Unrepresented model errors – effect on estimated soil hydraulic material properties

S. Jaumann[1,2] and K. Roth[1,3]

[1]Institute of Environmental Physics, Heidelberg University, Im Neuenheimer Feld 229, 69120 Heidelberg, Germany
[2]HGSMathComp, Heidelberg University, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany
[3]Interdisciplinary Center for Scientific Computing, Heidelberg University, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany

Correspondence to: S. Jaumann (stefan.jaumann@iup.uni-heidelberg.de)

Abstract. Unrepresented model errors influence the estimation of effective soil hydraulic material properties. As the required model complexity for a consistent description of the measurement data is application–dependent and unknown a priori, we implemented a structural error analysis based on the inversion of increasingly complex models. We show that the method can indicate unrepresented model errors and quantify their effects on the resulting materials properties. To this end, a complicated 2D subsurface architecture (ASSESS) was forced with a fluctuating groundwater table while Time Domain Reflectometry (TDR) and hydraulic potential measurement devices monitored the hydraulic state. In this work, we analyze the quantitative effect of unrepresented (i) sensor position uncertainty, (ii) small scale-heterogeneity, and (iii) 2D flow phenomena on estimated soil hydraulic material properties with a 1D and a 2D study. The results of these studies demonstrate three main points: (i) The fewer sensors are available per material, the larger is the effect of unrepresented model errors on the resulting material properties. (ii) The 1D study is likely to yield biased parameters due to unrepresented lateral flow. (iii) Representing and estimating sensor positions as well as small–scale heterogeneity decreased the mean absolute error of the water content data by more than a factor of 2 to 0.004.

1 Introduction

Soil hydraulic material properties are essential to advance quantitative understanding of soil water dynamics. Despite decades of research, direct identification of these properties is time–consuming and near to impossible at larger scales. Therefore, indirect identification methods, such as inversion (Hopmans et al., 2002; Vrugt et al., 2008), have been successfully applied to evaluate experiments starting from lab–scale (e.g., Parker et al., 1985; Van Dam et al., 1994; Šimůnek et al., 1998; Schneider et al., 2006) up to field–scale studies (e.g., Wollschläger et al., 2009; Huisman et al., 2010). Due to the multi-scale heterogeneity of the soil hydraulic material properties (Nielsen et al., 1973; Gelhar, 1986; Cushman, 1990; Vogel and Roth, 2003), effective material properties have to be identified directly at the scale of interest. Yet, most studies focus on 1D subsurface architectures with homogeneous layers, e.g., Abbaspour et al. (2000); Ritter et al. (2003); Mertens et al. (2006); Wöhling et al. (2008); Wollschläger et al. (2009). Only a few studies, e.g., Abbasi et al. (2004); Palla et al. (2009); Huisman et al. (2010), estimate material properties of effectively 2d subsurface architectures. Abbasi et al. (2004) conducted an irrigation experiment

to estimate soil hydraulic and solute transport properties for a 2D furrow architecture. Palla et al. (2009) estimated effective soil hydraulic properties for a 2D layered coarse grained green roof based on hydrographs. Huisman et al. (2010) estimated soil hydraulic properties of a homogeneous dike exploiting flat wire Time Domain Reflectometry (TDR) and electrical resistance tomography (ERT) data recorded during a fluctuating groundwater table experiment. With increasing computational power in recent years, 1D subsurface architectures were analyzed with ensemble–based parameter estimation methods reaching from Markov Chain Monte Carlo (MCMC) (e.g., Vrugt et al., 2008; Scharnagl et al., 2011; Wöhling and Vrugt, 2011) and data assimilation (e.g., Wu and Margulis, 2011; Li and Ren, 2011; Erdal et al., 2014) to data driven modeling (e.g., Over et al., 2015).

Most of these studies describe the given data with models chosen upfront with restricted complexity and a minimum number of parameters. If the models are too simple, critical uncertainties and processes may be neglected, leading to suboptimal results. If the models are too complex, the resulting material properties are likely to be application–dependent. In general, the required model complexity is unknown a priori (Vereecken et al., 2015). Quantitative learning about complicated systems is an iterative process (Gupta et al., 2008; Box et al., 2015). It starts from the current understanding of the system that is represented with a model (Clark et al., 2011; Gupta et al., 2012). The optimal experimental design is then based on the model and the resulting data are, figuratively speaking, answer of reality to the questions asked through the experiment. Disagreement between the model and the data reveals incorrect understanding of the system. Consequently, the concepts, decisions, and hypotheses integrated into the model (including evaluation procedures of the data) and the data themselves are revised. If the model predicts the data accurately and precisely enough, the research objectives are expanded, such that the data cover a larger part of the state space. Ultimately, this iterative procedure leads to data covering the whole state space and a statistical model–data mismatch corresponding to the data error model. In general, such data are not available and the application merely requires a limited accuracy and precision. Hence, determining the sufficient complexity of the model and the data for the required accuracy and precision is the crux.

This problem can be quantified with a Bayesian total error analysis (BATEA) (Kavetski et al., 2002, 2006) investigating the total uncertainty space which includes uncertainty in the observed input and responses as well as uncertainty in the model hypothesis. However, this analysis is computationally intensive if the number of uncertainties is large and required models may not be available, e.g., for hysteresis. For instance, Bauser et al. (2016) categorized the uncertainties a priori and estimated the most important ones along with effective material properties using an Ensemble Kalman Filter (EnKF) aiming for a consistent representation of reality. The temporal fluctuation of the estimated hydraulic parameters was used to identify a situation in which the representation of the dynamics is inconsistent. Hence, measurement data acquired during this period of time were merely used for state estimation and excluded from parameter estimation to prevent the incorporation of uncertainties in the dynamics into the estimated parameters.

In this work, we change the perspective and associate the model with our quantitative understanding of reality that is tested against the given measurement data. To analyze the required model complexity, we prescribe temporally constant material properties, calculate the maximum likelihood of increasingly complex models and analyze the corresponding structural model–data mismatch. We show that this structural error analysis indicates limitations of these models and quantifies the effect of the re-

Figure 1. View of ASSESS site with tensiometer access tube, weatherstation, and groundwater well along the left boundary. The jump in color reveals different sands that crop out at the surface (figure adapted from Jaumann (2012)).

spective unrepresented model errors on the material properties. Specifically, we analyze measurement data acquired at the test site (ASSESS) while it as forced with a fluctuating groundwater table which ensures a high dynamical range of the hydraulic state. We setup a basic representation accounting for uncertainties of the hydraulic material properties and the forcing. Following an uncertainty analysis, we additionally estimate the sensor position and small–scale heterogeneity. These increasingly complex models are applied to (i) three 1D profiles in ASSESS with an increasing number of sensors per material and (ii) the full 2D profile to additionally analyze the implications of the restriction to a 1D subsurface architecture and to few sensors per material.

2 Methods

2.1 ASSESS

The approximately $2\ \mathrm{m} \times 20\ \mathrm{m} \times 4\ \mathrm{m}$ large test site *ASSESS* (Fig. 1) is located near Heidelberg, Germany, and consists of three different kinds of sand (material A, B, and C) which are arranged to an effective 2D subsurface architecture (Fig. 2). The grain size distributions of these materials are presented in Table 1. A geotextile separates the sand from an approximately $0.1\ \mathrm{m}$ thick gravel layer below, which ensures a rapid water pressure distribution and connects a groundwater well with the rest of the test site. Below this gravel layer, a concrete layer confines the site. As the test site is built into a former fodder–silo, a concrete L–element serves as additional wall. In order to stabilize the material during the construction, it was compacted. Additional to the compaction interfaces shown in Fig. 2, Ground Penetrating Radar (GPR) measurements indicate even more compaction interfaces (Klenk et al., 2015, Fig. 1b and 6).

The test site is equipped with a weatherstation, a tensiometer (UMS T4-191), and 32 soil temperature and TDR sensors. Each TDR sensor has three cylindrical rods (length: $0.20\ \mathrm{m}$, diameter: $0.004\ \mathrm{m}$) which are separated by $0.03\ \mathrm{m}$. They are operated by a Campbell Scientific TDR100.

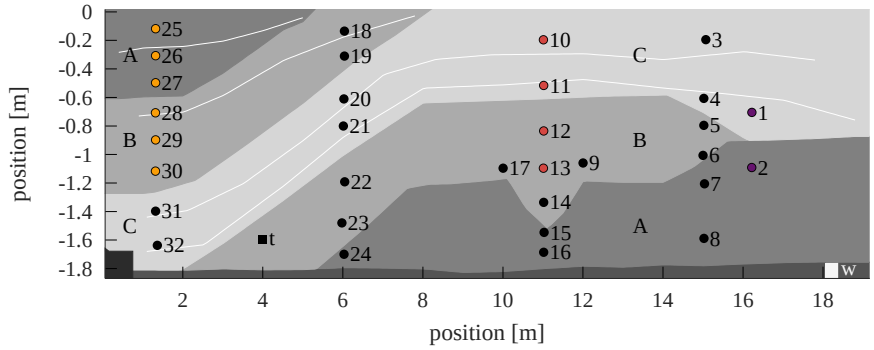

Figure 2. ASSESS features an effective 2D architecture with three different kinds of sand (A, B, and C). The hydraulic state can be manipulated with a groundwater well (white square, at 18.2 m) and is automatically monitored with 32 TDR sensors (dots) and one tensiometer (black square, at 4.0 m). The color of the dots associates some of the TDR sensors with different cases of the 1D study discussed in Sect. 3.1. The gravel layer at the bottom of the site ensures a rapid water pressure distribution over the site. An L–element (black polygon, at 0.4 m) and compaction interfaces (white lines) were introduced during the construction. Note the different scales on the horizontal and the vertical axis.

2.2 Representation

For representing the soil water dynamics in ASSESS during the experiment, we follow the lines presented by Bauser et al. (2016) and define the *representation of a system* as a set consisting of: dynamics (mathematical description), subscale physics (material properties), forcing (superscale physics), and states.

The representation of the hydraulic system also comprises its implementation. In order to separate the more general theoretical considerations from the application–dependent details, these are not directly given in this section but are gathered in the appendix A1.

2.2.1 Dynamics

The Richards equation (Richards, 1931) is the standard model to describe soil water dynamics

$$\partial_t \theta - \nabla \cdot [K(\theta)[\nabla h_\mathrm{m}(\theta) - \boldsymbol{e}_z]] = 0, \tag{1}$$

with the volumetric water content θ $[-]$, matric head h_m $[\mathrm{m}]$, time t $[\mathrm{s}]$, unit vector in z-direction $\boldsymbol{e}_z$ indicating the direction of gravity, soil water characteristic $\theta(h_\mathrm{m})$, and hydraulic conductivity function $K(\theta)$. The material properties $\theta(h_\mathrm{m})$ and $K(\theta)$ are required to solve this partial differential equation. Generally, these material properties are non–linear and vary over many orders of magnitude.

2.2.2 Subscale physics

We choose the Brooks–Corey parameterization (Brooks and Corey, 1966) for the soil water characteristic $\theta(h_\mathrm{m})$, since it has been found to describe the materials in ASSESS well (Dagenbach et al., 2013). This parameterization has four parameters: The saturated water content θ_s $[-]$, the residual water content θ_r $[-]$, a scaling parameter h_0 $[\mathrm{m}]$ related to the air entry pressure ($h_0 < 0$ m), and a shape parameter λ $[-]$ related to the pore size distribution ($\lambda > 0$). In general, $\theta(h_\mathrm{m})$ shows hysteretic behavior (Topp and Miller, 1966). Neglecting hysteresis, the parameterization may be inverted for $\theta_\mathrm{r} \leq \theta \leq \theta_\mathrm{s}$. This leads to

$$h_\mathrm{m}(\theta) = h_0 \left(\frac{\theta - \theta_\mathrm{r}}{\theta_\mathrm{s} - \theta_\mathrm{r}} \right)^{-1/\lambda}. \tag{2}$$

Inserting the Brooks–Corey parameterization into the hydraulic conductivity model of Mualem (1976) yields the parameterization

$$K(\theta) = K_\mathrm{s} \left(\frac{\theta - \theta_\mathrm{r}}{\theta_\mathrm{s} - \theta_\mathrm{r}} \right)^{\tau+2+2/\lambda} \tag{3}$$

for the hydraulic conductivity function where K_s $[\mathrm{m\,s^{-1}}]$ is the saturated hydraulic conductivity and τ $[-]$ a heuristic turtuosity factor.

