# Information Entropy to Measure Temporal and Spatial Complexity of Unsaturated Flow in Heterogeneous Media

David C. Mays,[1,2] Boris A. Faybishenko[1] and Stefan Finsterle[1]

19 November 2001

[1]Earth Sciences Division
Lawrence Berkeley National Laboratory

[2]Department of Civil and Environmental Engineering
University of California, Berkeley

# Abstract

Geometric heterogeneity, coupled with the nonlinear interplay of gravity, capillarity, and applied pressure gradients, results in a rich variety of flow behaviors in unsaturated fractured rocks and porous media. In this paper, we develop a procedure to measure the complexity of these behaviors using *information entropy*, a statistical quantity which indicates the unpredictability, or complexity, of a physical process. We create an empirical probability distribution function directly from the data set, then apply Shannon's definition of information entropy to quantify its complexity. As an example, we use entropy to evaluate the temporal and spatial complexity of simulated flow processes invoked by ponded infiltration into heterogeneous porous media. With an initial condition of constant saturation, we find that the entropy of saturation data increases as infiltration proceeds, while the entropy of capillary pressure data decreases as the system approaches equilibrium. Finally, we investigate the marginal value of additional data collection using randomly selected "virtual wells," which shows that computed entropy increases with additional virtual wells, but the rate of increase declines.

# 1 Introduction

There is considerable literature documenting the spatial and temporal variability in processes impacting unsaturated flow in fractured rocks and heterogeneous soils. Because the dynamics depend on a combination of conditions such as heterogeneity, moisture content, and chemistry, the resulting transient flow and transport are usually complex. Several field investigations [*Nativ et al.*, 1995; *Dahan et al.*, 2000; *Faybishenko et al.*, 2000; *Podgorney et al.*, 2000] and laboratory experiments [*Persoff and Pruess*, 1995; *Glass and Nicholl*, 1996; *Su et al.*, 1999] have highlighted this behavior.

In this paper, our goal is to use *information entropy* as a quantitative procedure to assess spatial and temporal complexity of unsaturated flow processes in heterogeneous media. We expect this method to give us insight into the underlying complexity of a variety of physical processes. Previous applications of information entropy in hydrology include the work of several authors [*Domenico*, 1973; *Chapman*, 1986; *Husain*, 1989; *Singh and Fiorentino*, 1992; *Woodbury and Ulrych*, 1996]. For this study, we define two specific objectives:

- To employ this quantitative procedure to evaluate changes in complexity for an infiltration process in heterogeneous media, and
- To investigate the marginal value of additional data collection using virtual boreholes.

# 2 Calculating Entropy from Data Sets

## 2.1 General Concept

We begin by recalling Shannon's definition [*Shannon and Weaver*, 1949]: For a variable with probability density function $f_X(x)$, the information entropy is the negative expected value of the log-probability,

$$H = -\int_{-\infty}^{\infty} f_X(x)\log(f_X(x))dx \tag{1}$$

where $H$ is the information entropy and the function log() can be taken with base 2, $e$, or 10, resulting in entropy units of bits, napiers or decibels, respectively [*Amorocho and Espildora*, 1973]. It is directly analogous to the thermodynamic definition of entropy. Like thermodynamic entropy, it explicitly depends on the enumeration of states, a point we will revisit in Section 2.4.

Information entropy measures the degree to which the probability distribution function (PDF) constructed from a data set matches the PDF corresponding to minimum information about the system. In our analysis, we take this minimum information PDF as the uniform distribution. If the data are predictable, corresponding to a peaked PDF, then the entropy is low. If the data are unpredictable, corresponding to a uniform PDF, then the entropy is high. Within this framework, the terms unpredictable, unstructured, and complex are analogous - we measure them with information entropy. For the remainder of this paper, the term "entropy" will refer to information entropy or relative information entropy, to be defined below.

