Including effects of watershed heterogeneity in the curve number method using variable initial abstraction

Vijay P. Santikari and Lawrence C. Murdoch
Department of Environmental Engineering and Earth Sciences, Clemson University, Clemson, SC 29634, USA

Correspondence: Vijay P. Santikari (vsantik@g.clemson.edu)

Received: 15 October 2017 – Discussion started: 8 November 2017
Revised: 21 May 2018 – Accepted: 25 June 2018 – Published: 10 September 2018

Abstract. The curve number (CN) method was developed more than half a century ago and is still used in many watershed and water-quality models to estimate direct runoff from a rainfall event. Despite its popularity, the method is plagued by a conceptual problem where CN is assumed to be constant for a given set of watershed conditions, but many field observations show that CN decreases with event rainfall ($P$). Recent studies indicate that heterogeneity within the watershed is the cause of this behavior, but the governing mechanism remains poorly understood. This study shows that heterogeneity in initial abstraction, $I_a$, can be used to explain how CN varies with $P$. By conventional definition, $I_a$ is equal to the cumulative rainfall before the onset of runoff and is assumed to be constant for a given set of watershed conditions. Our analysis shows that the total storage in $I_a$ ($I_aT$) is constant, but the effective $I_a$ varies with $P$, and is equal to the filled portion of $I_aT$, which we call $I_aF$. CN calculated using $I_aF$ varies with $P$ similar to published field observations. This motivated modifications to the CN method, called variable $I_a$ models (VIMs), which replace $I_a$ with $I_aF$. VIMs were evaluated against conventional models CM0.2 ($\lambda = 0.2$) and CM$\lambda$ (calibrated $\lambda$) in their ability to predict runoff data generated using a distributed parameter CN model. The performance of CM0.2 was the poorest, whereas those of the VIMs were the best in predicting overall runoff and watershed heterogeneity. VIMs also predicted the runoff from smaller events better than the CMs and eliminated the false prediction of zero-runoffs, which is a common shortcoming of the CMs. We conclude that including variable $I_a$ accounts for heterogeneity and improves the performance of the CN method while retaining its simplicity.

1 Introduction

The estimation of runoff from a rainfall event is of primary importance in applied hydrology. It is necessary in the engineering design of small structures, post-event appraisals, environmental impact work, and other applications (Hawkins, 1993). One of the most popular techniques used for this purpose is the curve number (CN) method, which has been in use for more than half a century (D’Asaro and Grillone, 2012; Hawkins et al., 2008; Kent, 1968; Ponce and Hawkins, 1996; Rallison and Miller, 1982; Soil Conservation Service, 1956, 1972). The method uses a parameter called curve number, which is assumed to depend mainly on land cover, soil types, and antecedent conditions within a watershed.

Curve number varies spatially due to watershed heterogeneity and temporally due to changes in soil moisture, land cover, temperature, and other processes (Hawkins et al., 2008; Ponce and Hawkins, 1996; Rallison and Miller, 1982). CN also varies with the magnitude (D’Asaro and Grillone, 2012; Hawkins, 1993; Hjelmfelt Jr. et al., 2001) and spatiotemporal distribution of rainfall (Hawkins et al., 2008; Van Mullem, 1997). When heterogeneity is known at sufficient detail, CN variation can be accounted for by using a distributed parameter model, e.g., SWAT (Gassman et al., 2007). Otherwise this approach can introduce more parameters than can be reliably estimated from the available data (Soulis and Valiantzas, 2013) and can potentially cause large uncertainties in the predicted runoff. There are several ways to account for temporal variation of CN, each with its own advantages and shortcomings (Santikari, 2017). CN variation with the distribution of rainfall is usually ignored (Hawkins et al., 2008). The CN method is most commonly applied as an event-scale lumped-parameter model, which is simple but also limited in its ability to account for the variations of CN.

Published by Copernicus Publications on behalf of the European Geosciences Union.
This diminishes the accuracy of its runoff predictions (e.g., Soulis and Valianzas, 2012).

The objective of this work is to improve the event-scale lumped-parameter application of the CN method by describing an approach for incorporating the spatiotemporal variations of CN. The investigation is described in two papers, which build on Santikari (2017). In this paper, effects of spatial variation of CN (heterogeneity) at the watershed scale are analyzed. Insights gained from this analysis are used to create modified models that account for heterogeneity. The modified models are evaluated using the runoff generated by a distributed parameter model applied to a hypothetical heterogeneous watershed. In a companion paper (Santikari and Murdoch, 2018) and in Santikari (2017), the modified models are refined by including an approach that accounts for the temporal variation of CN using antecedent moisture. The refined models, which account for spatial and temporal variability, are then evaluated using data from real watersheds.

1.1 Background

The CN method assumes that a rainfall event produces runoff ($Q$) when the event rainfall ($P$) exceeds the initial abstraction ($I_a$). $I_a$ includes interception storage (by tree canopy, roof tops, and such), early infiltration, and surface depression storage. The effective rainfall, $P - I_a$, is partitioned between $Q$ and further infiltration ($F$). This is given by mass balance as

$$P - I_a = F + Q \forall P \geq I_a. \quad (1)$$

Both $F$ and $Q$ are zero when $P \leq I_a$, and both increase with $P$ when $P > I_a$. It is assumed that $F$ has an upper limit, which is referred to as the potential maximum retention ($S$). In other words, $S$ is the total storage available for infiltration after the runoff begins.

The conceptual basis that defines the curve number method comes from the following assumption (Hawkins et al., 2008; NRCS, 2003; Ponce and Hawkins, 1996; Rallison and Miller, 1982; Woodward et al., 2002):

$$\frac{Q}{P - I_a} = \frac{F}{S}. \quad (2)$$

i.e., the runoff coefficient (left-hand side) is equal to the fraction of storage filled in $S$ (right-hand side). Equation (2) is developed using the reasoning that the equality holds at the end points ($P \leq I_a$ and $P \to \infty$) (Hawkins et al., 2008; Rallison and Miller, 1982; Woodward et al., 2002) and that the behavior of both ratios in the intermediate range is essentially the same (Fig. 1). When $P \leq I_a$, both $Q$ and $F$ are zero and therefore the ratios on either side of Eq. (2) are zero. When $P > I_a$, both ratios increase with $P$, whereas their rate of increase diminishes. At the limit of $P \to \infty$, both ratios approach unity.

To eliminate the need for an independent estimation of $I_a$ (Ponce and Hawkins, 1996; Rallison and Miller, 1982), it is assumed that

$$I_a = \lambda S, \quad (3)$$

where $\lambda$ is a dimensionless parameter called the initial abstraction ratio. Early field data suggested an optimum value of $\lambda = 0.2$ (Soil Conservation Service, 1956). However, more recent studies (Hawkins et al., 2008; Woodward et al., 2003) suggest that $\lambda = 0.05$ is more appropriate. Using Eqs. (1)–(3), $I_a$ and $F$ can be eliminated to give

$$Q = 0 \forall P \leq \lambda S, \quad (4)$$

$$Q = \frac{[P - \lambda S]^2}{P + (1 - \lambda)S} \forall P > \lambda S.$$

Since the value of $\lambda$ is usually fixed (at 0.2 or 0.05), Eq. (4) requires only one parameter, $S$, which varies within the range $0 \leq S < \infty$.

For convenience (Hawkins et al., 2008; Ponce and Hawkins, 1996), $S$ (units in mm) is mapped onto a dimensionless parameter called the curve number as

$$CN = \frac{25 400}{254 + S}, \quad (5)$$

so that CN is 100 when $S$ is zero, but approaches zero as $S$ approaches infinity. In practice, when $\lambda = 0.2$, CN ranges from around 30 (for vegetated surfaces with highly permeable soils) to close to 100 (for impermeable surfaces or soils) (USDA, 1986). Tabulated CN values for various land uses, soil types, and management scenarios are available in handbooks and manuals (NRCS, 2003; USDA, 1986). CN can also be determined from field data by solving Eq. (4) for $S$ as

$$S = \frac{1}{2 \lambda^2} \left[2\lambda P + (1 - \lambda)Q - \sqrt{(1 - \lambda)^2 Q^2 + 4\lambda P Q}\right]$$

and then using Eq. (5). Conversely, when the CN of a watershed is known, $Q$ can be estimated for a rainfall event using Eqs. (4) and (5).
The curve number method is appealing because it is based on an intuitive concept (Eq. 2), relies on only one parameter, has a large body of literature (Hawkins et al., 2008), and has a comprehensive database (NRCS, 2003; USDA, 1986). It has been included in many watershed and water-quality models such as SWAT (Soil and Water Assessment Tool) (Neitsch et al., 2005), CREAMS (Chemicals, Runoff and Erosion from Agricultural Management Systems), GLEAMS (Groundwater Loading Effects of Agricultural Management Systems) (Knisel and Douglas-Mankin, 2012), AnnAGNPS (Annualized Agricultural Non-point Source Pollution Model) (Bingner et al., 2011), EPIC (Environmental Policy Integrated Climate), APEX (Agricultural Policy/Environmental Extender) (Wang et al., 2012), and HydroCAD (HydroCAD, 2015). A physically based modeling framework, such as the diffusive-wave approximation for overland flow coupled with the Richard’s equation for unsaturated subsurface flow, e.g., Panday and Huyakorn (2004), may improve accuracy and resolution of model predictions compared to the CN method, when the necessary input data, expertise, and computing resources are available. However, the CN method will likely remain popular for many applications in runoff modeling because of its ease of use, wide knowledge base, and less demand on computational resources than many physically based models.

