Advances in Geosciences, 5, 37–41, 2005
SRef-ID: 1680-7359/adgeo/2005-5-37
European Geosciences Union

Advances in Geosciences

Optimising soil-hydrological predictions using effective CART models

B. Selle and B. Huwe

Soil Physics Group, University of Bayreuth, 95440 Bayreuth, Germany

Received: 7 January 2005 – Revised: 1 August 2005 – Accepted: 1 September 2005 – Published: 16 December 2005

Abstract. There are various problems with process-based models at the landscape scale, including substantial computational requirements, a multitude of uncertain input parameters and the limited parameter identificability. Classification And Regression Trees (CART) is a recent data-based approach that is likely to yield advantages both over process-based models and simple empirical models. This non-parametric regression technique can be used to simplify process-based models by extracting key variables, which govern the process of interest at a specified scale. In other words, the model complexity can be fitted to the information content in the data. CART is applied to model spatially distributed percolation in soils using weather data and the groundwater depths specific to the site. The training data was obtained by numerical experiments with Hydrus1D. Percolation is effectively predicted using CART but the model performance is highly dependant on the available data and the boundary conditions. However, the effective CART models possess an optimal complexity that corresponds to the information content in the data and hence, are particularly suited for environmental management purposes.

1 Introduction

1.1 Motivation

The modelling of soil-hydrological processes at landscape scales is seeing an increased need for water related management purposes, such as for reducing the environmental impact of agricultural irrigation or improving water quality in drinking water reservoirs.

However, the application of process-based models at the landscape scale is difficult due to requirements of both large computational efforts and many uncertain input parameters. A common situation in soil-hydrological modelling at the plot or landscape scale is the use of too complex models and the existence of problems in the ability to properly identify model parameters, since the information content of field measurements for calibration is often limited (Vrugt et al., 2004; Schulz and Jarvis, 2004).

The non-parametric regression technique Classification And Regression Trees (CART) (Breiman et al., 1984) may be used to simplify process-based models, i.e. extracting the important information and key variables, that govern the process of interest at a specified scale. In other words, the model complexity can be fitted to the information content in the available data.

In this study we have applied CART for modelling spatially distributed percolation in soils. CART was used to simplify detailed predictions of percolation previously obtained by process-based modelling. Furthermore we have investigated, how dependant these effective CART models are on the available data and its boundary conditions. CART is used to assess the model complexity required for spatially distributed modelling of percolation. Furthermore key variables may be provided that govern the variance of soil water fluxes at the landscape scale.

1.2 Background

The relationship between model complexity and the uncertainty of predictions is illustrated in Fig. 1, where increasing the model complexity leads to a decrease of the model error (i.e. bias of the model with respect to reality) due to the resulting structural improvements in the model. The model error results from the various assumptions and simplifications that are made to make models manageable. Simultaneously, the input error (i.e. uncertainty in model parameters/ inputs) rises as a consequence of the increasing number of uncertain model inputs. Thus, the prediction error (overall uncertainty) increases at some level of model complexity since the error consists of the two error components: the model error and the input error. At some intermediate complexity

Correspondence to: B. Selle
(benny.selle@dpi.vic.gov.au)

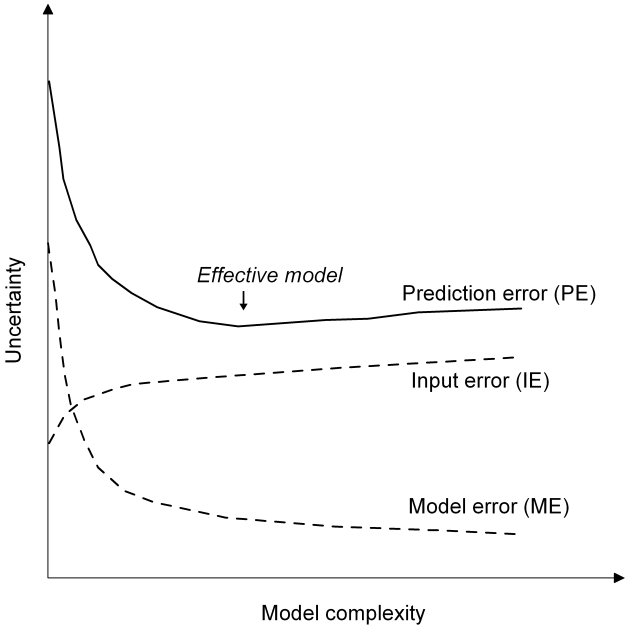

Fig. 1. Schematically: uncertainty of model predictions vs. model complexity.