## 2.2 Discrete Formulation

The definition of entropy used in Equation (1) is appropriate for variables $f_X(x)$ of a known distribution that can be fitted to data [*Chapman*, 1986; *Husain*, 1989]. In the case of laboratory or field data, measurements are discrete, representing data sets that are limited in time and space. Rather than fitting an analytical function to the data, we can establish a bin specification to construct a probability distribution directly from the data. To calculate entropy for a discrete function, we use the discrete analog of Equation (1) given by

$$H = -\sum_{i=1}^{N} P_i \log_2(P_i) \tag{2}$$

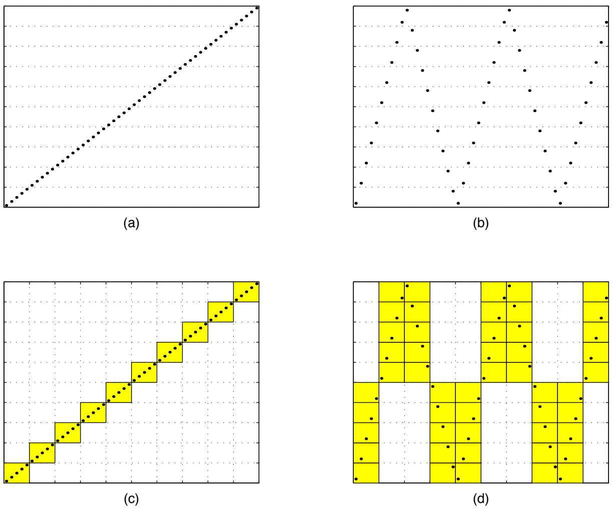


Figure 1: Illustration of entropy calculation in multiple bins. When we divide the data only along the $y$-axis, as in (a) and (b), the entropy for each example is the same. In (c) and (d), we divide the data along both the $x$- and $y$-axes, so we can differentiate the different structures in the data.

where $i$ is the bin number, $N$ is the number of populated bins, and $P_i$ is the proportion of data falling into the $i$th populated bin subject to the condition $\sum_i P_i = 1$. $N$ is determined by the size of the histogram bins used to compute $P_i$, as discussed in Section 2.4. In our calculations, we use base 2 logarithms, so entropy has units of bits.

Minimum entropy corresponds to a sharply peaked probability distribution. Maximum entropy corresponds to a uniform probability distribution. In this case, every bin is populated with an equal number of data points, such that $P_i = \frac{1}{N_b}$ for all $i$, where $N_b$ is the total number of bins. According to Equation (2) this gives $H_{max} = \log_2(N_b)$. We define relative entropy as entropy normalized by maximum entropy, $H_R = H/H_{max}$.

## 2.3 Multiple Dimensions

We generalize our definition of entropy to multiple dimensions by allowing bin divisions along each of the axes. To see why this is useful, consider the 2D sketches in Figure 1. In (a) and (b), we divide the data along the $y$-axis into rows, making a 1D bin division. For either linear or oscillating data, the measurements are uniformly distributed among 10 bins, producing relative entropy of 100%. Note these results are indistinguishable from uniformly distributed random data.

In (c) and (d), we provide a 2D bin division by dividing both the $x$- and $y$-axes into 10 bins each. Thus, in (c), we have 10 occupied bins, in which the data are uniformly distributed, producing $H = \log_2(10)$. Since there are 100 bins in all, we have $H_{max} = \log_2(100) = 2\log_2(10)$, so the relative entropy is $H_R = \frac{1}{2} = 50\%$. This is in contrast to independent, uniformly distributed random data, which would evenly fill all 100 bins and generate relative entropy of 100%. In (d), the data are uniformly distributed among 50 bins, so we have $H = \log_2(50) \approx 1.7\log_2(10)$, so the relative entropy is $H_R \approx \frac{1.7}{2} = 85\%$. This result is intuitively correct: linear data - 50% relative entropy; oscillating data - 85% relative entropy; and uniformly random data - 100% relative entropy.