1.2 CN variation with \( P \)

Curve number is assumed to be a watershed property that depends on the current conditions, but it also varies with \( P \) (e.g., Fig. 2a and b). This behavior was also observed in several previous studies (D’Asaro and Grillone, 2012; Hawkins, 1993; Hjelmfelt et al., 2001; Soulis et al., 2009), and it appears to be a common phenomenon. In 75% of the watersheds, CN decreased with increasing \( P \) and asymptotically approached a constant value. Hawkins (1993) referred to this as standard behavior. In 20% of the watersheds, CN decreased with \( P \) but an asymptote was not attained within the range of the observed \( P \). This was referred to as complacent behavior. In about 5% of the watersheds, the CN increased with \( P \) and asymptotically approached an apparent constant value. This behavior, referred to as violent, was often preceded by complacent behavior at smaller rainfalls. Hawkins (1993) hypothesized that the inverse relationship between CN and \( P \) may be due to some spurious correlation between them or due to a bias that inherently results from the selective omission of data from small storm events that failed to produce runoff. The reasoning is that large rainfalls always produce runoff but small rainfalls produce runoff only under wet conditions, when the CN is large. Therefore, small CN values for small rainfalls go unrecorded.
In watersheds showing a standard behavior, CN was treated as an asymptotic function of $P$ as

$$CN = CN_\infty + (100 - CN_\infty) e^{-kP}, \tag{7}$$

where $CN = CN_\infty$ is the asymptote and $k$ is a calibration parameter (Hawkins, 1993). $CN_\infty$ is the smallest possible value of CN for a watershed and is approached only at large values of $P$. To develop Eq. (7), measured values of $Q$, ideally for a large range of values of $P$, are needed. The usual procedure involves “frequency matching” the data (Hawkins, 1993), i.e., sorting the values of $P$ and $Q$ separately, and pairing them according to their rank. CN for each pair is then calculated using Eqs. (5) and (6). Frequency matching reduces the scatter of data points around the best-fit curve in a CN vs. $P$ plot.

A standard behavior of CN was also observed in two watersheds (BC5 and BC1) near Greenville, South Carolina, USA (Fig. 2a and b). In these watersheds, CN (calculated using $\lambda = 0.2$) decreased from 97 to 50 as $P$ increased from 2 to 128 mm. The data were characterized by a modest scatter ($R^2 = 0.9$) about the best-fit curve based on a quadratic function of $P$. Description of these watersheds is given by Santikari (2017) and Santikari and Murdoch (2018). The justification for using quadratic functions follows from the analysis of heterogeneity presented in Sect. 2.

The approach used in Fig. 2a and b avoids the commonly used frequency matching (e.g., Hawkins, 1993). Each CN value in the plot was calculated using the $P - Q$ pair from the same storm event. Frequency matching would significantly reduce the scatter in the plot, but it would also downplay the importance of CN variation due to antecedent conditions. Reducing the scatter by accounting for antecedent conditions, e.g., using antecedent moisture (Mishra et al., 2006), is a better approach.

The hypotheses given by Hawkins (1993) are valid, but insufficient to explain the standard and complacent behaviors. It may be true that small rainfalls produce runoff only under wet (large CN) conditions and therefore only the large CN values are recorded. However, if one has a large enough sample of storms, some of the larger storms also must have occurred during wet conditions. For the larger storms, therefore, one would expect to see the whole spectrum of CN values ranging from the largest to the smallest. However, this is not the case. As $P$ increases, the values of CN decrease consistently (Fig. 2a and b).

1.3 Heterogeneity as a cause of CN variation with $P$

Soulis and Valiantzas (2012) hypothesized that the observed variation of CN with $P$ in the standard and complacent cases is a consequence of watershed heterogeneity. They assumed a hypothetical heterogeneous watershed with two subareas characterized by different CNs. They then calculated the watershed runoff, for a range of values of $P$, as the area-weighted average of the runoffs from the subareas. Watershed CN calculated using this runoff varied with $P$ akin to the standard behavior. The shape of the synthetically generated CN vs. $P$ curve could be matched with the observations by adjusting the areas of the subareas and their respective CNs. This idea can also be extended to multiple subareas so that the heterogeneity within a watershed can be represented more accurately. However, this could lead to problems of over-parameterization, non-uniqueness, and non-convergence as pointed out by Soulis and Valiantzas (2012).

In a later paper, Soulis and Valiantzas (2013) suggested using spatial information on land cover and soils to delineate the areal extent of subareas and constrain their respective CNs. This approach would reduce the number of calibrated parameters by half because it only requires the calibration of the CNs for the subareas. In essence, the multiple-subarea approach is similar to a distributed modeling approach that calculates the watershed runoff as the area-weighted average of the runoffs from the subareas, e.g., SWAT (Gassman et al., 2007). The approach used by Soulis and Valiantzas (2013) attempts to match the observed and simulated values of CN, whereas that used by SWAT attempts to match the observed and simulated values of $Q$. Since CN and $Q$ are uniquely related for given values of $P$ and $\lambda$, these approaches are equivalent. A major implication of the work of Soulis and Valiantzas (2013) is that a distributed modeling approach can account for the standard and complacent behaviors of CN.

Using a single value of CN independent of $P$ in a heterogeneous watershed can cause a systematic error in $Q$ and lead to poor predictive ability of the method. This is because when CN is constant $Q$ may be underestimated for small $P$ and overestimated for large $P$ (e.g., Soulis and Valianzas, 2012, 2013). This problem can be addressed either by treating CN as a function of $P$, e.g., asymptotic fitting (Hawkins, 1993), or by using a distributed modeling approach that accounts for heterogeneity in sufficient detail, e.g., SWAT (Gassman et al., 2007) or Soulis and Valianzas (2013). An understanding of the mechanism of how watershed heterogeneity leads to the variation of CN with $P$ is also important. It could help in accounting for this variation without resorting to finer discretization or over-parameterization of the CN method. To accomplish this, an analysis of the effect of heterogeneity on $I_a$ and $S$ is performed, which can then be used to understand the effect on CN.

2 Reevaluation of initial abstraction

The quantities CN, $I_a$, and $S$ are considered to be the properties of a watershed that depend on current conditions. In usual practice, CN estimated for a certain set of conditions is applicable to any rainfall event occurring in those conditions irrespective of the magnitude of $P$. However, in every watershed evaluated by the previous studies (D’Asaro and Grillon, 2012; Hawkins, 1993) the CN varied with $P$. If so, since $I_a$ and $S$ are inversely related to CN (Eqs. 3 and 5), one
can expect that they too vary with $P$ but inversely to that of CN. The calculated values of $I_a$ (using Eqs. 3, 6, and $\lambda = 0.2$) for watersheds BC5 and BC1 near Greenville, SC, increase with $P$ and appear to approach a constant at large values of $P$ (Fig. 2c and d). A plot of $S$ vs. $P$ would be similar to the $I_a$ vs. $P$ plot, with the $y$ coordinate scaled by $1/\lambda$.

To evaluate the link between heterogeneity in $I_a$ and its variation with $P$, we looked at how the effective $I_a$ of a heterogeneous watershed is determined and whether it is affected by the magnitude of $P$. Our analysis shows that there is an inconsistency between the theoretical definition of $I_a$ and its calculated value at the watershed scale. It also shows how heterogeneity can cause $I_a$ to vary with $P$ and how this relates to variations of $S$ and CN with $P$.

2.1 Problems with the current usage of $I_a$

By the theoretical definition of $I_a$, if runoff is detected in the hydrograph, it is assumed that $I_a$ has been met for the watershed. Watersheds are heterogeneous combinations of various land-use–soil–slope complexes. These are referred to as hydrologic response units (HRUs) in SWAT (Gassman et al., 2007), and the same term is also used here. Each HRU is assumed to be homogeneous and is characterized by representative values of CN ($CN_i$) and $I_a$ ($I_{ai}$). During a rainfall event, the HRU with the smallest of the $I_{ai}$ values will be the first to generate runoff. Assuming that this runoff reaches the watershed outlet, by definition, the $I_a$ of the watershed should be equal to the smallest of the $I_{ai}$ values. This could even be zero if the watershed has surfaces such as open water bodies that cannot abstract the rainfall.

However, it is difficult to detect the exact moment of generation of runoff and determine the corresponding value of $I_a$, which is equal to the cumulative precipitation at that moment. There have been studies (Shi et al., 2009; Woodward et al., 2003) that tried to determine $I_a$ from hydrographs. A problem with this approach is that there can be a time lag between runoff generation in headwaters and its detection at gauging station. Rainfall that occurs during this time lag is also included in $I_a$, leading to its overestimation. Another possible approach would be to collect observations from a large number of rainfall events and take $I_a$ to be equal to the smallest $P$ that produced runoff. This would eliminate the problem with the lag time, but $Q$ needs to be insignificant to reduce the error in $I_a$. It should also be noted that $I_a$ determined this way is only representative of the antecedent conditions of the smallest event that produced runoff.

