot, or only those active near the sediment collection trough. Study methods have frequently excluded or suppressed processes active on natural hillslopes, or even on agricultural land. The small plots used in many studies (Bryan, 1991) often preclude rill development, which is particularly important as erosion rates usually increase after rill incision, significantly affecting erodibility ranking.

A second problem is that most of the detailed soil erosion information comes from studies on agricultural soils, where natural soil profile features have been homogenized, macroporosity largely obliterated, and new features such as plough pans created. Such disturbance changes water partitioning and hydrologic response during storm events, and alters the recurrence frequency of some erosion processes. Hortonian overland flow and related erosion processes, for example, are exaggerated on agricultural erosion plots, while subsurface processes tend to be suppressed. Prolonged agricultural disturbance also changes soil structure and organic content, often significantly reducing soil resistance. The limitations of small runoff plots for hydrologic prediction are well known, and it is somewhat surprising that the comparable limitations of the use of agricultural soils to predict natural soil behaviour have not been equally widely recognized.

The third problem is temporal variability of the soil properties which control erodibility. Soil erodibility was originally thought to be controlled by properties such as texture (e.g., Middleton, 1930; Bouyoucos, 1935) which change slowly, so that soil erodibility would remain essentially constant for a given soil. In fact, the properties that dominate erosional response, such as aggregation and shear strength, are highly dynamic, changing over short and long-term cycles of varying magnitude and predictability. This has often been recognized, but the implications for soil erodibility prediction generally have not.

Difficulties experienced in applying available soil erodibility information to prediction of erosion on natural hillslopes, or even on agricultural lands, have lead to a search for more explicit physically based understanding of erosional processes, which has dominated soil erosion research over the past thirty
years. While this has produced major advances in understanding of erosion processes, formidable research questions still remain.

2. Water erosion processes on hillslopes

Soil erodibility can be defined only in relation to specific erosion processes and erosive forces, so discussion of critical process characteristics is essential in understanding the effect of the dynamics of related soil properties. Rainsplash and runoff energy are the active erosive agents and if gully erosion, which often involves complex regional geomorphic factors, is excluded (as it is from this paper), they produce five more or less distinct sub-processes: splash erosion, sheetwash, rainflow, rill erosion and piping or tunnel erosion. Each can act in isolation, but all are commonly active on hillslopes, either sequentially or simultaneously. These sub-processes can be classified in various ways, but for geomorphic purposes the most useful distinction is between interrill processes in which entrainment is primarily caused by rainsplash energy, and rill erosion or piping, in which it is caused by runoff.

2.1. Interrill processes

Splash erosion is driven by rainsplash kinetic energy, which depends on raindrop characteristics (mass and impact velocity; $KE = 1/2mv^2$) closely linked to rainfall type and intensity. Effective availability of this energy depends on soil conditions, some is expended in deformation, wetting and disruption of soil particles, while the remainder is converted to a reactive force acting upwards and away from the point of impact. The distribution of these forces has been analysed by, e.g., Rose (1960), De Ploey and Savat (1968), Ghadiri and Payne (1981), Poesen and Savat (1981), Savat and Poesen (1981) and Park et al. (1982). The reactive force, which can entrain and transport soil particles, depends on rainfall characteristics, as modified by wind and canopy disturbance, and on soil water conditions. Wind disturbance can increase impact velocity and change impact angle, while canopy disturbance alters drop size spectra and impact velocities. Soil water is an important control and the splash process typically occurs in three stages (Yariv, 1976). On dry, loose soils, much energy is expended in particle disruption or deformation, but as soil water increases, shear strength drops, the soil becomes fluidized, and increasingly vulnerable to entrainment. The final stage follows ponding, as rainsplash starts to interact with overland flow. The effect of a surface water layer on splash detachment is controversial, Palmer (1963) and Mutchler and Larson (1971) reported increased detachment up to a threshold depth approximately equal to the median raindrop diameter ($d$), but other workers have found little increase with layers up to $3d$ in depth, followed by marked decrease (Moss and Green, 1983; Kinnell, 1990). Microtopographic roughness is sufficient on most slopes to produce discontinuous ponding and to suppress splash entrainment, but enhance particle transport in rainflow erosion. Raindrop impact also affects erosion processes by modifying flow hydraulics, as discussed below.

The importance of rainsplash energy has been confirmed by many empirical studies (e.g., Young and Weirsma, 1973), and is reflected by the dominant role assigned to rainfall erosivity in most soil erosion models. This can result in underprediction of soil erosion rates in areas where high-energy rainfall is rare, but overland flow is still frequent, as shown by De Ploey (1972) in Belgium and Bryan (1981) in Tanzania. Drop size has traditionally been measured by “flour tray” or stain methods (e.g., Laws, 1941; Laws and Parsons, 1943). However, recent studies in Canada (e.g., Sheppard and Joe, 1994) with more precise drop measurement by particulate measurement spectrometers, disdrometers and doppler radar, indicate that these early studies overestimated natural raindrop size, and therefore rainsplash energy. Nevertheless, rainsplash energy is an important erosive agent in splash and rainflow erosion which, by modifying soil surface properties and flow hydraulics, can strongly influence interrill and rill processes.

Shallow interrill flow has little entrainment capacity without raindrop impact, but runoff energy is critical for transport in sheetwash and rainflow. Runoff energy reflects flow discharge and hydraulics, which are strongly influenced by soil surface properties, microtopography and vegetation, and are therefore highly variable. These interactions are vital in physical definition and modelling of runoff.
erosion processes, and have been intensively studied. Studies of the applicability of deep channel hydraulic relationships to overland flow (e.g., Horton et al., 1934; Emmett, 1970; Phelps, 1975; Savat, 1977) have found significant differences. Deep channel flow is usually subcritical, hydraulically smooth and turbulent, with perimeter roughness elements entirely submerged by the viscous sub-layer. Overland flow, usually on significantly steeper slopes, is much thinner and often discontinuous, with depths that can vary by an order of magnitude over a few centimeters. Flow is frequently supercritical and conditions can vary between fully rough, transitional and hydraulically smooth over very short distances, as roughness elements often penetrate the viscous sub-layer, or the complete flow. Rainsplash energy has little effect on deep channel flow, but can strongly influence shallow flow hydraulics.

While the hydraulic conditions of overland flow determine the erosive forces acting on the soil, soil properties can also modify these conditions, most notably through their influence on surface roughness. Roughness relationships in overland flow are complex, and relative roughness (roughness element height:flow depth) is usually large. Hydraulic conditions vary significantly over short distances with changes in flow depth, or surface alteration during and between flow events. Both can be strongly affected by, for example, selective disintegration and transport of soil aggregates, differential swelling of clay minerals, surface sealing and crusting, and desiccation cracking. When roughness elements are fully submerged, hydraulic roughness declines with flow depth (shown by an inverse relationship between the Darcy–Weisbach friction factor \( ff = 8 g R S u^2 \) and flow Reynolds number \( Re = 4 \mu R / \nu \) (e.g., Savat, 1980, Gilley et al., 1990), with the relationship slope depending on flow regime. Flow resistance can be partitioned into grain and form components but in hydraulically smooth flow, grain resistance is dominant. When roughness elements equal or exceed flow depth, the \( ff:Re \) relationship becomes positive or convex (Abrahams et al., 1986; Abrahams and Parsons, 1991; Nearing et al., 1998), the significance of grain resistance declines (often to < 10% of total resistance) (Govers and Rauws, 1986; Rauws, 1988; Abrahams et al., 1996), and sediment movement is controlled by form resistance. However, the transition varies with soil characteristics. Fig. 1 shows data from simulated rainfall flume experiments in the University of Toronto Soil Erosion Laboratory; the well-aggregated Gobles silt loam (from S.W. Ontario, Canada) shows a clear positive trend above \( Re \) values around 5000, but the more erodible Swinton silt loam (from Saskatchewan, Canada) shows no
similar trend with $Re$ values up to 10000. The clay fraction of the chernozemic Swinton soil is dominated by smectite, and when the soil is saturated unstable, low-density aggregates disintegrate easily, leaving a smooth surface with few protrusions. Form resistance may be influenced by large stable aggregates, as on the Gobles soil, and also by stones, microtopographic roughness, scour features and knick points, vegetation fragments, and by raindrop impact disturbance.