When each of the dimensions in a data set is variable, the relative entropy varies from 0%, corresponding to a perfectly peaked distribution of data, to 100%, corresponding to uniformly distributed random data. However, if one of the dimensions is an independent variable, which is usually the case for time or space coordinates, then the relative entropy will fall into a smaller range. For example, if 2D data are regularly spaced on the independent axis, and there is an equal number of bins along each axis, the range of relative entropy in 2D will be $50\% \leq H_R \leq 100\%$. It would be straightforward to rescale $H_R$ such that 0% would indicate minimum relative entropy, for a specific set of independent variables in a given dimension, but instead we will continue to report $H_R$ as defined above, noting that the range of $H_R$ depends on the application.

## 2.4 Bin Selection

The calculated entropy depends on the complexity of the data set, on the dimensionality of the problem and on the choice of bins. If we have too few bins, we cannot resolve structure at small scales, and our calculated relative entropy will be artificially high. If we have too many bins, the data will be indistinguishable from a constant signal, and our calculated relative entropy will be artificially low. We can also think of the lower limit on bin size as an ergodicity requirement. Because entropy depends on the function $P_i$, a statistical description of the observed data, we need to ensure that $P_i$ properly expresses the variability of the underlying process. In the case of a time series, if we make our bins too short, we will be unable to sample the full distribution of states in the physical system, which will underestimate the variability of the process and lead to artificially low relative entropy.

For reliable entropy calculations, we use a number of bins below the point where subsequent divisions cease to differentiate the data set. This ensures we do not use too many bins. On the other hand, to ensure we do not use too few bins, our arbitrary rule is to use at least five bins in each direction, or for a $D$-dimensional data set, to use $5^D$ bins.

To evaluate changes in entropy as a function of time or space, we break the data set into segments along one of the independent variables, and then calculate the entropy for each using a constant set of bins. This "segmented entropy analysis" allows us to quantify changes in complexity according to time or depth. Examples of both are discussed below. Analysis of three synthetic example data sets (deterministic, chaotic, and stochastic) allows us to draw several conclusions about this procedure. First, if there is a sudden and significant change in the nature of the physical process, this will be reflected in the segmented entropy analysis. Second, if the data are chaotic, the entropy calculated for subsequent segments will show persistent variability. Third, the changes in entropy must be analyzed carefully, because they may "detect" changes in system dynamics resulting from random fluctuation.

# 3 Application to Ponded Infiltration Data

In this section, we employ the segmented entropy analysis described above to investigate the temporal and spatial complexity of ponded infiltration into initially dry, heterogeneous soil.

Here, we consider entropy as an integrated variable over space and time, a physical analog for the distribution of saturation ($S$) or capillary pressure ($P_c$) throughout a three-dimensional domain. First we consider entropy as a function of time, and then as a function of depth and time. Finally, we investigate the number of boreholes which would be required to fully describe the evolution of $S$ and $P_c$.

Our analysis is based on a numerical simulation of heterogeneous soil, produced by the software GSLIB [*Deutsch and Journel*, 1992], for which the transient response under ponded infiltration is calculated by the integral finite difference code TOUGH2 [*Pruess et al.*, 1999]. We consider a cubic domain, 3 m along each side, discretized into 10 cm grid blocks for a total of $30^3 = 27,000$ nodes. The permeability of each node is determined according to a bimodal distribution and a spherical variogram with a correlation length of 1 m. The soil characteristic curve for each node is specified by the van Genuchten model with $n = 2$ and mean $\frac{1}{\alpha} = 1$ kPa, with Leverett scaling such that $\alpha = \alpha_R \sqrt{k/k_R}$, where $\alpha_R$ and $k_R$ are the reference values. Our analysis is based on a single realization of such a permeability model.