It may be difficult to measure $I_a$ directly, but it can be calculated for any event using Eqs. (6) and (3). However, in medium to large rainfall events, even the HRUs with larger values of $I_{ai}$ will contribute to $Q$. Therefore, the calculated value of $I_a$ in these events will also be influenced by the larger values of $I_{ai}$. So, the calculated $I_a$ tends to be greater than the smallest of the $I_{ai}$ values. Moreover, it can be expected to increase with $P$ as increasingly larger rainfalls generate runoff from HRUs with increasingly larger values of $I_{ai}$. Thus, there is an inconsistency between the definition of $I_a$ and its calculated value at the watershed scale.

**Spatial-scale effect on $\lambda$**

Strictly adhering to the definition of $I_a$ at the watershed scale may also cause a spatial-scale effect on $\lambda$. Let us refer to the CN of the watershed as $CN_W$ and to $I_a$ as $I_{aw}$. One of the common ways to determine $CN_W$ is to calculate it as the area-weighted average of the $CN_i$ values (NRCS, 2003) as

$$CN_W = \sum_{i=1}^{n} a_i CN_i,$$

where $a_i$ is the fractional area of the $i$th HRU. Note that the fractional areas must add up to unity. By definition, $I_{aw}$ is equal to the smallest of the $I_{ai}$ values. Therefore, if $I_{a1} < I_{a2} < \ldots < I_{an}$, then

$$I_{aw} = I_{a1}.$$  (9)

From Eqs. (3) and (5) it can be shown that CN and $I_a$ are related as

$$CN = \frac{25400}{254 + \left(\frac{I_a}{\lambda}\right)},$$

(10)

If all the HRUs are assumed to have the same $\lambda = \lambda_i$, Eqs. (8)–(10) lead to

$$CN_1 > CN_2 > \ldots > CN_n$$

$$CN_W < CN_1$$

$$I_{aw} = I_{a1}$$

$$\frac{\lambda_w}{\lambda_i} > 1,$$

$$\lambda_w < \lambda_i,$$

(11)

where $\lambda_w$ is the effective initial abstraction ratio of the watershed. Therefore, if $\lambda$ is assumed to be the same among the component HRUs, it will have a smaller value at the watershed scale. This implies that $\lambda$ decreases with increasing spatial scale. Therefore, setting $\lambda$ constant, equal to 0.2 or 0.05, for all the spatial scales contradicts the definition of $I_a$. In any case, it is probably more accurate to calculate runoff at the HRU scale ($Q_i$) and take the area-weighted average of $Q_i$ values, rather than take the area-weighted average of the $CN_i$ values and calculate $Q$ at the watershed scale. It is also more appropriate because $Q$ is runoff per unit area, whereas CN is a dimensionless index variable.

The inconsistencies in the usage of $I_a$ are a direct result of heterogeneity in a watershed. Moreover, heterogeneity also appears to be responsible for the variation of $I_{aw}$ with $P$ (Fig. 2c and d). To verify this, a relationship between $I_{aw}$ and the magnitude and areal distribution of $I_{ai}$ values needs to be developed.
In a rainfall event, all the HRUs with $I_{ai}$ values of various HRUs mainly characterized by their land use types ($I_{a0} < I_{a1} < I_{a2} < I_{a3}$); (b) conceptual model in which each HRU is represented by a bin with height $= a_i$, width $= a_i$, and unit thickness; shaded area indicates the filled portion during an event with rainfall $= P$.

2.2 $I_a$ in a heterogeneous watershed

Consider a watershed with four HRUs mainly characterized by their land use types, viz. open waterbody ($I_{a0}$), urban area ($I_{a1}$), park ($I_{a2}$), and forest ($I_{a3}$) (Fig. 3a), such that $I_{a0} = 0 < I_{a1} < I_{a2} < I_{a3}$. An open waterbody generates runoff during every rainfall event. Other land use types generate runoff depending on the magnitude of the rainfall, with land uses of larger $I_{ai}$ values requiring larger magnitudes. If each land use type is assumed to be directly connected to the drainage network, the number of land use types contributing to the runoff, in other words the runoff contributing area, increases with rainfall. This process can be conceptualized by representing the storage distribution of $I_a$ as a series of bins where each bin corresponds to a HRU (Fig. 3b). The height and the width of a bin are given by $I_{ai}$ and $a_i$, respectively, and all bins have unit thickness. In a rainfall event, only the bins with $I_{ai} ≤ P$ are fully filled and contribute to runoff, whereas the others are partially filled and do not contribute to runoff. The total amount of filled storage in $I_a$ (shaded area in Fig. 3b) increases with $P$ until it reaches a constant value when all the bins are fully filled and the whole watershed is contributing to the runoff.

Consider a general case of a heterogeneous watershed with $n + 1$ HRUs such that

\[ I_{a0} = 0 < I_{a1} < I_{a2} < \ldots < I_{an}, \]

where $I_{a0}$ represents open water bodies and other surfaces that cannot abstract rainfall. The areal average of the total initial abstraction ($I_{aT}$) is given by

\[ I_{aT} = \sum_{i=0}^{n} a_i I_{ai}. \]

In a rainfall event, all the HRUs with $I_{ai} ≤ P$ have their initial abstractions completely filled while the others are partially filled. Just by analyzing the runoff for that event, it is impossible to quantify the magnitudes of $I_{ai}$ in HRUs that are partially filled. Because these HRUs have not contributed to the runoff, all that can be said is that their $I_{ai}$ values are greater than $P$, but their magnitudes remain unknown. However, the information on the magnitudes of $I_{ai}$ in HRUs that are completely filled should be present in the runoff data. In other words, it takes larger rainfalls to fill larger values of $I_{ai}$ and gather information about their magnitude.

Then what is the effective initial abstraction of the watershed for a given rainfall event? Consider an event where the rainfall falls within the range: $I_{am} ≤ P < I_{am+1}$. HRUs with $I_{ai} ≤ I_{am}$ have their initial abstractions completely filled and produce runoff, whereas HRUs with $I_{ai} ≥ I_{am+1}$ have their initial abstractions partially filled up to the level of $P$ and do not produce runoff. The areal average of the filled portion (includes completely filled as well as partially filled HRUs) of the initial abstraction is given by

\[ I_{af} = \sum_{i=0}^{m} a_i I_{ai} + \left(1 - \sum_{i=0}^{m} a_i\right) P. \]  

(14)

The first term on the right-hand side of Eq. (14) represents completely filled HRUs. The second term represents partially filled HRUs, all of which are filled to the level of $P$. Note that $I_{af}$ is the areal average of total initial abstraction, whereas $I_{af}$ is the areal average of the filled portion. Therefore,

\[ I_{af} < I_{af} \forall P < I_{am}, \]

\[ I_{af} = I_{af} \forall P \geq I_{am}. \]  

(15)

The conceptual model presented in Fig. 3 as well as in Eqs. (14) and (15) is intuitively appealing and also hints at the possibility that $I_{aw}$ may be equal to $I_{af}$. This is because $I_{af}$ increases with $P$ and approaches a constant value ($I_{af}$), similar to the observations in Fig. 2c and d. Equation (14) is also consistent with a distributed parameter model application of the CN method as described in Sect. 3.

2.3 Variation of $I_{af}$ with $P$

To investigate the variation of $I_{af}$ with $P$, Eqs. (14) and (15) are applied to the scenario presented in Fig. 3, where $n = 3$. A
plot of \( I_{aF} \) vs. \( P \) (Fig. 4) shows that \( I_{aF} \) increases with \( P \) and becomes constant (\( I_{aF} = I_{aT} \)) at large values of \( P \) (\( P \geq I_{a3} \)). The kink points joining the line segments occur when the initial abstraction of one of the HRUs becomes completely filled. At these points, \( P \) is equal to one of the \( I_a \) values. In between these points (\( I_{a0} < P < I_{a(m+1)} \)), the relationship between \( I_{aF} \) and \( P \) is linear with a slope of \( \left( 1 - \sum_{i=0}^{m} d_i \right) \). The slope abruptly changes across the kink points. It decreases with \( m \) and becomes zero when \( m = n \). The maximum value the slope can take is unity. This occurs with the line segment passing through the origin, when HRUs with zero initial abstraction are absent (i.e., \( a_0 = 0 \)). When these HRUs are present, however, the origin itself is a kink point where the slope abruptly jumps from unity to \( 1 - a_0 \).