Small–scale heterogeneities, i.e. the texture of the porous medium, can be represented with Miller scaling if the pore spaces at any two points are assumed geometrically similar (Miller and Miller, 1956). Scaling the macroscopic reference state $h^*_\mathrm{m}(\theta)$, $K^*(\theta)$ with a local ratio of characteristic lengths ξ $[-]$, leads to locally scaled material functions (Roth, 1995):

$$h_\mathrm{m}(\theta) = h^*_\mathrm{m}(\theta) \cdot \xi, \quad K(\theta) = K^*(\theta)/\xi^2. \tag{4}$$

2.2.3 Forcing

The hydraulic state was forced with a fluctuating groundwater table by pumping water in or out of a groundwater well. The experiment is arranged in three different phases: (i) initial drainage phase, (ii) multistep imbibition phase, and (iii) multistep drainage phase. The detailed forcing is presented in Table 2. Throughout the forcing, equilibration steps were included in between, such that the relaxation of the capillary fringe happened within the measurement volume of the TDR sensors where possible. We neglect evaporation in the following, since the experiment took place at the end of November and the weather was cloudy with 2–7 °C air temperature. The last precipitation was measured approximately 10 days before the experiment.

2.2.4 State

The hydraulic state was monitored in particular with hydraulic potential and water content measurements during the experiment. The hydraulic potential was assessed via the position of the fluctuating groundwater table. This position was measured (i) manually in the groundwater well and (ii) automatically with the tensiometer (Fig. 3). The gradient between the hydraulic potential in the groundwater well and the hydraulic potential in the test site drives the water flux. The largest part of this gradient equilibrates approximately within 5 minutes. Afterwards, the position of the groundwater table still changes which is due

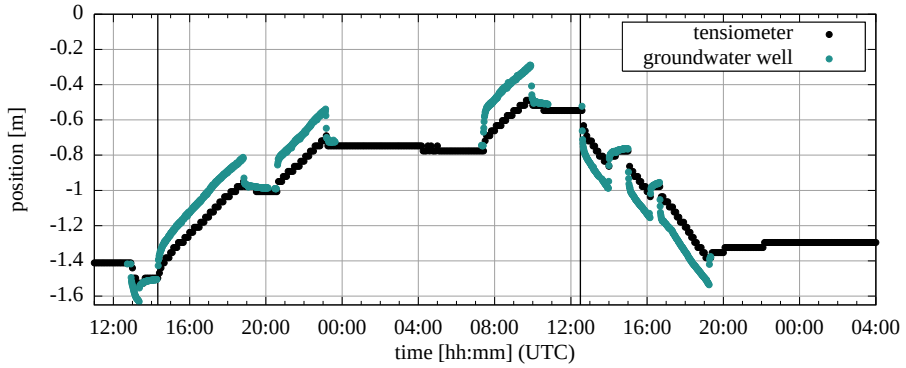

Figure 3. The position of the groundwater table was measured manually in the groundwater well and automatically with the tensiometer (Fig. 2) during three different phases (initial drainage, multistep imbibition, and multistep drainage – separated by the vertical black lines in the figure) of the experiment. Note that the discrete measurement steps reflect the resolution of the tensiometer.

to the long-term equilibration of the hydraulic state.

The water content data is based on measured TDR traces which yield the relative permittivity of the soil ε_b (Sect. A1.3). This permittivity is converted to water content θ using the Complex Refractive Index Model (CRIM) (Birchak et al., 1974):

$$\varepsilon_\mathrm{b}(\theta,T,\phi)^\alpha = \theta \cdot \varepsilon_\mathrm{w}(T)^\alpha + (\phi-\theta)\cdot\varepsilon_\mathrm{a}^\alpha + (1-\phi)\cdot\varepsilon_\mathrm{s}^\alpha, \tag{5}$$

with the geometry parameter $\alpha = 0.5$. In order to apply the CRIM, the porosity ϕ, the relative permittivity of water ε_w, the relative permittivity of air ε_a, and the relative permittivity of the soil matrix ε_s have to be known. The relative permittivity of air ε_a was set to 1.0. Assuming that the sand matrix consists mainly of quartz (SiO_2) grains, the relative permittivity of the soil matrix ε_s was set to 5.0 (Carmichael, 1989). Core samples of the materials A, B, and C yielded the porosity 0.41, 0.36, and 0.38, respectively. These values will be assumed for the saturated water content θ_s of the respective materials in the remainder of this paper. Following Kaatze (1989), we parameterize the dependency of the relative permittivity of water ε_w on the soil temperature T [°C] with

$$\varepsilon_\mathrm{w}(T) = 10.0^{1.94404 - T\cdot 1.991\cdot 10^{-3}} \tag{6}$$

and use soil temperature measurements near each TDR sensor to determine the according ε_w.

The evaluated water content data of those TDR sensors that were desaturated during the experiment are displayed in Fig. 4. The data show that the experiment is sensitive to complicated flow phenomena. The measured water content increases fast during the imbibition steps as the groundwater table reaches the TDR sensor because of the narrow transition zone of sandy materials during imbibition (Dagenbach et al., 2013; Klenk et al., 2015) and the small measurement volume of the TDR sensors (Robinson et al., 2003). During the equilibration phases, for example after the last drainage phase (19:15), the measured water content in the unsaturated material either decreases (e.g., sensor 27) or increases (e.g., sensor 2), depending on the hydraulic state at this position with respect to static hydraulic equilibrium. This effect is used in the following evaluation (Sect. 3.1.3).

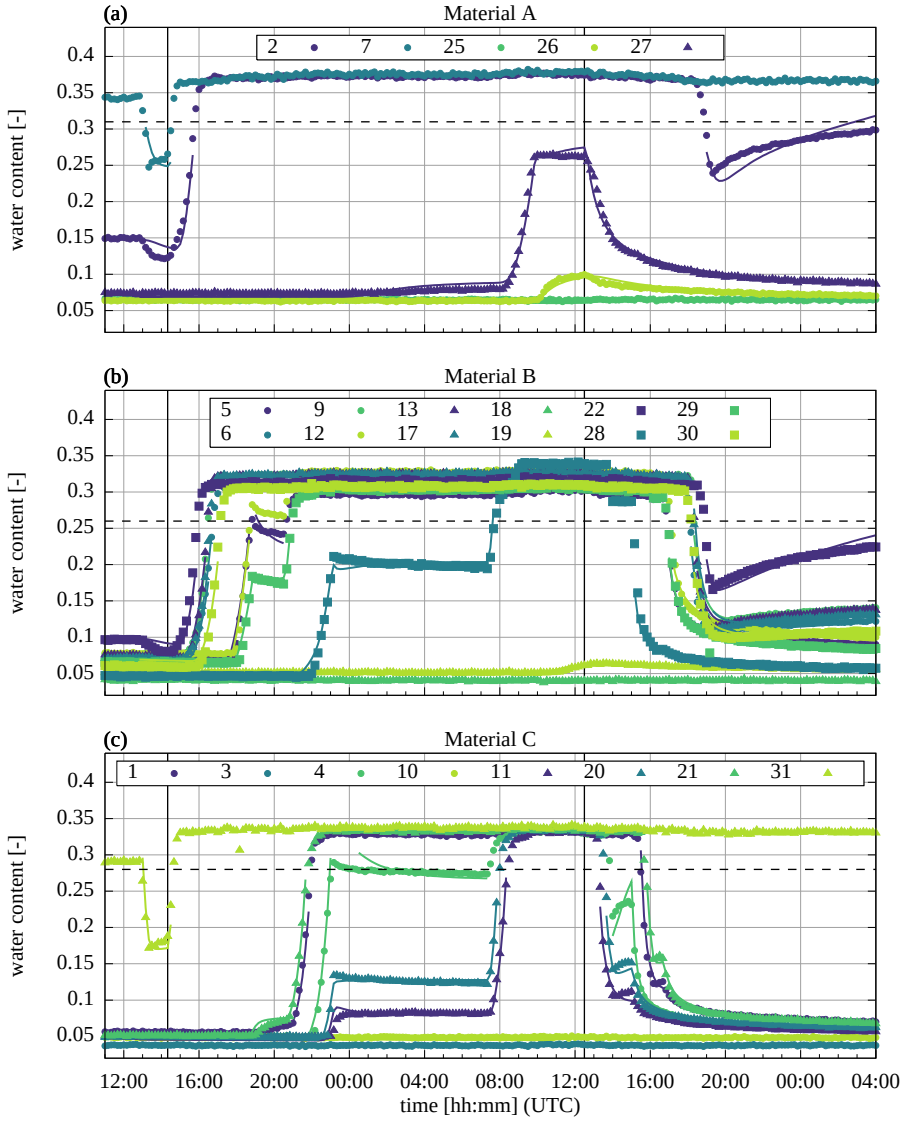

Figure 4. The measured water content data for the three different phases (initial drainage, multistep imbibition, and multistep drainage – separated by the solid vertical black lines in the figure) show a high variability up to and beyond the validity limits of the Richards equation due to the fluctuating groundwater table (Fig. 3). Hence, in order to avoid effects related to entrapped air and two-phase flow phenomena, we neglect all data with a volumetric air content smaller than 0.1 (all values above the dashed horizontal lines) based on measured porosities from core samples. The colored solid lines show the results of the setup *miller and position* of the 2D study (Sect. 3.2). The data measured before 12:50 are only used for the initial state estimation (Sect. A1.6).

We attribute the spread of the water content during saturation mainly to small–scale heterogeneity and quasi saturation due to

entrapped air (Christiansen, 1944). In order to avoid effects related to entrapped air and also two-phase flow, all TDR data with an air content below 0.1 (Faybishenko, 1995) are neglected subsequently.

2.3 Structural error analysis

As outlined in Sect. 1, the structural error analysis rests on a basic representation and a general assessment of the respective representation errors. Those representation errors, which are investigated in detail, are parameterized and implemented leading to a number of distinct representations with increasing complexity. Using inversion to estimate optimal parameters for each of the representations allows to analyze (i) the resulting residuals to improve the representations and (ii) effect of unrepresented model errors on the resulting material properties.

Preparing the tools for the method, we start this section with the Levenberg–Marquardt algorithm (Sect. 2.3.1) and discuss the assessment of the representation errors (Sect. 2.3.2) as well as the analysis of the resulting residuals (Sect. 2.3.3) afterwards.

2.3.1 Levenberg–Marquardt

We employ the Levenberg-Marquardt algorithm for parameter estimation. Our implementation is based on Moré (1978), Press (2007), and Transtrum and Sethna (2012) together with some further modifications.

Assuming (i) M data points m_μ $(1,\ldots,M)$ measured at position $\boldsymbol{x}_\mu$ featuring a white Gaussian measurement error with standard deviation σ_μ and (ii) a model f with P parameters p_π $(1,\ldots,P)$, then the χ^2 cost function is defined as

$$\chi^2(\boldsymbol{p}) = \frac{1}{2}\sum_{\mu=1}^{M}\left(\frac{m_\mu - f(\boldsymbol{x}_\mu,\boldsymbol{p})}{\sigma_\mu}\right)^2 = \frac{1}{2}\sum_{\mu=1}^{M} r_\mu(\boldsymbol{p})^2, \tag{7}$$

which assumes statistically independent representation errors that are normally distributed with zero mean and standard deviations σ_μ (perfect model assumption). The standardized residuals r_μ can be expanded

$$r_\mu(\boldsymbol{p}+\delta\boldsymbol{p}) \approx r_\mu(\boldsymbol{p}) + \sum_{\pi=1}^{P} J_{\mu\pi}\delta p_\pi \tag{8}$$

with the Jacobi matrix $J_{\mu\pi} = \partial r_\mu/\partial p_\pi$. The Jacobi matrix is assembled numerically with the finite differences method. Following Press (2007), the Hessian is approximated ($\mathbf{H} \approx \mathbf{J}^\top\mathbf{J}$), assuming that the second term in the derivative cancels out as $f(\boldsymbol{x}_\mu,\boldsymbol{p}) \to m_\mu$ with increasing number of iterations. For the Gauss-Newton algorithm then follows

$$\delta\boldsymbol{p} = -(\mathbf{J}^\top\mathbf{J})^{-1}\cdot\nabla\chi^2(\boldsymbol{p}). \tag{9}$$

Since $\mathbf{J}^\top\mathbf{J}$ does not always have full rank, the inversion may be ill–conditioned leading to uncontrolled large steps. One possibility to cope with this issue, is to regularize $\mathbf{J}^\top\mathbf{J}$ by adding a diagonal damping matrix $\mathbf{D}^\top\mathbf{D}$. We follow Transtrum and Sethna (2012) and choose this damping matrix, such that the diagonal entry for p_π contains the corresponding maximal diagonal entry of $\mathbf{J}^\top\mathbf{J}$ from all previous iterations if this value is larger than a predefined minimal value (1.0) which is used otherwise. The resulting damping matrix is scaled with a parameter λ which tunes both the amount of regularization and the

step size of the parameter update.

