Analysis of the nonlinearity in the hillslope runoff response to precipitation through numerical modeling

Jozsef Szilagyi

Department of Hydraulic and Hydrologic Engineering, Budapest University of Technology and Economics, 1111 Muegyetem Rkp. 3-9, Hungary
Conservation and Survey Division, Institute of Agriculture and Natural Resources, University of Nebraska-Lincoln, Lincoln, NE 68588-0517, United States

Received 31 January 2006; received in revised form 27 November 2006; accepted 2 February 2007

KEYWORDS
Nonlinear runoff response; Richards equation; Kinematic wave equation; Surface flow; Subsurface flow; Infiltration excess; Saturation excess

Summary  Coupled, finite-element models are applied for hillslope runoff investigations. Subsurface flow is modeled by the 2-D Richards equation extended for the saturated zone. Surface runoff is described by the linear and also by the nonlinear kinematic wave equations coupled to the subsurface model through infiltration and/or saturation excess. It is concluded that the typical nonlinearity in the runoff response is the result of not only the surface runoff process and the antecedent soil moisture condition but also of the way precipitation intensity affects the subsurface process. Namely, higher intensity rains produce surface runoff sooner through infiltration excess and bring the soil/aquifer closer to saturation in a shorter time leading to taller unit hydrographs of total runoff and to shorter time-to-peak periods, even if surface runoff is kept linear in nature.

© 2007 Elsevier B.V. All rights reserved.

Introduction

Since the publication of Minshall’s (1960) classical work on storm runoff from small experimental watersheds, hydrologists agree that the rainfall–runoff response is fundamentally nonlinear. Nonlinearity is meant in a dynamic sense (Sivapalan et al., 2002), that is, proportionality and additivity between effective precipitation and quick-storm or total storm response (as the sum of quick and delayed storm responses) is typically violated. This statement is especially true for small watersheds, where it is agreed that runoff is dominated by hillslope processes (Wooding, 1965; Kirkby, 1976; Wang et al., 1981; Beven and Wood, 1993; Robinson et al., 1995; Sivapalan et al., 2002; D’Odorico and Rigon, 2003) as opposed to the channel network response. The phrase “hillslope processes” is used in a general sense here, referring to water movement that takes place outside the stream network, i.e., on the ground or under the
surface, since the experimental watersheds in Illinois (USA) where Minshall (1960) documented this nonlinear behavior of runoff generally have mild slopes (1–12%).

While Minshall’s (1960) work is often cited in the hydrological literature when referring to nonlinearity of the watershed response, no one has attempted, to the best of the author’s knowledge, to quantitatively test at least the most plausible causes of this observed nonlinearity. Beven (2001) in his book on watershed hydrology provides some qualitative summarizing explanations (p. 46–47). In his view, the nonlinearity of the runoff response to precipitation is caused mainly by two factors: chief being the antecedent soil moisture condition, and secondarily, the change in flow velocity with discharge for both, surface and subsurface (most probably interflow), runoff.

This present study investigates whether the nonlinearity of the runoff response, as was reported by Minshall (1960), can be reproduced by relatively simple physically based modeling or not. If it can then the next goal is to identify the cause or causes of this nonlinearity similar to Beven (2001). It is hoped that the current approach will help in understanding and conceptualizing the most basic physical processes that can lead to the often quoted nonlinear behavior of the runoff response in small catchments. It is in the vein of Weiler and McDonnell (2004) that ‘hillslope models should be simple’ so that they can ‘serve ultimately as useful hypothesis testing tools’, which is being demonstrated in this study.

Model description

Subsurface dynamics are described by the application of the extended Richards equation (Lam et al., 1987) in a 2-D vertical plane

$$\frac{\partial}{\partial x} [K(\psi) \frac{\partial h}{\partial x}] + \frac{\partial}{\partial y} [K(\psi) \frac{\partial h}{\partial y}] = m c \frac{\partial h}{\partial t}$$  \hspace{1cm} (1)

where $K$ is the hydraulic conductivity, a function of the suction/pressure head, $\psi$; $h$ is the total hydraulic head; $m$ is the slope of the water retention curve which becomes the coefficient of volume change in the saturated zone; $c$ is the unit weight of water; and $x$, $y$, and $t$ denote the horizontal, vertical and temporal coordinates, respectively. The computational domain that represents the soil–aquifer system employed in this study is depicted in Fig. 1.