The relationships described depend on surface roughness and flow depth, and relate directly to soil erodibility assessment. In hydraulically smooth flow, sediment transport is controlled by grain shear stress (Singhal et al., 1980), and is resisted by the complete bed surface, but in hydraulically rough flow, shear stress is concentrated on, and resisted by roughness elements. In the example cited, the high erodibility of the Swinton soil results in hydraulically smooth conditions at higher Reynolds numbers than on the resistant Gobles soil. On actual hillslopes, of course, conditions can vary greatly over very short distances, reflecting a complex interplay of factors. Flow depth at any location reflects the balance between rainfall rate, flow delivery from upslope and infiltration rate. This changes continuously during storms, but depth generally increases with storm intensity and duration, producing a systematic temporal change from hydraulically rough to smooth conditions.

At any instance, a similar systematic spatial change would be expected with downslope increase in flow depth. Interaction of other factors, such as differences in infiltration characteristics, often related to surface sealing can make this difficult to demonstrate (Bryan and Poesen, 1989), but it has been shown on rough, semi-arid hillslopes at Walnut Gulch, AZ (Abrahams et al., 1989; Parsons et al., 1990). Relative roughness at any location also varies over time with changes in the soil surface, such as sealing or crusting, hydrocompaction, swelling, selective erosion or deposition (Slattery and Bryan, 1992a), but the most significant change is usually rill initiation. Conditions in interrill overland flow vary, but most flows are hydraulically rough so that form resistance dominates. Once flow concentrates into rill channels many roughness elements are submerged, and depth increases downchannel, producing hydraulically smooth flow (Gilley et al., 1990; Parsons et al., 1990). A clear exception to this generalization is the incidence of cyclic scouring or rilling, where there is no systematic downslope increase in flow depth. This has been demonstrated in several experimental flume studies at the University of Toronto Soil Erosion Laboratory (Bryan and Poesen, 1989; Bryan and Oostwoud Wijdenes, 1992). In these experiments, cyclic scouring on a sealing soil resulted in intermittent infiltration loss, limiting downslope increase in discharge. However, in other experiments, Bryan and Brun (1999) have shown a similar effect, due to cyclic deposition of miniature alluvial fans. These cyclic scour and rilling processes are discussed further below.

Hydraulic differences between interrill and rill flow not only reflect surface roughness:flow depth relationships, but also raindrop impact. Many studies have reported increased erosive capacity of shallow interrill flow due to rainsplash (e.g., De Ploey, 1971; Walker et al., 1978; Kinnell, 1985; Guy et al., 1987; Moss, 1988; Moss et al., 1979). Interrill processes include a complex mixture of rainsplash, sheetwash and rainflow erosion, which varies spatially with flow depth and temporally with variations in rainfall intensity and raindrop size. The effect of raindrop disturbance on flow hydraulics is less clear. Yoon and Wenzel (1971) and Shen and Li (1973) found significant increase in flow depth and Darcy-Weisbach friction factors, but Smerdon (1964) and Savat (1977) reported only modest changes, particularly on steeper slopes.

### 2.2. Rill erosion

The transition from interrill to rill processes is critical both for erosion rates and the geomorphic evolution of hillslopes. Interrill processes act intermittently over most parts of the hillslope, and are strongly influenced by rainsplash energy, while rill processes are concentrated and not directly affected by rainsplash. Because of the importance of the interrill:rill transition, many studies have examined threshold conditions for rill initiation (Bryan, 1987), though most have focused almost exclusively on hydraulic conditions with little or no attention to the influence of soil properties. Rill erosion involves flow concentration, often caused on natural hillslopes by microtopography, vegetation, or animal tracking,
and on agricultural soils by tillage. The resulting rills may be randomly or systematically distributed. Flow concentration alone does not necessarily cause rill incision (Emmett, 1970; Dunne and Dietrich, 1980), and if threshold hydraulic conditions are reached, is not an essential prerequisite. Some examination of hydraulic conditions is therefore a necessary basis for discussion of the role of soil properties.

Horton’s (1945) concept of drainage basin evolution linked rill initiation to the threshold tractive force for particle entrainment and transportation, which depends on both flow conditions and soil surface properties. This can occur in any flow, but the necessary flow depths are usually reached only in concentrated flow. Many proposed hydraulic indices of rill initiation are related directly to critical tractive force. These include shear velocity (Govers, 1985), bed shear stress (Chisci et al., 1985; Torri et al., 1987), stream power (Rose, 1985), unit stream power (Govers and Rauws, 1986; Moore and Burch, 1986), and either unit or total discharge (Meyer et al., 1975). Other authors have invoked moulding of mobile beds by supercritical flow (Savat and De Ploey, 1982; De Ploey, 1983; Bryan, 1990), or flow instability features (Rauws, 1987), linked with roll waves (Ishihara et al., 1953; Karcz and Kersey, 1980) or secondary flow cells (Moss et al., 1982; Merritt, 1984).

It is difficult to reconcile all results of the studies cited because of varied study conditions, including field plots and laboratory flumes of diverse design and dimensions, slope gradient and slope shape, both with and without rainfall. Not surprisingly, hydraulic indicators, which are effective in some studies, prove irrelevant in others. In many cases, information necessary to determine whether differences are real or artifacts of experimental procedures is not available. Nevertheless, some general conclusions can be suggested. The most consistently useful hydraulic indices appear to shear velocity \((u^* = gRS^{0.5})\), stream power \((w = pggS)\) and unit stream power \((y = r\alpha)\).

Govers (1985) identified a threshold \(u^*\) value of 0.03–0.035 m s\(^{-1}\) for calcareous loess in Belgium. Above this value, which Govers linked to a selective soil particle transport, flow sediment concentrations increased rapidly. The critical \(u^*\) will depend on soil properties. The Belgian calcareous loess tested by Govers is exceptionally erodible; in comparative laboratory tests of 12 soils from Canada, Belgium and Israel, Bryan and De Ploey (1983) found this soil to be much less resistant than even erodible loess soils from Israel and from the Rocky Mountain foothills in Canada. Significantly higher threshold \(u^*\) values have, for example, been reported for a range Canadian soils, by Bryan (1990), Slattery and Bryan (1992a) and Merz and Bryan (1993), including both sealing sandy loams and more aggregated clay-rich soils. In Fig. 2a, Govers’ \(u^*:O\) relationship is compared with experimental relationships established in the University of Toronto Soil Erosion Laboratory for the Gobles and Swinton silt loams, from Ontario and Saskatchewan, respectively. Even the Swinton soil, shown in flume experiments to be quite erodible (Bryan, 1996), shows a critical \(u^*\) value significantly above Govers’ value, which appears to be a minimum for the least resistant soils.

Rose (1985) used stream power \((w)\), proposed as a bedload formula by Bagnold (1960) and defined as the power per unit area of streambed, as an index for rill incision with a threshold value 0.5 W m\(^{-2}\). Nearing et al. (1998) found \(w\) to be the most consistently reliable indicator of unit sediment load across a range of experimental conditions and soil types. However, Govers and Rauws (1986) found unit stream power \((y)\), proposed as a total load formula by Yang (1973) and defined as power per unit weight of water, to be a good predictor of sediment concentration in thin flows on non-cohesive beds, with threshold values of 0.006–0.008 in. s\(^{-1}\). Moore and Burch (1986) also found that \(y\) fitted results from several experiments well and suggested a threshold value of 0.002 m s\(^{-1}\) as a possible constant for many soils.

The studies cited indicate that \(u^*\), \(w\) and \(y\) are all useful hydraulic indices for rill initiation, but data for the Gobles and Swinton silt loams in Fig. 2 support the Govers and Rauws’ (1986) use of \(y\) as a preferred index. The considerable scatter of data reflect
some system variability, but also the difficulty of obtaining accurate hydraulic measurements in very shallow flows and of identifying the initial stages of rill incision. Another variable linked with rill initiation is the Froude number \( Fr = u / gR^{0.5} \) indicating flow condition. Supercritical flows \( (Fr > 1) \) are associated with flow instability features, such as standing waves, which cause localized increase in bed shear stress. \( Fr \) values in rivers are usually \(< 1\), but supercritical flow is frequent in shallow flows. Several authors (e.g., Ishihara et al., 1953; Savat, 1976; Hodges, 1982; Bryan, 1990; Bryan and Oostwoud Wijdenes, 1992) have reported standing waves and roll waves at \( Fr < 1\), which can increase local bed shear stress by three- to fivefold (Horton, 1938), generate bed deformation and trigger rill incision. Bed deformation is usually associated with a most cohesionless soils, such as highly erodible loess from Belgium (Savat, 1976) or China (Luk et al., 1989). However, Hodges (1982) reported rill scouring caused by roll waves in smectite-rich depositional soils in Dinosaur Park Badlands, Alberta, and Bryan and Oostwoud Wijdenes (1992) observed similar features in field experiments on clay-rich lacustrine soils near Lake Baringo in northern Kenya. In these cases, soil cohesion apparently dropped below criti-

Fig. 3. Cyclic rill scour patterns in runon flume experiments on the Pontypool/Peel sandy loam (after Bryan and Brun, 1999).
cal levels due to high soil water contents. In virtually all experiments where flow instability triggered rill scouring test conditions produced positive soil water potential for most of the test. More detailed examinations of these conditions by Bryan and Rockwell (1998) are discussed below.