Finally, the parameter update $\delta \boldsymbol{p}$ is calculated via

$$\delta \boldsymbol{p} = -(\mathbf{J}^\top \mathbf{J} + \lambda \cdot \mathbf{D}^\top \mathbf{D})^{-1} \cdot \nabla \chi^2(\boldsymbol{p}), \tag{10}$$

where the linear problem is solved with a Singular Value Decomposition (SVD). If the condition number of the sensitivity matrix $S = \mathbf{J}^\top \mathbf{J} + \lambda \cdot \mathbf{D}^\top \mathbf{D}$ is larger than a threshold (10^{12}), the linear problem is solved approximately with the Conjugate Gradient algorithm by choosing the maximal number of iterations smaller than the number of parameters P. The proposed parameters at iteration i are given as

$$\boldsymbol{p}^{i+1} = \boldsymbol{p}^i + \delta \boldsymbol{p}^i. \tag{11}$$

The convergence path of the Levenberg–Marquardt algorithm is influenced by both the size of the scaling parameter λ_{initial} and the choice how to adapt λ after each iteration. In this work, we choose $\lambda_{\text{initial}} = 5.0$ and apply the *delayed gratification* strategy proposed by Transtrum and Sethna (2012). According to this strategy, λ is decreased by a previously chosen factor (2.0) if the parameter update is successful and increased by a larger factor (3.0) if the update is not successful.

The described gradient–based algorithm heuristically balances performance and stability. Expanding the stability measures, we introduce a damping vector $\boldsymbol{d}$ with entries $\in (0, 1]$ to decrease the correction of particular parameters via

$$\boldsymbol{p}^{i+1} = \boldsymbol{p}^i + \boldsymbol{d} \odot \delta \boldsymbol{p}^i, \tag{12}$$

where $\odot$ denotes the element–wise Hadamard product. Generally, the entries of the damping vector are set to 1. In order to delay the improvement for parameters which represent additional model components, we choose the according entries < 1. We use this approach in particular to estimate sensor positions and Miller scaling factors along with effective soil hydraulic properties (Sect. A1.4). First, these parameters are initialized to neutral values: The modeled sensor positions are initialized to the measured sensor positions and the Miller scaling factors to 1.0. Subsequently, the damping vector for the associated parameters is set to 0.1, reducing the applied correction of these parameters to 10% of the proposed correction by the Levenberg–Marquardt algorithm. Hence, the main focus of the algorithm is to estimate consistent effective soil hydraulic properties, whereas the sensor positions and Miller scaling factors are adjusted more gradually.

2.3.2 Assessment of representation errors

By applying the χ^2 cost function (Eq. (7)), it is implicitly assumed that the model is perfect aside from white Gaussian noise. This corresponds to complete quantitative understanding of reality and a Gaussian error model for the measurement data. Structural model–data mismatch indicates that this assumption is invalid. In our case, a Bayesian analysis of the total uncertainty space is not feasible, primarily due to a lack of models, e.g., for hysteresis. Hence, we have to neglect such representation errors and trust that the structural model–data mismatch will reveal any inadequacy. Table 3 gives an overview over the treatment of the representation errors considered in this work. The contribution of representation errors, which could not be quantified or

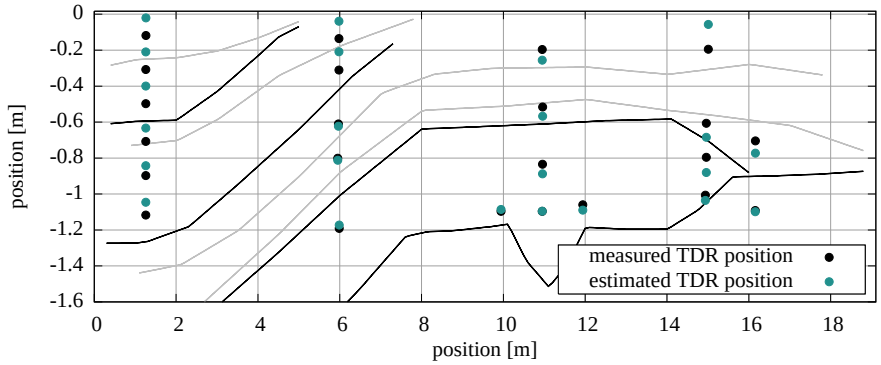

Figure 5. The subsurface architecture of ASSESS (Fig. 2) is shown with a comparison of measured and estimated TDR sensor positions based on a first evaluation of the hydraulic measurement data. The consistent deviation of the estimated TDR sensor positions reveal an unrepresented model error: The orientation of ASSESS (Sect. A1.2).

excluded from the measurement data a priori, is parameterized and explicitly estimated. Remaining structural model–data mismatch or deviation from the prior for the parameters hint at representation errors which should be corrected in the subsequent iteration of the analysis.

The structural error analysis and the assessment of uncertainties results from iterative evaluations. To illustrate the method, we present an iteration where the orientation of ASSESS was not yet compensated by rotating the geometry and the gravitation vector (Sect. A1.2). Considering the structural error analysis, we parameterized and estimated uncertain components in the representation. Hence, not only the Mualem–Brooks–Corey parameters, an offset to the Dirichlet boundary condition (Sect. A1.5) and the saturated hydraulic conductivity of the gravel layer, but also the position of the TDR sensors were estimated (Sect. A1.4). The results presented in Fig. 5 show that the estimated TDR positions display a consistent deviation from the positions, which were measured relative to the site's walls, as they compensate for the orientation of ASSESS. Thus, the position of most TDR sensors on the right is estimated to be higher and the position of most TDR sensors on the left is estimated to be lower than the measured ones. By estimating the TDR sensor position, we also incorporated other representation errors into the resulting parameters, such as small–scale heterogeneities and eventually a non–represented evaporation front mostly affecting the estimated position of the upper TDR sensors (3, 11, 18, and 25). Hence, this analysis (i) demonstrates the difficulty to separate representation errors and (ii) is able to identify representation errors which have to be improved subsequently.

2.3.3 Residual analysis

A visual analysis of the standardized residual increases the intuitive understanding of the model–data mismatch (e.g., Legates and McCabe, 1999; Ritter and Muñoz-Carpena, 2013). We analyze the standardized residual in two ways: (i) The visualization over time highlights the temporal development of the structural model–data mismatch. (ii) The visualization over theoretical quantiles corresponding to a Gaussian distribution with the standard deviation of the measurement data facilitates the com-

parison of the standardized residual distribution to the expected Gaussian distribution of the measurement data. Hence, if the perfect model assumption is true, the probability plot will show a straight line with slope 1. Yet, probability plots often show a characteristic *S*–shape (e.g., Fig. 7f): The slope < 1 for small residuals indicates that these residuals are smaller than expected for a Gaussian distribution with the standard deviation of the measurements. The slope > 1 for large residuals shows that these residuals are larger than expected for the presumed Gaussian distribution. Since in this work the theoretical quantiles are based on a Gaussian distribution, the *S*–shape generally indicates non-Gaussian distributions.

Additionally to the visual analysis of the standardized residual, statistical measures help to benchmark the model–data mismatch. As a single measure might be misleading (Legates and McCabe, 1999), we calculate the root mean square error (e_{RMS}) and the mean absolute error (e_{MA}).

2.4 Setup

The setup of the parameter estimation is explained with Fig. 6. For each of the three materials, we estimate the Mualem–Brooks–Corey parameters h_0, λ, K_s, τ, and θ_r (Sect. 2.2.2). The saturated water content θ_s is assumed to be equal to an estimate for the porosity ϕ based on core samples (Sect. 2.2.4). In order to avoid parameter bias due to representation errors, we (i) neglect measurement values with volumetric air content smaller 0.1 (Sect. 2.2.4), (ii) estimate a constant offset to the Dirichlet boundary condition (Sect. A1.5) and the saturated hydraulic conductivity of the gravel layer, and (iii) developed a method to estimate the initial water content distribution based on TDR measurement data (Sect. A1.6), because a spin-up phase would increase the computation time by up to a factor of 17. The details concerning the implementation of the TDR sensors and the small–scale heterogeneity with Miller scaling factors at the position of the TDR sensors are explained in Sect. A1.4.

In order to analyze the effect of the uncertainty of the sensor position, small–scale heterogeneity, and lateral flow on the estimated material properties along the lines presented in Sect. 2.3, we implemented a 1D and a 2D study with four different setups: (i) ***basic***: We estimate the hydraulic material properties, an offset to the Dirichlet boundary condition, and the saturated hydraulic conductivity of the gravel layer. (ii) ***position***: In addition to the parameters estimated in the *basic* setup, we also estimate the sensor positions. (iii) ***miller***: In addition to the parameters estimated in the *basic* setup, we estimate one Miller scaling factor for each TDR sensor. (iv) ***miller and position***: In addition to the parameters estimated in the *basic* setup, we estimate both the sensor positions and one Miller scaling factor for each TDR sensor.

2.4.1 1D study

In order to investigate the extent to which the experiment at ASSESS can be described with a 1D model, we set up three different cases with an increasing number of TDR sensors per material (Table 4): ***Case I*** includes the measurement data of sensor 1 in material C and sensor 2 in material A, and thus comprises one sensor per material. ***Case II*** includes two sensors per material, sensors 10 and 11 in material C and sensors 12 and 13 in material B. ***Case III*** includes three sensors per material, sensors 25, 26, 27 in material A and sensors 28, 29, 30 in material B. Note (i) that the cases are located at different positions in ASSESS (Fig. 2) and (ii) that since the hydraulic potential is not measured in the domain covered with these 1D studies, the

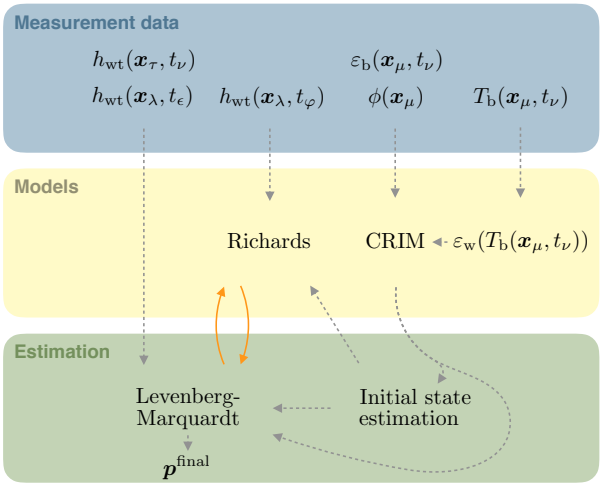

Figure 6. The available hydraulic potential h_{wt} is measured at the bottom of the groundwater well $\boldsymbol{x}_\lambda$ and at the position of the tensiometer $\boldsymbol{x}_\tau$. The data set, which is measured in the groundwater well, is split according to the measurement times: The data measured during the forcing phases t_φ enter the Levenberg–Marquardt algorithm (Sect. 2.3.1) directly, whereas the data measured during the equilibration phases t_ϵ are only used as boundary condition for the Richards equation (Sect. 2.2.1). The bulk relative permittivity $\varepsilon_{\mathrm{b}}(\boldsymbol{x}_\mu, t_\nu)$ and the bulk soil temperature $T_{\mathrm{b}}(\boldsymbol{x}_\mu, t_\nu)$ are measured at the position of the TDR sensors $\boldsymbol{x}_\mu$ at times t_ν. Together with the porosity $\phi(\boldsymbol{x}_\mu)$, these data are transferred to water content data (Sect. 2.2.4), which enter the initial state estimation (Sect. A1.6) yielding an initial water content distribution and optional initial parameter values for the Levenberg–Marquardt algorithm. Additionally, the water content data are also directly used in the Levenberg–Marquardt algorithm. Dashed grey arrows represent one–time preparation steps, whereas solid orange arrows represent the iterative steps of the Levenberg–Marquardt algorithm yielding the final material parameters $\boldsymbol{p}^{\mathrm{final}}$.

respective inversions are only based on the TDR water content measurements.