The soil–aquifer system is assumed to be drained by a fully penetrating stream (not marked in Fig. 1) from the right. Consequently, the right side of the computational domain serves as an effective value (Dirichlet) boundary condition (i.e. $h = \gamma$) provided the suction/pressure head, $\psi$, is positive, otherwise it is a no-flow flux boundary (Neumann). The bottom (location of the impervious layer) and left side (the groundwater divide) of the rectangle act as no-flow boundaries, while at the top, representing the ground surface, water can enter the domain at the precipitation intensity rate (Neumann) as long as $\psi$ is negative there, and it becomes an effective value boundary (i.e. $h = \gamma$) otherwise. Note that infiltration excess overland flow can form, because the topmost layer of the soil/aquifer can become saturated very fast with the prescribed boundary conditions, even though layers deeper below remain unsaturated for a while. Since switching boundary conditions is not explicitly allowed in the finite element software environment, FLEXPDE (www.pdesolution.com), the flux boundary condition is the default setting in this study. When necessary (i.e., when $-\psi \leq \varepsilon$, an arbitrary small positive number, at the boundary), an effective value boundary condition is invoked by turning the flux rate proportional to the magnitude of $\psi$ within the tolerance level, $\varepsilon$ (Gitirana et al.,

![Figure 1](image-url)
This ensures that the top layers stay saturated as long as the layers deeper become capable of distributing this extra moisture with an efficiency surpassing the infiltration rate at the top.

Initially \((t = 0)\) the soil/aquifer, containing a horizontal groundwater table at \(y = 0.1 \text{ m}\), is in a static equilibrium with the stream (Fig. 1). Drainage starts for \(t > 0\) by dropping the water level in the stream from 0.1 m to 0 m instantaneously.

Recharge of the soil/aquifer is achieved in two different ways. Either, in the form of a single rain event, or as a Monte-Carlo generated time series of precipitation. In both cases precipitation intensities \((P)\) during a rain event follow a squared sine-curve for a half period as

\[
P = cK_s \sin^2(5.102t) \tag{2}
\]

where \(K_s\) is the saturated hydraulic conductivity of the isotropic and homogeneous aquifer material, \(c\) is a constant multiplier, and \(t\) is measured in hours. In the time series case arrival times between precipitation events follow a Poisson process, with a parameter value of 3 h, while the multiplier, \(c\), is taken from a uniform distribution on the interval \((0–0.1)\). Prescribing a smoothly changing precipitation intensity curve rather than a step function is necessary for numerical stability and faster numerical integration of (1).

The soil–aquifer system is supposed to be made up of well-sorted sand. See Table 1 for detailed hydraulic properties of the chosen physical soil texture meant to assure high infiltration rates characteristic of the small experimental catchments in Illinois that were investigated by Minshall (1960). The prescribed 1% slope of the computational domain, as well as its thickness (50 cm), are also typical of those watersheds (especially of catchment W-1). Because of the sharp hydraulic head gradients that develop during infiltration, numerical integration of (1) is rather time-consuming on a standard personal computer. Therefore, the width of the aquifer (i.e. its horizontal extent) was limited to 10 m instead of the 90 m required by, e.g., catchment W-1, which has an area of 0.11 km² and a total estimated stream length of about 600 m. Note that in this study the aim was not to parallel the storm response of the experimental watersheds with the model values, for which matching the mean aquifer width between stream channels with the prescribed model value would most probably be necessary, but rather, to possibly replicate the general nonlinearity in the hillslope runoff response to precipitation through numerical modeling.

