Author's response

This authors response includes the

- reply to referee comment 1
- reply to referee comment 2
- reply to short comment 1
- change list of the most important changes
- comment concerning the marked up version of the manuscript
- marked up version of the manuscript

Since the replies were uploaded, minor improvements were added to the manuscript and to the replies. Particularly, the line numbers of the references in the included replies were adjusted to keep this response consistent.

Reply to referee comment 1

This paper deals with the use of inverse modelling of soil water content and soil pressure measurements for estimating effective hydraulic parameters. Data are obtained from the ASSESS test site, which is an advanced experimental facility with well-known but complex soil layering and well-controlled boundary conditions. In particular, the effect of unrepresented model errors is investigated, and more importantly procedures are proposed to account for these model errors within the inversion process. The representation errors that are considered include uncertain sensor positions, uncertainty in boundary conditions, local heterogeneity, and dimensionality of the model (here: 1D vs. 2D). For the selected boundary condition (multi-step drainage and imbibition from below), it is found that small representation errors in sensor position can significantly affect the inverted material properties. I am strongly supportive of the idea of this study. Many studies typically stop after a single inverse modelling run. Sometimes the residuals are inspected, but very rarely the results of inverse modelling are used to improve the model concept or the system representation. This study explores several representation errors, and the results seem to indicate that reasonably small changes in system representation can significantly improve the data fit and the properties of the residuals. However, I have a few general concerns and specific comments that I would like to see addressed. Addressing these comments likely involves moderate to major revisions. In addition, grammar and spelling should be improved in the revised version.
Reply: We thank the reviewer for the constructive comments and suggestions. The manuscript was revised accordingly, hence we refer to the revised manuscript.

General Comments
1. The introduction is rather unambitious and does do full justice to the content of the manuscript. The authors decided to include a second introduction in section 4.3 where the structural error analysis is introduced. I strongly encourage bringing the idea of structural error analysis in the beginning of the manuscript to better prepare the reader for what is coming. The general stance of this extended introduction could be: Analysis of inverse modelling results to improve models. As already indicated above, I think there are too few studies that pursue this idea.
Reply: We agree and revised the introduction accordingly.

2. A general concern with the chosen approach is that the same data are used for inverse modelling and evaluation of the results. Would it not be much stronger when the

inversely estimated parameters are tested on an independent dataset? Are such independent datasets available for the ASSESS test site? In the current manuscript, improvements in data fit are reported, but this is fully expected because the amount of parameters was increased at the same time.
Reply: Independent datasets can either be achieved by changing the measurement method or the experiment setup. The former leads to different model errors, e.g., due to different measurement volumes of different instruments, and the latter leads to a different sensitivity of the data on the estimated parameters. Hence, we decided to analyze datasets of different instruments separately and to compare the results. Please also note the reply to comment 5 of SC1.
The improvements are reported so that the readers can judge whether the size of improvement is worth the associated additional effort.

3. A short discussion about the transferability of the results to other soil types would also be useful for the readers. Of course, gradients in water content are steep in ASSESS and this may significantly impact the importance of accurate sensor positioning. Would the same insights be obtained when the ASSESS test would have consisted of different loam soils? Please comment.
Reply: We agree, that this is an interesting question. Beyond general comments, we cannot answer it with the given data of the presented case study.

4. The authors decided to not take the classical structure of Introduction, Materials and Methods, Results and Discussion, Conclusions. For me, the alternative structure is not really working. For example, part of the results are presented in section 4.3 where the used methods have not yet been clearly explained. Although I may be purist in this matter, I would say that this paper would benefit from an organization following the classical scheme.
Reply: We revised the structure of the manuscript, bringing it closer to the classical scheme.

Specific Comments
Page 1, Line 1. Abstract should be a single paragraph. In addition, it is customary to provide the scope of the manuscript with an opening statement. Here, the authors immediately jump to the aims of the study.
Reply: We revised the abstract and added an introductory sentence.

Page 1, Line 19. Is direct determination really expensive? I would prefer time-consuming here.
Reply: We changed the wording here.

Page 2, Line 19. Huisman et al. (2010) considered a soil layer on top of the dike material.
Reply: We checked the paper again and found that the dike consists of the investigated material (Fig. 2, 4, 6, 9, and 10).

Page 2, Line 21. I would like to see more information about the TDR system that was used. Did the authors rely on automatic waveform analysis, or was this done manually to obtain more accurate results?
Reply: We added this information in Sect. 2.1 (Page 3, Line 20), Sect. 2.2.4 (Page 6, Line 2), and Sect. A1.3 (Page 23, Line 1).

Page 2, footnotes. I find it very unusual that the authors use footnotes. Is this possible and common in HESS? In any case, it seemed to me that much of the information provided in the footnotes could have easily been integrated in the main text. Please reduce the amount of footnotes to a minimum.
Reply: We integrated the footnotes in the text.

Page 4, Line 14. One-sentence paragraphs should be avoided.
Reply: We improved the section, such that the one-sentence paragraph is avoided.

Page 6, Line 19. I am not so convinced that a separate section on the implementation is a good idea. In particular, I do not really like the three very short subsections that now follow. It makes the text unpleasant to read.
Reply: We decided to separate the more general theory from the case dependent implementation such that the readers can skip or flip through the more general theory and just read the details on the implementation and do not have to do the sorting themselves. The three short subsections were introduced for precise referencing.

Page 9, Line 5. I could not follow your implementation of small-scale heterogeneity. Are you using heterogeneous parameters fields throughout the domain, or is this heterogeneity only introduced locally? Please clarify.
Reply: We clarified the Sect. A1.4 (Page 23, Line 18).

Page 10, line 12. I know this as global-local approach.
Reply: We updated the description of the 1D setup (Sect. 2.4.1) and don't use the wording anymore.

Page 10, line 21. Not sure that standard deviation is appropriate here? Is this not the expected standard deviation of the residuals (e.g. measurement error).
Reply: We made the sentence more precise (Sect. 2.3.1, Page 8, Line 14).

Figure 7. This figure did not make things clearer for me. Consider deleting.
Reply: We still think that graphically representing the flow of information is useful.

Page 12, Line 5. The start of this section seems out of place. For me, this clearly belongs to the general introduction (see general comments).
Reply: We revised the introduction accordingly.

Page 13, Line 20-32. Perhaps I am a purist, but for me this is a result and this is not a good position in the paper to discuss a result. I would bring this later.
Reply: This is intended as an example to show that the method works. It is thus a methods piece, not a result.

Figure 9. It would be good to show measured and modelled data in at least one figure. Here, a third column could be added to the left in addition to the residuals.
Reply: We added the results of the *miller and position* setup from the 2D case study to the data in Fig. 4.