As described above, the analysis is organized in four different setups (*basic*, *position*, *miller*, and *miller and position*). The *basic* setup is adjusted for the 1D studies, such that not only the material functions of the materials with sensors, but also the saturated conductivity of the third material (material A in *case II* and material C in *case III*) are estimated for *case II* and *case III*. The other setups remain accordingly. Further details concerning the implementation of the 1D study are given in Sect. A2.1.

For each of the different setups, we ran an ensemble of 20 inversions starting from Latin–Hypercube sampled initial parameter sets in order to analyze the convergence behavior. The sampling algorithm was implemented with the help of the pyDOE package (https://github.com/tisimst/pyDOE). For each setup, we only analyze the ensemble member with minimal χ^2 in the subsequent discussion (Sect. 3.1).

2.4.2 2D study

In this study, we expand the investigated domain to 2D and analyze the performance of the improved representation. To this end, we set up four different setups *basic*, *position*, *miller*, and *miller and position* as described above. Since the position of both the tensiometer and the groundwater well is in the modeled domain, we use the hydraulic potential measurement data as well as the TDR measurement data in this study. Thus, the *position* setup is adjusted, such that both the positions of TDR sensors and the tensiometer are estimated. All inversions for the 2D study are initialized with the initial state material functions (Sect. A1.6) in order to bring out the quantitative effect of the different representations on the resulting material properties. Further details concerning the implementation of the 2D study are given in Sect. A2.2.

3 Results and discussion

In order to improve the quantitative understanding of the hydraulic behavior of ASSESS (Sect. 2.1), we evaluate a basic representation (Sect. 2.2) with a structural error analysis (Sect. 2.3) that is implemented as outlined in Sect. 2.4. The discussion of the results is done separately for the 1D study (Sect. 3.1) and the 2D study (Sect. 3.2).

3.1 1D study

3.1.1 Objectivity of the measurement data

The standardized residual for each case is presented in Fig. 7 combining the resulting data of all applied TDR sensors. Investigating them for *case I*, it is striking that all setups describe the data qualitatively equally well. Since the estimation of the material properties is only based on one sensor per material in this case, the parameterization offers enough freedom to describe the data. Hence, it also accommodates unrepresented model errors, such as the sensor position and small–scale heterogeneities. Therefore, additional representation and estimation of TDR sensor positions or Miller scaling factors do not lead to further improvement. The largest residuals occur during highly transient phases. Compared to the data, the simulated imbibition phase is too slow for sensor 1 and too fast for sensor 2. Also the simulated drainage phase is too slow for sensor 1 and drainage behavior of sensor 2 is consistently wrong. This structural model–data mismatch hints at unrepresented model errors due to the restriction to a 1D domain, which is further discussed in Sect. 3.1.3. Still, the residuals of all setups are smaller than 5 standard deviations, which translates to a volumetric water content of 0.035.

The large residuals are not random and preferably occur in transient phases. We attribute them to missing processes in the dynamics or to biased parameters. As the curves in the probability plot are basically centered at the origin, a significant constant bias in the residuum can be excluded. The according statistical measures are given in Table 5.

The e_{MA} of the *basic* setup increases in *case II*, because there are two sensors per material and the effective material parameterization can not completely compensate for the small–scale heterogeneity at the position of both sensors simultaneously. Consequently, representing the small–scale heterogeneity improves the description of the data. As before, the largest residuals occur during highly transient phases, especially during the drainage phase. Except for two outliers, the residuals stay smaller

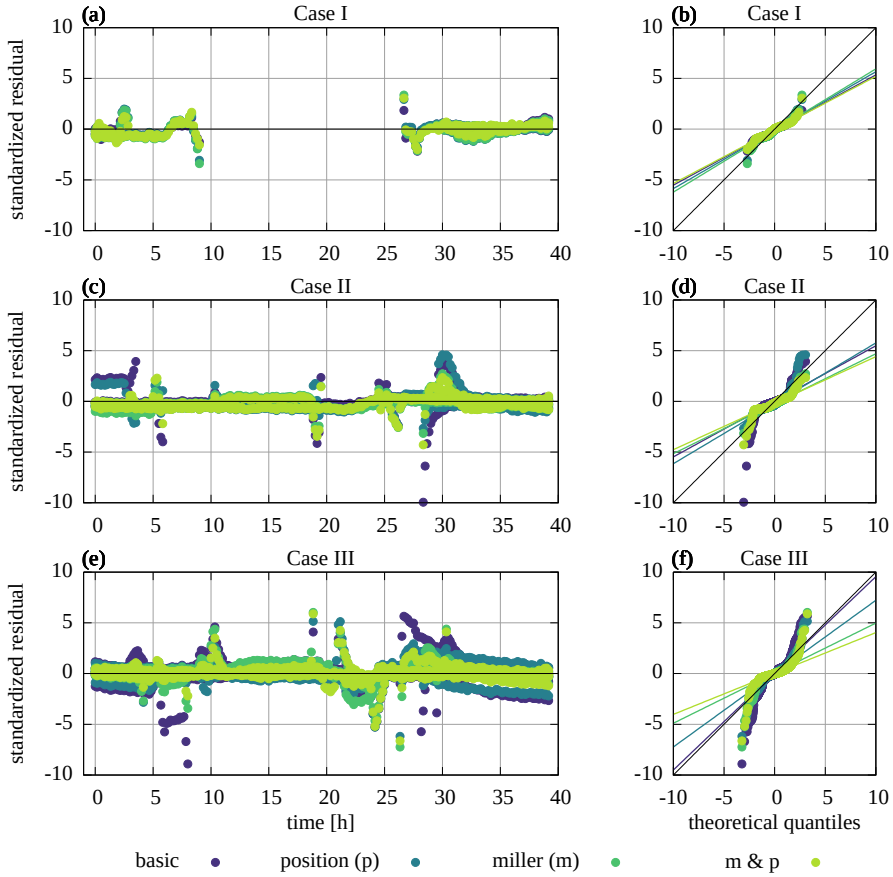

Figure 7. For the 1D study, the standardized residuals of the best ensemble member are visualized over time (left) and over the theoretical quantiles of a Gaussian with the estimated standard deviation of the TDR measurements (0.007) (right). The cases are analyzed with four setups ***basic***, ***position***, ***miller***, and ***miller and position***. The more sensors per material are used in the inversion, the worse the representation of the *basic* setup gets. In this case, representing uncertainties with respect to the sensor position and small–scale heterogeneities improves the representation substantially. The decreasing slope of a linear fit (thin lines in the probability plots), which is based on the standardized residuals within $[-2, 2]$ theoretical quantiles, also indicates this improvement.

than 5 standard deviations here as well. Considering three sensors per material in *case III*, the e_{MA} increases even further in the *basic* setup. Consequently, representing small–scale heterogeneities and uncertainties in the sensor position in the *miller and position* setup improves the e_{MA} by more than a factor of 2.

3.1.2 Separation of uncertain model components

Comparing the resulting material properties of the evaluated ensemble members for the different cases and setups (Fig. 8), we notice that the resulting soil water characteristic functions are shifted within each material. During static phases and if only few

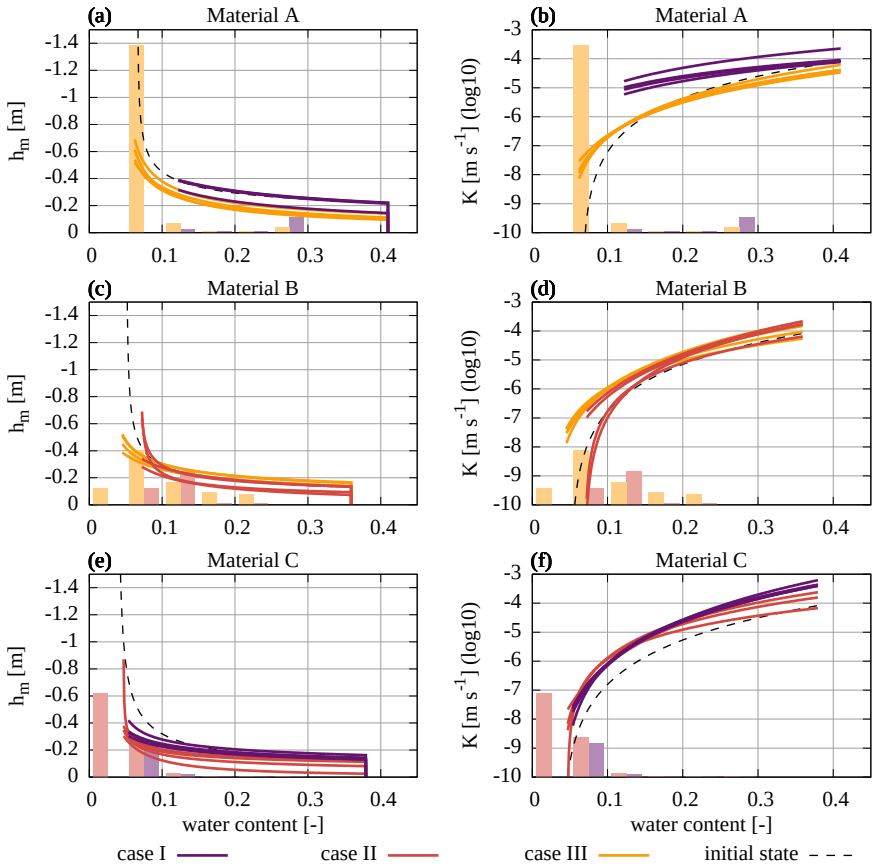

Figure 8. The estimated material functions of the best ensemble member are shown for each of the three cases (***case I***, ***case II***, and ***case III***) and the four setups of the 1D study. Additionally, we present the material functions resulting from the ***initial state estimation*** (Sect. A1.6). The plot range is adjusted to the available water content range for all inversion results. The number of water content measurements within intervals of 0.05 is indicated with histogram bars for each case. The height of these bars is normalized over all figures in this work. The main message of this figure is, that unrepresented model errors may lead to biased hydraulic parameters. In particular, this can be seen by comparing the hydraulic conductivity K of material A for the cases I and III.

measurement sensors are available, the parameters for the estimated uncertain model components (Sect. 2.4) can be correlated. However, during transient phases and if a larger number of measurement sensors is available, the distinct properties of these uncertain model components are more clearly pronounced (Fig. 9 and Sect. 3.2.3).

We also ran the inversions without estimating the offset to the Dirichlet boundary condition (Sect. A1.5). Besides destabilizing the convergence of the Levenberg–Marquardt algorithm, this fully transfers the uncertainty in the boundary condition to the sensor position. Hence, setups that estimate the sensor position clearly outperform the others. Additionally, this does not remove the shift of the soil water characteristics.

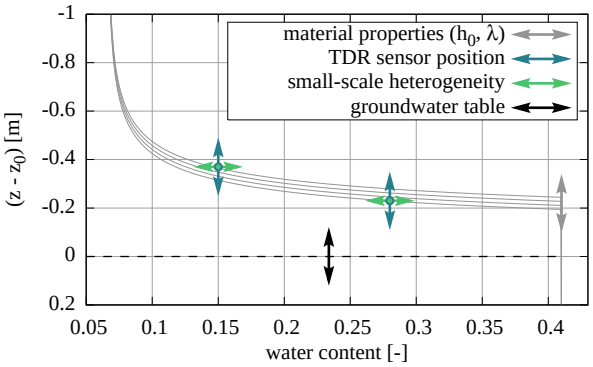

Figure 9. The estimation of uncertain model components can lead to correlated estimated parameters, e.g., as an incorrect position of the groundwater table (z_0) can be compensated by changing h_0 *and* λ during static phases. During transient phases, however, the components have distinct effects, e.g., as λ also changes the conductivity function. Hence, the ability of the parameter estimation algorithm to separate these uncertainties depends on the available measurement data. Also, the more sensors are available, the fewer uncertain model components can be compensated simultaneously by the parameterization.

3.1.3 Lateral flow

The three cases cover the three materials at different locations in ASSESS and are based on distinct data with respect to both quantity and data range.

This is most evident for material A which is located at the bottom of ASSESS and nearly saturated in *case I* whereas it is at the top and rather dry in *case III* (colored dots in Fig. 2). To illustrate that this leads to a different sensitivity on the unrepresented model errors, we highlight one example which is most pronounced during the final equilibration phase. In *case III*, the water content at position of the TDR sensors 25, 26, and 27 is higher than in static hydraulic equilibrium, leading to a drainage flux and a decrease in water content (Fig. 4). However, in *case I*, at the position of TDR sensor 2, the water content increases as the sensor monitors the relaxation of the capillary fringe. Due to the different hydraulic properties of the materials in ASSESS, this relaxation also includes unrepresented lateral flow.