The effect of hydraulic conditions on rill initiation and has been extensively studied, and the hydraulic geometry of evolving rill networks has received some attention (Schumm, 1956; Kashiwaya, 1978; Rown-tree, 1982; Schumm et al., 1987), but the impact of network geometry on sediment delivery has been almost entirely ignored. This is an important omission as entrained soil rarely moves directly to the outlet. As a result measurements at a terminal weir may bear little relationship to upslope erosion patterns. Even in simple rill systems, intermittent storage and remobilisation occurs, leading to cyclic rill patterns, such as those shown in Fig. 3 developed in Bryan and Brun’s (1999) flume experiments. Bryan and Poesen (1989) showed that such intermittent scour and deposition can complicate hydrologic response, as localised scour below erosional knickpoints produces short-term peaks in infiltration rates. The dramatic effect of different soil conditions on rill network configuration has been shown by experiments in the University of Toronto Soil Erosion Laboratory (Fig. 4). Fig. 4a shows limited network development on a sealing soil by 2 h of intense rainfall, strongly influenced by subsurface moisture and seepage (Bryan et al., 1998). Fig. 4b shows a very different high-density network developed on the

Fig. 4. (a) Rill network development on the Pontypool/Peel sandy loam after 120 min of simulated rainfall at 60 mm h \(^{-1}\). (Note complete flume measures 7.1 × 2.4 m. Transverse markers are at 1-m intervals.) (b) Rill network development on the Gobles silt loam after 120 min of simulated rainfall at 60 mm h \(^{-1}\). (Note complete flume measures 7.1 × 2.4 m. Transverse markers are at 0.5-m intervals.)
much more cohesive, resistant Gobles silt loam under similar rainfall conditions (Brunton and Bryan, 1999).

The clear differences in rill network geometry with soil type, shown in Fig. 4 are of importance because the potential for temporary sediment storage within the rill system increases with the network complexity, reflecting the effect of confluence characteristics on local hydraulic conditions. Sediment transport patterns have received little attention in previous studies such as those of Schumm et al. (1987), which focused primarily on rills as potential analogues for large river systems. Most work on the effect of confluences on sediment transport has been carried out in large river channels (e.g., Best, 1988; Roy and Bergeron, 1990; Roy et al., 1988), focusing particularly on secondary flow circulation patterns (e.g., Rhoads and Kenworthy, 1998, 1999; Lane et al., 1999).

The applicability of data from the deep-channel confluence studies to rill-scale confluences where the standing waves produced may greatly exceed flow depth (Fig. 5) is not clear. In one of the very few rill-scale studies available Mosley (1976) showed that tributary confluence angle and relative discharge can strongly influence bed scour and deposition both at the confluence and downchannel. This study involved artificial sediments in a very low angle flume with run-on discharge. Brunton and Bryan (2000) studied a more natural rill system, developed on Gobles silt loam in a 5° inclination flume under simulated rainfall (Fig. 4b), and compiled preliminary sediment budgets. These showed marked, but very complex effects of confluence evolution on sediment storage patterns. Considerable further experimental work, with comparable data from a wider range of soils is essential to establish a general model of the effect of rill network geometry.

Fig. 5. Hydraulic disturbance and standing wave at artificial rill confluence in experiments at University of Toronto Soil Erosion Laboratory.
on hillslope sediment production, and the way in which this is linked to soil erodibility. The situation is complicated as rill system characteristics are often ephemeral, changing significantly during individual rainstorms. Dunne and Aubry (1986) described expansion and contraction of rill system on sandy soils at Amboseli, Kenya, in response to the balance between rill and interrill processes. Interrill erosion was dominant in low and moderate intensity storms, delivering more sediment to rills than could be evacuated, causing network shrinkage. In prolonged or high intensity storms, rill evacuation exceeded interrill supply, and the network expanded. A detailed sediment budget for a rill network can therefore provide a sensitive index of the influence of different soils on hillslope response to variations in storm conditions.

2.3. Pipe and tunnel erosion

While research in soil erosion by water has concentrated overwhelmingly on analysis of surface processes, subsurface erosion in pipes or tunnels has long been recognised as a significant agent in sediment transport and hillslope development, particularly in dryland regions. Subsurface erosion is also very important in temperate mountainous peatlands and in deep highly weathered humid tropical soils, but several reviews (Gilman and Newson, 1980; Jones, 1981; Higgins and Coates, 1990; Bryan and Jones, 1997) indicate that small-scale pipes or tunnels are probably almost as ubiquitous as rills. Piping develops spontaneously by outlet sapping where soil water potential is positive and high hydraulic gradients produce seepage forces that can eject particles and enlarge fabric macropores. Forces close to the surface (analyzed in detail by Iverson and Major, 1986; Howard and MacLane, 1988; Dunne, 1990) can usually eject only small particles from soils of low cohesion, so true piping is usually found only in saturated dispersed clays, loess and organic soils. Tunnel erosion exploits existing large macropores such as animal burrows, root channels and desiccation cracks. These can occur in any soils, but are particularly common where swelling smectites clays coincide with intense seasonal drying, and where soil chemistry favours dispersion. Piping and tunnelling are physically distinct, but are often functionally indistinguishable, and are usually grouped as piping.

Descriptions of many piping systems have been published, and linkages to soil properties such as clay mineralogy, sodium adsorption ratio and exchangeable sodium percentage are well-established. Very few measurements of flow conditions or sediment movement within systems are available. Bryan and Harvey (1985) instrumented two large pipe systems in the Dinosaur Badlands, Alberta, Canada. While piping slightly delayed storm hydrograph peaks, sediment and solute concentrations were significantly higher than in surface ephemeral channels. Bryan et al. (1978, 1983), in the same badlands, and Yair et al. (1980) in the Zin Badlands in Israel, found that sediment concentrations and hydrologic response of micropipe flow was indistinguishable from rill flow. However, Torri et al. (1994) found that sediment concentrations in micropipe flow were higher than in rills on clay biancane in Tuscany, Italy. Direct hydraulic measurements in micropipes are very scarce, but values in Table 1, measured in adjacent rills and micropipes formed in flume experiments at the University of Toronto Soil Erosion Laboratory (Fig. 6) are very similar.

Few attempts have been made to analyze linkages between piping and rill initiation, except on sodic smectites in badlands, where rills develop by col-

<table>
<thead>
<tr>
<th>Time</th>
<th>( Q_y )</th>
<th>( u^* )</th>
<th>Re</th>
<th>Fr</th>
<th>( f_f )</th>
<th>w</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rill</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>81</td>
<td>29.36</td>
<td>6.15</td>
<td>1555</td>
<td>0.989</td>
<td>0.498</td>
<td>1.179</td>
<td>1.502</td>
</tr>
<tr>
<td>83</td>
<td>23.20</td>
<td>5.17</td>
<td>1054</td>
<td>1.125</td>
<td>0.385</td>
<td>0.705</td>
<td>1.438</td>
</tr>
<tr>
<td>94</td>
<td>40.75</td>
<td>5.16</td>
<td>940</td>
<td>1.013</td>
<td>0.475</td>
<td>0.633</td>
<td>1.290</td>
</tr>
<tr>
<td>107</td>
<td>34.06</td>
<td>5.14</td>
<td>970</td>
<td>1.058</td>
<td>0.436</td>
<td>0.658</td>
<td>1.342</td>
</tr>
<tr>
<td>116</td>
<td>37.27</td>
<td>5.59</td>
<td>978</td>
<td>0.829</td>
<td>0.710</td>
<td>0.673</td>
<td>1.144</td>
</tr>
<tr>
<td>Micropipe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>63</td>
<td>25.33</td>
<td>5.47</td>
<td>736</td>
<td>0.655</td>
<td>1.102</td>
<td>0.882</td>
<td>1.342</td>
</tr>
<tr>
<td>73</td>
<td>23.14</td>
<td>6.09</td>
<td>975</td>
<td>0.637</td>
<td>1.202</td>
<td>0.846</td>
<td>0.959</td>
</tr>
<tr>
<td>87</td>
<td>39.13</td>
<td>6.44</td>
<td>1262</td>
<td>0.700</td>
<td>0.996</td>
<td>1.091</td>
<td>1.112</td>
</tr>
<tr>
<td>101</td>
<td>6.09</td>
<td>994</td>
<td>0.649</td>
<td>1.157</td>
<td>0.863</td>
<td>0.977</td>
<td></td>
</tr>
<tr>
<td>113</td>
<td>25.98</td>
<td>6.32</td>
<td>1202</td>
<td>0.705</td>
<td>0.980</td>
<td>1.079</td>
<td>1.100</td>
</tr>
</tbody>
</table>
Fig. 6. Micropipes and rills formed in laboratory experiments on Swinton silt loam. Micropipes developed along interstorm desiccation cracks.
lapse of micropipes formed along desiccation cracks (Bryan et al., 1978; Hodges, 1982; Gerits et al., 1987). Torri and Bryan (1997) found a very close interaction between rill development and micropiping on clay biancane in Tuscany (Fig. 7), but this piping was caused by soil dissolution due to capillary moisture diffusion, rather than positive soil water potentials. Until recently it was generally believed that vulnerability to rilling and piping is quite distinct, and the processes rarely coincide, except on extreme badland soils. Recent studies suggest, however, that close linkage is actually quite common on other soils, particularly when near-surface drainage is impeded due to natural soil profile features or artificial plough pan development. Recent experimental studies at the University of Toronto Soil Erosion Laboratory which linked rill network development to soil water conditions monitored by micro time domain reflectometer probes (Bryan et al., 1998), have shown that both surface and subsurface processes simultaneously influenced rill initiation, and neither the processes nor their morphological effects could be clearly separated. On field soils, partitioning between rill and pipe processes will usually reflect short-term weather conditions. Bryan (1996) showed that both rills and micropipes could develop on the Swinton silt loam chernozem (Fig. 6) depending on precise antecedent soil water conditions, and Govers (1987) also identified close links between rills and piping caused by mole burrows on loess soils at Huldenberg, Belgium.