### Table 1

<table>
<thead>
<tr>
<th>Soil property</th>
<th>Value/equation</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\phi) ([-])</td>
<td>0.395</td>
<td>(\phi) is the total porosity</td>
</tr>
<tr>
<td>(K_s) ([\text{m/d}])</td>
<td>1.25</td>
<td>(K_s) is the saturated hydraulic conductivity</td>
</tr>
<tr>
<td>(\Psi_{ae}) ([\text{kPa}])</td>
<td>1.18</td>
<td>(\Psi_{ae}) is the air entry pressure</td>
</tr>
<tr>
<td>(\Psi(\phi))</td>
<td>(\Psi_{ae}(\phi)^b)</td>
<td>(\phi) is the volumetric water content ([-])</td>
</tr>
<tr>
<td>(K(\phi))</td>
<td>(K_s(\phi)^{2b-1})</td>
<td>(b(=4.05)) is the pore size distribution index</td>
</tr>
</tbody>
</table>


Initially \((t = 0)\) the soil/aquifer, containing a horizontal groundwater table at \(y = 0.1 \text{ m}\), is in a static equilibrium with the stream (Fig. 1). Drainage starts for \(t > 0\) by dropping the water level in the stream from 0.1 m to 0 m instantaneously.

Recharge of the soil/aquifer is achieved in two different ways. Either, in the form of a single rain event, or as a Monte-Carlo generated time series of precipitation. In both cases precipitation intensities \((P)\) during a rain event follow a squared sine-curve for a half period as

\[
P = cK_s \sin^2(5.102t) \tag{2}
\]

where \(K_s\) is the saturated hydraulic conductivity of the isotropic and homogeneous aquifer material, \(c\) is a constant multiplier, and \(t\) is measured in hours. In the time series case arrival times between precipitation events follow a Poisson process, with a parameter value of 3 h, while the multiplier, \(c\), is taken from a uniform distribution on the interval \((0–0.1)\). Prescribing a smoothly changing precipitation intensity curve rather than a step function is necessary for numerical stability and faster numerical integration of (1).

The soil–aquifer system is supposed to be made up of well-sorted sand. See Table 1 for detailed hydraulic properties of the chosen physical soil texture meant to assure high infiltration rates characteristic of the small experimental catchments in Illinois that were investigated by Minshall (1960). The prescribed 1% slope of the computational domain, as well as its thickness (50 cm), are also typical of those watersheds (especially of catchment W-1). Because of the sharp hydraulic head gradients that develop during infiltration, numerical integration of (1) is rather time-consuming on a standard personal computer. Therefore, the width of the aquifer (i.e. its horizontal extent) was limited to 10 m instead of the 90 m required by, e.g., catchment W-1, which has an area of 0.11 km² and a total estimated stream length of about 600 m. Note that in this study the aim was not to parallel the storm response of the experimental watersheds with the model values, for which matching the mean aquifer width between stream channels with the prescribed model value would most probably be necessary, but rather, to possibly replicate the general nonlinearity in the hillslope runoff response to precipitation through numerical modeling.

The coupled models were first run in a single precipitation event mode. Fig. 2 displays the five precipitation events employed, each described by Eq. (2) with different values of the constant multiplier. Modeled subsurface and total runoff responses for the smallest and largest precipitation events are illustrated in Fig. 3. As expected, the two runoff responses (total and subsurface) are proportional to the triggering precipitation events in their volumes and corresponding peak values but not linearly. For the light rain a significant portion of the total runoff response is generated from the subsurface. It is somewhat contrary to the assumption of Minshall (1960) who concludes that subsurface runoff can only play an insignificant part of the runoff generation in the Illinois experimental watersheds, because the flow returns to about 1% of the peak value in about 3 h after the peak. In the present numerical study the same occurs, however, with a not so negligible subsurface contribution. The relatively high saturated hydraulic conductivity value of the model aquifer makes it possible. This reinforces the potential significance of the preferential flow process in runoff generation for soils/aquifers that otherwise exhibit low laboratory-derived permeability values, and underlines the importance of aquifer parameter estimation methods (Brutsaert and Nieber, 1977; Brutsaert and Lopez, 1998; Szilagyi et al., 1998; Parlange et al., 2001; Szilagyi, 2003; Brooks et al., 2004; Rupp et al., 2004) that can account for such preferential flow effects.
By further investigation of Fig. 3 it may seem a bit surprising that the larger peaks occur sooner for both types of runoff, i.e., it takes less time to reach the peak value when it is larger. Figs. 4 and 5 corroborate this finding, separately for both, surface and subsurface, runoff. The range in the time-to-peak value between the largest and smallest hydrographs is about 400 s for subsurface and 375 s for surface runoff. Note that while subsurface runoff
otherwise seems to be linearly proportional (in peak and total volume) to the triggering precipitation event, the surface runoff response is in a clear violation of proportionality and, therefore, linearity. Note also that linearity of the subsurface runoff response can be maintained only because the aquifer does not become completely (i.e., \( \Psi = 0 \) everywhere never happens) saturated during the selected precipitation events. Naturally, once complete saturation of the