In order to minimize the structural model–data mismatch during this equilibration phase, the parameter estimation algorithm increases the hydraulic conductivity to compensate for the non–represented lateral flow with additional vertical flow from above the sensor. Hence, the hydraulic conductivity of *case I* is larger than the hydraulic conductivity for both the *case III* and the 2D study, which is discussed in Sect. 3.2.4.

The measurement data of material B used in the inversions of *case II* and *case III* do not emphasize the relaxation of the capillary fringe strongly. Hence, we expect that the effect of the unrepresented lateral flow is not as significant as for material A leading to relatively congruent resulting material functions. This expectation is confirmed by the results, except for those setups of *case II*, in which no Miller scaling factor was estimated. These setups show a larger curvature of the soil water characteristic and of the hydraulic conductivity function which is explained in Sect. 3.2.4 in more detail. Additionally, we can identify the

previously discussed shift of the soil water characteristic (Sect. 3.1.2).

Similarly as for material B, the inversions for material C are not strongly influenced by the relaxation of capillary fringe. The large uncertainty in the saturated hydraulic conductivity reflects the low sensitivity of the measurement data on this parameter due to the lack of measurements influenced by the saturated material C.

3.1.4 Quality of the initial state material functions

The curvature of the soil water characteristic for the inversion results is reasonably close the initial state material functions (Sect. A1.6), although the initial parameter sets for the 1D inversions were Latin Hypercube sampled. This allows to use the the initial state material functions to initialize gradient–based inversion methods. The estimate of the initial state material function for material C deviates strongest from the inversion results compared to the other two materials, since in material C only few sensors are available to assess the form of the capillary fringe. Naturally, the better the available number of TDR sensors is spread over the water content range, the better the fit of the initial state parameters gets. Iteratively restarting the inversion using the previous inversion results as initial state material functions is likely to improve the representation. Since K_s and τ are not estimated along with the initial water content distribution but prescribed a priori, the hydraulic conductivity functions associated with the initial state show large deviations from the inversion results.

3.2 2D study

3.2.1 Objectivity of the measurement data

For the 2D study, the number of sensors is comparable to the number of hydraulic material parameters. Therefore, estimating sensor positions and Miller scaling factors increases the total number of parameters and thus the computational cost considerably (*basic*: 17, *position*: 41, *miller*: 41, *miller and position*: 65). The total number of analyzed TDR sensors increased to 25, corresponding to 5, 12, 8 TDR sensors for the materials A, B, C, respectively. In the 1D study, the residuals increased considerably during transient phases reaching up to 5 standard deviations in the *miller and position* setup (except for 3 outliers). Due to the larger number of considered TDR sensors in the 2D study, the measurement data cover more architectural situations and thus more complicated flow phenomena. In particular there are more transient phases observed than in the 1D studies. Therefore, we expect that (i) the resulting parameters are more objective (not shown, however), (ii) the standardized residuals at least in the *basic* setup increase, and (iii) estimating sensor positions and Miller scaling factors improve the description of the TDR data significantly. The standardized residuals visualized in Fig. 10 confirm the last two expectations. However, similar to the 1D study, even the residuals of the *miller and position* setup still reach more than 5 standard deviations for the 2D representation.

In order to understand this deviation in more detail, we investigate the remaining structural model–data mismatch during the final drainage and equilibration phases between 30–40 h. The largest residuals occurring during the drainage phase around 30 h come from the TDR sensors 6, 9, 13, and 17. We identified that these sensors are located close to a compaction interface (Sect. A1.6). Hence, the large residuals indicate that this horizontal compaction layer is not correctly represented with a point–scale

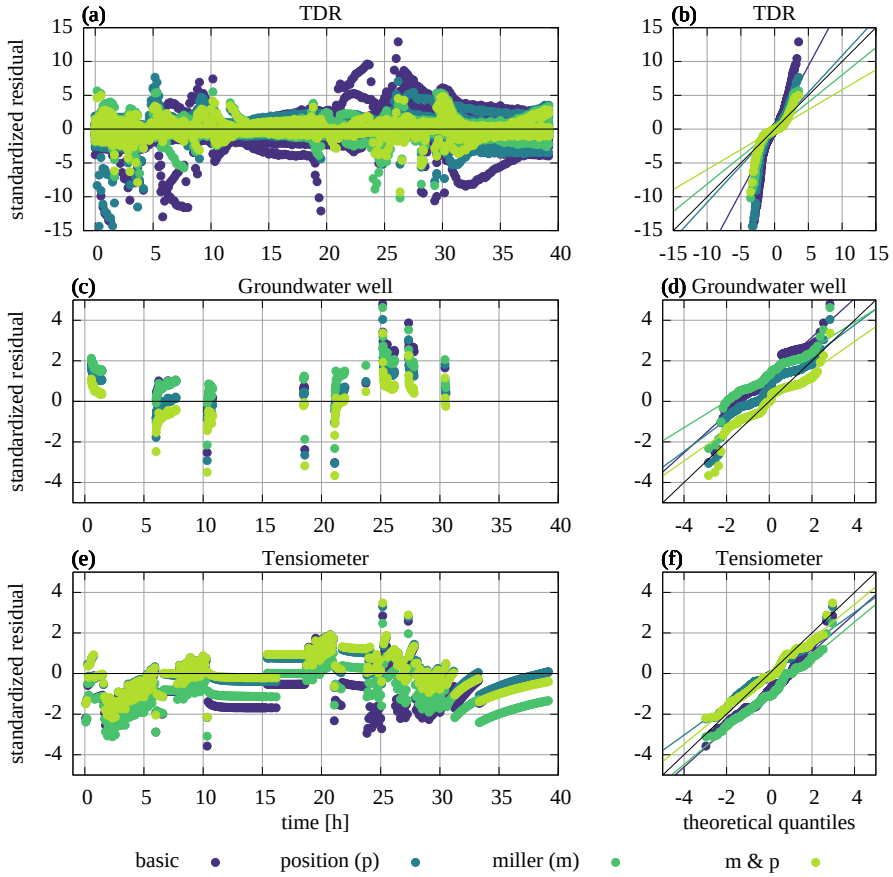

Figure 10. The standardized residuals of the 2D study are visualized over time (left) and in a probability plot (right) for all TDR and hydraulic potential sensors. The color associates the results with the four setups of the study (***basic***, ***position***, ***miller***, and ***miller and position***). Same as for the 1D study, the standard deviation for the TDR measurement data is chosen as 0.007. We choose the standard deviation for both the manual measurements in the groundwater well and the tensiometer measurement data as $0.025\ \mathrm{m}$. The representation of uncertainties with respect to the sensor positions and small–scale heterogeneities improves the description of the TDR data quantitatively. The decreasing slope of a linear fit (thin lines in the probability plots), which is based on the standardized residuals within $[-2, 2]$ theoretical quantiles, also indicates this improvement. The structural model–data mismatch for the hydraulic potential data is mainly due to (i) uncertainties concerning the position of the tensiometer and (ii) unrepresented 3D flow phenomena.

representation of the small–scale heterogeneity. The largest residuals during the final equilibration phase between 30–40 h come from TDR sensors 2 and 22 close to the capillary fringe. We attribute them to unrepresented processes in the dynamics, such as hysteresis or 3D flow (Sect. 3.2.2).

Due to the persisting large residuals during transient phases, the probability plot (Fig. 10b) displays a characteristic *S*–shaped curve for the TDR data (Sect. 2.3.3). The large residuals during transient phases are evidently different from the small residuals

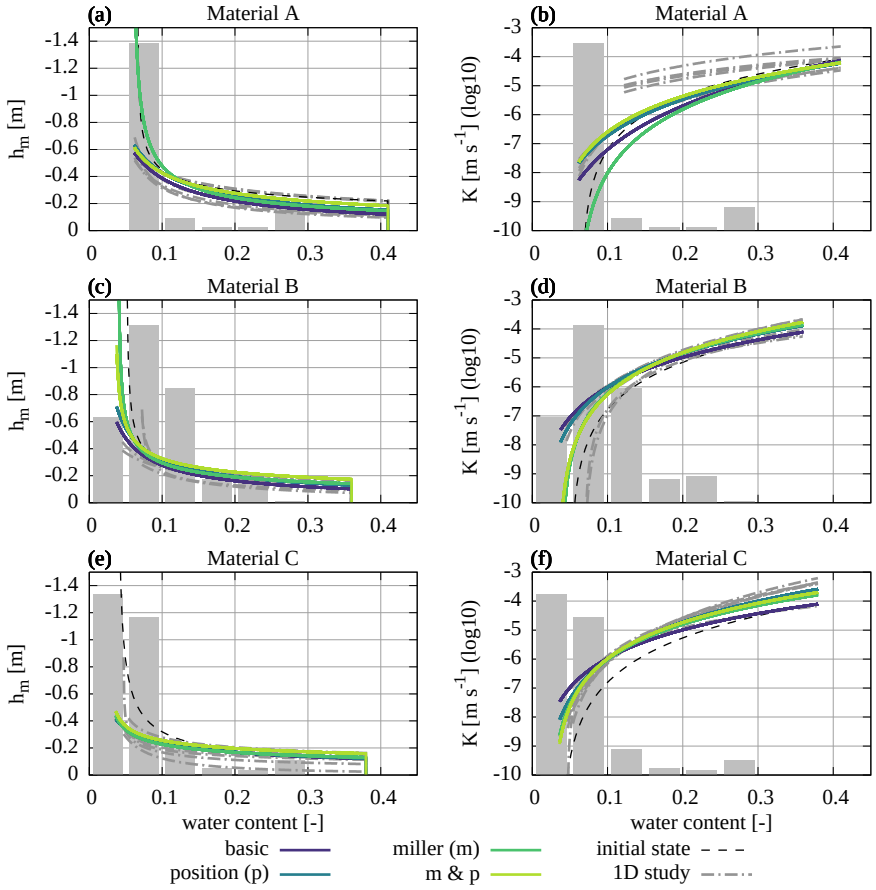

Figure 11. We show the resulting material functions for all three materials involved in the 2D study which is analyzed with the four setups ***basic***, ***position***, ***miller***, and ***miller and position***. The plot range is adjusted to the available water content range for each material. The height of the histogram bars denotes the number of available water content measurements and is normalized over all figures in this work. Since the inversions for all setups are initialized with the material functions resulting from the ***initial state estimation*** (Sect. A1.6), the difference between the results is directly linked to the estimation of sensor positions and small–scale heterogeneities. For direct comparison, the results of the *1D study* are also shown.

during static phases. This is corroborated by a linear fit based on the residuals within $[-2, 2]$ theoretical quantiles. For both the *miller* and the *miller and position* setup, the fits yield a slope < 1, indicating that distribution of the small residuals is more narrow than a Gaussian with a standard deviation of 0.007. This standard deviation is a measure that includes both precision and accuracy. We calculated the precision of the evaluated measurement data with a cubic spline fit yielding a precision of 0.001, 0.007 m, 0.006 m for the water content, tensiometer, manual groundwater position data, respectively. With complete quantitative understanding (Sect. 2.3), the standard deviation of the residuals would correspond to this precision. Lacking ground truth, the accuracy of the measurement data is unknown a priori and may depend on the hydraulic state. In this study,

its estimated contribution dominates the size of the standard deviations. Our results show that the model can represent static phases better than highly transient phases and that the accuracy of the measurement data is higher than estimated a priori. The statistical measures for the water content data given in Table 6 reveal that the e_{MA} of the *basic* setup merely increases by less than a factor of 2 compared to *case III* of the 1D study. Estimating sensor positions and Miller scaling factors improves the description of the TDR measurement data by more than a factor of 2 leading to a e_{MA} of 0.004.

3.2.2 Hydraulic potential

The description of the hydraulic potential data only improves in those setups, in which the sensor position is estimated (Fig. 10 and Table 6). Also the temporal structure of the model–mismatch does not change significantly with the different setups. The data show a gradient of the hydraulic pressure between the tensiometer and the groundwater well during the forcing phases (Fig. 3). Considering symmetry, we also assume this gradient of the hydraulic potential in the neglected third dimension. Hence, the forcing via the groundwater well leads to a 3D water flux during the experiment. This makes a correct representation of the groundwater table impossible in 2D. Consequently, the simulation should predict a higher position of the groundwater table in the well during imbibition phases and a lower groundwater table during the drainage phases. This expectation is confirmed by the standardized residuals shown in Fig. 10. Thus, the structural model–data mismatch of the tensiometer data indicates that employing the groundwater table as Dirichlet boundary condition overestimates the forcing in the simulation. Therefore, the simulated hydraulic pressure during the imbibition is larger than the measured one which leads to negative residuals. As expected, this behavior reverses during drainage phases.

