Geoderma 217–218 (2014) 144–160

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

What determines the strength of preferential transport in undisturbed soil under steady-state flow?

John Koestel *, Helena Jorda

Department of Soil and Environment, Swedish University of Agricultural Sciences (SLU), P.O. Box 7014, 750 07 Uppsala, Sweden

ARTICLE INFO

Article history:
Received 22 December 2012
Received in revised form 17 October 2013
Accepted 15 November 2013
Available online 22 December 2013

Keywords:
Solute transport
Preferential flow
Random forest
Pedotranfer function
Breakthrough curve
Machine learning

ABSTRACT

Preferential flow and transport has to be taken into account to successfully predict solute transport through the vadose zone. The relative 5%-arrival time of inert tracer can serve as a measure for the strength of preferential transport. As direct measurements of solute transport are not practical at large scales, soil susceptibility to preferential flow and transport has to be estimated using proxy variables. In this study we investigated how well the relative 5%-arrival time of inert tracer could be inferred from soil properties, site factors, scale and hydrologic conditions for 442 breakthrough curve experiments on undisturbed soil columns under steady state irrigation. Using a random forest as a global regression tool, we found a coefficient of determination of 0.561 in a ten-fold cross-validation. When predicting relative 5%-arrival times on a completely independent benchmark dataset of 149 experiments we obtained a still reasonable coefficient of determination of 0.336. When the soil columns had not been sampled from the same site and soil horizon, the random forest was able to rank the experiments correctly according to their relative 5%-arrival time, apart from one exception. Our study demonstrates that soil susceptibility to preferential flow and transport occurring under steady state initial and boundary conditions is to a large part predictable from proxy variables. We furthermore found evidence that the prediction performance should be considerably increased if information on the water saturation state during the experiment could be included into the random forest. An investigation of the importance of the predictors for estimating the relative 5%-arrival time yielded that the clay content was fundamental. Next important were the ratio between clay content and organic carbon, the lateral observation scale and whether the column had been slowly saturated from the bottom prior to the experiment or not. Flow rate, soil management and bulk density were found useful to further refine the predictions. A caveat has to be given that the investigated dataset includes few experiments on large columns and no experiments under natural transient hydrologic boundary conditions, since such experiments are scarce. Availability of such experiments is crucial to account for additional important preferential flow transport mechanisms caused by hydrophobicity, instabilities at infiltration fronts or funneling at soil horizon boundaries.

1. Introduction

The standard modeling approach for water flow and solute transport through soil is the Richards equation in combination with the convective-dispersion equation. It has been shown that the assumption of homogeneity underlying both equations is not met for structured soil with preferential flow, especially near saturation when a large fraction of the flow takes place in macropores (Beven and Germann, 1982; Jarvis, 2007). In such cases, dual-domain models like MACRO (Larsbo et al., 2005) provide more accurate modeling results. It is therefore necessary to decide for which soils the classical modeling approaches can be used and for which soils preferential transport has to be taken into account explicitly (Jury et al., 2011). Since flow and transport experiments are hardly practical at large scales, soil susceptibility to preferential flow and transport has to be inferred from proxy variables, such as soil properties, land use and management, scale and hydrologic conditions.

One difficulty in benchmarking such a classification scheme is that a measureable quantity or indicator must be defined to infer the strength of preferential flow. The shape of breakthrough curves (BTCs) from inert tracer experiments can be used for this purpose. An early tracer arrival and long tailing are indicators for preferential flow and transport (Brusseau and Rao, 1990). Knudby and Carrera (2005) showed that the relative 5%-arrival time of inert tracer is a reliable indicator of the presence of connected preferential flow and transport paths. Koestel et al. (2011) found that the relative 5%-arrival time was especially suitable in cases when the BTC raw data are not available and model parameters must be used to estimate early tracer arrival times.