Page 15, Line 5. Avoid repetitions. This has already been described four lines ago.
Reply: This comment is unclear to us. We rechecked the paragraph and could not identify any repetition.

Figure 10. This figure is too complicated. I am not sure how to read it. I am particularly unsure about the green.
Reply: We removed the indication of the setups in order to simplify the figure.

Page 19, Line 32. It is not so clear how you reached this conclusion. Perhaps this needs to be emphasized better when discussing the results.
Reply: We separated the Sect. 3 in subsections and clarified the analysis in Sect. 3.1.3.

Reply to referee comment 2

Dear Editor:
The study is interesting and demonstrates a huge work. However, before it can be transferred to the HESS step of the journal, I suggest the authors should discuss some key points and possibly make some changes in the text. I apologize for having been a bit late with my appraisal, but this also gave me the opportunity to read the comments from another referee and one discussant. I have listed below one general comment and several specific remarks, the most significant of which are starred (*).
Reply: We thank the reviewer for the constructive comments and suggestions. The manuscript was revised accordingly. Hence, we refer to the revised manuscript.

General Comments
As a referee, but also as a reader of studies dealing, among various sources of uncertainties, also with those associated with the locations of sensors that monitor a flow process, there is always something causing me some concern. When setting up an experimental test, efforts are made reducing errors (especially the systematic errors) and, among other things, one measures the positions of the various sensors as accurately as possible. I also understand that this task can be a bit more complicated under field conditions, especially when inserting the sensors at the greatest soil depths. Therefore and to the benefit of a wider readership, the authors should justify more why they are interested in this type of uncertainty. Moreover, I have the feeling that the error in sensor location should be viewed more as a systematic error rather than a random error. I think that the method employed by the authors might not be adequate to treat the presence of systematic errors. Some clarifications and a discussion on this point seem deserving.
Reply: We agree, that efforts are made to measure the positions of the various sensors as accurately as possible. Yet, the surface and/or the subsurface structure may change with time and requirements for accuracy and precision may change a posteriori. We clarified this in Sect. A1.4 (Page 23, Line 9).
We agree that the uncertainty in the sensor position is a systematic or structural error. This is the reason why this uncertainty was represented and the parameter estimation algorithm was used to propose more consistent positions of the sensors minimizing this systematic error.

Specific remarks
(*) P.1, L.13. The authors claim that the approximated soil water retention function is

reasonable close to the inversion results. *Actually and allowing for the types of water flow processes investigated, it would have been more interesting and effective that the favorable outcome is in terms of the unsaturated hydraulic conductivity function. From the results depicted in the right plots of Fig.10 and Fig.13, this does not seem the case.*
Reply: Lacking direct measurements of the unsaturated hydraulic conductivity at the position of the TDR sensors, the presented method merely yields an estimate of the initial hydraulic state and an approximation of the soil water characteristic. The remaining parameters for the initial hydraulic conductivity function (K_s and τ) are taken from Carsel and Parrish (1988, 10.1029/WR024i005p00755) and are independent of the presented measurement data. Hence, the presented method is not applicable to approximate the hydraulic conductivity function.

P.1, L.20-23. On the topic of inverse modeling applied to Soil Hydrology, I suggest citing the more recent and comprehensive papers by Hopmans et al. (2002) and/or by Vrugt and Dane (2006). Concerning the lab-scale experiment, the paper by Romano and Santini (1999) also treat types of errors of interest for the present study. As for the in-situ applications, the paper by Romano (1993) can also be in line with some aspects of the present study.
Reply: We revised the introduction accordingly. Please also note the reply to comment 1 of SC1.

P.1, L.22. The paper by Schneider et al. (2006) was published in HESS, not in Hess-D.
Reply: We corrected the reference.

(*) P.2, L.10-13. It is not clear (at least to me) which processes the authors are talking about. For example, the sensor position is definitely not a process. Moreover, as far as I am aware, the previous studies refer to minimum unknown parameters to be estimated mainly because they employed the classic χ^2 penalty criterion coupled with the Levenberg-Marquardt (LM) algorithm. Why do not compare the present results with those ones whether you use, for example, the DREAM tool developed by Vrugt (2016)? By doing that way, the paper would be even more interesting since the authors claim of having developed a modified LM algorithm.
Reply: We agree and changed the formulation (Page 2, Lines 9 – 12).
As the major focus of the manuscript, we investigate the effect of neglected structural errors which lead to suboptimal results using the χ^2 penalty criterion. Therefore, we also use the χ^2 penalty criterion coupled with the Levenberg-Marquardt algorithm and quantify the effect of unrepresented model errors by resulting residuals and material properties of the different setups (Sect. 2.3 and Sect. 2.4).
In order to compare the best result of the different setups, we are rather interested in maximum likelihood instead of its distribution in this work. The former is more efficiently found with the Levenberg-Marquardt compared to the DREAM algorithm. Additionally, if the χ^2 is used as likelihood function in DREAM, the discussed problem of neglected processes and uncertainties will remain the same as we use a flat prior in this study. Also, adding additional material would make the already long manuscript

even longer.

(*) P.4, L.8-10. Strictly speaking, the θ-based Richards equation describes the variations in space (x, y, and z coordinates) and time (t) of the volumetric soil water content. Then, due to the selected relationship between water content and matric pressure head, one can retrieve the corresponding variations in h.
Reply: We changed the wording in Sect. 2.2.1 (Page 4, Line 9).

(*) P.19, L.25-27. This is a quite common outcome when modeling of data with a maximum likelihood estimator and optimization techniques. I think that this problem should be addressed in another way. Namely, more in terms of the information content of the available input datasets. Does the initial information content increase when adding the additional data? Are the additional data not at all, or weakly, or strongly correlated among them and with the already available input datasets?
Reply: If the sensors monitor hydraulic dynamics which is not represented perfectly in the model, the residual will increase as the probability to monitor these model errors is increased with the number of sensors. In information theory, the information content of data is often quantified with measures such as the Shannon entropy. In order to apply these measures, the input data have to be transferred to random data. This requires knowledge about the general data structure which has to be gained from the data themselves. This implies massive practical issues in heterogeneous media. Since the TDR data monitor the same process at different positions, the Pearson correlation coefficient of the data is mainly positive and depends in particular on the recorded hydraulic dynamics. As the materials A and C which are flipped in case I and III, the characteristics of the monitored hydraulic dynamics changes. Hence, the correlation of these data is weak in general. The hydraulic state of material B is monitored at a similar position in cases II and III. Thus, the correlation of the according data increases.