3.2.3 Separation of uncertain model components

The 2D study is based on a larger number of water content measurements, additional hydraulic potential measurements, and more complicated flow phenomena compared to the previously discussed 1D study (Sect. 3.1). This improves the ability of the Levenberg–Marquardt algorithm to separate uncertain model components (Sect. 3.1.2) and decreases the shift in the soil water characteristics of the different setups compared to the 1D study (Fig. 11). Solely for material A the shift between the setups is comparably large. This can be explained with the relatively low number of water content measurements monitoring transient phases. Although the total number of evaluated measurements in the highly dynamical water content range (≈ 0.1–0.3) of material A is comparable to that of material C, Fig. 4 shows that fewer sensors monitor the transient phases in material A compared to material C.

3.2.4 Effect of unrepresented model errors

Each setup is started from the same initial material functions (Sect. A1.6). Therefore, the difference between the resulting material properties of the setups (Fig. 11) is a direct consequence of the representation of uncertainties in the sensor position and small–scale heterogeneities.

To investigate this, consider the initial state estimation for material B shown in Fig. A2. The measurement data of the sensors

5, 12, and 29 which are approximately $0.6\ \mathrm{m}$ above groundwater table deviate from the estimated function considerably. In order to cope with this deviation, the least–squares fit for the initial state draws the estimated soil water characteristic to higher water contents. Due to the rigidity of the Brooks–Corey parameterization, this causes an overestimation of the water content at the position of the sensors 0.8 and $1.4\ \mathrm{m}$ above the groundwater table (sensors 28 and 18). If the uncertainty in sensor position and small–scale heterogeneities are represented in the model, the outlying measurement data can be described without altering the effective material properties.

It is worth noting, that although the uncertainty of the measured grain size distribution (Table 1) is large, the resulting material properties confirm the measurements in that material A is the finest and the properties of materials B and C are similar. Our final best estimates for the effective hydraulic material properties for the *miller and position* setup are given in Table 7.

4 Summary and Conclusions

We applied a structural error analysis on a representation of the effectively 2D architecture ASSESS. This representation includes TDR and hydraulic potential measurement data which was acquired during a fluctuating groundwater table experiment. Based on the assumption that structural model–data mismatch indicates incomplete quantitative understanding of reality, we implemented a 1D and a 2D study organized in different setups with increasingly complex models. Starting with the estimation of effective hydraulic material properties and we added the estimation of sensor positions, small–scale heterogeneity, or both. It was demonstrated that the structural error analysis can indicate significant unrepresented model errors, such as the slope of the ASSESS test site or 3D water flow.

We showed that estimated material properties resulting from a 1D study are biased due to unrepresented lateral flow. Analyzing representations with increasing data quantity, it was also found that the fewer sensors are available per material, the stronger is the influence of the unrepresented model errors on the estimated material properties. We illustrated, that the more complicated flow phenomena are represented, the better uncertain model components can be separated by the parameter estimation algorithm leading to more reliable material properties. Generally, representing sensor position uncertainty and small–scale heterogeneity improved the description of the water content data quantitatively in setups with many sensors. Yet, the residuals of the water content data still reach more than 5 standard deviations during transient phases (Fig. 10). We attribute this to remaining representation errors in the forcing and compaction interfaces.

In order to minimize the error in the initial state, we developed a method to estimate the initial water content distribution based on TDR measurements and an approximation of hydraulic head which additionally yields an approximation of the soil water characteristic. We found, that this approximation is reasonably close to inversion results and that the according parameters can be used as initial parameters for gradient–based optimization. Since all the inversions of the 2D study are initialized with these parameters, the comparison of the results directly display the quantitative effect of the according unrepresented model errors on the estimated material properties.

Since the three approaches (i) initial state estimation, (ii) 1D inversion, and (iii) 2D inversion allow to estimate effective hydraulic material parameters, we finally discuss their levels of improving the quantitative understanding of soil water dynamics.

The initial state estimation requires at least three water content measurements per material over the full water content range and the position of the groundwater table to estimate the parameters for soil water characteristic for one specific equilibrated hydraulic state. Lacking direct measurements of the unsaturated hydraulic conductivity, the method cannot estimate the other parameters K_s and τ required to model soil water dynamics. Additionally, it is highly susceptible to uncertainties related to the sensor position and small–scale heterogeneities. Yet, the method is fast (few seconds on a local machine) and suitable to provide initial parameters for gradient–based inversion methods.

The 1D inversions are comparably fast (minutes up to hours on a local machine) and can represent transient states. In contrast to the initial state estimation, 1D inversions can estimate all parameters of the material functions. However, more complicated flow phenomena including lateral flow can not be represented. This leads to biased parameters.

The unique characteristics of the 2D inversions (days on a cluster with same number of cores as parameters) is the ability to represent lateral flow phenomena which are typically monitored with a high number of sensors. Hence, the consistency of the representation is implicitly checked. Therefore, we expect that of the three approaches discussed, this one yields the most reliable material properties. Still, unrepresented model errors including 3D flow phenomena influence the results.

5 Data availability

The underlying measurement data is available at http://ts.iup.uni-heidelberg.de/data/jaumann-roth-2017-hess.zip

6 Competing interests

The authors declare that they have no conflict of interest.

Appendix A: Details of the implementation

A1 Representation

A1.1 Richards equation solver

The Richards equation (Eq. 1) is solved numerically with $\mu\varphi$ (muPhi, Ippisch et al., 2006) on a rectangular structured grid using a cell centered finite volume scheme with full upwinding in space and an implicit Euler scheme in time. The nonlinear equations are linearized with an inexact Newton-Method with line search and the linear equations are solved with an algebraic multigrid solver.

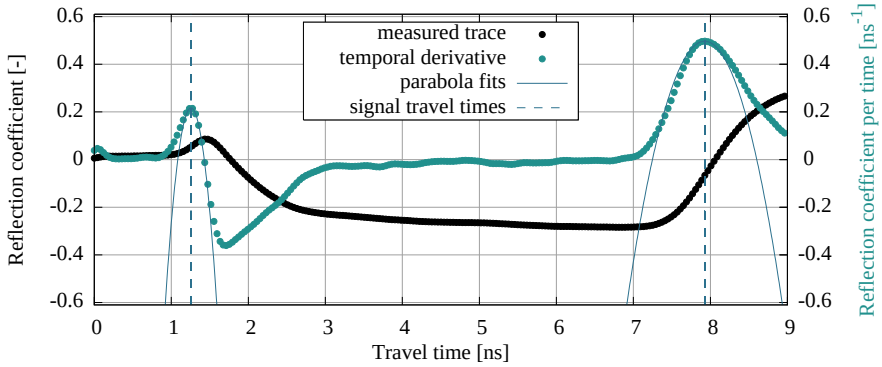

Figure A1. The evaluation of a TDR trace is based on the detection of the inflection points caused by the probe head and the end of the rod. This is done automatically after calculating of the first temporal derivative of the trace. Parabolas are fitted to the maxima of the temporal derivative to increase the precision of the evaluated signal travel time.

A1.2 Orientation of ASSESS

ASSESS is not built completely rectangular. Most importantly, both the surface and the ground are not horizontal but primarily inclined towards the groundwater well with a mean slope of $\approx -\frac{0.1}{20} = -0.005$. Since the applied Richards solver $\mu\varphi$ demands a rectangular structured grid, the geometry was rotated. This rotation was compensated by a counter-rotation of the gravity vector $\boldsymbol{g} \approx (0.0708, -9.8097)^\top$.

A1.3 Evaluation of TDR traces

The volumetric water content is evaluated from measured TDR traces (Fig. A1). As inflection points of the measured signal can be chosen to mark the reflections at the probe head and at the end of rods, the evaluation of the two-way signal travel time is based on detecting the maxima of the first temporal derivative of the recorded trace. To increase the precision of the evaluation, parabolas are fitted to the detected maxima. Finally, the maxima of the parabolas are employed to evaluate the two-way signal travel time. With the help of individual calibration data for each sensor comprising measurements in air and desalinated water, the travel time is converted into the relative permittivity ε_b of the bulk.

A1.4 Sensor position and small–scale heterogeneity

The numerical solution of the Richards equation (Eq. 1) is discretized in space with a rectangular structured grid (Sect. A1.1). Generally, the simulated value for the modeled position of a sensor is bilinearly interpolated from the simulated values at the center of the surrounding grid cells. Due to measurement uncertainties and subsidence after the construction, Antz (2010) and Buchner et al. (2012) assess the uncertainty concerning positions of sensors and material interfaces in ASSESS to $\pm\, 0.05$ m with respect to the model. However, since imbibition fronts can be very steep in sandy soils (Dagenbach et al., 2013; Klenk et al., 2015) and the measurement volume of the applied sensors is small, fluctuating groundwater table experiments are

very sensitive to the sensor position. Hence, we (i) enable the parameter estimation algorithm (Sect. 2.3.1) to estimate the sensor positions and (ii) implement the measurement volume of the TDR sensors by averaging the simulation data within a measurement radius of $0.015\ \mathrm{m}$.

In order to represent the heterogeneity of ASSESS which is not covered by describing the different sand types with distinct material properties due to the small-scale variability of the pore space, the center of each grid cell is associated with a Miller scaling factor (Eq. 4) that is initialized to 1.0. As the information about this small–scale heterogeneity only enters via the TDR measurement data, the exact position of each TDR sensor in the grid is also associated with a Miller scaling factor. This scaling factor may be estimated with the parameter estimation algorithm (Sect. 2.3.1). The scaling factors in the neighborhood of the TDR sensor are determined with a bivariate Gaussian distribution in horizontal and in vertical direction. This distribution is centered at the position of the TDR sensor and its amplitude corresponds to the associated Miller scaling factor. With a standard deviation of $0.015\ \mathrm{m}$ in both directions, it approaches 1.0 with increasing distance from the TDR sensor. Finally, this distribution is projected on each grid cell yielding the applied scaling factors which are only different from 1.0 in the neighborhood of the TDR sensors.

A1.5 Boundary condition

Generally, the boundary of the simulation is implemented with a Neumann no–flow condition. However, during the forcing phases, we prescribe the measured groundwater table as Dirichlet boundary condition at the position of the groundwater well. Additionally to the orientation of ASSESS (Sect. A1.2), the uncertainty of the sensor positions (Sect. A1.4) directly translates to an uncertainty in the Dirichlet boundary condition. Since representation errors of the forcing have a large impact on the resulting parameters, we implemented an optional offset to the Dirichlet boundary condition which can be estimated (Sect. 2.4).

A1.6 Initial state estimation

Since we use an inversion method for parameter estimation (Sect. 2.4), starting as near as possible to the measured initial state is key. Usually, this is achieved with a spin–up phase, which is computationally very expensive, however. Hence, we developed a method to estimate the initial water content distribution based on TDR measurement data.