3. Soil properties and soil erodibility

The preceding review of rill, interrill and piping processes has identified numerous circumstances in which erosion incidence and intensity can be strongly influenced by soil properties. Many studies have examined the effect of soil properties on erosion, but with a wide range of methodologies, soil types, climatic conditions, and soil management histories, different properties have proven effective in different situations. Almost any soil property may influence erosion response, but, as Lal (1990) has pointed out, no single, simple, measurable soil property can fully represent the integrated response that constitutes soil erodibility. In practice, a few properties, particularly soil aggregation, consistency and shear strength, usu-
ally dominate erosional response, and other properties are only indirectly effective. These properties collectively influence water movement, the distribution of erosive forces, and resistance to entrainment. In most cases, the most important initial effect is on the way in which soils respond to rainfall.

3.1. Soil behaviour under rainfall

Water movement into, through and over the soil is of paramount importance in soil erosion. It encompasses infiltration, percolation and retention. These are largely controlled by the volume, size, distribution and continuity of pore space, and therefore, the skeletal framework or geometric arrangement of textural particles and aggregates. This framework is dynamic, changing at varying rates through physical and biochemical processes (Horn, 1988, 1994; Horn et al., 1994). Sometimes in undisturbed soils, textural particles and aggregates exist in a non-coherent "card-house" fabric, but most commonly they are joined by bonds of varying character and strength into a coherent structure. The fabric of soils disturbed by tillage, on the other hand, is often non-coherent. An important, and largely ignored, issue in soil erosion studies is the rate at which coherence is reestablished. The reestablished, coherent fabric differs significantly from the non-coherent fabric, but also from that of undisturbed soil. Coherence is not synonymous with cohesion and the structure of cohesive soils is typically non-coherent after disturbance. Transition from non-coherence to coherence, described as "welding" by Kwaad and Mucher (1994), involves re-establishment of bonds, which can take place slowly between, or more rapidly during rainstorms. On most disturbed soils coherence develops quite quickly, but soils such as highly flocculated calcareous soils or tropical oxisols or nitosols (Ahn, 1979), can remain non-coherent almost indefinitely.

The complete soil profile affects water relationships and changes at depth due to swelling and shrinkage, dispersion, hydrocompaction and clay migration, can affect erosional response. However, most attention has been focused on the effect of surface sealing or crusting on infiltration, ponding and runoff generation. Many studies have examined seal structure and micromorphology, formation processes and rates, relationship to soil properties and rainfall characteristics, and impact on soil water relations (e.g., McIntyre, 1958a,b; Farres, 1978; Moore, 1981; Callebaut et al., 1985; Norton, 1987; Mualem et al., 1990; Römkens et al. 1990; Sumner and Stewart, 1992; Valentin and Bresson, 1992; Le Bissonnais, 1996; Assouline and Mualem, 1997). It is impossible to review all relevant studies comprehensively here, but research on surface sealing has been notably prominent in Canada, particularly on smectite-rich solonetzic and chernozemic soils developed over Cretaceous marine shales in western Alberta (Bryan, 1973; Arefad and Mermut, 1988; Römkens et al., 1995). Despite differences in soil properties and experimental conditions, there is agreement about the processes involved (raindrop compaction, clay migration and filtration, vesicle formation, and, possibly, clay skin orientation), the impact on infiltration (marked decline), and linkage of sealing susceptibility to soil aggregation. Moss (1991) has described a "seismic wave" effect as raindrops hit the surface causing collapse of pores, especially in silty soils. Unstable aggregates break down rapidly under rain-splash and wetting stress, providing particles for seal formation, resulting in ponding and runoff generation. The degree of breakdown varies greatly with aggregate characteristics, but is also strongly influenced by initial soil moisture conditions, typically being most rapid on dry soils (Bryan, 1971a; Luk, 1985; LeBissonnais, 1990, 1996; Loch and Foley, 1994). Slattery and Bryan (1992b) have described a typical sequence of seal development from experiments in the University of Toronto Soil Erosion Laboratory. Seal properties and formation rates vary with soil and rainfall characteristics, but they usually form almost completely in the first 5–10 min of rainfall when most loose particles are available. Virtually all studies of surface sealing have involved disturbed soils, and the significance of sealing processes on undisturbed, coherent soils is largely unknown. Most studies have also used very small soil samples, and so provide little information on the spatial continuity of sealing on hillslopes or its effects on microtopography and flow hydraulics. It should also be noted that virtually all experimental studies have taken place under high intensity rainfall, and the processes of soil sealing may be quite different under low intensity rainfall, particularly on swelling-clay soils. Because of its hydrologic impact,
sealing susceptibility is generally interpreted as a direct measure of vulnerability to interrill erosion, but, in fact, the effect of a ponded layer on rain-splash entrainment is ambivalent and the increase in bed shear stress with flow depth can be offset by increased seal strength. On a sandy loam soil from southern Ontario, Canada, Slattery and Bryan (1992b) found that seal strength increased only during seal formation in the initial minutes of a rainstorm, and peak splash entrainment actually coincided with peak soil strength. Bradford et al. (1986) also reported positive correlation between splash and soil strength. However, different relationships can be expected on clay-rich soils, particularly when seal development is also influenced by runoff.

Most studies have not differentiated among seal types, but Arshad and Mermut (1988) separated in situ disruptional seals, formed by raindrop compaction and filtration of fines, from depositional seals caused by selective deposition, while Sumner (1995) identified four categories of seals or crusts: chemical crusts; structural crusts; depositional crusts, and cryptogamic crusts. Surface roughness is often increased by structural sealing, with formation of micro hoodoos and depressions (Slattery and Bryan, 1994; Helming et al., 1998), but is decreased by depositional sealing. Both structural and depositional seals can occur on both interrills and rills, but because of the microtopographic amplitude (± 5 mm), the hydraulic impact is greatest on interrill flow. Smooth depositional seals can establish critical conditions for flow instability and rill scour, and play a key role in cyclic rilling (Bryan and Oostwood Wijdenes, 1992; Bryan and Brun, 1999). Bryan and Hodges (1984) and Bryan et al. (1978) also showed that depositional seals play a critical role as the initial loci for runoff and rill generation on smectite-rich mudstones in the Dinosaur Badlands, Alberta.

The change from non-coherence to coherence affects entrainment resistance as well as soil hydrology. On non-coherent surfaces, resistance to entrainment by flow (Hjulstrom, 1935) or splash (Poesen and Savat, 1981) is determined by particle size and mass. As coherence increases, individual particle properties become less important and entrainment is resisted by the shear strength of the coherent soil fabric. Recognition of this transition helps to explain the prominence of certain properties as erodibility indices, and also their apparently erratic performance. Many indices which have proved effective relate to aggregation, either directly (e.g., Yoder, 1936; Adams et al., 1958; Rose, 1960; Bryan, 1969, 1974; Luk, 1979), or through composite measures such as Middleton’s (1930) erosion and dispersion ratios, or Andre and Anderson’s (1961) surface-aggregation ratio. In theory, one would expect aggregation to directly determine the resistance of disturbed, non-coherent soils, but measures of soil strength or consistency to be more effective on coherent soils. No single measure appears to be able to capture fully the complex evolution to a coherent fabric.