As general and final comment, I should say that the English usage is very good. Nevertheless, the text is hard to follow. I do not have suggestions on this point, but the authors should make any effort to improve this aspect of the manuscript. Also, sub-section 4.1 might be left out from the manuscript, whereas I do not see the need to have so many small sub-sections in Section 3. Section 6, albeit being a summary, seems pointless and ineffective, chiefly because it also contains many repetitions. A real concluding remark section would be more effective, if necessary. Footnotes are rare or even absent in our scientific literature.
Reply: We revised the general structure of the manuscript. Please note the reply to comment regarding Page 6, Line 19 of RC1. We also revised Sect. 3 and Sect. 4 to make them more concise and generally integrated the footnotes into the text.

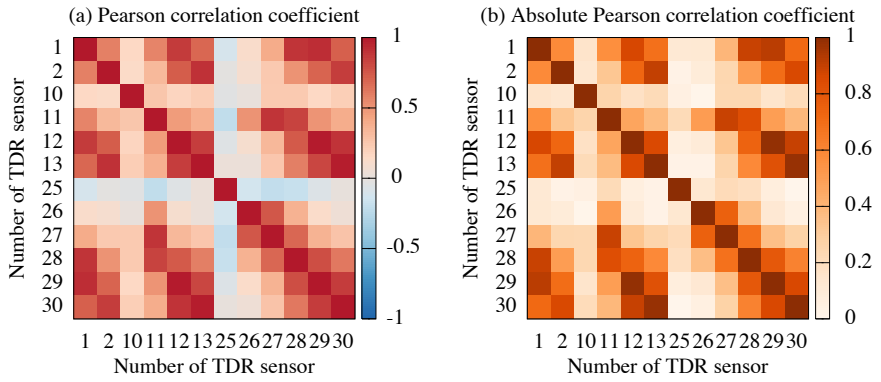

Figure 1: Pearson correlation coefficient for the data used in the 1D study

Reply to short comment 1

I congratulate the authors to an interesting study at the ASSESS experimental site. I consider the topic and the discussion manuscript highly relevant and worth to be published in HESS. Because of this, I would like to contribute some comments for a revision.
Reply: We thank Conrad Jackisch for the constructive comments and suggestions. The manuscript was revised accordingly. Hence, we refer to the revised manuscript.

1. If I understand correctly, the authors argue for a retention-dynamics-based identification of soil hydraulic material properties based on inverse modelling of an imbibition and outflow experiment. There have been many studies on the issue of inverse parameter estimation, which I consider relevant for the MS. This also holds for the discussion of heterogeneity and unrepresented model errors*. I.e. the authors name the* validity limits of the Richards equation *but I do not see the conceptual basis of the argumentation for their approach. Moreover, I suggest to present an independent reference for the found parameters (e.g. from laboratory analysis) and to include a critical view on the TDR inferred soil moisture values.*
Reply: The manuscript combines two main lines of thoughts: One is concerned with the estimation of hydraulic material properties on the basis of TDR measurement data acquired in a complicated subsurface architecture, which was forced with a fluctuating water table. We agree that this approach is not new and was applied already for many one-dimensional systems in the laboratory and also some in the field. The introduction cannot list all of the available literature. Rather it connects to the related literature we deem most relevant for this manuscript. The other line of thought considers a general problem in modeling, namely the investigation, which physical processes and uncertainties have to be represented in order to describe the measurement data adequately. As the true behavior of the system of interest is unknown and since required adequacy depends on the application at hand, we choose to test different hypotheses (realized by increasingly complicated models) and analyze their results. We improved the introduction to better reflect these two lines.
The Richards equation is only valid where water and air phase decouple, i.e. at intermediate saturation. At high saturation, water– and air–flow become coupled and a two–phase formulation is required. Conversely, at low saturation, vapor transport in the air phase is no more negligible and at least a two–components model is required. Richards equation is a single–phase model.

The transfer of laboratory data to field situations is notoriously difficult. Major challenges are (i) bringing an undisturbed sample into the laboratory, (ii) representing structures that are larger than the sample. In our opinion, there thus cannot be such a thing as an independent reference for a field site.
We assessed the precision from TDR data close to saturation and the accuracy with error propagation considering uncertainties in porosity and in bulk permittivity (Jaumann, 2012) yielding an uncertainty of 0.007 volumetric water content (Sect. A2.1). This result is of the same order as the evaluation of Roth et al., (1990, 10.1029/WR026i010p02267).
A major point of critique of the Complex Refractive Index Model (CRIM) concerns that it is a physically-motivated and not a physically-based model (e.g., Brovelli and Cassiani, 2008, 10.1111/j.1365-2478.2008.00724.x). Additionally, other uncertainties such as the influence of the electrical conductivity on the evaluated water content and on the temperature model for the permittivity of water as well as the spatial distribution of the relative permittivity of the soil bulk are neglected in the manuscript.

2. Despite my appreciation of the logical intention of the structure of the MS, I find it very difficult to follow. Especially, I could not trace answers to my expectations from the title and abstract – probably because they became obscured by many detailed side-tracks and because some promised elements (like GPR data or elaboration on what are model errors) are not really followed. Maybe a fundamental revision and exhibition of the main story line could clarify most of the forthcoming points.
Reply: We revised the structure of the manuscript accordingly.

3. What is the reason to use own models, solvers and the LM least squares optimizer instead of established and tested toolboxes? Is it really matter of the MS to present the technical details and equations although they are not developed further, taken up or discussed later on? How can be assured that numerical errors in the code do not bias the results (see also Clark and Kavetski 2010, 10.1029/2009WR008896)? I can imagine that the details suit well as appendix and that an explanation of the concept and intention to use these tools can clarify much of my second concern.
Reply: The solver for the Richards equation (muPhi) is tested, published (Ippisch et al., 2006, 10.1016/j.advwatres.2005.12.011), and it is, to the best of our knowledge, the numerically most efficient solver. The Levenberg-Marquardt algorithm was implemented according to published literature, because some of the required approaches are not implemented in available toolboxes.
We present only those technical details in the manuscript that are necessary to understand the evaluation procedure, such that the methods are traceable and reconstructible. The position of the methods section depends on the philosophy of the journal.
Due to the discretization of the problem in space and time, numerical errors are always existent, essentially balancing computational effort and numerical accuracy. We chose the grid resolution and meta–parameters given in the manuscript based on a grid convergence analysis.
We adjusted the structure of the manuscript accordingly.

4. Since heterogeneity is also an issue of scale and conceptual deficiency, I find the arguments not yet well drawn. What support of the TDR sensors is integrated by the measurements? How exactly are the estimated positions of the TDR sensors calculated and how precisely are the real positions known?
Reply: We clarified this issue in Sect. A1.4.
The support of the TDR sensors depends in particular on the sensor design and can be calculated (Robinson, 2003, 10.2136/vzj2003.4440). For the TDR sensors in ASSESS, the measurement volume contains a cylinder with a radius of approximately half the rod distance around the central rod in homogeneous electrical permittivity distribution. The rod distance of the TDR sensors used in ASSESS is 0.03 m.