Figure 4  Subsurface runoff responses (over a unit stream length) to the five precipitation events displayed in Fig. 2.

Figure 5  Surface runoff responses (over a unit stream length) to the five precipitation events displayed in Fig. 2.
aquifer is reached, additional increase in the precipitation intensity cannot cause any further boost to subsurface runoff, therefore, at that point linearity in aquifer response would also be violated.

Surface runoff in the model is generated in two ways: (a) as infiltration excess, and; (b) as saturation excess overland flow. The first occurs when the soil–aquifer system becomes saturated from above (i.e. when water infiltrates in faster than can be redistributed), while the second happens when the groundwater table rises to the surface, as illustrated in Fig. 6. The location of ‘groundwater ridging’ observable near the stream in Fig. 6 is not by chance. The groundwater ridge develops because the horizontal groundwater table initially is closer to the surface near the stream than farther away of it due to a sloping surface. Thus, the vadose zone retains more moisture near the stream initially and can become saturated faster during an actual rain event. The model this way can account for the variable source concept of runoff generation formulated by Hewlett and Hibbert (1967) and Dunne and Leopold (1978). Since these source areas where saturation excess overland flow occurs change their location and extent during a typical storm event, it is questionable if one can tell the beginning of rainfall excess as was attempted by Minshall (1960). In lieu of that, time to peak with model data are calculated as the time elapsed between the initial rise and peak in the total runoff hydrograph. (When the total runoff hydrograph was replaced by the surface runoff hydrograph in the time-to-peak calculations the difference was negligible). The most striking and the most frequently quoted part of Minshall’s (1960) work is his computed unit hydrographs (UH) for watershed W-1 in Illinois. The hydrographs have two characteristics: the peak UH value decreases with decreasing ‘excess precipitation’ intensity while at the same time the UH not only becomes more spread out (as follows from the previous property since mass must be conserved) but also the time-to-peak interval becomes longer. If watershed response was linear the unit hydrographs ought to be virtually overlapping independently of the triggering rainfall intensity. The question is what causes this clear deviation from linearity. Beven (2001) mentions the effect of differing antecedent soil moisture condition for an explanation. With the present numerical model, this can be easily tested, since model runoff can be started with exactly the same initial (i.e., antecedent) soil moisture condition, as was indeed performed in the single rain event mode.

Unit hydrographs were created by dividing the model runoff values with the corresponding total precipitation amounts of the five different rain events considered. Fig. 7 displays the so-derived UHs of total runoff. The UHs express the nonlinear features mentioned above: they become more flattened with their peaks delayed as the mean rain intensity of the corresponding precipitation events decreases. The range in the time-to-peak value between largest and smallest UHs in Fig. 7 is 235 s, which is about 4 min. Any possible difference in initial soil moisture conditions as an explanation for the observed nonlinearity can be discarded in this experiment. Can the nonlinear nature of surface runoff, as was suggested by Beven (2001), then explain this nonlinearity? As Fig. 8 suggests, not entirely. When switching from the nonlinear kinematic wave equation to its linear version, the range in the time-to-peak values (Fig. 8) reduces to about 120 s, which is about half (2 min)
of what was observed previously. Also, the spreading of the UHs become somewhat reduced, but it still remains significant.