Simple compaction and interlocking due to mechanical pressure can produce a coherent fabric, but usually aggregate disintegration and the development of new bonds between released textural particles or aggregate fragments also occurs. This partially explains the complex behaviour of aggregate stability as erodibility index. On the one hand, stable aggregates resist entrainment by flow or splash more effectively, but they also obstruct the development of coherence, which ultimately provides greater strength and resistance against the higher shear stresses generated in channel flow. Young (1974), for example, found that soils with particularly high aggregate breakdown were resistant to rill incision. The ultimate effect on erosion rates will depend on the relative resistance of the soil in coherent and non-coherent states, and on runoff hydraulics which determine partitioning between grain and form resistance. The situation is complicated by the bonding mechanisms, both those responsible for original aggregates and those that form as the soil becomes coherent. These include electrostatic bonds between clay crystals, cation bridges, humic acids or metallic humates, various microbial products, and water cohesion and surface tension bonds. Some bonds, including those of metallic humates, develop slowly and resist change. Others, such as the electrostatic bonds between clay particles, are weaker, and can change rapidly with pore water electrolyte concentration during rainstorms. It is useful to distinguish between relatively strong long-term bonds that develop slowly, and dominate the aggregation and structure of the original, undisturbed soil, and short-term weaker bonds, which can form very quickly and dominate reestablishment of fabric coherence.
3.2. Aggregation

Aggregation is a particularly important factor in erosion resistance and sealing susceptibility of non-coherent soils. This presents some practical problems in assessing soil erodibility. Aggregation is a complex, composite property and there is no general agreement about the most significant aggregation characteristics for soil erosion, or the most appropriate analytical methods. Some appreciation of the processes involved in aggregation processes is necessary to understand the problem. A number of good comprehensive reviews are available (e.g., Harris et al., 1966; Tisdall and Oades, 1982; Swift, 1991; Waters and Oades, 1991; Tisdall, 1994), so these are summarized very briefly.

Aggregate formation involves physical stresses, which can force particles together, or apart, and binding agents, which cement particles with varying strength. Physical stresses include frost action, root action, compaction and shrinkage. Binding agents include humic acids, microbial muscilages, organic metallic compounds, mineral deposits and various electrostatic bonds related to clay crystalline structure, diffuse double layer charges, moisture content and electrolyte concentration (Gerits et al., 1990). Aggregation represents the integrated response to all the complex, dynamic interactions between these stresses and binding agents.

It is now generally accepted that soil aggregation involves a hierarchy of particle sizes characterised by different structure and bonding agents (Tisdall and Oades, 1982; Oades and Waters, 1991). This model, in which organic matter is critical, applies to all soils except those of the humid tropics where iron and aluminum oxides dominate. The basic blocks are clay platelets, bonded by aluminosilicates and organic polymers in units up to 0.2 \( \mu \text{m} \) diameter. These units bond around organic kernels of microbial debris and persistent polysaccharides to form extremely stable, resistant aggregates ranging up to 90 \( \mu \text{m} \) diameter. Larger aggregates up to 200 \( \mu \text{m} \) diameter combine these units with clay flocs, and sand and silt particles around stable cores of plant debris, which are physically shielded from decomposition. These in turn are bound into aggregates up to 250 \( \mu \text{m} \) diameter by roots, fungal hyphae and transient organic matter acting as a “sticky string bag” (Oades, 1993). They may also be strongly cemented by hydrous oxides and carbonates. The terms microaggregate and macroaggregate have been used variably to identify the units described. In soil erosion research, it is most useful to distinguish units up to 250 \( \mu \text{m} \) diameter as microaggregates. These are tightly bonded dense particles of low porosity, distinct from the larger, loosely bonded macroaggregates which incorporate microaggregates, stones, unprotected organic debris and abundant pore space in low density units up to 10 mm or more in diameter.

The splash and runoff entrainment relationships presented by Hjulstrom (1935) and Poesen (1981) would indicate aggregate size as the characteristic of most importance in erosion. However, these relationships were developed with material of uniform specific gravity where size and mass are directly related, and more complex relationships would be expected for composite heterogeneous aggregates. Preferential entrainment of large aggregates of low specific gravity appears particularly common in shallow, hydraulically rough flows (Farenhorst and Bryan, 1995). In flume experiments at the University of Toronto Soil Erosion Laboratory, Bryan and Brun (1999) found that large aggregates move preferentially in shallow flows, and play a critical role in initiating cyclic deposition. In any case, aggregates vary greatly in stability, both between soils and within the same soil, so aggregation properties measured on isolated samples do not necessarily characterize accurately the material affected by erosive stress. Aggregate stability is a complex property and measured resistance, rate of disintegration and character of breakdown products all vary with methods of analysis and particularly the methods of wetting used.

Imeson and Vis (1984a) and Loch (1994), among many others, have comprehensively reviewed methods of aggregate stability analysis. Aggregate stability is defined by the balance between applied stress and material resistance. As moisture content affects both, measured values are strongly influenced by antecedent moisture and wetting conditions. It is therefore important to attempt to match the method used with the specific erosion processes active. On wetting, aggregates are stressed by pore air compression (slaking), differential swelling of components and rainsplash energy. All reach peak intensity and produce maximum breakdown when rain falls on dry
erosion research. An exception is the engineering, but have not been widely used in soil tion. Atterberg limits are routinely measured in soil also be significant for rill and pipe network genera-
tion. Atterberg consistency limits, which empirically de-
indices for sealing and fabric dynamics are probably
strength of the coherent soil. The most effective
tend to disintegrate fairly quickly under wetting
stresses, but the constituent microaggregates are
more much resistant, and will determine the nature
of surface sealing and welding processes and therefore
effective erodibility. It should be noted that
while aggregate size and stability are important in-
fluences on erosional response, aggregate shape and
size heterogeneity, which also influence pore space
geometry in the evolving soil fabric, and surface roughness in shallow flows, have seldom been con-
sidered in erodibility assessment. It is possible that
predictive capacity could be increased by incorpora-
tion of a wider range of aggregation characteristics.

3.3. Soil consistency and shear strength

Aggregate properties influence many aspects of
erosional response, but do not directly determine either the evolution of surface coherence or the shear strength of the coherent soil. The most effective indices for sealing and fabric dynamics are probably Atterberg consistency limits, which empirically de-
fine soil behaviour as a function of changing soil moisture content. The limits immediately relevant are the cohesion limit: the moisture content at which soil becomes cohesive; the plastic limit at which it deforms plastically under stress; and the liquid limit at which it flows under defined, moderate stress. The shrinkage limit, below which soil cracks form may also be significant for rill and pipe network generation. Atterberg limits are routinely measured in soil engineering, but have not been widely used in soil erosion research. An exception is the $C_{S-10}$ ratio which De Ploey and Mucher (1981) introduced as an effective index of sealing susceptibility of Belgian loess soils. This is an extension of the standard liquid limit test that identifies soil behaviour close to satu-
ration. Bryan and De Ploey (1983) found that this ratio correlated quite well with field behaviour in
comparative tests of the erodibility of Canadian, Belgian and Israeli soils.

Soil shear strength appears to be the most important control on the entrainment resistance of coherent soils, but it is difficult to obtain useful measurements for the portion of the soil directly affected by erosive forces. Soil shear strength is a measure of resistance to failure under applied force, and is defined by the Coulomb equation as:

$$\tau = c + \zeta \tan \phi$$

The active components in resistance are cohesion ($c$), the summation of the effect of all the bonding agents effects discussed, and internal friction ($\phi$), which integrates both surface and interlocking friction. Both are affected by soil moisture content and pore pressure conditions, and $\phi$ is also determined by the normal stress ($\zeta$) or overburden pressure. Soil strength is therefore not a unique property, but varies with position in the soil and with stress orientation, and is defined by test conditions. No standard soil strength test is really suitable for soil erosion re-
search. Samples for triaxial and direct shear tests are
removed for laboratory testing, with obvious prob-
lems of fabric disturbance, particularly when stones or roots are present. The vane shear test, which can be used on soil in situ, is preferable, though accurate replication of stress conditions is difficult. Many soil erosion studies have used vane shear tests, with blades of varying depth and diameter (e.g., Luk and Hamilton, 1986; Torri et al., 1987; Coote et al., 1988; Brunori et al., 1989; Govers and Loch, 1993; Merz and Bryan, 1993), and the EUROSEM soil erosion model still employs torvane measurements as a standard index of soil cohesion (Morgan et al., 1998). Vane tests can give useful data on temporal strength variations linked to soil water, but it is difficult to relate strength measurements directly to interrill shear stress.