The analysis presented so far represents a discrete case where each HRU is homogeneous and has a finite area. The values of \( I_{a} \) vary discontinuously across the HRUs. Their areal distribution can be represented by a plot of \( a_i \) vs. \( I_a \) (Fig. 5a). The smaller the area of HRUs, the more numerous they are and the more accurate the representation of the heterogeneity within the watershed is. The most ideal representation would occur when the HRUs shrink to points. Then the values of \( I_{a} \) within the watershed vary continuously and therefore can be represented by a probabilistic distribution of areal occurrence (Fig. 5b). It is impractical to characterize the watershed at such fine scale, but it is worth understanding the properties of the initial abstraction at the finest resolution first and then making assumptions or simplifications later to suit the practical needs.

For the case of a continuous distribution of \( I_a \), Eq. (14) takes the form

\[
I_{aF} = \int_0^P I_a a(I_a) \, dI_a + \left( 1 - \int_0^P a(I_a) \, dI_a \right) \, P, \tag{16}
\]

where \( a(I_a) \) is the probability density function of areal occurrence of \( I_a \). The fractional area with initial abstraction \( = I_a \) is given by \( a(I_a) \, dI_a \). The upper limit of the integrals is set to \( P \) because the last initial abstraction to completely fill up would be equal to \( P \). The areal average of total initial abstraction, \( I_{aT} \), is given by

\[
I_{aT} = \int_0^{1} I_a a(I_a) \, dI_a, \tag{17}
\]

where \( I_{a,\text{max}} \) is the maximum value of \( I_a \) within the watershed. Thus, \( I_{aT} \) is equal to the mean of the distribution (Fig. 5b). Equation (15) then becomes

\[
I_{aF} < I_{aT} \quad \forall \quad P < I_{a,\text{max}},
I_{aF} = I_{aT} \quad \forall \quad P \geq I_{a,\text{max}}. \tag{18}
\]

Unlike the discrete case, the slope of the \( I_{aF} \) curve for the continuous case decreases smoothly with increasing \( P \) (Fig. 6). This is because the line segments in the discrete case (Fig. 4) shrink to points in the continuous case. It follows from Eq. (16) that

\[
\frac{dI_{aF}}{dP} \bigg|_{P=0} = 1,
\frac{dI_{aF}}{dP} \bigg|_{P\geq I_{a,\text{max}}} = 0. \tag{19}
\]

Thus, the \( I_{aF} \) curve is bounded by a line of slope \( = 1 \) passing through the origin and a line of slope \( = 0 \) with the intercept equal to \( I_{aT} \) (Fig. 6). The line of slope \( = 1 \) is referred to

![Figure 4. Variation of \( I_{aF} \) (Eqs. 14 and 15) with \( P \) for the scenario presented in Fig. 3.](image)

![Figure 5. Representing areal distribution of \( I_a \) within a watershed (a) discrete case and (b) continuous case.](image)
as the no-runoff line because along this line $I_{aF} = P$. When the whole watershed is represented by a single HRU, the $I_{aF}$ curve coincides with the no-runoff line until $I_{aF} = I_{aT}$. A comparison of Figs. 6 to 2c and d strengthens the case that $I_{aW}$ is equal to $I_{aF}$.

### 2.4 Variation of $CN_W$ with $P$

Let us hypothesize that $I_{aW} = I_{aF}$; i.e., the effective $I_a$ of a watershed is equal to the area-weighted average of the filled portion of the initial abstraction. Then, if Eq. (10) is written for $CN_W$, $I_a$ can be replaced by $I_{aF}$. Substituting Eq. (16) in Eq. (10) gives $CN_W$ as a function of $P$. When plotted against $P$, $CN_W$ starts at 100 when $P = 0$ and then decreases with increasing $P$ (Fig. 7).

Differentiating Eq. (10) and using Eq. (19) gives

$$\frac{d(CN_W)}{dP} \bigg|_{P=0} = -\frac{10}{\lambda},$$

$$\frac{d(CN_W)}{dP} \bigg|_{P \geq I_{a,max}} = 0,$$

where the constant 10 has units of $1/\text{in.}$ Thus, the $CN_W$ vs. $P$ curve is at its steepest at $P = 0$ and flattens with increasing $P$, and it becomes constant when $P \geq I_{a,max}$. This constant, $CN_T$, is the smallest value $CN_W$ can take and corresponds to the case $I_{aF} = I_{aT}$, when the initial abstractions of all the HRUs are fully filled. $CN_W$ as a function of $P$ is bounded by a curve corresponding to the condition $P = I_{aF}$, the no-runoff line, and a line of slope 0 with the intercept equal to $CN_T$ (Fig. 7).

The shape of the $CN_W$ vs. $P$ curve (Fig. 7) generated using Eqs. (10) and (16) is quite similar to the best-fit curves from field observations (Fig. 2a and b). Nearly 95% of the watersheds evaluated in the previous studies (D’Asaro and Grillone, 2012; Hawkins, 1993) also had responses identical to Fig. 7, supporting the hypothesis that $I_{aW} = I_{aF}$. Thus, as also concluded by Soulis and Valiantzas (2012, 2013), the observed complacent and standard behaviors are caused by the inevitable presence of heterogeneity in a watershed. Moreover, complacent behavior appears to be a special case of standard behavior (Soulis and Valiantzas, 2012), where observations from larger rainfalls are unavailable. Therefore, it is probably more appropriate to refer to any “$CN$ decreasing with $P$” trend as standard behavior. It also shows that assuming a partial source area whenever a complacent behavior is observed (D’Asaro and Grillone, 2012, 2015) can be misleading.

### 2.5 $I_{aF}$ and $CN_W$ curves for various distributions of $I_a$

The functional form of $a(I_a)$ defines the areal distribution of $I_a$ within a watershed. We considered idealized functional forms of $a(I_a)$ that correspond to uniform, normal, triangular, and bimodal distributions (Table 1). In each $a(I_a)$, the maximum or other key value was constrained so that the total area under the distribution was unity. For example, the $y$ coordinate of the apex in the triangular distribution must be equal to $2/I_{a,max}$ (Table 1). In the case of normal distribution, however, the area under the curve is unity only when the limits are infinite. Therefore, a standard deviation ($\sigma$) much less than $I_{a,max}$ was used so that the area under the curve within the range $0 \leq I_a \leq I_{a,max}$ is approximately equal to unity.

For each distribution, the corresponding functional form of $I_{aF}$ was determined using Eq. (16) and the results are presented in Table 1. For the general case of $a(I_a)$ as a polynomial, the corresponding $I_{aF}$ is a polynomial two degrees higher than $a(I_a)$. For the normal distribution, $I_{aF}$ is a combination of Gaussian and error functions (Table 1).

For the purpose of comparison, symmetrical versions of the distributions were considered such that all of them have the same minimum, mean, and maximum values of $I_a$ (Fig. 8a). The minimum value of $I_a$ was set to zero and the maximum value was $I_{a,max}$. Therefore, the mean for all the distributions was $I_{a,max}/2$.

The kurtosis (peakedness) of $a(I_a)$ has a major influence on the shapes of $I_{aF}$ and $CN_W$ plotted as functions of $P$ (Fig. 8). The normal distribution has the greatest kurtosis,
whereas the bimodal distribution has the least. As the kurtosis decreases, the $I_aF$ and $CN_W$ curves deviate further from the bounding lines (Fig. 8). When there is a gap in the distribution, as in the case of the bimodal distribution, the corresponding $I_aF$ curve is linear for the range spanning the gap. This is consistent with the discrete case where $I_a$ was represented by line segments for the gaps in between the discrete values of $I_a$ (Fig. 4).

Skewness of $a(I_a)$ also affects $I_aF$, and this is illustrated by an idealized case where an initially uniform distribution is positively skewed (Fig. 9a and b). The mean of $a(I_a)$, which is equal to $I_aT$ (Eq. 17), decreases with increasing positive skewness. This is important because a land use change such as conversion of forest to urban land is expected to increase the positive skewness (i.e., more low values of $I_a$). During the conversion, $I_{a,max}$ remains unchanged while some forested land remains. When the entire forest is converted, $I_{a,max}$ drops to a lower value.

The analysis also shows that a watershed cannot be characterized or compared with other watersheds using a single value of $CN$ (such as $CN_{\infty}$ used in asymptotic fitting, Eq. 7). Depending on the distribution of heterogeneity, the relative runoff potential of a watershed can be $P$ dependent. This is illustrated by considering two uniform distributions, uni1 and uni2, where uni2 has a narrower range and a smaller mean than uni1 (Fig. 9c). For smaller values of $P$, $I_{aF,uni1} < I_{aF,uni2}$ (Fig. 9d) and therefore $CN_{W,uni1} > CN_{W,uni2}$. However, for larger values of $P$, the converse is true. Thus, the watershed with uni1 generates more runoff for smaller values of $P$, whereas the watershed with uni2 generates more runoff for larger values of $P$.