The effect of nonlinearity versus linearity of the surface runoff can be best analyzed in Figs. 9 and 10. The celerity, changing with discharge, accelerates the peak value (therefore, broadening the range of the time-to-peak interval between the largest and smallest precipitation events) while delays the small ones (creating long UH tails) when compared to the linear case. The range in the time-to-peak

**Figure 7**  Unit hydrographs of modeled total runoff using the nonlinear kinematic wave equation.

**Figure 8**  Unit hydrographs of modeled total runoff using the linear kinematic wave equation.
value this way is increased from about 150 s in the linear case to the already mentioned 375 s for the nonlinear one, a 2.5-times growth. Here the UHs were obtained by normalizing the surface runoff values with the total surface runoff volume rather than the total precipitation volume, since precipitation now is not the direct source (i.e., the model input) of surface runoff, the difference (q) in precipitation and infiltration intensities is. Note that even the linear kinematic wave would not produce a linear surface runoff response, since the UHs do not overlap completely, rather
they are shifted in time. The reason being that surface run-
off is produced sooner due to infiltration excess plus the
aquifer becomes closer to saturation faster (Fig. 4) during
high intensity rain events, since then more water can ini-
tially infiltrate into the soil—aquifer system. This dynamic
behavior of the aquifer and its effect on runoff generation
is often overlooked. It is an important feature, however,
since without accounting for its effect (exerted through
infiltration and saturation excesses) in runoff generation
the apparent nonlinearity of the modeled storm response,
and so that of hillslope or watershed response in general,
could not be understood and interpreted correctly.

In the second set of numerical experiments, applying
Monte-Carlo generated precipitation time series, the addi-
tional effect of varying antecedent soil moisture status to
the ensuing runoff was investigated. As has been mentioned,
all individual rain events (separated by waiting times from a
Poisson process) within the time series keep the functional
form of Eq. (2). Similar to Minshall (1960), the smallest storm
events (i.e., when the peak total runoff rate was smaller
than $10^{-5}$ m$^3$ s$^{-1}$) were left out from the analysis.

Figs. 11 and 12 display the model data together with Min-
shall’s (1960) data for catchment W-1. The two data sets ex-
hibit very similar properties: the time-to-peak period
decreases with increasing mean precipitation intensities
(Fig. 11) while peak values of the unit hydrographs for runoff
increase (Fig. 12). This is in full agreement with Fig. 7. Differ-
ences in antecedent soil moisture status led to a wider range
in the time-to-peak interval (from 4 min to 13 min) compared
with the single-rain-event model results, even though the range in the mean precipitation intensities remained the
same 21 mm h$^{-1}$ (i.e. 13–34 vs. 48–69 mm h$^{-1}$) as previously
(and most likely led also to a larger scatter in Fig. 11 since a
less intense rain event now with an initially moist soil—aqui-
fer system could produce the same runoff as before with a
drier moisture condition). Note that overall smaller precipi-
tation intensities were used in the present time series case
because the finite element program slows down significantly
when complete saturation of the aquifer is achieved. This
must be due to the requirement of virtually switching bound-
ary conditions simultaneously at the top and at the right side
(representing the stream bank) of the computational domain
as the aquifer is drained by the stream while simultaneously
being recharged to full saturation. Therefore, it was at-
ttempted to limit the number of occurrence of such a condi-
tion during modeling in order to avoid literally tens of
hours of computer iterations just for a single rain event. Note
that this does not mean that the soil/aquifer could not be-
come fully saturated over even 99% of its area.

As a summary it can be stated that the nonlinearity of the
storm response, typical of small watersheds, could be suc-
cessfully recreated by coupling the kinematic wave equa-
tion to the extended Richards equation (Lam et al., 1987)
in a finite element model. It was demonstrated that differ-
ces in the antecedent soil moisture condition and the
nonlinear nature of the employed kinematic wave equation,
the two most commonly cited causes (Beven, 2001), cannot
completely explain the nonlinearity of the storm response
to precipitation detectable in the numerical results. By
eliminating both of these causes, modeled total runoff still
expressed a considerable degree of nonlinearity (Fig. 8).
This nonlinearity can be attributed to the way the soil—aqui-
fer system reacts to effective precipitation, i.e., the por-
tion of precipitation that actually reaches the soil surface.