5. Since GPR data of the experiment appears to be existing (Klenk et al. 2015 under review in HESSD doi:10.5194/hessd-12-12215-2015) I do not understand, why it is not used for the study (although mentioned in the abstract and introduction)? I suppose that the TDR and GPR data could be a very valuable pair of observations to be compared directly (as both rely on the rel. electrical permittivity). The strong advantage of GPR as spatially continuous technique could be related to the local measurement with higher absolute precision of the TDRs.
Reply: Three single channel time–lapse GPR radargrams were acquired during the experiment and are currently evaluated for a separate publication. The measurement data presented in Klenk et al. (2015) were recorded during a different imbibition and drainage experiment. The main focus of this manuscript is to quantify the effect of unrepresented model errors on the soil hydraulic material properties and to find consistent description of TDR measurement data. These data are characterized by a point–scale measurement volume of the sensors, which is the main reason for the described effect of uncertainties concerning the sensor position and small-scale heterogeneity. Since such point–measurements are rather the rule then the exception in most large–scale studies, the related issues require critical consideration. A rather complementary analysis is required for the GPR data taking into account the larger measurement volume and GPR related representation errors. This would blow the limits of a single paper. Please also note the reply to comment 2 of RC1.

6. Figures 10 and 13 suggest to me, that the observations relate to the portion of the (sandy) retention curve which is rather linear (and that the strongly non-linear part is actually only of importance at low matric potential). How is a transfer of the found parameters to the full retention spectrum validated? Since the ASSESS site is an artificial, well-defined test bed I would assume that the actual retention properties are known and that local deviations are mainly due to differences in bulk density. Hence I could imagine that the authors could use fig. 11 in the methods section to explain their approach in much more detail and related to specific research hypotheses referring to the retention properties. At the moment, I find it very difficult to read figure 9 and 12 and to compare the 1D and 2D case.
Reply: A transfer of the results to the full retention spectrum can neither be made nor validated with the available water content data and missing hydraulic potential mea-

surements. We explained in the reply to comment 1, why no laboratory–based reference retention properties are known for ASSESS. We think that Fig. 9 is required for the discussion of the results and is best understood with a direct reference to the application. We improved the description, how to read these figures in Sect. 2.3.3 (Page 10, Line 17). Please also note the reply to the comment of RC1 concerning Fig. 7 (Fig. 9 in the unrevised manuscript).

Please find minor comments highlighted in the attached MS file.
Reply: We revised the manuscript considering these comments.

List of most important changes

(Reference to revised version)

- Major revision of the text
- Footnotes were included in the text
- Introduction: Clarified storyline and perspective
- Restructured Section 2:
  Moved application-dependent details of the representation to the appendix A1
- Figure 2: Added face color to indicate 1D studies
- Figure 4: Added simulation of the 2D *miller & position* setup
- Restructured Section 2.3: Structural error analysis
  (includes old Section *Parameter estimation*)
- Restructured Section 3.1: 1D study: Discuss results in subsubsections
- Restructured Section 3.2: 2D study: Discuss results in subsubsections
- Figures 10 and 11 changed slightly due fixed error in data preprocessing
- Added Section 6: Competing interests
- Added Section A1.3: Evaluation of TDR traces
- Added Section A2: Setup
- Tables 5 and 6: Decreased number of significant digits
- Added Table 7: Parameters of the 2D *miller & position* setup

Comment to the marked up version of the manuscript

As the manuscript was revised completely, the structure and sections changed considerably. Hence, the standard markup tool for latex (latexdiff) has major problems generating a marked up version of the manuscript. Please excuse format errors, broken references to figures, sections, and tables and use the marked up version as a general indicator for changes.

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. We investigate the quantitative effect of unrepresented (i) sensor position uncertainty, (ii) small scale-heterogeneity, and (iii) 2D flow phenomena on estimated Unrepresented model errors influence the estimation of effective soil hydraulic material properties. Therefore, a 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) , Ground Penetrating Radar (GPR), and hydraulic potential measurement devices monitored the hydraulic stateduring the experiment. Since the measurement data are analyzed with an inversion method, starting close to the measurement data is key. Therefore, we developed a method which estimates the initial water content distribution by approximating the soil water characteristic on the basis of TDR measurement data and the position of the groundwater table. In order to reduce parameter bias due to unrepresented model errors, we implemented a structural error analysis to identify uncertain model components which have to be included in the parameter estimation. Hence, focussing on TDR and hydraulic potential data, we realized . 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 studywith increasingly complex setups: Starting with estimating effective hydraulic material properties, we added the estimation of sensor positions, the estimation of small-scale heterogeneity, or both. The analysis . The results of these studies with a modified Levenberg-Marquardt algorithm demonstrates demonstrate three main points: (i) The approximated soil water characteristic for the initial water content distribution is reasonably close to inversion results. (ii) Although the material properties resulting from fewer sensors are available per material, the larger is the effect of unrepresented model errors on the resulting material properties. (ii) The 1D and 2D studies are similar, the 1D studies are likely to yield study yields biased parameters due to unrepresented lateral flow. (iii) Representing and estimating sensor positions as well as small-scale heterogeneity improves 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 expensive time–consuming and near to impossible at larger scales. Therefore, indirect identification methods, such as inversion methods (**??**)(Hopmans et al., 2002; Vrugt et al., 2008a), have been successfully applied to evaluate many experiments starting from lab-scale with One-Step Outflow (Parker et al., 1985), Multistep Outflow (Van Dam et al., 1994), and evaporation (Šimůnek et al., 1998; Schneider et al., 2006), up to field scale studies (Wollschläger et al., 2009; Huisman et al., 2010)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 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 2d subsurface architectures. Abbasi et al. (2004) conducted an irrigation experiment to estimate soil hydraulic and solute transport properties for a 2D furrow architecture. Based on subsurface flow hydrographs for eight rain events, Palla et al. (2009) estimated effective soil hydraulic properties for a 2D layered coarse grained green roof . Exploiting 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) measurement data recorded during a fluctuating groundwater table experiment, Huisman et al. (2010) estimated soil hydraulic properties of a homogeneous dike. 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., 2008b; 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).

Being common practice, these studies neglect critical uncertainties , e. g., concerning the input error or small-scale heterogeneity, and restrict the number of estimated material parameters to a minimal amount. Our main hypothesis is that this 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 biased estimates for effective soil hydraulic properties due to neglected processes. 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.