### 3 Effect of heterogeneity on $S$

Similar to the case of $I_a$, the presence of heterogeneity also causes the effective $S$ of a watershed ($S_W$) to vary with $P$. The functional form of $S_W$ depends not only on the potential maximum retentions of the HRUs (values of $S_i$) but also on the values of $I_{ai}$. $S_W$ can be estimated using Eq. (2) if the quantities $I_{ai}$, $Q_i$, and $F_i$ are known. A distributed modeling approach can be used to calculate these quantities for a heterogeneous watershed. Distributed CN models, e.g., SWAT (Gassman et al., 2007), commonly calculate $Q$ as the area-weighted average of $Q_i$s, and this assumption can also be extended to $F_W$. Thus,

$$Q_W = \sum_{i=0}^{n} a_i Q_i,$$

$$F_W = \sum_{i=0}^{n} a_i F_i . \tag{21}$$

Using Eq. (21) and applying mass balance (Eq. 1) at watershed and HRU scales gives Eq. (14) for $I_{ai}W$. This shows that $I_{ai}W$ calculated using a distributed model is equal to $I_{ai}F$.

Writing an expression for $S_W$ in terms of $I_{ai}$ and $S_i$ values for a general case of a heterogeneous watershed is cumbersome. Therefore, it is only presented graphically for an example of a heterogeneous watershed. However, an expression for $S_W$ can be presented in a compact form for a special case where all the values of $I_{ai}$ are zero as

if $I_{ai} = 0 \forall i \in {0, 1, \ldots n}$

$$S_W = \frac{\sum_{i=0}^{n} a_i S_i}{P + \sum_{i=0}^{n} a_i} , \tag{22}$$
Thus, $S_W$ varies from the area-weighted harmonic mean 
\[
\left(\sum_{j=0}^{n} a_i S_j \right)^{-1}
\] 
when $P = 0$ to the area-weighted arithmetic mean 
\[
\left(\sum_{j=0}^{n} a_i S_j \right)
\] 
when $P \gg S$.

To illustrate the effect of heterogeneity on $S_W$, an example watershed with the storage distribution shown in Table 2 was considered. The variation of $S_W$ with $P$ was analyzed for the cases of $\lambda_i = 0$ and $\lambda_i = 0.2$ (Fig. 10). In both cases, $S_W$ increases with $P$ and approaches the area-weighted arithmetic mean, $S_\infty$, for large values of $P$. In the case of $\lambda_i = 0$, the slope of the curve is maximum at the origin and decreases monotonically with $P$. In the case of $\lambda_i = 0.2$, however, the slope is zero at the origin and generally increases with $P$ until $P \approx I_{an} = 40$ mm ($P \approx I_{a,max}$ for the continuous case), where it reaches its maximum value. Thereafter the slope decreases monotonically with $P$, giving an $S$-shaped curve. In other words, the slope generally increases with $P$ until the entire watershed area contributes to the runoff and decreases thereafter.

The similarities between $I_{AW}$ and $S_W$ are that they both increase with $P$ and have an upper limit equal to the area-weighted arithmetic mean of their respective components.

The difference is that $I_{AW}$ reaches its upper limit of $I_{AT}$ for a finite value of $P$ ($P = I_{an}$ or $P = I_{a,max}$), whereas $S_W$ requires large values of $P$ ($P \gg S$) to reach its upper limit of $S_\infty$. Moreover, $S_W$ vs. $P$ is an $S$-shaped curve when $\lambda_i > 0$. This shows that $I_{AW}$ and $S_W$ are not proportional, i.e., $\lambda_W$ is not a constant even though $\lambda_i$ values are assumed to be equal and constant.

4 Application

The analysis from previous sections shows that $I_{AW}$ and $S_W$ are functions of $P$ and gives their functional forms. Incorporating these functions in the lumped-parameter application can potentially improve the performance of the CN method.

4.1 $I_{AW}$ as a function of $P$

The distributed parameter modeling approach, Eq. (21) with the application of mass balance (Eq. 1) at watershed and HRU scales, shows that $I_{AW} = I_{AF}$. $I_{AF}$ is given by Eq. (14) for the discrete case and Eq. (16) for the continuous case. All the distributions in Table 1, except the normal distribution, gave a zero-intercept polynomial for $I_{AF}$. Therefore, using a quadratic function of the form...
Table 1. Functional forms of \( a(I_a) \) and \( I_{aF} \) for various synthetic distributions.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Graph</th>
<th>( a(I_a) )</th>
<th>( I_{aF} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td><img src="image" alt="Uniform Graph" /></td>
<td>( \frac{1}{I_{a,\text{max}}} )</td>
<td>( p - \frac{p^3}{2I_{a,\text{max}}} )</td>
</tr>
<tr>
<td>Normal</td>
<td><img src="image" alt="Normal Graph" /></td>
<td>( \frac{1}{\sigma \sqrt{2\pi}} e^{-\left(I_a - \mu\right)^2/(2\sigma^2)} )</td>
<td>( p - \frac{\sigma}{\sqrt{2\pi}} \left[ e^{-\left(P - \mu\right)^2/(2\sigma^2)} - e^{-\mu^2/(2\sigma^2)} \right] - \left(\frac{P - \mu}{\sigma} - \frac{P - \mu}{\sqrt{2\pi}} \right) \left[ \text{erf} \left( \frac{P - \mu}{\sigma \sqrt{2\pi}} \right) + \text{erf} \left( \frac{-\mu}{\sigma \sqrt{2\pi}} \right) \right] )</td>
</tr>
<tr>
<td>Triangular</td>
<td><img src="image" alt="Triangular Graph" /></td>
<td>( \frac{2I_{a,\text{max}}}{a I_{a,\text{max}} - a} ) if ( I_a \leq a ) ( \frac{2I_{a,\text{max}}}{a I_{a,\text{max}} - a} \left( 1 - \frac{I_a}{I_{a,\text{max}}} \right) ) if ( a &lt; I_a )</td>
<td>( p - \frac{p^3}{2I_{a,\text{max}}} ) if ( P \leq a ) ( \frac{1}{I_{a,\text{max}} - a} \left[ \frac{a^2}{2} + I_{a,\text{max}} P - P^2 + \frac{p^3}{2I_{a,\text{max}}} \right] ) if ( a &lt; P )</td>
</tr>
<tr>
<td>Bimodal</td>
<td><img src="image" alt="Bimodal Graph" /></td>
<td>( u ) if ( I_a \leq a ) ( v = \frac{1 - au}{I_{a,\text{max}} - b} ) if ( a &lt; I_a &lt; b ) ( v ) if ( b \leq I_a )</td>
<td>( p - \frac{u p^2}{2} ) if ( P \leq a ) ( \frac{u p^2}{2} + (1 - u a) P ) if ( a &lt; P &lt; b ) ( \frac{u p^2}{2} + (1 - u a + v b) P - \frac{v p^2}{2} ) if ( b \leq P )</td>
</tr>
</tbody>
</table>

Figure 10. Variation of \( S_{aw} \) with \( P \) in a heterogeneous watershed with the storage distribution shown in Table 2.

\[
I_{aw} = c_1 P - c_2 P^2 \quad \forall \quad P \leq I_{a,\text{max}}
\]

\[
I_{aw} = I_{aT} = c_1 \left( I_{a,\text{max}} \right) - c_2 \left( I_{a,\text{max}} \right)^2 \quad \forall \quad P > I_{a,\text{max}}
\]

is an efficient way to describe \( I_{aw} \). In Eq. (23), \( c_1 \) and \( c_2 \) are calibration parameters such that \( 0 \leq c_1 \leq 1 \) and \( c_2 \geq 0 \). Since the slope of \( I_{aw} \) is zero at \( P = I_{a,\text{max}} \) (Eq. 19), it follows from Eq. (23) that

\[
I_{a,\text{max}} = \frac{c_1}{2c_2}
\]

\[
I_{aT} = \frac{c_1^2}{4c_2}
\]

Similarly, the slope of \( I_{aw} \) is unity at \( P = 0 \) so \( c_1 \) should be unity. However, it was kept as a free parameter in Eq. (23) to

www.hydrol-earth-syst-sci.net/22/4725/2018/
allow for the approximation of piecewise functions (e.g., \( I_{aF} \) for triangular and bimodal distributions in Table 1). Moreover, the analysis for the discrete case shows that when HRUs with zero initial abstraction are present, the origin is a kink point where the slope abruptly jumps from unity to \( 1 - a_0 \). To avoid over-parameterization of the model, a polynomial of degree \( > 2 \) for \( I_{aW} \) was not considered.

### 4.2 \( S_W \) as a function of \( P \)

The sigmoid-shaped function of \( S_W \), with the conditions that \( S_W = 0 \) when \( P = 0 \) and that the maximum slope occurs at \( P = I_{a,max} \), requires at least two parameters to describe it. However, this along with Eq. (23) would also increase the number of calibrated parameters in the CN method, increasing its complexity and potentially causing non-uniqueness. A relatively simple approach is to assume that \( S_W \) is constant similar to the conventional CN method. Another approach is to assume that \( S_W \) is proportional to \( I_{aW} \); i.e., Eq. (3) is applicable for a heterogeneous watershed.

Here the emphasis is placed on treating \( I_{aW} \) as a function of \( P \) while offering some flexibility on how \( S_W \) is treated. This is because the variation of \( I_{aW} \) with \( P \) had a significant impact on the model performance, whereas including the variation of \( S_W \) with \( P \) showed only marginal or no improvement. This may be because \( I_{aW} \) is a component of mass balance (Eq. 1) while \( S_W \) is not. \( F_W \), which is the filled portion of \( S_W \), is a component of mass balance and varies with \( P \) even if \( S_W \) is assumed to be a constant. Therefore, to maintain the simplicity of the CN method and avoid the problems of over-parameterization and non-uniqueness, modeling the sigmoid-shaped function of \( S_W \) is omitted.