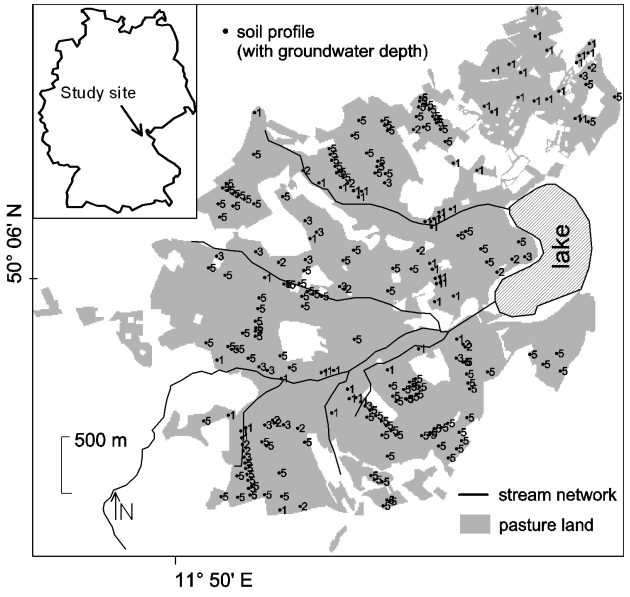

Fig. 2. The study site "Weißenstädter Becken" in Germany. Spatial distribution of the sampled sites with groundwater depth.

the minimal-prediction-error model exists, i.e. an effective model supported by the available data.

The assumed relationship between uncertainty and model complexity is known from statistics as the bias variance trade-off and can be formulated as follows. The prediction error, expressed as Mean Squared Error (MSE), can be decomposed into variance and bias

$$\overbrace{E\left[\hat{y}-y\right]^2}^{MSE} = \underbrace{E\left[\hat{y}-E\left(\hat{y}\right)\right]^2}_{\text{Variance}} + \underbrace{\left[E\left(\hat{y}\right)-E\left(y\right)\right]^2}_{\text{Bias}^2} \qquad (1)$$

where the variance term $E\left[\hat{y}-E\left(\hat{y}\right)\right]^2$ measures how much the predictions $\hat{y}$ vary when using different data samples of a population. If the variance is high, the predictions (and the model) are changing significantly from one data sample to another. In other words, it assesses the sensitivity of $\hat{y}$ to the noise on the data and thus, describes the prediction error due to uncertainties in the model input/parameters. This input error is typically increasing when using more complex models. The bias term $\left[E\left(\hat{y}\right)-E\left(y\right)\right]$ is the systematic error in the predictions and typically decreases as the model complexity grows. It is the distance of the average prediction $E\left(\hat{y}\right)$ using different data samples from the unknown true average value $E\left(y\right)$ of the population and it usually occurs with a restricted flexibility that can not properly model the observed data (model error). More generally, increasing the model complexity potentially leads to an increase of the variance term but typically the bias tends to decrease. Thus, both terms cannot be minimised at the same time and the prediction error has to be minimised in order to find a trade off between bias and variance.

However, if we would have a perfect model P but incorrect model input/parameters $f_j, ..., f_m$, $\hat{y}$ may be expressed as

$$\hat{y} = P\left(f_j, ..., f_m\right) \qquad (2)$$

The MSE may be approximated by the Taylor method supposing that the model inputs/parameters are uncorrelated:

$$MSE = E\left(\hat{y}-y\right)^2 \approx \sum_{j=1}^{m} E\left(e_j\right)^2 \left(\frac{\partial P}{\partial f_j}\right)^2 \qquad (3)$$

where e_j is the error of the jth model input/parameter and $\partial P / \partial f_j$ is the partial derivative of P with respect to f_j, i.e. the sensitivity of the model output $\hat{y}$ with respect to changes in f_j. If the MSE is, as usually, composed from two components: model error (ME) and input error (IE)

$$MSE = IE + ME \qquad (4)$$

and IE can be estimated from (3), the MSE can be estimated predicting on a independent validation sample and the model error ME can be calculated by subtracting the former from the latter. The input error IE typically increases with more and more uncertain input parameters, respectively. However, the different error components are hard to assess since no perfect model exists in practise.