We show for a 1D and a 2D study, that representing and estimating uncertain model components improves the quality of the representation significantly (Sect.3) . These studies are setup according to an uncertainty analysis indicating which uncertainties to represent (Sect.**??**) . Providing the measurement data (Sect. **??**) for this analysis, Time Domain Reflectometry (TDR), Ground Penetrating Radar (GPR), and hydraulic potential measurement devices monitored the hydraulic system (Sect. **??**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 respective unrepresented model errors on the material properties. Specifically, we analyze measurement data acquired at the test site (ASSESS) while it was as forced with a fluctuating groundwater table

2 Methods

Our

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) with different grain size distributions (Table 1). Its which are arranged in an effective 2D subsurface architecture is visualized in Fig. **??**. The approximately $2\ \mathrm{m} \times 20\ \mathrm{m} \times 4\ \mathrm{m}$ large site is equipped with

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)).

a weatherstation[1], 32 TDR sensors[1], one tensiometer (UMS T4-191), and a well to monitor and manipulate the groundwater table(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 is the only connection of the 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-silofodder–silo, a concrete L-element L–element serves as additional wall. In order to stabilize the material during the construction, it was compacted. Beyond Additional to the compaction interfaces shown in Fig. **??**, GPR measurements, e.g., presented by Klenk et al. (2015), 2, Ground Penetrating Radar (GPR) measurements indicate even more compaction interfaces (Klenk et al., 2015, Fig. 1b and 6).

We use this site to improve and develop GPR measurement and evaluation methods which increase the quantitative understanding of soil water dynamics. These methods comprise water content measurement (**?**), estimation of the position of material interfaces as well as the effective relative permittivity distribution (Buchner et al., 2012), identification of the appropriate parameterization type for the hydraulic material properties (Dagenbach et al., 2013), and high precision monitoring of fluctuating groundwater table and infiltration experiments (Klenk et al., 2015; **?**).

The view over ASSESS from $0-19\ \mathrm{m}$ shows the tensiometer, the weatherstation, and the groundwater well (left to right) as well as the color of the different sand types (figure adapted from Jaumann (2012)).

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

[1]The weatherstation measures precipitation, relative humidity, radiation, wind direction, and wind velocity.

[1]Each TDR sensor has three rods (length: $0.20\ \mathrm{m}$, diameter: $0.005\ \mathrm{m}$) and is associated with a soil temperature sensor.

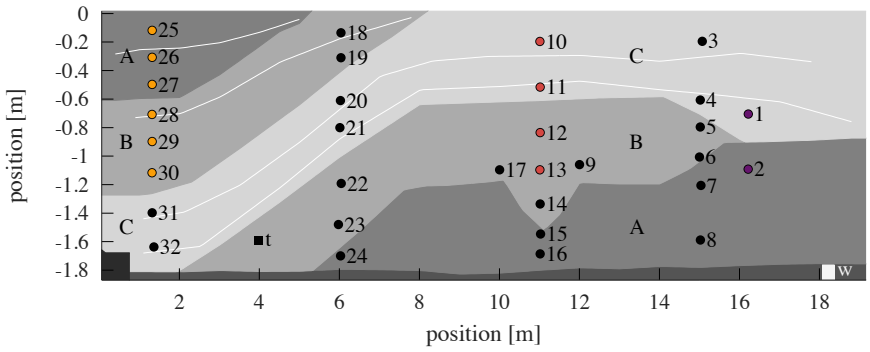

Figure 2. ASSESS features an effective 2D architecture with three different kinds of sand (A, B, and C). The hydraulic state can be manipulated with a groundwater well (white square, at 18.2 m) and is automatically monitored with 32 TDR sensors (dots) and one tensiometer (black square, at 4.0 m). The 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 L–element (black polygon, at 0.4 m) and compaction interfaces (white lines) were introduced during the construction. Additionally to those shown, GPR evidence indicates additional compaction interfaces. Note the different scales on the horizontal and the vertical axis.

(material properties), forcing (superscale physics), and states.We close this section by discussing the implementation of the representation

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 the propagation of the soil water dynamics

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

with the time t [s], volumetric water content θ_w [−] in time t [s] with respect to the θ [−], matric head h_m [m]. The solution of this partial differential equation requires the specification of material properties, namely the , unit vector in z-direction $\boldsymbol{e}_z$ indicating the direction of gravity, soil water characteristic $\theta_\mathrm{w}(h_\mathrm{m})$ and the $\theta(h_\mathrm{m})$, and hydraulic conductivity function $K_\mathrm{w}(\theta_\mathrm{w})$, which are (i) highly non-linear, (ii) varying $K(\theta)$. 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, (iii) showing hysteretic behavior, (iv) impossible to determine a priori, and (v) very expensive to measure directly. The unit vector in z-direction $\boldsymbol{e}_z$ indicates the direction of gravity, typically pointing downwards. .

2.2.2 Subscale physics

Many heuristic parameterization models exist for the soil hydraulic material properties. We choose the Mualem-Brooks-Corey parameterization (Brooks and Corey, 1966; Mualem, 1976)Brooks–Corey parameterization (Brooks and Corey, 1966) for the soil water characteristic $\theta(h_\mathrm{m})$, since it describes has been found to describe the materials in ASSESS appropriately (Dagenbach et al., 2013). Brooks and Corey (1966) parameterized the soil water characteristic $\theta_\mathrm{w}(h_\mathrm{m})$ with a saturated water content $\theta_\mathrm{w,s}$ $[-]$, a residual water content $\theta_\mathrm{w,r}$ $[-]$, a scaling parameter h_0 $[m]$ well (Dagenbach et al., 2013). This parameterization has four parameters: A scaling parameter h_0 $[\mathrm{m}]$ related to the air entry pressure ($h_0 < 0$ m), the saturated water content θ_s $[-]$, the residual water content θ_r $[-]$, 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, this the parameterization may be inverted for $\theta_\mathrm{w,r} \leq \theta_\mathrm{w} \leq \theta_\mathrm{w,s}$, leading to Inserting the Brooks-Corey $\theta_\mathrm{r} \leq \theta \leq \theta_\mathrm{s}$. 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 Mualem-Brooks-Corey parameterization 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

which includes where K_s $[\mathrm{m\ s^{-1}}]$ is the saturated hydraulic conductivity $K_\mathrm{w,0}$ $[\mathrm{m\ s^{-1}}]$ and a fudge factorτ $[-]$ in addition to the parameters $\theta_\mathrm{w,r}$, $\theta_\mathrm{w,s}$, and $\lambda$and τ $[-]$ a heuristic turtuosity factor.