It is known that a multitude of factors determine tracer arrival times. Among them are not only soil properties but also hydrologic initial and boundary conditions, site factors and the scale of experiment. Below we give a short overview of the known relevant variables. In a preliminary evaluation of relative 5%-arrival times on a meta-database containing 733 steady-state experiments with inert tracers collated from the peer-reviewed literature, Koestel et al. (2012) observed that a clay content of 0.08–0.09 $cm^3\ cm^{-3}$ had a threshold-like impact on early tracer

* Corresponding author.
E-mail address: john.koestel@slu.se (J. Koestel).

0016-7061/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.geoderma.2013.11.009

arrival. Among the 733 experiments, there was not a single experiment with strong preferential characteristics that was conducted on soil that contained 8% clay or less. It is noteworthy that Quisenberry et al. (1993) also identified a clay content of 8% as the most important indicator for whether preferential transport would occur in a soil, although this conclusion was based on a mere 11 samples. Apart from larger clay contents, it has also been found that for arable topsoil, larger organic carbon contents tend to reduce the strength of preferential transport (e.g. Jarvis et al., 2007). In this context, it has been recently hypothesized that the ratio between clay content and organic carbon content is especially important as an indicator of structural development and should therefore be a suitable predictor for preferential flow and transport (de Jonge et al., 2009; Dexter et al., 2008). However, this hypothesis has not yet been tested on larger datasets. In another study, Koestel et al. (2013) found a strong negative correlation between relative 5%-arrival times and bulk density, ρ (g cm^{-3}), on 65 soil columns sampled from loamy arable topsoil. The influence of clay and organic carbon content on the relative 5%-arrival times was apparently not important, but both variables showed only limited variation at this field site. Recently, Jarvis et al. (2009) presented an empirical decision tree for predicting four distinct classes of soil susceptibility to preferential macropore flow and transport. Besides soil texture and organic carbon content, the prediction scheme is also based on information on site factors like land use and tillage practices and pedological descriptors (e.g. soil horizon designations according to the world reference base (WRB) classification scheme). Jarvis et al. (2009) used the relative concentration peak arrival-time of breakthrough curve experiments with inert tracers to quantify the strength of preferential flow. They found that their decision tree explained 30% of the variability in relative peak-arrival times ($R^2 = 0.3$) when benchmarking their decision tree on a dataset of 52 BTCs collected for near-saturated or saturated conditions on undisturbed soil columns assembled from the peer-reviewed literature. A later test of a slightly modified version of the decision tree on a larger, independent dataset (N = 203) gave improved results ($R^2 = 0.44$, Jarvis et al., 2012).

The decision trees in Jarvis et al. (2009 and 2012) do not directly predict the strength of preferential transport but only soil susceptibility to preferential transport. The reason for this is that the hydrologic initial and boundary conditions are highly relevant for the generation of preferential flow and transport (Beven and Germann, 1982; Hendrickx and Flury, 2001; Jarvis, 2007; Kluitenberg and Horton, 1990). For example, it is known that dry initial conditions will be favorable for preferential flow and transport for several reasons. Firstly, under such conditions unstable infiltration fronts may develop which lead to finger flow (Ghesmat and Azaiez, 2008; Raats, 1973; Scheidegger, 1960). Secondly, soils may become water repellent when they are dry which may also lead to preferential transport (Carrick et al., 2011; Jarvis et al., 2008; Ritsema et al., 1993). Finally, less antecedent water has to be displaced when the soil is dry (Kluitenberg and Horton, 1990). Another important factor determining the occurrence or absence of preferential flow and transport is the degree of water saturation (Hendrickx and Flury, 2001). Under saturated and near-saturated conditions, strong preferential transport will be triggered in soil macropores (Jarvis, 2007; Koestel et al., 2012).