2 Material and methods

2.1 General modelling approach

The CART model was derived from a training set obtained from different Hydrus1D runs (Simunek et al., 1998) which simulated water flow in one-dimensional soil columns. Hydrus1D was applied to a representative sample of soils and

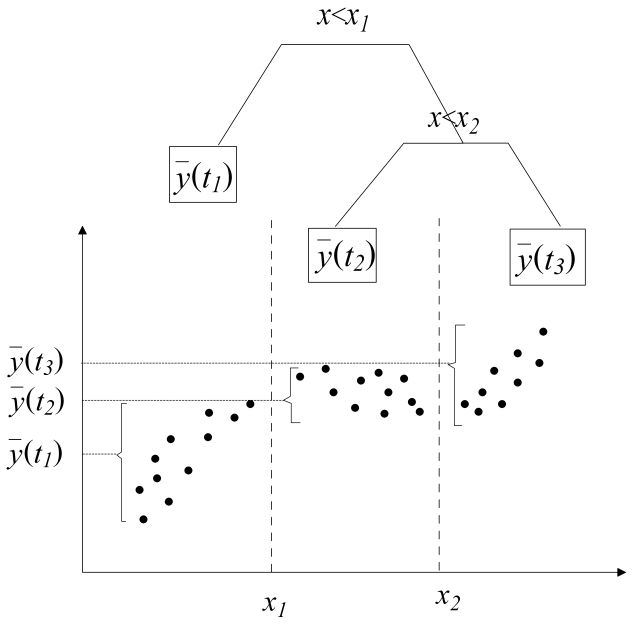

Fig. 3. Principles of a regression tree with response variable y, the predictive variable x and terminal nodes $t_1 ... t_3$.

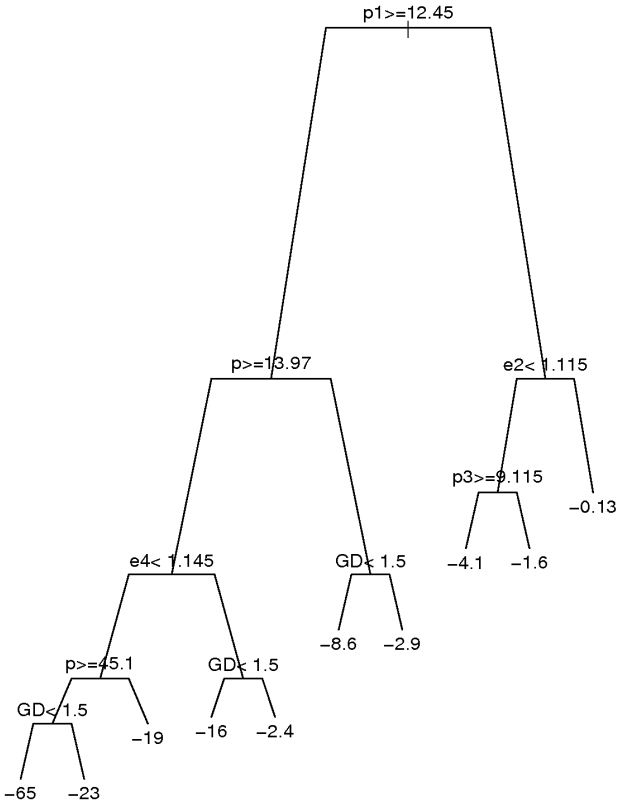

Fig. 4. Sub-tree of the CART model used to predict percolation from precipitation p, potential evapotranspiration e and groundwater depth GD. Numbers behind the predictive variables denote lag times.

site conditions within the study site "Weißenstädter Becken" and for the period 1998–2001. The daily percolation q at time t and place x was modelled with CART using weather data (precipitation p, potential evapotranspiration e) at different lag times (maximum time lag of 4 days) and groundwater depths specific to the site GD_x.

$$q_{x,t} = f\left(e_{x,t-i},\, p_{x,t-i},\, GD_x\right) \text{ for } i=0,\, 1...,\, 4 \qquad (5)$$

In the CART-model we considered the landscape as a number of non-interacting soil columns.