In the first step, we assume static hydraulic equilibrium and approximate the matric potential at the measured position of the TDR sensors with the negative distance of this position to the groundwater table. Subsequently, the approximated matric potential is associated with the measured water content for each sensor. Further, we assume spatially homogeneous and temporally constant material properties which allows us to group the data of the TDR sensors by material, together with the approximated matric potential and the measured water content. For each material, we then fit the parameters h_0, λ, and θ_r of the Brooks–Corey parameterization to the approximated matric potential and the measured water content (Fig. A2). The saturated water content θ_s is assumed to be known from core samples. This yields an approximation for the initial water content distribution between the TDR sensors. With the resulting parameter values for each material, the subsurface material distribution, and the position of the groundwater table, we can calculate an estimation of the initial water content distribution in ASSESS (Fig. A3). As the parameters for the Brooks–Corey parameterization are derived from static measurement data, we may use them as

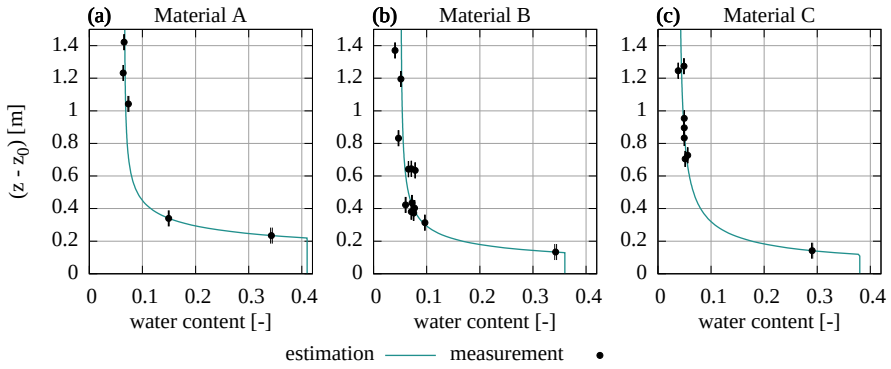

Figure A2. We use the Brooks–Corey parameterization to estimate the initial water content distribution between the TDR sensors. Assuming hydraulic equilibrium, we approximate the matric potential h_m with the negative distance to the groundwater table position z_0: $h_\mathrm{m} \approx -(z - z_0)$. For each material, we then use the approximated matric potential at the position of the TDR sensors and the corresponding water content measurement data to fit the Brooks–Corey parameters. Each dot depicts the mean of 15 subsequent data points measured in the $4\ \mathrm{h}$ preceding the experiment. The according standard deviations are all smaller than 0.002, which indicates (i) that the hydraulic system is relatively equilibrated at the beginning of the experiment and (ii) that the deviations from the estimation are statistically significant.

initial parameter values for computationally expensive gradient–based inversions of dynamic measurement data (Sect. 2.4.2). The missing initial parameter values $\tau = 0.5$ and $K_\mathrm{s} = 8.3 \cdot 10^{-5}\ \mathrm{m\,s^{-1}}$ are taken from Carsel and Parrish (1988). We refer to these parameter sets as *initial state material functions* in this work.

In particular due to (i) a limited number of TDR sensors, (ii) missing hydraulic potential measurements at the position of the TDR sensors, and (iii) spatial small–scale heterogeneity present in the materials, structural deviations between the estimation and the measurements occur, which indicate limitations of describing ASSESS with effective soil hydraulic material properties.

A2 Setup

A2.1 1D study

The forward simulations were calculated with a grid resolution of $0.005\ \mathrm{m}$ and 10^{-8} as limit of the Newton solver (Sect. A1.1). Following Jaumann (2012), the standard deviation of the TDR measurements is assumed as 0.007. We use the manually measured groundwater table data as Dirichlet boundary condition. Uncertainties concerning the position of the sensors and the subsurface material interfaces directly translate to uncertainties in the boundary condition (Sect. A1.5). Accounting for the orientation of ASSESS (Sect. A1.2), we add a constant offset to the Dirichlet boundary condition for each case (*case I*: $-0.02\ \mathrm{m}$, *case II*: $-0.05\ \mathrm{m}$, *case III*: $-0.12\ \mathrm{m}$). In order to minimize the input error, we also estimate this offset in the inversion. If TDR sensor positions are estimated, these are initialized to the measured position. Similarly, the Miller scaling factors are initialized to 1.0.

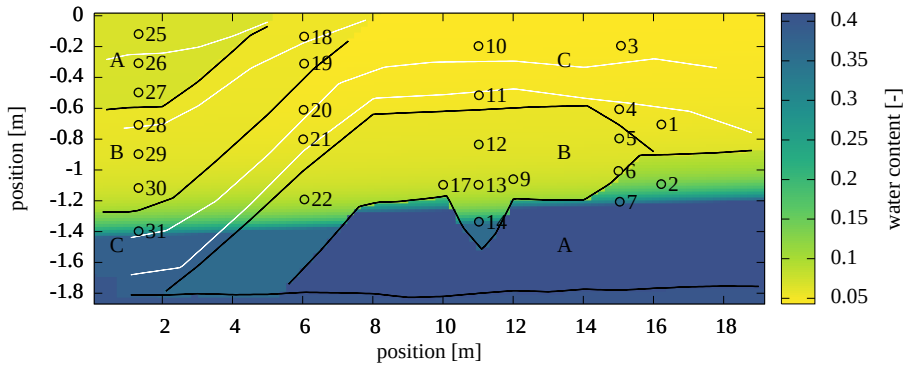

Figure A3. The estimated initial water content distribution is based on the TDR measurement data (Fig. A2, shown as face color of the circled dots). Since the saturated water content θ_s is fixed for each material a priori, only TDR sensors in unsaturated material are shown. Due to the orientation of ASSESS (Sect. A1.2), the groundwater table is slightly slanted. The black lines indicate material interfaces, whereas the white lines indicate compaction interfaces, which were introduced during the construction of ASSESS. Note the different scales on the horizontal and the vertical axis.

A2.2 2D study

The 2D simulations in this work are calculated with a grid resolution of $0.2\ \mathrm{m} \times 0.02\ \mathrm{m}$. The limit of the Newton solver is set to 10^{-8} (Sect. A1.1). Like for the 1D studies, we choose 0.007 as the standard deviation of the TDR measurements. The standard deviation of the tensiometer ($0.025\ \mathrm{m}$) is assessed from the accuracy ($\pm\ 5\ \mathrm{hPa}$) as specified by the manufacturer. In order to transfer the given uniform distribution with range $\pm\ 5\ \mathrm{hPa} \approx \pm\ 0.05\ \mathrm{m}$ to a Gaussian distribution, we associate this range with the 2σ interval of a Gaussian ($5\ \%$ to $95\ \%$). This leads to an approximate standard deviation of $(0.05\ \mathrm{m} \cdot 2)/4 = 0.025\ \mathrm{m}$. Lacking an independent estimate for the accuracy of the manual groundwater table position measurement, we employ the accuracy of material interfaces in ASSESS (Sect. A1.5). Same as for the tensiometer, this leads to a standard deviation of $0.025\ \mathrm{m}$. Some TDR sensors are located close to or even below the groundwater table. Therefore, the position and the Miller scaling factor could not be estimated for TDR sensors. Hence, no position was estimated for sensors 7, 8, 14, 15, 16, 23, 24, 31, and 32 and no Miller scaling factor was estimated for sensors 8, 14, 15, 16, 17, 23, 24, 31, and 32.

Author contributions. S. Jaumann designed and conducted the experiment, developed the main ideas, implemented the algorithms, and analyzed the measurement data. K. Roth contributed with guiding discussions. S. Jaumann prepared the manuscript with contributions of both authors.

Acknowledgements. We thank Jens S. Buchner for the code to process the ASSESS architecture raw data, Angelika Gassama for technical assistance with respect to ASSESS, and Andreas Dörr for helping to set up a beowulf cluster. Additionally, we thank Hannes H. Bauser,

Andreas Dörr, and Patrick Klenk for discussions that improved the quality of the manuscript. We especially thank Patrick Klenk and Elwira Zur for assistance during the experiment. The authors acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grants INST 35/1134-1 FUGG and RO 1080/12-1. We are also grateful to the editor Roberto Greco, two anonymous reviewers, and Conrad Jackisch, who all helped to improve the manuscript significantly.

References

Abbasi, F., Feyen, J., and Van Genuchten, M. T.: Two-dimensional simulation of water flow and solute transport below furrows: model calibration and validation, Journal of Hydrology, 290, 63–79, doi:10.1016/j.jhydrol.2003.11.028, 2004.

Abbaspour, K., Kasteel, R., and Schulin, R.: Inverse parameter estimation in a layered unsaturated field soil, Soil Science, 165, 109–123, 2000.

Antz, B.: Entwicklung und Modellierung der Hydraulik eines Testfeldes für geophysikalische Messungen, Diploma Thesis, Heidelberg University, 2010.

Bauser, H. H., Jaumann, S., Berg, D., and Roth, K.: EnKF with closed-eye period – towards a consistent aggregation of information in soil hydrology, Hydrology and Earth System Sciences, 20, 4999–5014, doi:10.5194/hess-20-4999-2016, 2016.

Birchak, J. R., Gardner, C. G., Hipp, J. E., and Victor, J. M.: High dielectric constant microwave probes for sensing soil moisture, Proceedings of the IEEE, 62, 93–98, doi:10.1109/PROC.1974.9388, 1974.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M.: Time Series Analysis: Forecasting and Control, John Wiley & Sons, 2015.

Brooks, R. H. and Corey, A. T.: Properties of porous media affecting fluid flow, Journal of the Irrigation and Drainage Division, 92, 61–90, 1966.

Buchner, J. S., Wollschläger, U., and Roth, K.: Inverting surface GPR data using FDTD simulation and automatic detection of reflections to estimate subsurface water content and geometry, Geophysics, 77, H45–H55, doi:10.1190/geo2011-0467.1, 2012.

Carmichael, R. S.: Physical Properties of Rocks and Minerals, CRC press Boca Raton, 1989.

Carsel, R. F. and Parrish, R. S.: Developing joint probability distributions of soil water retention characteristics, Water Resources Research, 24, 755–769, doi:10.1029/WR024i005p00755, 1988.

Christiansen, J.: Effect of entrapped air upon the permeability of soils, Soil Science, 58, 355–366, 1944.

Clark, M. P., Kavetski, D., and Fenicia, F.: Pursuing the method of multiple working hypotheses for hydrological modeling, Water Resources Research, 47, doi:10.1029/2010WR009827, 2011.

Cushman, J.: An introduction to hierarchical porous media, in: Dynamics of Fluids in Hierarchical Porous Media. Academic Press, Inc., San Diego, California., pp. 1–6, 1990.

Dagenbach, A., Buchner, J. S., Klenk, P., and Roth, K.: Identifying a parameterisation of the soil water retention curve from on-ground GPR measurements, Hydrology and Earth System Sciences, 17, 611–618, doi:10.5194/hess-17-611-2013, 2013.

Erdal, D., Neuweiler, I., and Wollschläger, U.: Using a bias aware EnKF to account for unresolved structure in an unsaturated zone model, Water Resources Research, 50, 132–147, doi:10.1002/2012WR013443, 2014.

Faybishenko, B. A.: Hydraulic behavior of quasi-saturated soils in the presence of entrapped air: Laboratory Experiments, Water Resources Research, 31, 2421–2435, doi:10.1029/95WR01654, 1995.

Gelhar, L. W.: Stochastic subsurface hydrology from theory to applications, Water Resources Research, 22, doi:10.1029/WR022i09Sp0135S, 1986.

Gupta, H. V., Wagener, T., and Liu, Y.: Reconciling theory with observations: elements of a diagnostic approach to model evaluation, Hydrological Processes, 22, 3802–3813, doi:10.1002/hyp.6989, 2008.

Gupta, H. V., Clark, M. P., Vrugt, J. A., Abramowitz, G., and Ye, M.: Towards a comprehensive assessment of model structural adequacy, Water Resources Research, 48, doi:10.1029/2011WR011044, 2012.

Hopmans, J. W., Šimůnek, J., Nunzio, R., and Wolfgang, D.: Simultaneous determination of water transmission and retention properties. Inverse Methods., in: Methods of Soil Analysis. Part 4. Physical Methods, edited by Dane, J. and Topp, G. C., pp. 963–1008, Soil Science Society of America Book Series, 2002.

Huisman, J., Rings, J., Vrugt, J., Sorg, J., and Vereecken, H.: Hydraulic properties of a model dike from coupled Bayesian and multi-criteria hydrogeophysical inversion, Journal of Hydrology, 380, 62–73, doi:10.1016/j.jhydrol.2009.10.023, 2010.

Ippisch, O., Vogel, H.-J., and Bastian, P.: Validity limits for the van Genuchten-Mualem model and implications for parameter estimation and numerical simulation, Advances in Water Resources, 29, 1780–1789, doi:10.1016/j.advwatres.2005.12.011, 2006.

Jaumann, S.: Estimation of effective hydraulic parameters and reconstruction of the natural evaporative boundary forcing on the basis of TDR measurements, Diploma Thesis, Heidelberg University, 2012.

Kaatze, U.: Complex permittivity of water as a function of frequency and temperature, Journal of Chemical and Engineering Data, 34, 371–374, doi:10.1021/je00058a001, 1989.

Kavetski, D., Franks, S. W., and Kuczera, G.: Confronting input uncertainty in environmental modelling, Calibration of Watershed Models, pp. 49–68, doi:10.1029/WS006p0049, 2002.

Kavetski, D., Kuczera, G., and Franks, S. W.: Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, Water Resources Research, 42, doi:10.1029/2005WR004368, 2006.