The greatest difficulty with vane testing is to
to ensure that the soil layer tested is actually the critical zone in soil erosion. In standard tests, the soil depth tested is usually at least several centimeters, while on interrills shear stress affects a much thinner soil layer, usually only a few millimeters thick. Stress conditions in this layer differ from those at depth, and, when surface sealing occurs, the soil fabric is also completely different from that of the underlying
It is also difficult to measure shear strength over the very small areas where stress is concentrated, particularly in hydraulically rough flow, when form resistance may be dominated by individual aggregates or small protuberances. The layer affected by shear stress is always saturated when runoff occurs, but standard instruments cannot measure the very low values typical of saturated soils accurately. Slattery and Bryan (1992b) developed a very sensitive vane shear instrument with 3.5 mm deep blades and a specially constructed helicoidal spring to measure these very low strengths. This instrument accurately reflected temporal changes in soil strength during sealing, but measured strength values still substantially exceeded calculated flow shear stress for demonstrably erosive runoff. A new multi-bladed vane, with individual blades 1.5 mm in height, used with a modified Wykeham Farrance laboratory vane shear apparatus with highly sensitive calibrated springs, has now been developed (Bryan and Kuhn, 2000, under review). This is more sensitive to subtle changes in soil strength, than the original instrument, and experiences considerably less spring fatigue. Despite improved instrumentation, routine direct, reliable measurements of shear strength in the thin surface soil layer immediately vulnerable to entrainment will remain difficult and measured data will provide, at best, an index of actual conditions.

As the surface soil layer directly affected by erosion is very thin, the frictional component of strength is usually assumed to be insignificant. However, recently developed instruments including microtensiometers, micro time domain reflectometers and microautomated standpipes have made accurate monitoring of soil water conditions at very small temporal and spatial scales possible (Bryan et al., 1999). Using these instruments, Bryan and Rockwell (1998) showed that soil water potential, which primarily affects frictional strength, strongly influenced erosion rates in flume experiments. Tests were carried out over an impermeable flume bed at 10–12 cm depth, so that a perched water table invariably developed. In each case, erosion rates increased sharply immediately the perched water table formed, before positive soil water potential occurred at the surface or flow hydraulics changed measurably. The precise process linkage is not yet entirely understood, but the results show that any moisture impeding layer close to the surface can affect erosion processes. The most important immediate implication of these results is that soil erodibility cannot be assessed by the properties of the surface soil alone. Some consideration of the upper soil profile properties is also essential.
4. Tentorial and spatial change in soil erodibility

The properties that affect erosion resistance most directly all vary significantly over time. On many hillslopes, systematic spatial variations also occur caused by slope shape, aspect or microclimate. These can produce major differences in erosional response during rainstorms, as shown by Bryan (1996) in controlled simulated rainfall flume experiments with the Gobles and Swinton silt loams (Fig. 9). The properties reflect many factors, and some variation is essentially random, with effects that cancel out over long time periods. This was recognized when the USLE “K” factor was defined as a long-term variable rather than an index of response in specific rainstorm events. However, some variations follow predictable trends or cycles and understanding of these is important both for physical event-based erosion models and long-term prediction of hillslope evolution. All the factors influencing these trends or cycles are not fully understood, but the most important appear to be frost action, soil water dynamics, microbial action and organic decomposition.

4.1. Frost action

Soil erodibility has usually been treated as a constant, but systematic seasonal variations caused by frost action have long been recognized. Young et al. (1983) noted marked seasonal variations on erosion plots on fallow agricultural soils, and similar patterns occur on natural hillslopes. Schumm and Lusby (1963) and Schumm (1964), for example, described seasonal changes in surface soil properties, and appearance and disappearance of rills on smectite-rich Mancos shale hillslopes in Colorado, reflecting water content changes and frost action in swelling soils. Bryan and Price (1980) described similar patterns, with the seasonal obliteration of rills on the lacustrine Scarborough Bluffs in Toronto, Canada, due to frost action and solifluction.

Frost action is particularly effective in disrupting aggregates when exposed soils are subjected in severe winter climates and a considerable amount of research on the Canadian Prairies has documented the effects on soil erodibility (e.g., Sillanpaa and Webber, 1961; Bisal and Nielsen, 1967; Benoit, 1973). Most commonly frost action disrupts aggregates and increases erodibility, but the effect varies with soil texture, antecedent moisture and rate of temperature change (Mostaghimi et al., 1988). Rapid freezing always causes aggregate breakdown, but Bryan (1971b) found that slow temperature drop can produce ice segregation on “frost susceptible” soils and actually increase aggregate stability. This effect tends to disappear with increased freeze–thaw frequency and Williams (1991) showed peak macroaggregate breakdown with four to five freeze–thaw cycles, with coincident increase in microaggregation. The most severe impact on soil erodibility occurs when an insulating snow cover is thin or absent. In experimental work in southern Ontario, Canada, Pall et al. (1982) and Coote et al. (1988) found that frost action is a particularly critical factor causing high soil erosion rates during spring thaw. Bryan and Rockwell (1998) identified the importance of perched water tables caused by frozen subsoil as a critical factor in spring rill erosion on sandy loam soils near Toronto. Data from agricultural research stations in the United States generally show maximum K values immediately after snowmelt, followed by exponential decline to the end of the growing season (Young et al., 1990). Imeson and Vis (1984b) found a similar pattern with monthly aggregation in forested and grassland soils in the Luxembourg Ardennes.

4.2. Soil water conditions

The single most important control on erosional response during rainstorms is probably soil water and, globally, rainfall variations cause the most marked seasonal changes in credibility. Soil water balance profoundly affects soil structural and hydraulic properties and largely determines erosional response (Cresswell et al., 1992). Effects are particularly marked in arid and semi-arid tropics where both rainfall and drying rates are often intense. Very high erosion rates are frequently observed early in the wet season, though it is often difficult to separate the effect of soil properties from those of storm characteristics and reduced vegetation cover. Dangler and El-Swaify (1977) found that seasonal credibility differences in Hawaii were sufficient to require separate wet and dry season K values. The effect of soil water changes on soil erodibility is complex but, as discussed above, all the critical soil properties are strongly influenced. Antecedent soil water at the start
Fig. 9. (a) Relationship between soil moisture content, soil shear strength and erosion response during sequential flume experiments under simulated rainfall (Gobles silt loam). (b) Relationship between soil moisture content, soil shear strength and erosion response during sequential flume experiments under simulated rainfall (Swinton silt loam).
of rainstorms is particularly important and several studies (e.g., Grissinger, 1966; Luk, 1985; Rauws and Auzet, 1989; Govers et al., 1990; Truman et al., 1990) have shown the effect of antecedent water content on resistance to runoff or splash entrainment. Govers and Loch (1993) and Bryan (1996) have reported up to sevenfold and ninefold differences in mean sediment concentration between dry and wet antecedent conditions. These figures can mask complex relationships, however, as antecedent water content usually varies significantly on hillslopes, affecting sealing and hydrocompaction processes, as well as particle detachability. Aggregate breakdown peaks with sudden wetting of dry soils (Le Bissonnais et al., 1989), but soil shear strength typically declines progressively as water content increases. The net effects can be complex and vary significantly between different soils. Most commonly soil erodibility increases with rainfall duration, but the reverse can also occur. The latter effect is usually attributed to surface structural sealing (De Ploey, 1981; Govers, 1991), but we have also observed another very subtle form of sealing during flume experiments under simulated rainfall. This can occur under sheet or rillwash where raindrop impact is reduced, and can obstruct rill incision. This is neither a structural nor a depositional seal, but appears to form by plastic deformation of clay-rich soils, particularly vertisols, in shallow flows that are sufficiently deep to limit raindrop impact, but which do not produce threshold hydraulic conditions for entrainment of cohesive soils.

While a lot of attention has been paid to the progressive change in soil properties during the wetting cycle, potentially equally important changes during the drying cycle have been largely ignored. Particularly significant is “hard-setting” of aggregates weakened or disrupted during wetting into a hard, structured mass during drying (Mullins and Lei, 1995). Desiccation cracking can also determine partitioning of flow between rills and pipes (Bryan, 1996) and strongly influence rill network development. Rill incision is critical, affecting lower layers of different moisture content, strength and bulk density and producing varied time-dependent changes in sediment concentration (Govers and Loch, 1993). Fundamental controls of cracking behaviour are not yet well-understood. Clay mineralogy is important (Bryan, 1973), but some cracking occurs on all cohesive soils. It is maximal on very dry soils, but can occur with even a small drop in moisture content. There also appears to be a link with aggregate fracture patterns but fractal scaling relationships between disintegrating aggregates and cracked surface seals have not yet been fully explained (Perfect and Kay, 1991; Young and Crawford, 1991; Preston et al., 1997).