Small-scale 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 $h^*_\mathrm{m}(\theta_\mathrm{w})$, $K^*_\mathrm{w}(\theta_\mathrm{w})$ $h^*_\mathrm{m}(\theta)$, $K^*(\theta)$ with a local characteristic length ratio of characteristic lengths ξ $[-]$, leads to locally scaled material functions (Roth, 1995):

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

2.2.3 Forcing

The experiment presented in this work investigates the evolving hydraulic state which is hydraulic state was forced with a fluctuating groundwater table . The boundary condition is separated into three by pumping water in or out of a groundwater

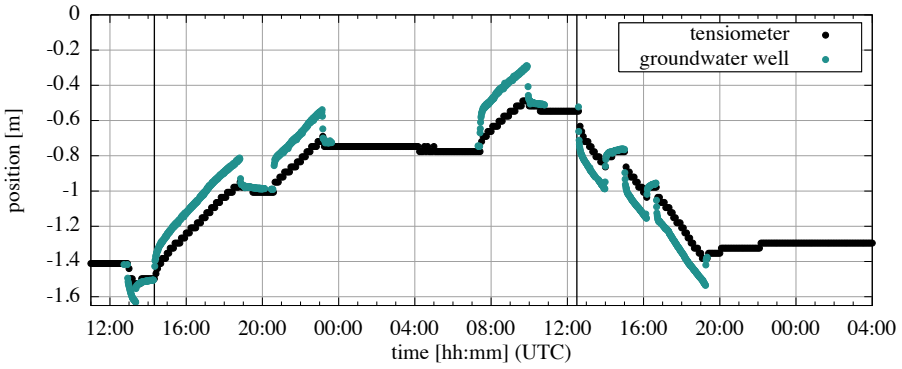

Figure 3. During The position of the experiment groundwater table was measured manually in the groundwater well and automatically with the tensiometer (Fig. 2) during three different phases (initial drainage, multistep imbibition, and multistep drainage – separated by the vertical black lines in the figure) , the position of the groundwater table was measured manually in the groundwater well (at 18.2 m) and automatically with the tensiometer (at 4.0 m)experiment. The pressure gradient between the groundwater well and the test site, i.e. the tensiometer, drives the water flux. The largest part of this pressure gradient equilibrates approximately within 5 minutes. Afterwards, the position of the groundwater table still changes, which is due to the long-term equilibration of water content distribution. Note that the discrete measurement steps reflect the resolution of the tensiometer.

well. The experiment was arranged in three different phases: (i) initial drainage phase, (ii) multistep imbibition phase, and (iii) multistep drainage phase(Fig. **??**). The position of the fluctuating groundwater table is measured manually[1] in the groundwater well (at 18.2 m) and with the tensiometer (at 4.0 m). During the multistep imbibition phase, 17.8 m^3 water were pumped into the groundwater well in 9.6 h. The equilibration steps in between were included . 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 range volume of the TDR sensors . During the multistep drainage phase, 13.9 m^3 were pumped out of the groundwater well in 5.2 h. The detailed setup of the forcing is presented in Table 2.

2.2.4 State

The experiment

The hydraulic state was monitored in particular with soil temperature, hydraulic potential , TDR, and GPR measurements . In this work, we focus on TDR and hydraulic potential measurement data. 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

[1]The position of the groundwater table was measured with a measurement band at the rim of the groundwater well.

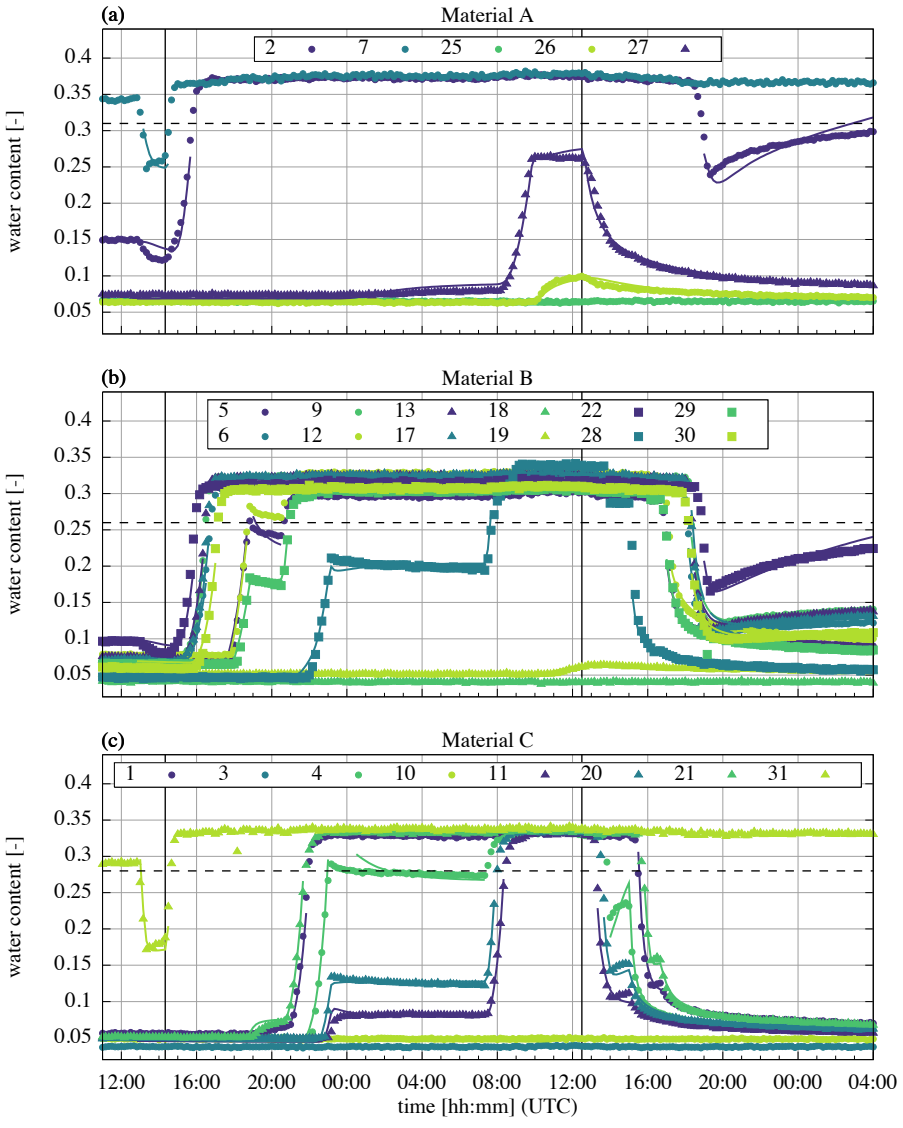

Figure 4. The measured water content data for the three different phases (initial drainage, multistep imbibition, and multistep drainage – separated by the solid vertical black lines in the figure) show a high variability up to and beyond the validity limits of the Richards equation due to the fluctuating 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).

which is due to the long-term equilibration of the hydraulic state.