2.2 Site characteristics

The study site "Weißenstädter Becken" is an approx. 10 km^2 plateau with a subdued relief at approx. 600–700 m a.s.l. in the "Fichtelgebirge" (Northern Bavaria, Germany) (Fig. 2). The primary land use is as pastureland. The water balance is positive as a result of the high annual precipitation (ranging from 900 to 1000 mm) and also due to the low mean annual temperature which ranges from 5° to 6°C. Cambisols occur most frequently in the area, having developed on periglacial muds from the acid bed rock parent material (granite, mica slate) while Histosols and Gleysols occur only in consequence of the topography. The database is very detailed due to the large amount of previously completed studies (weather data, groundwater depth, etc.).

2.3 Determination of the training set

First, 242 one meter deep soil profiles were obtained by soil augering. Their bulk density and soil texture were examined for each horizon using field identification methods. Second, the soil-hydraulic parameters according to Van Genuchten (1980) were obtained using these parameters and the pedo-transfer functions from Rosetta (Schaap et al., 1998). Third, percolation dynamics 1998–2001 (daily time steps) for each soil profile were computed using Hydrus1D. The upper boundary condition (weather data) was uniform for all simulated soil profiles. Groundwater depth was taken for the lower boundary condition. Lastly, percolation was calculated for each profile at an uniform depth of 1 m.

2.4 Classification and regression trees

The CART method (Classification And Regression Trees) is a recent non-parametric statistical technique, which can be used to solve both regression and classification problems. CART is so termed because the model can be displayed in the form of a binary decision tree (Fig. 3). The decision tree is obtained by recursive data partitioning, thereby splitting the data set into increasingly smaller subsets based on the predictive variables. Different types of predictive variables (categorical and continuous) can be integrated into the model. The predictions of the CART model are provided by either the average of the response variable (regression) or the most frequent class of the targets (classification) at the terminal nodes, depending on the type of problem to be solved. The theoretical background of CART is described in

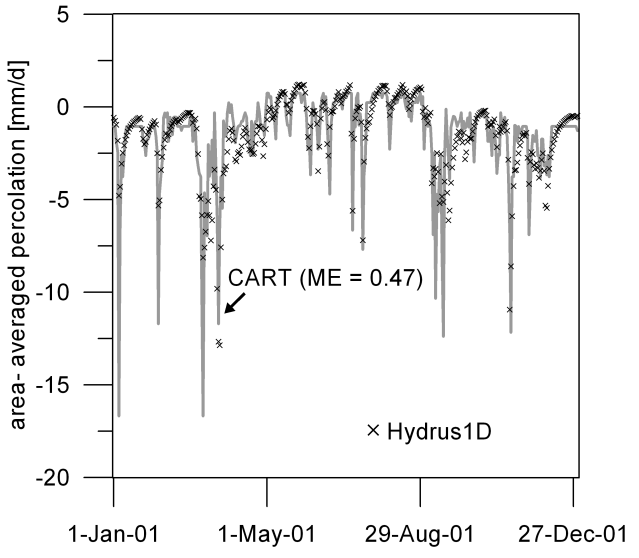

Fig. 5. Prediction of the area-averaged percolation 2001 using Hydrus1D (crosses) and CART (shaded line). ME: Modelling Efficiency (Nash and Sutcliffe, 1970).

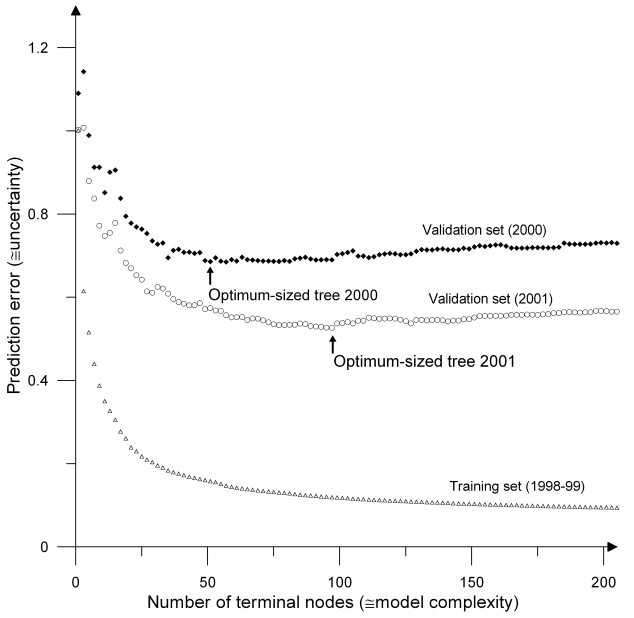

Fig. 6. Prediction errors (uncertainty of model predictions) vs. model complexity for the CART model. Prediction error was evaluated for the training data (1998–1999) and different validation samples (2000 and 2001).