While antecedent soil water content during individual storms is most critical for erosional response, soil water regime over longer periods also influences erodibility. Soulides and Allison (1961), Tisdall et al. (1978) and Shiel et al. (1988) have all shown that repeated wetting–drying cycles cause progressive decline in aggregate stability, while Utomo and Dexter (1982) reported a more complex response with both increased and decreased stability. Stability in undisturbed soils invariably declined, but in tilled soils it increased for some days after tillage before declining. The short term influence of antecedent water in individual rainstorms is believed to be almost entirely physical, but the long term effects of repeated wetting–drying cycles may involve both physical and organic effects. Organic effects, including both microbial action and complex decomposition of organic matter, can profoundly affect aggregation characteristics, but Utomo and Dexter’s (1982) results occurred with both non-sterile and sterile soils, and were attributed primarily to microcracking due to shrinkage.

The soil water regime is strongly influenced by rainfall patterns, and also by soil water retention characteristics, which should therefore be incorporated in assessment of temporal variations in soil erodibility. The standard laboratory technique tests very small disturbed samples which have been passed through a 2-mm sieve in porous plate vacuum apparatus. These data are very relevant to the behaviour of agricultural soils, particularly seed beds, but as natural structure and macroporosity is eliminated in these tests, they do not really provide useful information for hillslope geomorphology and hydrology. There is clearly a need for a more appropriate test, which uses much larger block or in situ samples, in which macroporosity is intact, and the effect of profile layers of low permeability can also be identified.
Frost action and soil water variations strongly affect soil physical conditions, and certainly contribute strongly to temporal change in erodibility. However, interpretation of changes in purely climatic terms is difficult because disturbed soils can also undergo progressive change, referred to as age-hardening, curing or thixotropy, without frost action or water content fluctuations (Utomo and Dexter, 1981). These changes have been attributed to progressive particle re-orientation with creation of new bonds and strengthening of old bonds. Dexter et al. (1988) suggested that age-hardening reflects compression of the diffuse double layer with changes in electrolyte concentration or changes in the pH at which net charge on colloidal particles is zero as the main bonding factor. In experiments with Australian, Israeli and American soils, Dexter et al. (1988) showed that age-hardening causes up to fourfold increases in soil penetration strength over periods of 50 days. The effect of these changes on soil erodibility has not been examined, but could be very complex. Increased hardness linked with reduced pore size would probably reduce aggregate size and entrainment resistance, but could also increase resistance to breakdown, and therefore reduce the possibility of surface sealing. While shear strength and penetration resistance are not identical, increased shear strength with direct effects on both splash and runoff resistance would be expected.

4.3. Soil organic matter and microorganisms

Soil erosion response is strongly influenced by antecedent soil water content and the recent history of exposure to physical stresses such as tillage, frost action and wetting–drying cycles. These short-term physical stresses are primarily effective in disrupting macroaggregates. Response is also affected by long-term trends that reflect the effect of organic matter and microorganisms on aggregation, as shown in the Tisdall and Oades model. These long-term trends govern the relative proportions of macro and microaggregates, which depend on the amount and activity of various organic binding agents present. Gabriels and Michiels (1991) distinguished three categories of organic binding agents: transient agents such as polysaccharides, produced and decomposed rapidly; temporary agents such as thin roots and fungal hyphae, developed over weeks or months and decomposing over months to years; and persistent agents such as strongly sorbed polymers which form slowly as end-products of organic fractionation, and persist over several thousand years. Typical accumulation and decay patterns and relationship to aggregation are shown schematically in Fig. 10.

Proportions of organic binding agents vary with the amount, nature and location of organic inputs, and with time-dependent decomposition and fractionation processes. Inputs are largely controlled by vegetation conditions and dynamics. They may be strongly seasonal and localized, as with leaf-fall in temperate forests, or more continuous and dispersed throughout the soil, as in grasslands. Decomposition and fractionation depend on microbial activity, which in turn depends on microclimate, particularly as it affects soil moisture, temperature and aeration. Inevitably the amount and composition of soil organic matter vary greatly, but under undisturbed conditions they reach an equilibrium level which, at any location, may remain fairly constant or fluctuate through predictable seasonal, annual or multi-annual cycles. The net effect is to produce “characteristic” temporal aggregation patterns for natural, undisturbed soils, with specific proportions of macro and microaggregates. Temporal variation in aggregation characteristics may be minimal, or highly significant, depending on vegetation and climate dynamics and is typically largest in arid and semi-arid regions, where vegetation is patchy and biomass production and microbial activity are highly opportunistic. Lavee et al. (1996, 1998) have shown such fluctuations along a bioclimatic gradient from the Judaean Hills to the Dead Sea, in Israel, with minimum aggregate stability at all locations in March, and maximum seasonal variation on stability in intermediate moisture zones.

Natural soil aggregation behaviour has some analogies to the stability and resilience of plants and plant communities. Natural aggregation can be strongly affected by human activity. Some soils, particularly those that are highly flocculated or rich in stable humic acids, can withstand considerable stresses that disrupt or eliminate aggregation in less stable or resilient soils. The stresses linked to human
activity typically involve vegetation change, alteration of organic inputs, and physical disruption of the soil, which transforms soil climate and organic decomposition rates. While human activity is by far the most dominant cause of disruption, natural occurrences such as forest fire (Ternan and Neller, 1999) or animal trampling can also significantly change aggregation characteristics by, for example, reducing organic material or precipitating hydrophobic substances which temporarily "water-proof" aggregates. The effects of human disruption can include short-term increases in macroaggregation due to addition of manures, but probably the most common and important effect is the exposure by tillage of organic material formerly protected in almost anaerobic conditions in microaggregates to active decomposition. Human activity and particularly intensive agriculture, almost inevitably result in long-term reduction of soil organic matter, and particularly the persistent binding agents. While documentary evidence is scant, reduction in persistent binding agents would become particularly apparent in declining microaggregate content.

Recognition of the linkage of characteristic aggregation patterns to organic matter dynamics suggests a possible new approach for modelling and prediction of soil erodibility. With a few exceptions, undisturbed soils are almost entirely aggregated, with either specific proportions of macro and microaggregates, or proportions fluctuating within identifiable ranges. Macroaggregates are generally less stable and resistant to entrainment, and experience most disruption during rainstorms, so their proportion should provide a direct index of potential erodibility. This proportion will decline during a given rainstorm, and, depending on storm intensity and duration, may be eliminated entirely. Soil erodibility in ensuing storms would be reduced, unless macroaggregate proportions are restored by organic incorporation, microbial activity or physical agents. Where the timing of organic inputs, the intensity of physical disruption can be predicted, and the controls on microbial activity are understood, it should be possible to predict the period necessary for restoration could be forecast, and the progressive effect on soil erodibility.

Fig. 10. Conceptual diagram showing relationships between input of organic matter to soil decomposition and fractionation of organic matter and soil aggregation.
5. Conclusions

Resistance to erosive forces is primarily determined by soil properties, which are therefore critical in determining spatial and temporal patterns of sediment transport on hillslopes, thereby affecting not only hillslope evolution, but sediment delivery patterns in drainage basins at all scales. The importance of soil properties has been recognized since the earliest days of erosion research, but it was originally thought that soil erodibility was an essentially constant characteristic that could be accurately identified by one or two easily measurable soil properties. While a slightly more complex understanding of variations in soil erodibility emerged as a result of early soil erosion research, it was essentially this concept of a constant soil erodibility or "K" factor, which was incorporated in the influential Universal Soil Loss Equation. It is clear, in the light of much related research over the last several decades that this concept is highly simplistic and obsolete, and that soil erodibility is not a single, simply identified property, but is more appropriately regarded as the summation of a highly complex response pattern, strongly influenced by intrinsic soil characteristics and extrinsic, macroenvironmental variables.

The USLE was designed as a practical tool to assist in agricultural management and has been quite successful for guiding land management practices, particularly when used in the region where the original data were acquired. It is a purely empirical tool derived from observations on disturbed agricultural soils on very gentle slopes, which was never intended for application to the more complex soils and slopes which are typically of interest in geomorphology and hydrology. It was also designed to identify long-term patterns and trends and was not intended to provide the spatially and temporally discrete information required for event-based predictive modelling. In the light of these limited objectives, the use of a very simplistic \( K \) factor was not a critical flaw.