The influence of alternative experimental boundary conditions on solute transport, like water ponding instead of irrigation (e.g. Ghodrati and Jury, 1990) or the presence of a seepage face at the base (e.g. Flury et al., 1999) is often debated. Both ponding and the presence of a seepage face imply at least partially water-saturated conditions. The impacts on the observed BTCs of entrapped air (Císlerová et al., 1988; Snehota et al., 2008) or flow along artificial macropores between the soil and column walls (e.g. Bergström, 1990) have also been discussed. Many studies therefore aim to prevent air entrapment by slowly saturating the soil column from the bottom before the start of the experiment (e.g. Goncalves et al., 2001; Kjaergaard et al., 2004). In other studies, a sealant is applied between the soil and column wall (Reungsang et al., 2001; Vervoort et al., 1999). However, the sealant may then enter the soil pore system and block preferential flow paths (Vanderborght et al., 2002). Further discussions arise about whether anionic tracers like chloride and bromide which are often used in predominantly negatively charged soils as a conservative tracer can be considered as inert. It has been argued that anion exclusion may in such cases lead to different results as compared to a non-charged tracer like deuterium or tritium (Rose et al., 2009).

Finally, it is also known that the scale of the experiment exerts a strong control on the shape of the BTCs. It has been for example shown that the apparent dispersivity increases with travel distance (Gelhar et al., 1992; Roth and Hammel, 1996; Vanderborght and Vereecken, 2007) as well as with lateral observation scale (Kolenbrander, 1970; Vanderborght and Vereecken, 2007; Vanderborght et al., 2001). The former suggests imperfect lateral mixing (Flühler et al., 1996) as with perfect lateral mixing, i.e. under a convective-dispersive mixing regime, the apparent dispersivity would remain constant with travel distance. In this case, the relative 5%-arrival time should increase with travel distance as the coefficient of variation of the travel-time PDF decreases (Jury and Roth, 1990).

It has been shown that the relationships between the relative 5%-arrival time and some soil properties, site factors and experimental conditions are highly non-linear (Koestel et al., 2012). Therefore, non-linear regression techniques have to be used for predicting the relative 5%-arrival time from proxy variables. These techniques are commonly referred to as 'machine learning' methods (Alpaydin, 2004; Hastie et al., 2009). Machine learning methods have already been applied in the soil science community. A prominent example is ROSETTA, a pedotransfer framework for predicting soil hydraulic properties which is based on artificial neural networks (Schaap et al., 2001). However, many machine learning approaches including k-nearest neighbors, artificial neural networks or support vector machines, perform well when making predictions, but do not provide information on which predictors were important to arrive at the prediction (Archer and Kirnes, 2008). This is amended by a relatively new machine learning approach (Breiman, 2001) that is referred to as 'random forest' (Hastie et al., 2009; Strobl et al., 2009).

In this study we used random forests to investigate to what degree the strength of preferential flow and transport as measured by the relative 5%-arrival time of inert tracer is predictable by proxy variables such as soil properties, site factors, scale and hydrologic conditions. A second goal of this study is to quantify and rank the importance of the above discussed factors for predicting the strength of preferential transport using the relative 5%-arrival time as a proxy. The results are discussed with respect to the possibility of generalizing them to larger scales and natural climatic conditions.

2. Material and methods

2.1. Investigated dataset

We investigated a dataset of 591 BTC experiments on undisturbed soil columns with inert tracers collected from 59 peer-reviewed articles. Only flux-concentration data were considered. The dataset was divided into a training set (442 BTCs from 51 source articles) and a benchmarking set (149 BTCs from 8 source articles). The training set is a subset of the BTC meta-study published by Koestel et al. (2012) for which soil texture data as well as the applied water flow rate were known. An overview on the training set is given in the Appendix A (Table A1). The BTC experiments in the benchmarking set were added by Jorda (2013). They are summarized in Table 1.

We derived the 5%-arrival time relative to the average arrival time, $p_{0.05}(-)$, for all 591 BTCs from model parameters of the convective-dispersion equation and of the mobile-immobile version of this model. A detailed description of the calculation of $p_{0.05}$ is published in Koestel et al. (2012). Fig. 1a shows a histogram of the relative 5%-arrival times, $p_{0.05}$, for the training set. It can be seen