Table 1. Dryness index (potential evapotranspiration/ precipitation ratio) for different weather data sets.

Data	1998–1999	2000	2001
Dryness index	0.47	0.65	0.49

Ripley (1995) and in the monograph of Breiman et al. (1984). The CART technique was conducted using the *rpart* package (Therneau and Atkinson, 1997) from the R-environment[1], which is programmed according to the algorithms suggested by Breiman et al. (1984).

The following steps must be taken in order to obtain the optimum sized tree, which is the goal of the CART method. Initially a maximum sized tree is generated through continuous binary splitting of the data, where splits are inequality conditions on the predictive variables. After which the maximum-sized tree is repeatedly pruned back to increasingly smaller trees until only the root node remains. The optimum sized tree, i.e. the optimum number of terminal nodes, is chosen from the sequence of sub-trees as the one that performs best on a validation sample and thus provides the best generalisation characteristics.

3 Results and discussion

CART was able to explain more than 40% of the spatial and temporal percolation variance using only precipitation, potential evapotranspiration and groundwater depths specific to the site (Fig. 4).

The area-averaged percolation, estimated using the arithmetic mean of all soil profiles, is satisfactorily predicted by CART. This can be visually evaluated from Fig. 5.

The optimum-sized tree is selected as the sub-tree that performs best on a validation set (Fig. 6). The model performance is assessed by the Mean Squared Error (MSE) of the root node divided by the MSE of each sub-tree. Thus, the model performance measures the proportion of the variance explained by the different-sized sub-trees.

It is interesting to note, that the optimum-sized CART-model depends on the validation data used, i.e. there are different minima and uncertainties associated with the validation sets 2000 and 2001, respectively (Fig. 6). Two theoretical reasons may explain this differences (Fig. 7).

First, the input errors may be different for 2000 and 2001 (Fig. 7, top), e.g. due to distinct measurement errors concerning the inputs precipitation, potential evapotranspiration and groundwater depths. However, the input error would be relevant only if measurements of percolation instead of the Hydrus1D calculations were used to determine the prediction error.

Second, the model errors may vary between 2000 and 2001 (Fig. 7, bottom). It may be assumed that the model structure depends on the climatic boundary conditions. Thus, climatic conditions 2001 are more similar to the conditions in the training period 1998–1999 than in the validation period 2000. This assumption can be confirmed by measures of climatic wetness (dryness index) for the different weather data (Table 1).

[1]R is a language and environment for statistical computing and graphics. It is an Open Source system written by a team of volunteers (www.r-project.org).