To simulate ponded infiltration, we begin with a uniform saturation of $S = 1\%$ throughout the domain, and impose $S = 100\%$ along the top boundary beginning at $t = 0$. The progression of infiltration is illustrated in Figures 2 and 3. After five days (Figure 2), the infiltration front is mostly confined to the upper 50 cm of soil, except for a distinct preferential flow path near the front edge of the cube. The saturation throughout the remainder of the domain remains at the initial value of $S = 1\%$. As time progresses (Figure 3), other preferential flow paths become distinct, while the bulk infiltration front continues to move downward. This results in a greater variety of saturation values.

## 3.1 Temporal Complexity

### 3.1.1 Saturation

First we consider the time evolution of relative entropy calculated from the saturation data for the whole 3D domain (Figure 4). In accordance with the discussion in Section 2.4, we divide the data into 5 bins along each of the $x$, $y$, $z$ and $S$ directions at each time step, where $S$ is saturation in percent. Of these variables, all but $S$ are uniformly distributed, so the minimum relative entropy for the whole set is $\frac{\log_2(125)}{\log_2(625)} = \frac{3\log_2(5)}{4\log_2(5)} = 75\%$. Thus, for early time, the data set has minimum entropy. As time progresses, the system approaches a new state of entropy, determined by a combination of boundary conditions, heterogeneous permeability distribution, and soil water retention characteristics. The monotonic increase in relative entropy indicates that the PDF of saturation is becoming broader with time. The small change in relative entropy of saturation between 45 and 60 days indicates that the system has nearly arrived at a new equilibrium. In this way, the relative entropy shows us the declining rate of change in saturation.

### 3.1.2 Capillary Pressure

Figure 5 shows the entropy analysis for capillary pressure, which exhibits the opposite trend to the case of saturation. At early times, when the saturation is constant, the large variability of the imposed capillary pressures reflect the heterogeneity of permeability, because capillary

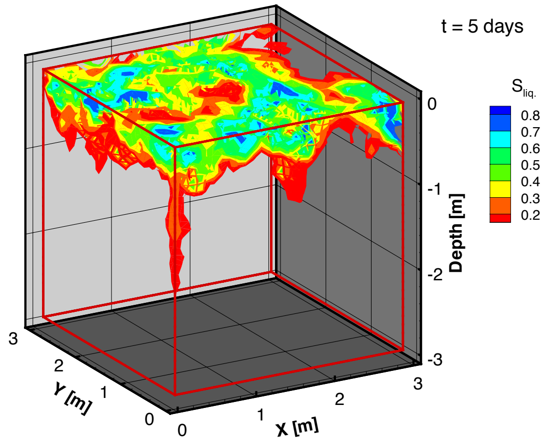


Figure 2: Saturation distribution at $t = 5$ days. The initial condition is 1% saturation throughout, with ponding at top side beginning at $t = 0$.

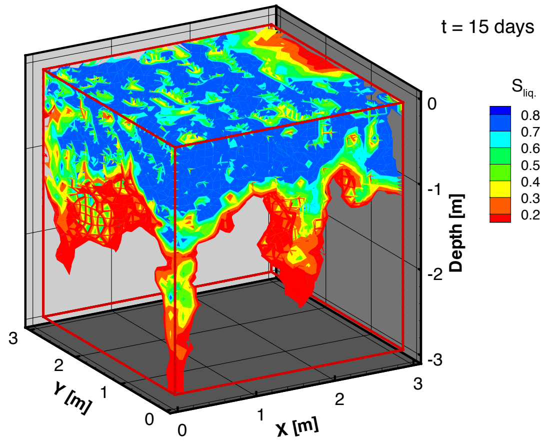


Figure 3: Saturation distribution at $t = 15$ days.

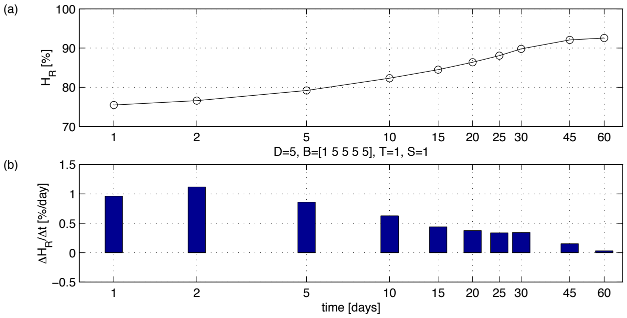


Figure 4: Temporal evolution of relative entropy of saturation for infiltration. The first data point is at $t = 1$ d. For earlier times, the relative entropy is constant at 75%.