![Figure 11](image-url)  
**Figure 11** Precipitation intensity versus time to peak runoff rate. $R^2$ is the explained variance of the linear regression line. The circles are from Minshall (1960) for catchment W-1, while the crosses designate the model data using Monte-Carlo simulated time series of precipitation events.
Note that in the present numerical experiment the shielding effect of vegetation (interception) was not considered, thus, precipitation acted as effective precipitation. Higher intensity rains produce surface runoff sooner through infiltration excess and bring the soil/aquifer closer to saturation in a shorter time (Fig. 4), therefore, leading to taller unit hydrographs of total runoff (Fig. 8) and to shorter time-to-peak periods (Figs. 8 and 10), even if surface runoff is kept linear in nature.

As exemplified in Fig. 3, subsurface flow can produce a significant portion (i.e., baseflow) of the total runoff and cannot be assumed to be negligible automatically by simply observing a short base of storm response (Minshall, 1960) even in a shallow aquifer case. Highly conductive soil types/aquifer materials (as in the present numerical study) and/or the presence of preferential flow can produce fast declining hydrographs. In cases where an increase in stream discharge entails only a negligibly small change in stream water depth (as in the present study where stream water depth is neglected for \( t > 0 \)), the peak in subsurface runoff can (contrary to general expectations) precede the peak in surface runoff and consequently, the peak in total runoff (Fig. 3). The present numerical results seem to corroborate the findings of earlier reports that were based on physical and isotope experiments by, e.g., Bonell et al. (1990), Turton et al. (1992), and Hinton et al. (1994).

When prioritizing the possible causes of nonlinearity in the rainfall–runoff relationship the following can be stated, based on the model results:

(a) The simulated nonlinearity of the hillslope runoff response to precipitation can be best explained by the nonlinear nature of the surface runoff. On an about 10 m long slope of 1%, switching from linear to nonlinear version of the kinematic wave equation doubled the difference of the time-to-peak value (and flattened the unit hydrographs considerably due to numerical dispersion and the diffusion term of [3]) between the slowest and fastest waves (Figs. 9 and 10) from about 2 min to 4 min. In the linear kinematic wave equation cases the difference in the time-to-peak values exists because the soil/aquifer system starts to produce surface runoff sooner during a high intensity precipitation event through infiltration excess and/or saturation excess processes. This means that celerity differences in the nonlinear wave cases accounted for about 2 min in the time-to-peak values, which would roughly translate to about 20 min had the slope have the same breadth characteristic of watershed W-1 in Minshall's (1960) study. This 20 min value compares favorably with Minshall’s time-to-peak difference of about 27 min even if the model simulation and Minshall’s time-to-peak values are defined somewhat differently, as has been mentioned. Note that while the time-to-peak differences resulting from the nonlinearity of surface runoff scale with the slope length, it is not so for the infiltration and/or saturation excess processes. In other words, the time-shift for the peaks of Fig. 10 would most probably not be affected by the length of the slope.

(b) The role of the antecedent soil moisture condition in runoff generation is similar to that of the precipitation intensity. An initially wetter/drier soil can produce more/less abundant surface runoff and in less/more time, mainly due to saturation excess, leading to the above mentioned nonlinearity than a drier/wetter one. Similarly, a higher/
lower intensity precipitation event can cause infiltration and/or saturation excess in more/less abundance and sooner/later in time than a more/less gentle rain. Since the two processes may amplify or cancel each other out, they certainly will jointly contribute to the modeled and observed (by Minshall) scatter in the mean precipitation intensity and unit-hydrograph peak vs. time-to-peak relationships of Figs. 11 and 12. It is thus suggested that whenever the antecedent soil moisture condition is mentioned as a possible source of the typical nonlinearity in the rainfall–runoff relationship, the intensity of the effective precipitation as it affects the ensuing infiltration and/or saturation excess processes ought to be mentioned as well. Naturally, in either case surface runoff occurs provided the precipitation intensity (1) exceeds the infiltration capacity at the surface for some time, or; (2) exceeds the overall drainage rate of the aquifer so that it can become saturated from below (i.e., the case of saturation excess).

Disclaimer: The views, conclusions and/or opinions expressed in this paper are solely those of the author and not the University of Nebraska, state of Nebraska or any political subdivision thereof.

References