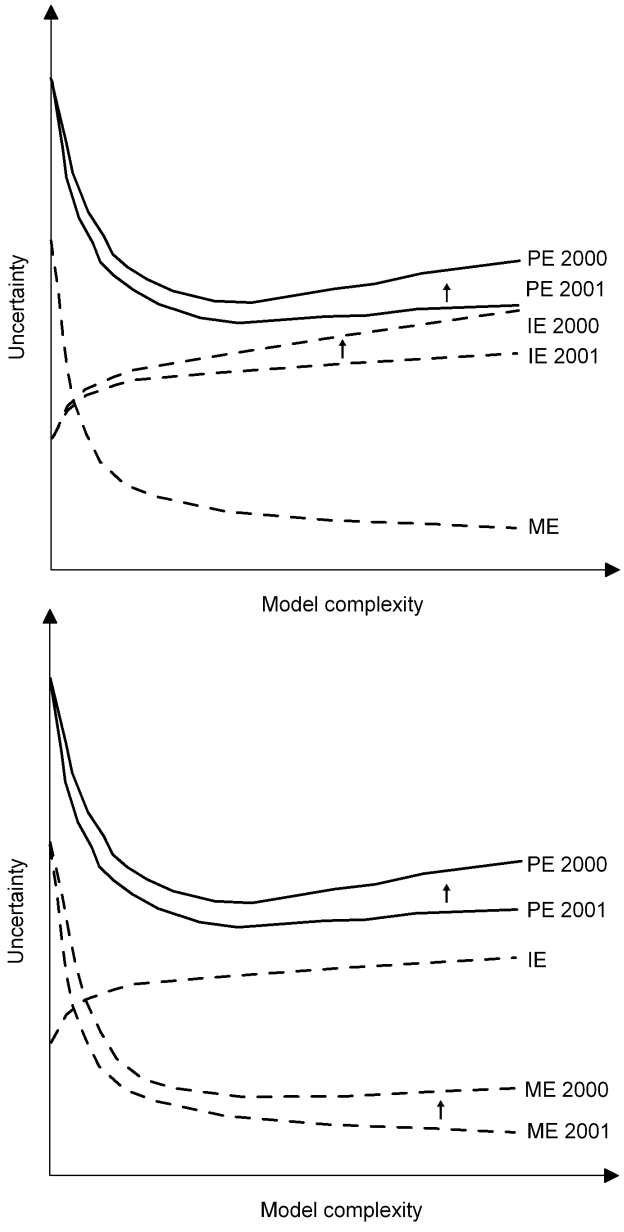

Fig. 7. Two reasons for obtaining different prediction errors when using different validation sets.

4 Conclusions

Percolation is effectively predicted with CART using weather data and the groundwater depths but the model performance is not robust with respect to changes in the data samples and the climatic boundary conditions. Therefore, CART is likely to yield satisfactory predictions only inside the range of variations in boundary conditions of the training sample (interpolation), while it seems not to be an appropriate method for extrapolation outside this range.

However, we have demonstrated, how soil hydrologic predictions can be optimised by choosing an appropriate model complexity with CART. We hypothesise, that these CART models may result in better predictions than calibrated process-based models, if the number of inputs and model parameters is significantly reduced (see Eq. 3). Furthermore, the effective models are particularly suited for environmental management purposes due to their simplicity, transparency and easy manageability On the other hand, process-based models describe the whole system providing predictions for more than just one process (e.g. percolation and actual evapotranspiration). Thus, one needs more variables and parameters to begin with, yet one also obtains more information out of the model. However, to obtain a detailed process understanding and for scientific purposes there appears to be no better alternative to process-based modelling whereas for practical applications effective models may provide a more appropriate process description.

Acknowledgements. The authors are grateful to the German Ministry of Science and Education for the financial support (BMBF 0339476 D). C. Tarn proof read the paper.

Edited by: P. Krause, K. Bongartz, and W.-A. Flügel
Reviewed by: anonymous referees

References

Breimann, L., Friedmann, J. H., Olshen, R. A., and Stone, C. J.: Classification and Regression Trees, Pacific Grove, Wadsworth, 385 p., 1984.

Jansen, M. J. W.: Prediction error through modelling concepts and uncertainty from basic data, Nutrient Cycling in Agroecosystems, 50, 247–253, 1998.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models. Part I. A discussion of principles, J. Hydrol., 10, 282–290, 1970.

Schaap, M. G., Leij, F. J., and Van Genuchten, M. T.: Neural network analysis for hierarchical prediction of soil water retention and saturated hydraulic conductivity, Soil Sci. Soc. Am. J., 62, 847–855, 1998.

Simunek, J., Huang, K., Sejna, M., and Van Genuchten, M. Th.: Hydrus1D, U.S. Salinity Laboratory, USDA/ARS, Riverside, Version 7.0., http://typhoon.mines.edu/software/igwmcsoft/, 2004.

Schulz, K. and Jarvis, A. J.: Environmental and biological controls on the seasonal variations in latent heat fluxes derived from flux data for three forest sites, Water Resour. Res., 40, W12501, doi:10.1029/2004WR003155, 2004.

Therneau, T. M. and Atkinson, E. J.: An introduction to recursive partitioning using the RPART routines, Department of Health Science Research, Mayo Clinic, Rochester, Technical Report Series No. 61, 52 p., 1997.

Van Genuchten, M. T.: A closed-form equation for predicting the hydraulic conductivity of unsaturated soils, Soil Sci. Soc. Am. J., 44, 892–898, 1980.

Venables, W. N. and Ripley B. D.: Modern Applied Statistics with S, Springer-Verlag, New York, Berlin and Heidelberg, 495 p., 2003.

Vrugt, J. A., Schoups, G., Hopmans, J. W., Young, C., and Wallender, W. W.: Inverse modeling of large-scale spatially distributed vadose zone properties, Water Resour. Res., 40, W06503, doi:10.1029/2003WR002706, 2004.