### 4.3 Lumped-parameter models

Lumped-parameter application of the CN method was modified by treating \( I_{aW} \) as a function of \( P \) as described in the previous section. Modified lumped-parameter CN models were evaluated by comparing their performance with that of the conventional lumped-parameter CN models.

#### 4.3.1 Conventional models (CMs)

Conventional CN models are defined by Eqs. (1) through (5) and by the assumption that \( I_{aW} \) and \( S_W \) are independent of \( P \). In this study two types of conventional models, referred to as CM0.2 and CM\( \lambda_W \), were used. In CM0.2, \( \lambda_W \) was fixed at 0.2, and in CM\( \lambda_W \), \( \lambda_W \) was determined by calibration. Thus, CM0.2 had one free parameter, \( S_W \), whereas CM\( \lambda_W \) had two free parameters, \( \lambda_W \) and \( S_W \).

#### 4.3.2 Variable initial abstraction models (VIMs)

VIMs are defined by Eqs. (1), (2), (4), (5), and (23), and they have three free parameters. If \( S_W \) is assumed to be independent of \( P \), then the model requires calibration of \( c_1 \), \( c_2 \), and \( \lambda_W \).

### 5 Evaluation

Lumped-parameter models described in the previous section were evaluated in their ability to predict runoff and account for watershed heterogeneity. Accounting for heterogeneity means that the model accurately predicts \( I_{aW} \) and \( S_W \), as well as runoff from smaller events. This is because (i) \( I_{aW} \) and \( S_W \) as functions of \( P \) are directly related to heterogeneity, and (ii) the model’s inability to account for their variation with \( P \) causes underestimation of runoff in smaller events.

Evaluation of lumped-parameter models requires the data for \( I_{aW} \), \( Q_W \), and \( S_W \). This is generated using a distributed parameter model application of the CN method. The assumption is that a distributed parameter model accounts for heterogeneity, and therefore its estimates of \( I_{aW}, Q_W, \) and \( S_W \) are accurate.

#### 5.1 Distributed parameter model

In a distributed parameter model, Eqs. (1) through (5) are applicable at the HRU scale, with the assumption that \( I_{aW} \) and \( S_i \) are independent of \( P \). Once \( Q_i \) and \( F_i \) are calculated for each HRU, watershed-scale quantities \( I_{aW} \), \( Q_W \), \( F_W \), and \( S_W \) are calculated using Eqs. (14), (21), and (2).

The distributed parameter model was applied to an idealized synthetic watershed with the storage distribution shown in Table 2, for the cases of \( \lambda_j = 0, 0.2, \) and 0.5. A range of values of \( P \) were synthetically generated such that they vary lognormally from 0.1 to 200 mm with a median of 8 mm. For each rainfall event, \( I_{aW}, Q_W, F_W, \) and \( S_W \) were calculated and used in the evaluation of the lumped-parameter models.

The reason for using a synthetic watershed here is that the heterogeneity can be precisely defined and used to evaluate the predictions of heterogeneity by the lumped-parameter models. In real watersheds, the heterogeneity has to be determined by calibration, and there can be non-uniqueness when multiple HRUs are present. Application of these modified models to data from real watersheds is discussed by Santikari (2017) and Santikari and Murdoch (2018).

### Table 2. Storage distribution in a hypothetical heterogeneous watershed used to illustrate the variation of \( S_W \) with \( P \).

<table>
<thead>
<tr>
<th>HRU</th>
<th>( a_i )</th>
<th>( S_i ) (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.05</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0.20</td>
<td>50</td>
</tr>
<tr>
<td>2</td>
<td>0.35</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>0.25</td>
<td>150</td>
</tr>
<tr>
<td>4</td>
<td>0.15</td>
<td>200</td>
</tr>
</tbody>
</table>

\( S_W \) and is referred to as VIMS. If Eq. (3) is also included, then the model requires calibration of \( c_1 \), \( c_2 \), and \( \lambda_W \) and is referred to as VIM\( \lambda \).
5.2 Model evaluation criteria

Each lumped-parameter model was calibrated by minimizing the sum of the squared residuals between its predicted runoff ($Q_{\text{P}}$) and the baseline from the distributed parameter model. All the models were evaluated using the Nash–Sutcliffe efficiency parameter (NSE), the standard error of estimate (SEE), and the percent bias (PB) (McCuen, 2003; Moriasi et al., 2007). NSE can vary from $-\infty$ to 1. The calculations and observations are exactly equal when NSE = 1. The calculations are only as good as the average observation when NSE = 0. SEE is the root-mean-square residual adjusted to the degrees of freedom (Santikari, 2017). A smaller SEE indicates a better performance, and its ideal value is zero. PB indicates whether the model is over- (PB < 0) or under-predicting (PB > 0) on average. The optimal value for PB is zero.

NSE values were calculated for the model predictions of runoff (NSE$_Q$), initial abstraction (NSE$_{Ia}$), potential maximum retention (NSE$_S$), and runoff from events with $P$ less than the median value (NSE$_{Q50}$). Relative NSE, rNSE (Krause et al., 2005), was used instead of NSE when the latter was strongly influenced by larger events. PB values were calculated for runoff from all the events (PB$_Q$) and runoff from events with $P$ less than the median value (PB$_{Q50}$). SEE was calculated for runoff from all the events (SEE$_Q$).

NSE$_I$ and NSE$_S$ indicate how accurately a lumped-parameter model predicts the watershed heterogeneity. NSE$_Q$, SEE$_Q$, and PB$_Q$ reflect the overall accuracy in a model prediction of runoff from all the events, whereas NSE$_{Q50}$ and PB$_{Q50}$ reflect the accuracy in predicting runoff from smaller events ($P < 8$ mm). Conventional models tend to under-predict runoffs from smaller events because of the usage of constant $I_a$ and $S$. They often falsely predict zero-runoffs because the runoff condition ($P > I_a$) cannot be overcome in smaller events. NSE$_{Q50}$ and PB$_{Q50}$ are used to expose this shortcoming.

6 Results and discussion

The results show that using variable initial abstraction improved the accuracy of model predictions of runoff and heterogeneity (Table 3). Based on their overall performance, the models can be arranged from the best to the worst as VIM$\lambda$ > VIMS > CM$\lambda$ > CM0.2. Results for the case of $\lambda_i = 0$ are not presented in Table 3 because VIM$\lambda$, VIMS, and CM$\lambda$ performed equally well while CM0.2 was the worst (i.e., VIM$\lambda$ = VIMS = CM$\lambda$ > CM0.2).

Variable $I_a$ models predicted runoff better than the conventional models. It was not possible to determine relative model performance using NSE$_Q$ because it was 1.0 for all the models. This was because NSE$_Q$ was strongly influenced by a few larger events, and a good fit in these events was sufficient to render NSE$_Q = 1.0$. Therefore, rNSE$_Q$ (Krause et al., 2005) was calculated instead and listed in Table 3. Larger events had greater influence on rNSE$_Q$ as well, but the values varied slightly between the models (Table 3). rNSE$_Q$ increased down the table whereas SEE$_Q$ decreased, with both indicating an improvement in model performance. PB$_Q$ was positive for all the models, indicating that they all under-predicted runoff. The extent of under-prediction, however, was smaller in variable $I_a$ models than the conventional models.

Variable $I_a$ models gave a better estimate of watershed heterogeneity than the conventional models as indicated by the higher values of NSE$_{Ia}$ and NSE$_S$ (Table 3). NSE$_{Ia}$ was zero or negative in the conventional models, whereas it varied from 0.2 to 0.7 in the variable $I_a$ models. NSE$_S$ was negative in all the models, indicating that their estimates of $S$ were poor. In the case of the conventional models this was due to using uniform $I_a$ and $S$ and thereby homogenizing the watershed. In the case of the variable $I_a$ models, this was due to their inability to model the $S$-shaped function of $S$. Based on NSE$_{Ia}$ and NSE$_S$, VIM$\lambda$ was the best model in estimating watershed heterogeneity.

Variable $I_a$ models also predicted runoff better than the conventional models in smaller rainfall events ($P < 8$ mm) as indicated by NSE$_{Q50}$ and PB$_{Q50}$. In both cases of $\lambda_i = 0.2$ and 0.5, only HRU no. 0 (Table 2) produced runoff when $P < 8$ mm. This was similar to the case of a partial source area. As CM0.2 and CM$\lambda$ predicted an $I_a > 8$ mm in both cases (Table 3), they falsely predicted zero-runoffs in all the events with $P < 8$ mm because the runoff condition ($P > I_a$) could not be overcome. Therefore, their PB$_{Q50} = 100$ in both cases, indicating a 100 % under-prediction in small events. Their NSE$_{Q50}$ was also poor with the same value in both cases. VIM$\lambda$ performed slightly better than the conventional models with 70–90 % under-predictions and with NSE$_{Q50}$ varying from −0.8 to −1.8 (Table 3). VIM$\lambda$ performed significantly better than all the other models with 30 % or less under-predictions and with NSE$_{Q50}$ varying from 0.6 to 0.9. Even though there were under-predictions, there was no false prediction of zero-runoffs for any of the events in the variable $I_a$ models.