The evaluated relative permittivity 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_w with θ using the Complex Refractive Index Model (CRIM) (Birchak et al., 1974):

According to **?**, we set

$$\varepsilon_b(\theta, T, \phi)^\alpha = \theta \cdot \varepsilon_w(T)^\alpha + (\phi - \theta) \cdot \varepsilon_a^\alpha + (1 - \phi) \cdot \varepsilon_s^\alpha, \quad (5)$$

with the geometry parameter α to $0.5\alpha = 0.5$. 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 1.01.0. Assuming that the sand matrix consists mainly of Quartz quartz (SiO_2) grains, the relative permittivity of the soil matrix ε_s was set to 5.0 (Carmichael, 1989). Corresponding to core-cutter measurements, the porosity 5.0 (Carmichael, 1989). Core samples of the materials A, B, and C was assumed as yielded the porosities 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

The measured

$$\varepsilon_w(T) = 10.0^{1.94404 - T \cdot 1.991 \cdot 10^{-3}} \quad (6)$$

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

The evaluated water content data of those sensors, which TDR sensors that were desaturated during the experiment , are displayed in Fig. **??**. Due to the small measurement volume (Robinson et al., 2003) and the narrow transition zone during imbibition (Dagenbach et al., 2013; Klenk et al., 2015), the 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 . It is worth noting that if the material is not saturated at the position 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 sensor, the sensors (Robinson et al., 2003). During the equilibration phases, for example after the last drainage phase (19:15), the measured water content either decreases or increases during the equilibration phases, 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 the static hydraulic equilibrium. This effect is used in the following evaluation (Sect. 3.1.3).

We attribute the spread of the water content during saturation mainly to small-scale heterogeneities and *quasi saturation* because of 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 measurement data with an air content below 0.1 0.1 (Faybishenko, 1995) are neglected subsequently.

Due to the fluctuating groundwater table (Fig. **??**), the water content measurement data for the three different phases (initial drainage, multistep imbibition, and multistep drainage – separated by the vertical black lines in the figure) show a high variability up to and beyond the validity limits of the Richards equation. 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 lines).

2.3 Structural error analysis

2.3.1 Richards equation solver

2.3.1 Orientation of ASSESS

2.3.1 Boundary condition

2.3.1 Initial state estimation

Since we will use inversion methods for parameter estimation, starting the as near as possible to the measured initial state is key. Usually, this is achieved with a spin-up phase. However, for some of our investigations, a spin-up phase would exceed the computational resources available to us[1]. 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 position of the TDR sensors with the negative distance of the sensor 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 TDR sensors – together with the approximated matric potential and the measured water content – by material. For each material, we then fit the parameters h_0, λ, and $\theta_{\mathrm{w,r}}$ of the Brooks-Corey parameterization[1] to the approximated matric potential and the measured water content (Fig. **??**) . 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

As outlined in Sect. 1, the groundwater table, we can calculate an estimation of the initial water content distribution in ASSESS (Fig.**??**).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 represen-

[1]Depending on the hydraulic material properties, the 45 h forward simulation of the 2D study presented in Sect. 3.2 took 0.25–1.0 h at low grid resolution. The parameter estimation for this case took about 3–4 days on a cluster with as many cores as parameters. A proper spin-up phase would at least cover a month, increasing the simulated time to $45\ \mathrm{h} + 30 \cdot 24\ \mathrm{h} = 765\ \mathrm{h}$. This would increase the computation time up to a factor of 17, namely 4.25–17.0 h per forward run and approximately 51–68 days for the inversion on a cluster with as many cores as parameters.

[1]The saturated water content $\theta_{\mathrm{w,s}}$ is assumed to be known from core-cutter measurement data.

tations allows to analyze (i) the resulting residuals to improve the representations and (ii) effect of unrepresented model errors on the resulting material properties.

As the parameters for the Brooks-Corey parameterization are derived from measurement data, we may also use them as initial parameter values for computationally expensive gradient-based inversions Preparing the tools for the method, we start this section with the Levenberg–Marquardt algorithm (Sect. 3.2) . The missing initial values for the parameters τ and $K_{\mathrm{w},0}$ are taken from Carsel and Parrish (1988) in this work[1]. We will refer to these parameters as *initial state material functions* in the remainder of 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 differences between the estimation and the measurements occur which indicate limitations of describing ASSESS with effective soil hydraulic material properties[1].

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 0.0017, 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.

The estimated initial water content distribution is based on the TDR measurement data (face color of the circled dots, Fig. **??**). Since the saturated water content $\theta_{\mathrm{w,s}}$ is fixed for each material a priori, only TDR sensors in unsaturated material are shown. Due to the orientation of ASSESS (Sect. **??**), 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. Additionally to those shown, GPR evidence indicates additional compaction interfaces. Note the different scales on the horizontal and the vertical axis.

2.3.1 Small-scale heterogeneity and TDR measurement volume

In order to represent the small-scale heterogeneity of the material properties, the center of each grid cell is associated with a Miller scaling factor that is initialized to 1.0. As the information about the small-scale heterogeneity only enters via the TDR measurement data, the exact position of each TDR sensor is also associated with a Miller scaling factor. For each TDR sensor, we implemented a bivariate Gaussian distribution, which determines the scaling factors in the neighborhood of this sensor. The distribution is centered at the position of the sensor, has a standard deviation of $0.015\ \mathrm{m}$ in horizontal 2.3.1) and discuss

[1] We used the parameter set *sand* with $\tau = 0.5$ and $K_{\mathrm{w},0} = 8.3 \cdot 10^{-5}\ \mathrm{m\,s^{-1}}$.

[1] Additional insight can be gained by closely investigating the structural deviation of the measured water content of TDR sensors 5, 12, and 29 from the estimation of the initial state for material B in Fig. **??**. Klenk et al. (2015, Fig. 1b and 6) presented GPR measurements, which indicate that at least TDR sensors 6, 9, 13, 17, and 22 are closely below a compaction interface and thus are experiencing a compacted pore structure. This can explain, why these TDR sensors measure smaller water content values compared to the ones measured by the TDR sensors 5, 12, and 29.

the assessment of the representation errors (Sect. 2.3.2) as well as in the vertical direction, and approaches 1.0 with increasing distance from the TDR sensor. Finally, these distributions determine the Miller scaling factor at the center of each grid cell.

2.3.1 Levenberg–Marquardt

For estimating parameters, we We employ the Levenberg-Marquardt algorithm . We include this locally convergent algorithm in a local-global approach, in order to analyze the convergence behavior or if no suitable initial parameters are available. Therefore, we generate an ensemble of the initial parameter sets with a Latin Hypercube algorithm[1]. As the sampled initial parameter sets are uniformly distributed in parameter space, the convergence path and the resulting parameter sets of the Levenberg-Marquardt algorithm contain much information regarding the convergence radius and the distribution of local minima.

Our implementation of the Levenberg-Marquardt algorithm[1] is mainly for parameter estimation. Our implementation is based on Moré (1978), Press (2007), and Transtrum and Sethna (2012) . As it additionally includes some modifications, it is sketched shortly in the followingtogether with some further modifications.