Klenk, P., Jaumann, S., and Roth, K.: Quantitative high-resolution observations of soil water dynamics in a complicated architecture using time-lapse ground-penetrating radar, Hydrology and Earth System Sciences, 19, 1125–1139, doi:10.5194/hess-19-1125-2015, 2015.

Legates, D. R. and McCabe, G. J.: Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation, Water Resources Research, 35, 233–241, doi:10.1029/1998WR900018, 1999.

Li, C. and Ren, L.: Estimation of unsaturated soil hydraulic parameters using the ensemble Kalman filter, Vadose Zone Journal, 10, 1205–1227, doi:10.2136/vzj2010.0159, 2011.

Mertens, J., Stenger, R., and Barkle, G.: Multiobjective inverse modeling for soil parameter estimation and model verification, Vadose Zone Journal, 5, 917–933, doi:10.2136/vzj2005.0117, 2006.

Miller, E. and Miller, R.: Physical theory for capillary flow phenomena, Journal of Applied Physics, 27, 324–332, doi:10.1063/1.1722370, 1956.

Moré, J. J.: The Levenberg-Marquardt algorithm: Implementation and theory, in: Numerical Analysis, pp. 105–116, Springer, doi:10.1007/BFb0067700, 1978.

Mualem, Y.: A new Model for predicting the hydraulic conductivity of unsaturated porous media, Water Resources Research, 12, 513–522, doi:10.1029/WR012i003p00513, 1976.

Nielsen, D. R., Biggar, J. W., and Erh, K. T.: Spatial variability of field-measured soil-water properties, University of California, Division of Agricultural Sciences, doi:10.3733/hilg.v42n07p215, 1973.

Over, M. W., Wollschläger, U., Osorio-Murillo, C. A., and Rubin, Y.: Bayesian inversion of Mualem-van Genuchten parameters in a multilayer soil profile: A data-driven, assumption-free likelihood function, Water Resources Research, 51, 861–884, doi:10.1002/2014WR015252, 2015.

Palla, A., Gnecco, I., and Lanza, L.: Unsaturated 2D modelling of subsurface water flow in the coarse-grained porous matrix of a green roof, Journal of Hydrology, 379, 193–204, doi:10.1016/j.jhydrol.2009.10.008, 2009.

Parker, J., Kool, J., and Van Genuchten, M. T.: Determining soil hydraulic properties from one-step outflow experiments by parameter estimation: II. Experimental studies, Soil Science Society of America Journal, 49, 1354–1359, doi:10.2136/sssaj1985.03615995004900060005x, 1985.

Press, W. H.: Numerical Recipes 3rd Edition: The Art of Scientific Computing, Cambridge University Press, 2007.

Richards, L. A.: Capillary conduction of liquids through porous mediums, Physics, 1, 318–333, doi:10.1063/1.1745010, 1931.

Ritter, A. and Muñoz-Carpena, R.: Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments, Journal of Hydrology, 480, 33–45, doi:10.1016/j.jhydrol.2012.12.004, 2013.

Ritter, A., Hupet, F., Muñoz-Carpena, R., Lambot, S., and Vanclooster, M.: Using inverse methods for estimating soil hydraulic properties from field data as an alternative to direct methods, Agricultural Water Management, 59, 77–96, doi:10.1016/S0378-3774(02)00160-9, 2003.

Robinson, D., Jones, S. B., Wraith, J., Or, D., and Friedman, S.: A review of advances in dielectric and electrical conductivity measurement in soils using time domain reflectometry, Vadose Zone Journal, 2, 444–475, doi:10.2136/vzj2003.4440, 2003.

Roth, K.: Steady state flow in an unsaturated, two-dimensional, macroscopically homogeneous, Miller-similar medium, Water Resources Research, 31, 2127–2140, doi:10.1029/95WR00946, 1995.

Scharnagl, B., Vrugt, J., Vereecken, H., and Herbst, M.: Inverse modelling of in situ soil water dynamics: Investigating the effect of different prior distributions of the soil hydraulic parameters, Hydrology and Earth System Sciences, 15, doi:10.5194/hess-15-3043-2011, 2011.

Schneider, K., Ippisch, O., and Roth, K.: Novel evaporation experiment to determine soil hydraulic properties, Hydrology and Earth System Sciences, 10, 817–827, doi:10.5194/hess-10-817-2006, 2006.

Šimůnek, J., van Genuchten, M. T., and Wendroth, O.: Parameter estimation analysis of the evaporation method for determining soil hydraulic properties, Soil Science Society of America Journal, 62, 894–905, doi:10.2136/sssaj1998.03615995006200040007x, 1998.

Topp, G. C. and Miller, E.: Hysteretic moisture characteristics and hydraulic conductivities for glass-bead media, Soil Science Society of America Journal, 30, 156–162, 1966.

Transtrum, M. K. and Sethna, J. P.: Improvements to the Levenberg-Marquardt algorithm for nonlinear least-squares minimization, arXiv:1201.5885 [physics.data-an], 2012.

Van Dam, J., Stricker, J., and Droogers, P.: Inverse method to determine soil hydraulic functions from multistep outflow experiments, Soil Science Society of America Journal, 58, 647–652, doi:10.2136/sssaj1994.03615995005800030002x, 1994.

Vereecken, H., Huisman, J. A., Hendricks Franssen, H. J., Brüggemann, N., Bogena, H. R., Kollet, S., Javaux, M., van der Kruk, J., and Vanderborght, J.: Soil hydrology: Recent methodological advances, challenges, and perspectives, Water Resources Research, 51, 2616–2633, doi:10.1002/2014WR016852, 2015.

Vogel, H.-J. and Roth, K.: Moving through scales of flow and transport in soil, Journal of Hydrology, 272, 95–106, doi:10.1016/S0022-1694(02)00257-3, 2003.

Vrugt, J. A., Stauffer, P. H., Wöhling, T., Robinson, B. A., and Vesselinov, V. V.: Inverse modeling of subsurface flow and transport properties: A review with new developments, Vadose Zone Journal, 7, 843–864, doi:10.2136/vzj2007.0078, 2008.

Wöhling, T., Vrugt, J. A., and Barkle, G. F.: Comparison of three multiobjective optimization algorithms for inverse modeling of vadose zone hydraulic properties, Soil Science Society of America Journal, 72, 305–319, doi:10.2136/sssaj2007.0176, 2008.

Wollschläger, U., Pfaff, T., and Roth, K.: Field-scale apparent hydraulic parameterisation obtained from TDR time series and inverse modelling, Hydrology and Earth System Sciences, 13, 1953–1966, doi:10.5194/hess-13-1953-2009, 2009.

Wu, C.-C. and Margulis, S. A.: Feasibility of real-time soil state and flux characterization for wastewater reuse using an embedded sensor network data assimilation approach, Journal of hydrology, 399, 313–325, doi:10.1016/j.jhydrol.2011.01.011, 2011.

Wöhling, T. and Vrugt, J. A.: Multiresponse multilayer vadose zone model calibration using Markov chain Monte Carlo simulation and field water retention data, Water Resources Research, 47, doi:10.1029/2010WR009265, 2011.

Table 1. The grain size distribution in percent by weight displays the different granularity of the materials A, B, and C of ASSESS (G. Schukraft, personal communication, Institute of Geography, Heidelberg University, 2010). Whereas the composition of the materials B and C is similar, material A features a higher percentage of fine sand. Since the mechanical wet analysis is time-consuming and laborious, only material B was sampled twice. Thus, $80\ \mathrm{g}$ out of approximately $400\ \mathrm{Mg}$ were sampled. Due to rounding, the sum of the values is not always 100.

		grain size range		A	B_1	B_2	C
gravel	total	$2-63\ \mathrm{mm}$	[%]	2	5	4	5
sand	total	$63-2000\ \mu\mathrm{m}$	[%]	97	96	95	95
	coarse	$630-2000\ \mu\mathrm{m}$	[%]	10	24	20	17
	medium	$200-630\ \mu\mathrm{m}$	[%]	65	64	68	72
	fine	$63-200\ \mu\mathrm{m}$	[%]	22	8	7	6
silt	total	$2-63\ \mu\mathrm{m}$	[%]	0	0	0	0
clay	total	$<2\ \mu\mathrm{m}$	[%]	0	0	0	0

Table 2. During the experiment, ASSESS was forced with a fluctuating groundwater table. Therefore, 17.8 m^3 of water were pumped in and 14.7 m^3 were pumped out of the groundwater well. For the calculation of the according flux and equivalent height of the water column Δh_{eq}, the surface area of ASSESS was approximated with 80 m^2. All times are given in UTC.

phase	time start	time end	duration [min]	water volume [m^3]	flux [$10^{-6}\ \mathrm{m\,s^{-1}}$]	Δh_{eq} [m]
initial drainage	12:55:00	13:20:00	25	−0.7649	−6.4	−0.01
multistep imbibition	14:20:00	18:50:00	270	8.3900	6.4	0.10
	20:35:00	23:10:00	155	4.7809	6.4	0.06
	07:25:00	09:55:00	150	4.6361	6.4	0.06
multistep drainage	12:35:00	14:00:00	85	−3.9970	−9.8	−0.05
	15:00:00	16:10:00	70	−3.1709	−9.4	−0.04
	16:40:00	19:15:00	155	−6.7299	−9.0	−0.08

Table 3. This overview includes specification whether the considered model error is represented and explicitly estimated within the scope of this study.

model error	represented	estimated
local non-equilibrium	✗	✗
hysteresis	✗	✗
numerical error	✗	✗
orientation of ASSESS	✓	✗
initial state	✓	✗
entrapped air	✓	✗
boundary condition	✓	✓
sensor position	✓	✓
small–scale heterogeneity	✓	✓
material properties	✓	✓

Table 4. The 1D study comprises three different cases which investigate the three materials with increasing number of TDR sensors per material at different locations in ASSESS (Fig. 2). Note that each material is covered twice.

case	sensors	materials	position [m]
I	1 & 2	C, A	16.16
II	10, 11 & 12, 13	C, B	10.95
III	25, 25, 27 & 28, 29, 30	A, B	1.26

Table 5. In order to analyze the results of the 1D study, the performance of the best ensemble members for each case and for each setup are benchmarked with statistical measures. With increasing numbers of included TDR sensors per material, the statistical measures for the *basic* setup indicate worse description of the measurement data. However, estimating the position and the Miller scaling factor for each TDR sensor, improves description of the measurement data significantly according to the statistical measures.

case	setup		e_{RMS}	e_{MA}
I	basic		0.004	0.003
	position	(p)	0.004	0.003
	miller	(m)	0.005	0.004
	m & p		0.004	0.003
II	basic		0.007	0.003
	position	(p)	0.005	0.003
	miller	(m)	0.004	0.003
	m & p		0.004	0.003
III	basic		0.009	0.006
	position	(p)	0.006	0.004
	miller	(m)	0.005	0.003
	m & p		0.004	0.002

Table 6. For each setup of the 2D study, the results are benchmarked with statistical measures. Similar to the 1D study, estimating the sensor position and the Miller scaling factors improves the statistical measures related to the water content significantly. The statistical measures for the position of the groundwater table including both the tensiometer and the groundwater well data improve only for setups in which the sensor positions are estimated.

		water content		water table	
setup		e_{RMS}	e_{MA}	e_{RMS}	e_{MA}
basic		0.017	0.011	0.04	0.03
position	(p)	0.011	0.006	0.02	0.02
miller	(m)	0.008	0.005	0.03	0.03
m & p		0.006	0.004	0.02	0.02

Table 7. We present the effective hydraulic material parameters obtained with the setup *miller and position* of the 2D study. The formal standard deviations of the parameter estimation are given with the understanding that these are specific to the applied algorithm and will change for different algorithm parameters. The estimation for the saturated hydraulic conductivity of the gravel layer and for the offset to the Dirichlet boundary condition are $10^{-0.728\pm0.006}\ \mathrm{m\ s^{-1}}$ and $-0.034 \pm 0.001\ \mathrm{m}$, respectively.

material	h_0 [m]	λ [$-$]	K_s [m s^{-1}]	τ [$-$]	θ_r [$-$]	θ_s [$-$]
A	-0.184 ± 0.005	1.94 ± 0.07	$10^{-4.212\pm0.004}$	0.33 ± 0.07	0.025 ± 0.004	0.41
B	-0.174 ± 0.004	2.54 ± 0.06	$10^{-3.77\pm0.02}$	0.78 ± 0.05	0.035 ± 0.001	0.36
C	-0.159 ± 0.004	3.28 ± 0.02	$10^{-3.70\pm0.02}$	0.74 ± 0.06	0.026 ± 0.002	0.38