In the models where $\lambda_W$ was calibrated (CM$\lambda$ and VIM$\lambda$), it was smaller than $\lambda_i$ (Table 3). This shows that $\lambda$ at the watershed scale tends to be smaller than that at the HRU scale in the lumped-parameter models. All the models under-predicted $I_a$ or $I_{\alpha T}$ with CM$\lambda$ being the most severe. There was also a corresponding over-prediction of $S$ or $S_{\infty}$ by all the models except for the case of $\lambda_i = 0.2$ in CM0.2. Again, the most over-prediction of $S$ occurred in CM$\lambda$. The under-prediction of $I_a$ and the corresponding over-prediction of $S$ is due to the transfer of storage from $I_a$ to $S$, which generally improves the performance in the conventional models.
6.1 Storage transfer from $I_a$ to $S$

The storage in a watershed is distributed between $I_a$ and $S$. $I_a$ is the part of the storage that does not produce runoff while being filled, whereas $S$ is the part that produces runoff while being filled. Using Eqs. (2) and (1) it can be shown that

$$S = \frac{(P - I_a)(P - I_a - Q)}{Q}.$$  \hspace{1cm} (25)

For an observed storm event, $P$ and $Q$ are known and therefore are constants in Eq. (25), so decreasing $I_a$ will increase $S$. However, the magnitude of increase in $S$ will be greater than the magnitude of decrease in $I_a$. This is illustrated by differentiating Eq. (25) and using Eq. (4) to give

$$\frac{dS}{dI_a} = -\left(1 + \frac{2S}{P - I_a}\right) \forall P > I_a.$$ \hspace{1cm} (26)

Thus, $dS/dI_a$ is always negative and less than or equal to $-1$. If $(P - I_a) \gg S$ or $S \approx 0$, then $dS/dI_a \approx -1$, implying an equal transfer in storage between $I_a$ and $S$. However, as $P$ decreases, $dS/dI_a$ becomes less than $-1$, implying that $S$ changes more rapidly than $I_a$. In other words, the relative change of magnitude in $S$ with respect to $I_a$ is large for smaller $P$, decreases with increasing $P$, and approaches unity for large values of $P$.

Storage transfer is evident when the values of $I_a$ and $S$ for the models CM0.2 and CM$\lambda$ are compared (Table 3). For the case of $\lambda_i = 0.2$, $I_a$ decreased from 19 mm in CM0.2 to 9 mm in CM$\lambda$, whereas $S$ increased from 97 to 132 mm, i.e., $dS/dI_a = -3.5$. Similarly, for the case of $\lambda_i = 0.5$, $dS/dI_a = -4.2$.

A transfer of storage from $I_a$ to $S$ improves the performance in the conventional models (i.e., CM$\lambda$ > CM0.2) because (i) a smaller $I_a$ reduces the percentage of events with falsely predicted zero-runoffs and (ii) it allows the model to mimic a variable $I_a$. Because of a larger $I_a$, CM0.2 falsely predicted zero-runoffs in 80% of the events for $\lambda_i = 0.2$ and in 85% of the events for $\lambda_i = 0.5$. In the case of CM$\lambda$, the percent of events with zero-runoffs dropped to 57% and 81%, respectively, because $I_a$ for CM$\lambda$ was smaller than CM0.2. Mimicking variable $I_a$ can be explained by considering $I_aF$ and $F$, which are the filled portions of $I_a$ and $S$, respectively. $I_aF$ and $F$ have similar functional relationships with $P$ (compare Fig. 6 to Fig. 1); i.e., they both increase with $P$ and approach a constant for large values of $P$. In the conventional CN models, there is no provision to represent $I_aF$ as a function of $P$. However, $F$ is understood to be a function of $P$ and is treated as such through Eq. (2) and Fig. 1. Therefore, by transferring the storage from $I_a$ to $S$, CM$\lambda$ uses $F$ as a surrogate for $I_aF$, thereby partly mimicking the variable nature of $I_aF$.

Storage transfer from $I_a$ to $S$ also causes a decrease in $\lambda_W$ (Table 3). Conversely, when $\lambda_W$ decreases, storage is transferred from $I_a$ to $S$. This is important because several studies (Baltas et al., 2007; D’Asaro and Grillone, 2012; Shi et al., 2009; Woodward et al., 2003) found that the optimal value of $\lambda_W$ was much less than 0.2 and even close to zero in many watersheds. This shows that there is a positive correlation between a decrease in $\lambda_W$, storage transfer from $I_a$ to $S$, and a general increase in model performance for the reasons mentioned above.

6.2 Application to real watersheds

The models were also evaluated using rainfall-runoff observations from 9 real watersheds located in different parts of the world (Santikari, 2017; Santikari and Murdoch, 2018). The models’ ability to predict the observed runoff was assessed using NSE$Q$. Results show that in all the watersheds VIMs performed better than CMs but the difference in performance, $\Delta$NSE$Q$, varied across the wa-
tersheds. Between VIMλ and CM0.2, ΔNSEQ < 0.05 in one watershed, 0.05 ≤ ΔNSEQ < 0.7 in six watersheds, and ΔNSEQ ≥ 0.7 in two watersheds. Between VIM2 and CMλ, ΔNSEQ < 0.05 in three watersheds, 0.05 ≤ ΔNSEQ < 0.1 in four watersheds, and ΔNSEQ ≥ 0.1 in two watersheds. Based on their performance, the models can be arranged from the best to the worst as VIM2 > VIMS > CMλ > CM0.2, which is consistent with results from their application to the synthetic watershed.

### 6.3 Effect of degree of heterogeneity

The degree of heterogeneity, defined as the sharpness of change in CN, Ia, or S between the HRUs, may affect the relative performance of the models. To verify this, the degree of heterogeneity of the synthetic watershed (Table 2) was increased by doubling the values of Si for HRUs 3 and 4.

<table>
<thead>
<tr>
<th>Lumped model</th>
<th>λi = 0.2</th>
<th>λi = 0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>rNSEQ</td>
<td>SEEQ</td>
</tr>
<tr>
<td>CM0.2</td>
<td>0.92</td>
<td>1.54</td>
</tr>
<tr>
<td>CMλ</td>
<td>0.95</td>
<td>0.19</td>
</tr>
<tr>
<td>VIMS</td>
<td>0.97</td>
<td>0.12</td>
</tr>
<tr>
<td>VIMλ</td>
<td>1.00</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Comparing the results (Tables 3 and 4) shows that the performance of VIMS remained nearly the same, whereas the performance of CM0.2 decreased and that of CMλ increased. The relative order of performance remained unchanged, i.e., VIMλ > VIMS > CMλ > CM0.2.

The results from real watersheds (Santikari, 2017; Santikari and Murdoch, 2018) also show that the performance of CM0.2 was poor, NSEQ < 0.25, in watersheds with a sharp change in CN. Therefore, CM0.2 is unsuitable when the degree of heterogeneity is large. CMλ performed moderately well on synthetic and real watersheds with a large degree of heterogeneity, possibly by transferring the storage (Sect. 6.1). So, CMλ is suitable for predicting overall runoff, but unreliable for predicting heterogeneity or runoff from small events. VIMS outperformed CM2 in synthetic (Table 4) as well as real watersheds (Santikari, 2017; Santikari and Murdoch, 2018) with a large degree of heterogeneity, and therefore they are more reliable.

### 6.4 Model suitability

One of the main objectives of this study was to improve the predictive ability of the CN method while maintaining its simplicity. Using the number of calibrated parameters as an indicator, the models can be arranged in the order of increasing complexity as CM0.2 (one) < CMλ (two) < VIMS = VIMλ (three). CM0.2 was the simplest but also had the poorest performance (Tables 3 and 4). Moreover, there is no justification in fixing λW at 0.2 or any other constant as its optimal value can vary from zero to one (Hawkins et al., 2008). Therefore, the usage of CM0.2 is not recommended.

CMλ predicted the overall runoff and the runoff from small events better than CM0.2. Often, the optimal λW is much smaller than 0.2 and this allows CMλ to partly mimic a variable IaF by transferring storage from Ia to S. A smaller λW also reduces the false prediction of zero-runoffs, which are completely eliminated when λW = 0. Compared to the variable Ia models, CMλ is a poor predictor of runoff and watershed heterogeneity (Table 3). However, in watersheds with negligible Ia values (or λi ≈ 0) CMλ can perform as well as the variable Ia models and therefore may be preferable because of its simplicity.

Variable Ia models show that significant improvement in the model prediction of overall runoff and heterogeneity can be achieved by using one extra parameter (Table 3). This is because the functional form of IaF (Eq. 23) is consistent with the observations (Fig. 2c and d) and the results from the theoretical analysis of heterogeneous watersheds (Eq. 16, Fig. 6, and Table 1). Using variable Ia also improved the runoff predictions in small events and eliminated the false prediction of zero-runoffs. Therefore, their application is recommended in heterogeneous watersheds with nonzero initial abstractions.