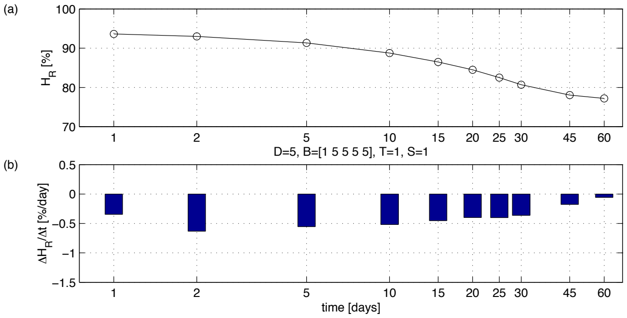


Figure 5: Temporal evolution for relative entropy of capillary pressure during infiltration. Note the behavior is opposite to that shown in Figure 4, because low saturation implies large variability in capillary pressure in heterogeneous media.

pressure is specified as a function of the permeability and saturation at each node. Indeed, the initial condition of $S = 1\%$ throughout is an artificial, non-equilibrium condition. As time progresses, the pressure distribution becomes more structured, showing low relative entropy. If the long-term equilibrium is no flow, then the pressure will display a hydrostatic profile. If the long-term equilibrium is constant infiltration, then we will observe a unit hydraulic gradient flow regime, where pressure is constant. Either way, the relative entropy will drop from its initial value as the distribution of $P_c$ becomes sharper.

### 3.1.3 Entropy Content of Water Retention

Because the entropy for both saturation and capillary pressure is based on a dimensionless probability distribution $P_i$, calculated for the same bin designation in each case, we can use their numerical values to find the entropy content of water retention, given by $\overline{H_R} = [H_R(S) + H_R(P_c)]/2 \approx$ CONSTANT, as shown in Table 1. This reflects the fact that, at each node, $S$ and $P_c$ are dependent variables, which are related by a one-to-one moisture retention curve whose parameters depend on the local value of intrinsic permeability. Therefore, an increase in $H_R(S)$ is matched by a decrease in $H_R(P_c)$ such that the information content of water retention in the system remains constant. In other words, given a complete description of $S$ or $P_c$, intrinsic permeability, and moisture retention relationships, the time evolution of

| Step Number | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $H_R(S)$ [%] | 75 | 75 | 75 | 75 | 75 | 75 | 77 | 79 | 82 | 85 | 86 | 88 | 90 | 92 | 93 |
| $H_R(P_c)$ [%] | 94 | 94 | 94 | 94 | 94 | 94 | 93 | 91 | 89 | 87 | 85 | 83 | 81 | 78 | 77 |
| $\overline{H_R}$ [%] | 84 | 84 | 84 | 84 | 84 | 85 | 85 | 85 | 86 | 86 | 86 | 85 | 85 | 85 | 85 |

Table 1: Calculation of average entropy content $\overline{H_R}$ for each time step. We attribute the slight variation in $\overline{H_R}$ to roundoff error.

the system is a strictly deterministic process which does not result in loss of information.

In a field application, we would not expect the average of $H_R(S)$ and $H_R(P_c)$ to be constant for three reasons. First, physical measurements of saturation and capillary pressure are based on different sample volumes and frequently must be taken at slightly different locations. Second, hysteretic effects mean there is no one-to-one relationship between $S$ and $P_c$. Third, the spatial distribution of permeability and moisture retention relationships are always uncertain. Thus, complete information about $S$ does not specify complete information about $P_c$ nor vice versa. Indeed, the purpose of a hydraulic test is to constrain these unknown relationships. Calculating the total entropy gives us a way to measure the dissipation of information in a physical system, which indicates how close we are to perfect characterization.