Attempts to expand the use of the USLE and to extrapolate from the original data to very different environmental circumstances lead to clear recognition of its limitations, and to many recent attempts to develop more effective physically based alternatives, of which the WEPP model in the United States and the EUROSEM model in Europe are recent prominent examples. These have generated many important research projects and much insight into the nature of erosional processes, though the body of precise experimental data is still dominated by studies of disturbed agricultural soils, and particularly sandy or loamy soils. There are still major gaps in our understanding of the behaviour of natural soils, including particularly stoney soils on steep slopes, or soils under natural undisturbed vegetation, where structural characteristics differ greatly from those of agricultural soils. EUROSEM has started to incorporate some of the extensive research carried out on stoney soils carried out in Europe during the past decade (e.g., Poesen et al., 1994) but the WEPP model is still strongly oriented towards agricultural soils. It is still too early to judge the performance of these physically based models, most of which are in an early stage of development. However, it is clear that a major limitation will be the availability of the necessary abundant, sophisticated data. While more extensive testing is clearly required, the initial results for the WEPP model, in existence for a decade, have been somewhat disappointing, in terms of more effective prediction than the USLE.

The preoccupation with development of physically based models has been important in generating research, particularly during the last decade, but it is not clear that such models will ultimately be really effective in predicting soil erosion response, particularly in geomorphologically context. While computing capacity is a rapidly declining constraint, the abundance and sophistication of the field data necessary to run even relatively simple models is daunting, particularly when spatial variability and temporal dynamics are considered. These would need to be greatly expanded and increased in precision to encompass many of the complex relationships discussed in this paper. In any case, it is not yet clear that all the processes and interactions involved in soil erodibility can be physically modelled. The complexity of, for example the microbial, physical, chemical and microclimatic factors which determine soil aggregation patterns, may well be more suited to stochastic modelling. Dunne (1991) and Benda and Dunne (1997) applied stochastic modelling to hillslope evolution in Kenya dominated by splash and rainflow, and to drainage basins in the Pacific northwest United States, dominated by periodic landslid-
ing. These applications were based primarily on the stochastic properties of rainfall as the dominant factor. This paper shows that critical soil properties such as aggregation, which affect erosional response by influencing the distribution of erosive forces, or soil resistance, also have characteristic probability distributions. Once probability distributions of critical factors are adequately defined, it should be possible to combine these with stochastic rainfall characteristics for better site-specific, temporally dynamic prediction of erodibility. It is clear that definition and prediction of erodibility cannot be separated from understanding of the controls on sediment transport and deposition across hillslopes. Accordingly, temporally dynamic models of soil resistance will have to be coupled with appropriate sediment transport routing models, in which the transition from interrill to rill or piping processes will be extremely critical. This transition and the subsequent evolution of rill and pipe networks may also involve stochastic elements as suggested by Moore and Foster (1990). However, the probability characteristics of rill and pipe network development have received very little detailed study, and must remain a high priority for future research.

6. List of symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u'$</td>
<td>shear velocity</td>
</tr>
<tr>
<td>$w$</td>
<td>stream power</td>
</tr>
<tr>
<td>$y$</td>
<td>unit stream power</td>
</tr>
<tr>
<td>$\phi$</td>
<td>angle of internal friction</td>
</tr>
<tr>
<td>$z$</td>
<td>normal stress</td>
</tr>
<tr>
<td>$c$</td>
<td>cohesion</td>
</tr>
<tr>
<td>$ss$</td>
<td>shear strength</td>
</tr>
<tr>
<td>$Fr$</td>
<td>Froude number</td>
</tr>
<tr>
<td>$Re$</td>
<td>Reynolds number</td>
</tr>
<tr>
<td>$\nu$</td>
<td>fluid viscosity</td>
</tr>
<tr>
<td>$p$</td>
<td>fluid density</td>
</tr>
<tr>
<td>$O_s$</td>
<td>sediment concentration</td>
</tr>
<tr>
<td>$q$</td>
<td>water discharge</td>
</tr>
<tr>
<td>$\tau$</td>
<td>shear stress</td>
</tr>
<tr>
<td>$d$</td>
<td>median drop diameter</td>
</tr>
<tr>
<td>$u$</td>
<td>flow velocity</td>
</tr>
<tr>
<td>$g$</td>
<td>gravitational acceleration</td>
</tr>
<tr>
<td>$R$</td>
<td>hydraulic radius</td>
</tr>
<tr>
<td>$S$</td>
<td>slope</td>
</tr>
</tbody>
</table>

Acknowledgements

Much of the research discussed in this paper has been supported by grants from the Natural Sciences and Engineering Research Council, Canada, which are gratefully acknowledged. Numerous friends and colleagues have assisted me in experiments and with ideas, but bear no responsibility for any errors. I am particularly indebted to Ross Sutherland, Dirk Oostwoud, Mike Slattery, Richard Hawke, Alex Brunton, Erik Brun and Niklaus Kuhn. I would also like to pay tribute particularly to my former student and colleague, Shiu Luk, who made numerous important contributions to soil erosion research, and whose career was tragically abbreviated by illness.

References

Bryan, R.B., Brun, S.E., 1999. Sequential scour and development
Bryan, J.L., 1988. Sediment transport and bed morphology at river
Bisal, F., Nielsen, K.F., 1967. Effect of frost action on the size of
Bouyoucos, G.J., 1935. The clay ratio as a criterion of the
surface sealing on splash from a single waterdrop. Soil Sci.
Brunori, F., Penzo, M.C., Torri, D., 1989. Soil shear strength: its
Brunton, D.A., Bryan, R.B., 2000. Rill network development and
sediment budgets. Earth Surface Processes Landforms, (submitted manuscript).
Bryan, R.B., 1968. The development, use and efficiency of indices
Bryan, R.B., 1969. The relative erodibility of soils developed in
Bryan, R.B., 1971a. The efficiency of aggregation indices in the
22, 166–178.
Bryan, R.B., 1971b. The influence of frost action on soil aggrega-
Bryan, R.B., 1974. Water erosion by splash and wash and the
Bryan, R.B., 1981. Soil erosion under simulated rainfall in the
field and laboratory: variability of erosion under controlled
(Ed.), Field Experiments and Measurement Programs in Geo-
169.
weathering conditions. In: Anderson, M.G., Brooks, S.A.
(Eds.), Advances in Hillslope Processes. Wiley, Chichester,
pp. 589–612.
Bryan, R.B., Brun, S.E., 1999. Sequential scour and development
of banded vegetation. Catena 37, 147–163.
measurements with different laboratory rainfall simulators.
significance of tunnel erosion in a semiarid ephemeral drainage
dynamics in Canadian badland microcatchments. In: Walling,
D.E., Burt, T.P. (Eds.), Catchment Experiments in Geomor-
processes: inventory and prospect. Geomorphology 20, 209–
219.
determination of soil shear strength in erosion studies. Soil
Bryan, R.B., Oostwoud Wijdenes, D., 1992. Field and laboratory
experiments on the evolution of microsteps and scoured chan-
Bryan, R.B., Poesen, J.W.A., 1989. Laboratory experiments on
the influence of slope length on runoff, percolation and rill
initiation and implications for erosional response. Geomor-
phology 23, 151–169.
the initiation of runoff and piping in Dinosaur Provincial Park
Bd.
and sediment entrainment on microcatchments in the Dinosaur
of subsurface moisture on rill system initiation. Earth Surface
microstandpipe System for soil erosion research. Soil Sci.
Callebaut, F., Gabriels, D., De Boodt, M. (Eds.), Assessment of
Soil Sealing and Crusting. State University of Ghent, Belgium.
Chisci, G., Sfalanga, M., Torri, D., 1985. An experimental model
for evaluating soil erosion on a single-rainstorm basis. In: El
Swaify, S.A., Moldenhauer, W.C., Lo, A. (Eds.), Soil Erosion
and Conservation, Soil Conservation of America. Ankeny,
Iowa, pp. 558–565.
Am. J. Sci.
Coote, R., Malcolm-Govern, A., Wall, G.J., Dickinson, W.T.,
Rudra, R.P., 1988. Seasonal variation in soil erodibility in-
dices based on shear strength and aggregate stability in some
Cresswell, H.P., Smiles, D.E., Williams, J., 1992. Soil structure,
soil hydraulic properties and the soil water balance. Aust. J.
Soil Res. 30, 265–283.
De Ploey, J., 1971. Liquefaction and rainwash erosion. Z. Geo-
morphol. 15, 491–496, Supp. Bd.


Horton, R.E., 1945. Erosional development of streams and their...


Sumner, M.E., Stewart, B.A. (Eds.), Soil Crusting: Chemical and Physical Processes. A.S. Lewis, Boca Raton.