When the watershed heterogeneity is known in great detail such that the number of calibrated parameters ≤ 3, a distributed modeling approach (e.g., SWAT; Gassman et al., 2007 or Soulsi and Valianzas, 2013) may be preferable over the variable Ia models. A distributed parameter model has advantages similar to the variable Ia models over the conventional models. It would inherently account for the variation of the CN method’s parameters spatially and with P. It would also avoid the false prediction of zero-runoffs in small events because HRUs with larger CNs, which generate runoff even in small events, are explicitly considered. When the heterogeneity is unknown, however, the number of calibrated parameters (for values of CNi and ai) in a distributed model with n HRUs is 2n – 1. This number would increase further if values of λi are also calibrated. Therefore, when the number of calibrated parameters > 3, application of a variable Ia model should be considered.

---

**Table 4. Performance of the models for the cases of λi = 0.2 and 0.5, when the degree of heterogeneity in the synthetic watershed (Table 2) was increased by doubling the values of Si for HRUs 3 and 4.**
6.5 Model limitation

A strength of the models proposed in this paper is that they provide a compact way to account for the spatial variation of CN, $I_a$, or $S$ (watershed heterogeneity), but a limitation is that they do not account for the temporal variation. During dry periods, $I_a$ and $S$ increase whereas, CN decreases. The behavior is opposite during the wet periods. Changes in land cover introduce additional temporal variations. Therefore, the calibrated model parameters in this paper can be considered as temporal averages. The models may underpredict runoff during wet periods and overpredict during dry periods. A procedure to account for temporal variations using antecedent moisture is described by Santikari (2017) and Santikari and Murdoch (2018).

Another limitation of VIMs is that the CN values calculated using Eqs. (5) or (10) are incompatible with the standard CN values (NRCS, 2003; USDA, 1986) derived using CM0.2. However, this limitation is not unique to VIMs because any method, including CM$\lambda$, which involves an altered relationship between $I_a$ and $S$ (i.e., $\lambda \neq 0.2$) leads to CN values that are incompatible with those derived from CM0.2. Given that (i) CM0.2 is a poor predictor of runoff (Tables 3 and 4; Santikari, 2017; Santikari and Murdoch, 2018) and (ii) the evidence contradicts $\lambda = 0.2$ (Baltas et al., 2007; D’Asaro and Grillone, 2012; Shi et al., 2009; Woodward et al., 2003), the above-mentioned limitation is an acceptable compromise.

7 Conclusions

Watershed heterogeneity causes calculated values of $I_a$, $S$, and CN to vary with $P$. Therefore, using a single effective value of these quantities at the watershed scale can lead to systematic errors in the predictions of $Q$. This problem can be mitigated by treating $I_a$, $S$, or CN as functions of $P$. A theoretical analysis assuming spatial variation of $I_a$ led to the following conclusions.

1. Effective $I_a$ of a watershed is equal to the filled portion of the total storage in $I_a$. The total storage (called $I_{aT}$) is constant, whereas the filled portion (called $I_{aF}$) is a function of $P$ (Eq. 16). Variation of $I_{aF}$ with $P$ (Fig. 6) is similar to the variation of calculated $I_a$ (also called effective $I_a$ or $I_{aW}$) with $P$ (Fig. 2c and d). This shows that $I_{aW} = I_{aF}$, which is also supported by a distributed model using many HRUs (Eq. 21). The form of $I_{aF}$ as a function of $P$ depends on the spatial distribution of $I_a$ within a watershed (Table 1, Figs. 8 and 9).

2. $\lambda$ decreases with increasing spatial scale. Using CN$_W$, calculated as the area-weighted average of CN$_i$ values (CNs of the HRUs), and the definition of $I_a$, it can be shown that $\lambda_W < \lambda_i$ (Eqs. 8 through 11). Even when $\lambda_W$ was calibrated using CM$\lambda$, the result was $\lambda_W < \lambda_i$ (Table 3). This shows that in conventional models $\lambda$ at the watershed scale tends to be smaller than that at the HRU scale; i.e., $\lambda$ decreases with increasing spatial scale.

3. Replacing $I_a$ with $I_{aF}$ can account for heterogeneity. Heterogeneity causes the effective $I_a$ of a watershed to vary with $P$, so to account for heterogeneity variable $I_a$ models (VIMs) replace $I_a$ with $I_{aF}$, which is a function of $P$ (Fig. 6). For practical purposes, $I_{aF}$ can be treated as a quadratic function of $P$ (Eq. 23) with two free parameters $c_1$ and $c_2$ that need to be calibrated. In addition, the model also requires the calibration of either $S$ (VIMS) or $\lambda$ (VIM$\lambda$).

4. Variable $I_a$ models perform better than the conventional models. Variable $I_a$ models predict runoff and heterogeneity better than the conventional models CM0.2 ($\lambda = 0.2$) and CM$\lambda$ (calibrated $\lambda$). They also eliminate the false prediction of zero-runoffs and improve runoff predictions in small events. Based on their overall performance, the models are arranged from the best to the worst as VIMS $> \text{VIM}$ $> \text{CM}\lambda > \text{CM0.2}$.

5. Storage transfer can improve model performance. Storage transfer from $I_a$ to $S$ generally improves the model performance because the filled portions of $I_a$ and $S$, $I_{aF}$ and $F$, respectively, have similar functional relationships with $P$ (compare Figs. 6 to 1). This enables a CN model to partly mimic a variable $L_{aF}$ by using $F$ as its surrogate. Storage transfer also lowers the threshold $P$ for runoff generation, thereby reducing the false prediction of zero-runoffs. Storage transfer decreases $\lambda_W$ (Eq. 3), and this can explain why the optimal value of $\lambda_W$ from published studies is much less than 0.2 or even zero in many watersheds.

Data availability. This paper uses synthetic data, which can be generated using Table 2 and the procedure described in Sect. 5.1.
Appendix A: List of symbols

- \(a\): fractional area of the \(i\)th HRU
- \(a(I_a)\): probability density function of areal occurrence of \(I_a\)
- CM0.2: conventional curve number model with \(\lambda = 0.2\)
- CM\(\lambda\): conventional curve number model with calibrated \(\lambda\)
- CN: curve number, applicable to any spatial scale
- CN\(_i\): curve number of the \(i\)th HRU
- CN\(_T\): curve number of a watershed when \(I_{aF} = I_{aT}\)
- CN\(_W\): curve number of a watershed
- \(F\): cumulative infiltration after runoff begins
- HRU: hydrologic response unit
- \(I_a\): initial abstraction, applicable to any spatial scale
- \(I_{aF}\): areal average of the filled portion of \(I_{aT}\)
- \(I_{ai}\): initial abstraction of the \(i\)th HRU
- \(I_{aT}\): areal average of the total initial abstraction
- \(I_{aW}\): effective initial abstraction of a watershed
- \(I_{a,max}\): maximum value of \(I_a\) within a watershed
- \(\lambda\): initial abstraction ratio, applicable to any spatial scale
- \(\lambda_i\): initial abstraction ratio at HRU scale
- \(\lambda_W\): initial abstraction ratio at watershed scale
- \(m\): number of fully filled HRUs in which \(I_{uw} \neq 0\)
- \(n\): number of HRUs in which \(I_{uw} \neq 0\)
- NSE: Nash–Sutcliffe efficiency parameter
- \(P\): event rainfall
- PB: percent bias
- \(Q\): event runoff
- \(R^2\): coefficient of determination
- \(S\): potential maximum retention, applicable to any spatial scale
- \(S_i\): potential maximum retention of \(i\)th HRU
- \(S_{\infty}\): maximum value of \(S_W\), which occurs when \(P\) is infinitely large
- \(S_W\): effective potential maximum retention of a watershed
- SEE: standard error of estimate
- VIM\(\lambda\): variable initial abstraction model in which \(\lambda\) is calibrated
- VIMS: variable initial abstraction model in which \(S\) is calibrated
Author contributions. VS conceived the idea and performed the analysis. LM supervised the analysis and influenced the overall direction and content of the work. Both VS and LM wrote the paper.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. Primary funding for this study was provided by the USDA Natural Resources Conservation Service (NRCS-69-4639-1-0010) through the Changing Land Use and Environment (CLUE) project at Clemson University. Additional support was provided by the USDA Cooperative State Research, Education, and Extension Service under project number SC-1700278.

Edited by: Thomas Kjeldsen
Reviewed by: two anonymous referees

References


HydroCAD: http://www.hydrocad.net, last access: August 2018.


Santikari, V. P. and Murdoch, L. C.: Accounting for spatiotemporal variations of Curve Number using variable initial abstraction and antecedent moisture, in review, 2018.


Souls, K. X., Valiantzas, J. D., Dercas, N. and Londra, P. A.: Investigation of the direct runoff generation mechanism for the analysis of the SCS-CN method applicability to a partial area