## 3.2 Spatial Complexity

### 3.2.1 Saturation

Here we consider the spatial variation of entropy resulting from the infiltration experiment. Think of breaking the 3 m cube into 10 horizontal slices, each 30 cm thick, and then calculating the relative entropy for each slice using 5 bins along the $x$, $y$ and $S$ directions and one bin in the $z$ direction. For data arranged as [$t$, $x$, $y$, $z$, $S$], we used the bin specification $B = [1, 5, 5, 1, 5]$. Because $x$ and $y$ are uniformly spaced, the relative entropy will range from $\frac{\log_2(25)}{\log_2(125)} = \frac{2\log_2(5)}{3\log_2(5)} = 67\%$ to 100%. Results for subsequent time steps are summarized in Figure 6.

Initially, when $S = 1\%$ almost everywhere, we see the minimum value of relative entropy at all depths. Then, as time progresses, and water infiltrates into the porous block, the variability of saturation values becomes greater, generating higher relative entropy. Note that the results are analogous to the transient behavior of saturation itself.

### 3.2.2 Capillary Pressure

As discussed in Section 3.1.2, the entropy of capillary pressure decreases as the infiltration system approaches equilibrium. This is illustrated in Figure 7, which shows that the infiltration process is reflected in a decline in relative entropy of capillary pressure. As in the previous figure, the limits of relative entropy in this case are 67% and 100%.

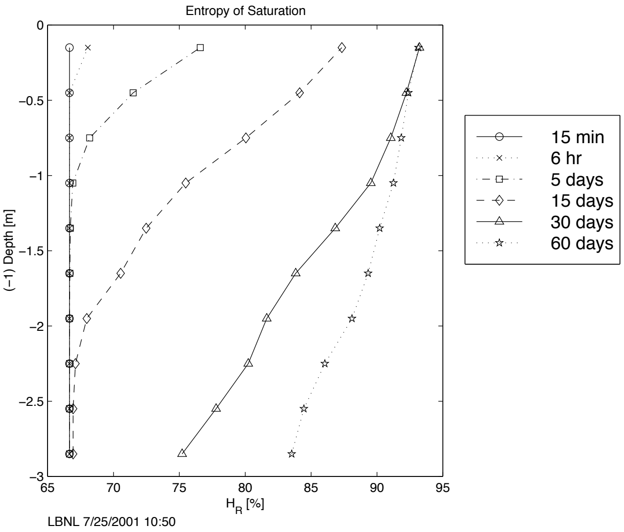


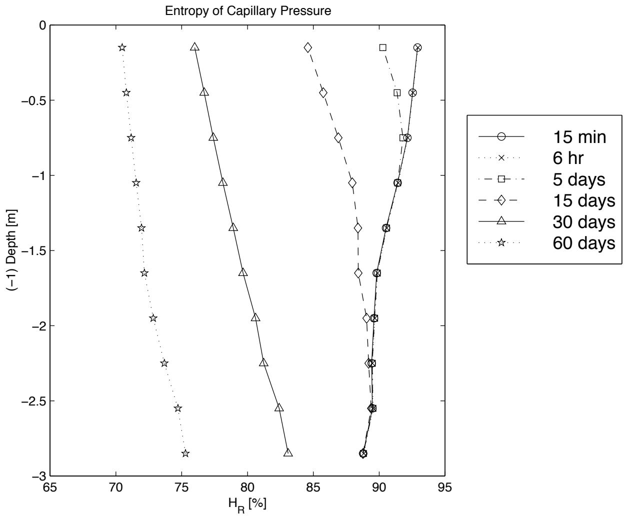
Figure 6: Entropy of saturation plotted against depth at various times. At early times, saturation is 1% throughout, so the relative entropy is minimum at 67%. At later times, saturation takes on a wide variety of values, so the relative entropy grows.

## 3.3 Uncertainty Caused by Borehole Sampling

In the previous sections we had 27,000 numerically generated data points at each time step. In reality we have only a subset of the $(t, x, y, z, S)$ data from limited measurements in boreholes. To see what happens in more realistic situations, we investigate the impact on entropy calculations when we use only a subset of the data.

This section uses the concept of "virtual wells." The simulated porous medium has an upper surface of 3 m by 3 m, which is divided into 30x30 = 900 grid blocks. A virtual well consists of the data from one of these grid blocks at a fixed time, so that $(t, x, y)$ are fixed but $(z, S)$ are variable. The simulation calculates the relative entropy for an increasing number of wells. In a first step, we may sample 5 wells, then 10 wells, then 15, and so on. Wells are randomly selected without replacement. This process allows us to track the incremental value, in terms of information gain (or uncertainty reduction) of each virtual well. Entropy, in bits, is calculated using only data from selected wells. The maximum entropy, in bits, is $H_{max} = \log_2(wB_zB_s)$, where $w$ is the number of virtual wells, $B_z$ is the number of bins in the $z$-direction, and $B_s$ is the number of bins in the $S$-direction. By assumption, $B_t = B_x = B_y = 1$.

Figure 8 shows the results for saturation data at $t = 5$ days. For the first 100 virtual wells, note how the entropy and relative entropy continue to grow as additional wells resolve a broader range of saturation values, corresponding to a more uniform PDF. After adding more virtual wells, although the entropy continues to grow, the relative entropy levels out between 400 and 500 wells. Because of the random well selection, results are slightly different for each run of the simulation, but we have not constructed confidence intervals using Monte Carlo methods since our intent is to illustrate the qualitative relationship between number of wells and calculated entropy. We observe similar results for data taken at other times



LBNL 7/25/2001 11:03

Figure 7: Entropy of capillary pressure plotted against depth at various times. Again, note the pattern is opposite to that for the saturation data.

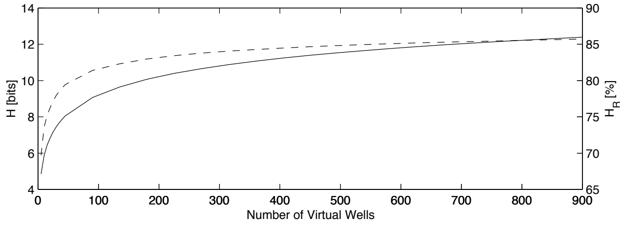


Figure 8: Entropy of saturation data at $t = 5$ days depends on the number of "virtual wells", which are added in groups of 5 up to 45 virtual wells, then in groups of 45 up to 900 virtual wells. Adding wells provides additional information, but the rate at which such information is added declines. The solid line is $H$ [bits] on the left axis; the dashed line is $H_R$ [%] on the right axis.

during the infiltration process.

From this analysis, we can draw the following conclusions. First, the calculated value of relative entropy is non-decreasing with respect to additional wells, as expected. Second, if the calculated value is going to level off, it does so only after several hundred virtual wells, which is clearly impractical. This confirms the notion that it is not possible to measure the full 3D spatial complexity of infiltration using borehole measurements and motivates the application of techniques such as cross-borehole geophysical tomography, which provide more continuous spatial information.

# 4 Summary and Conclusions

After reviewing the physical basis for complexity in subsurface flow processes and stating the definition of information entropy, we quantified entropy in a numerical simulation of infiltration. We see that infiltration results in increasing entropy, or less structure, in the saturation measurements and decreasing entropy, or more structure, in the capillary pressure measurements. This is because dynamic equilibrium depends on the pressure distribution, not saturation, so the system will progress from an artificially imposed pressure disequilibrium to a system with a complex saturation distribution needed to equilibrate pressures in the heterogeneous medium. Thus, there is a link between the level of organization of the system and the degree to which the system is approaching equilibrium.

We investigated the marginal value of additional data collection by calculating the spatial entropy of infiltration using randomly selected virtual wells. We saw that the relative entropy continues to increase with the addition of more virtual wells, but the rate of increase is declining. Results are twofold: 1) the marginal value of each well declines, and 2) it would be difficult to design a monitoring system that is both economical and able to capture the full amount of information about spatial complexity.

## 5 Acknowledgments

The first author was supported by the Office of Civilian Radioactive Waste Management Fellowship Program, administered by Oak Ridge Institute for Science and Education under a contract between the U.S. Department of Energy (DOE) and Oak Ridge Associated Universities. We appreciate partial support from Idaho National Engineering and Environmental Laboratory and from the Environmental Management Science Program, Office of Science and Technology, Office of Environmental Management, DOE under contract DE-AC03-76SF00098. We appreciate the constructive comments of R.P. Ewing of Iowa State University and T.J. Kneafsey of Lawrence Berkeley National Laboratory.

## 6 References

Amorocho, J. and B. Espildora, Entropy in the assessment of uncertainty in hydrologic systems and models, *Water Resour. Res.*, 9(6), 1511-22, 1973.

Chapman, T.G., Entropy as a measure of hydrologic data uncertainty, *Journal of Hydrology*, 85, 111-26, 1986.

Dahan, O., R. Nativ, E. Adar, B. Berkowitz, and N. Weisbrod, On fracture structure and preferential flow in unsaturated chalk, *Ground Water*, 38(3), 444-51, 2000.

Deutsch, C.V. and A.G. Journel, *GSLIB: Geostatistical Software Library and User's Guide*, New York: Oxford University Press, 1992.

Domenico, P.A., *Concepts and Models in Groundwater Hydrology.* New York: McGraw-Hill, 57-72, 1972.

Faybishenko, B.A., C. Doughty, M. Steiger, J. Long, T. Wood, J. Jacobsen, J. Lore and P. Zawislanski, Conceptual model of the geometry and physics of water flow in a fractured basalt vadose zone, *Water Resour. Res.*, 37(12), 3499-3522, 2000.

Glass, R.J. and M.J. Nicholl, Physics of gravity fingering of immiscible fluids within porous media: An overview of current understanding and selected complicating factors, *Geoderma*, 70(2-4), 133-63, 1996.

Husain, T., Hydrologic uncertainty measure and network design, *Water Resour. Res.*, 25(3), 527-34, 1989.

Nativ, R., E. Adar, O. Dahan and M. Geyh, Water recharge and solute transport through the vadose zone of fractured chalk under desert conditions, *Water Resour. Res.*, 31(2), 253-61, 1995.

Persoff, P. and K. Pruess, Two-phase flow visualization and relative permeability measurement in natural rough-walled rock fractures, *Water Resour. Res.*, 31(5), 1175-86, 1995.

Podgorney, R., T. Wood, B. Faybishenko and T. Stoops, Spatial and temporal instabilities in water flow through variably saturated fractured basalt on a one-meter scale, Geophysical Monograph No. 122, *Dynamics of Fluids in Fractured Rock*, 129-46, 2000.

Pruess, K., C. Oldenburg and G. Moridis, TOUGH2 User's Guide, Version 2.0, Lawrence Berkeley National Laboratory, Berkeley, CA, LBNL-43134, 1999.

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

Singh, V.P. and M. Fiorentino, A historical perspective of entropy application in water resources, in V.P. Singh and M. M Fiorentino (eds.), *Entropy and Energy Dissipation in Water Resources*, 21-61, Dordrecht, Netherlands: Kluwer Academic Publishers, 1992.

Su, G.W., J.T. Geller, K. Pruess, and F. Wen, Experimental studies of water seepage in intermittent flow in unsaturated, rough-walled fractures, *Water Resour. Res.*, 35(4), 1019-37, 1999.

Woodbury, A.D. and T.J. Ulrych, Minimum relative entropy inversion: Theory and application to recovering the release history of a groundwater contaminant, *Water Resour. Res.*, 32(9), 2671-81, 1996.