Assuming (i) M data points m_μ $(1,2,\ldots,\mu,\ldots,M)$ 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 $(1,2,\ldots,\pi,\ldots,P)$, $1,\ldots,P)$, then the χ^2 cost function is defined as

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

It implicitly This cost function assumes statistically independent random representation errors which residuals r_μ that are normally distributed with zero mean . The standardized residuals r_μ and standard deviations σ_μ (perfect model assumption). These residuals 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 $J_{\mu\pi} = \partial r_\mu/\partial p_\pi J_{\mu\pi} = \partial r_\mu/\partial p_\pi$. The Jacobi matrix is assembled numerically with the finite differences methodwhich allows for trivial parallelization of the required P forward runs. Following Press (2007), the Hessian is approximated $(\mathbf{H} \approx \mathbf{J}^\top\mathbf{J})(\mathbf{H} \approx \mathbf{J}^\top\mathbf{J})$, assuming that the second term in the derivative cancels out as $f(\boldsymbol{x}_\mu,\boldsymbol{p}) \to m_\mu$ $f(\boldsymbol{x}_\mu,\boldsymbol{p}) \to m_\mu$ with increasing number of iterations. For the Gauss-Newton algorithm then follows

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

[1]The sampling algorithm was implemented with the help of the pyDOE package: .

[1]Our implementation of the Levenberg-Marquardt algorithm is written in C++ and employs the Eigen library (**?**).

Since $\mathbf{J}^\top\mathbf{J}$ does not always have full rank, the inversion may be ill conditioned 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 $\mathbf{D}^\top\mathbf{D}$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 (1.01.0) 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 finally given as

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

The convergence path of the Levenberg-Marquardt 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. For In this work, we chose choose $\lambda_{\text{initial}} = 5.0$ and applied apply the *delayed gratification* strategy proposed by Transtrum and Sethna (2012). According to this strategy, λ is decreased by a previously chosen factor (2.02.0) if the parameter update is successful and increased by a larger factor (3.03.0) if the update is not successful.

The described gradient-based gradient–based algorithm heuristically balances performance and stability. Expanding the stability measures, we add an optional damping factor which decreases introduce a damping vector $\boldsymbol{d}$ with entries $\in (0, 1]$ to decrease the correction of certain parameters 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. This damping factor is intended for parameters representing higher order uncertainties. We use this approach in particular to estimate sensor positions and Miller scaling factors along with effective soil hydraulic properties . Therefore, we initialize sensor positions and (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 neutral values and set the damping factor for these parameters to 0.1. This reduces 1.0. Subsequently, the damping vector for the associated parameters is set to 0.1, reducing the applied correction of these parameters to 1010% of the proposed correction by the Levenberg-Marquardt Levenberg–Marquardt algorithm. Hence, the main focus of the algorithm is to estimate consistent effective soil hydraulic properties, whereas

The general setup of the parameter estimation for ASSESS (Sect. **??**) is explained with Fig. **??**. For each of the three materials, we estimate the Mualem-Brooks-Corey parameters h_0, λ, $K_{\text{w,0}}$, τ, and $\theta_{\text{w,r}}$ (Sect. 2.2.2). The saturated water content

$\theta_{w,s}$ is assumed to be equal to an estimate for the porosity ϕ based on core-cutter measurements (Sect. 2.2.4). In order to avoid parameter bias due to input errors, we estimate (i) a constant offset to the Dirichlet boundary condition (Sect. **??**) and (ii) the saturated hydraulic conductivity of the gravel layer. Depending on the setup (Sect. 3), we also estimate TDR and tensiometer sensor positions as well as the sensor positions and Miller scaling factors at the position of the TDR sensors (**??**).

The available hydraulic potential h_{wt} is measured at the position 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 $\varepsilon_b(\boldsymbol{x}_\mu, t_\nu)$ and the bulk soil temperature $T_b(\boldsymbol{x}_\mu, t_\nu)$ 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. **??**) 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 $\boldsymbol{p}^{\text{final}}$.

2.3.2 Assessment of representation errors

Quantitative learning about complicated systems is an iterative process (Box et al., 2015; Gupta et al., 2008). Starting from conceptual ideas, the modeler represents the current understanding of the system with a model incorporating decisions and underlying hypotheses (Clark et al., 2011; Gupta et al., 2012). The optimal experimental design addressing specific research objectives is based on the model and thereby on the current understanding of the system. The resulting measurement data reveal the answer of reality to specific questions posed by the experimentator. By comparing the forecast of the model with the measurement data, it can be investigated, how well the questioned behavior of the system is understood quantitatively. Thus, disagreement between the model and the measurement data reveals incorrect understanding of the system. Consequently, the concepts, decisions, and hypotheses with respect to the model (including measurement data evaluation procedures) and the measurement data themselves have to be revised. This leads to an improved model as well as improved measurement data acquisition and evaluation procedures. If the model predicts the measurement data accurately and precisely enough, the research objectives have to be expanded, such that the measurement data cover a larger part of the state space. This step is necessary, because high model complexity admittedly yields an accurate description of the measurement data, which, however, is forcedly based on biased and case dependent parameters. Ultimately, this iterative procedure leads to measurement data covering the whole state space and a statistical model-data mismatch corresponding to the measurement data error model – indicating complete understanding of reality. In general, however, such measurement data are not available and the application merely requires a limited accuracy and precision. Hence, determining the sufficient complexity of the model and the measurement data for the required accuracy and precision is the crux.By By applying the χ^2 cost function (Eq. (7)), it is implicitly assumed that

the model is perfect aside from a white Gaussian noise. This corresponds to complete quantitative understanding of reality and a Gaussian measurement data error model . Structural model-data error model for the measurement data. Structural model–data mismatch indicates that this assumption is invalid. One way to quantify this problem is to analyze the total uncertainty space with a Bayesian total error analysis (BATEA) (Kavetski et al., 2002, 2006). 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. Instead, we neglect highly complicated representation errors in the hope that if their representation is necessary, the structural model-data Hence, we have to neglect such representation errors and trust that the structural model–data mismatch will reveal thisany 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 excluded from the measurement data a priori, is parameterized and explicitly estimated. Table 3 gives an overview over the treatment of the considered representation errors. Structural deviations from the measurement data or prior estimates Remaining structural model–data mismatch or deviation from the prior for the parameters , which remain after the optimization, 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 preliminary evaluations. In order to showcase the power of the methodand the sensitivity of the fluctuating groundwater table experiment, we shortly present the results of one of those preliminary evaluations. In this case, the evaluations. To illustrate the method, we present an iteration where the orientation of ASSESS (Sect. **??**) was not yet compensated for by rotating the geometry and the gravitation vector . (Sect. A1.2). Considering the structural error analysis, we parameterized and estimated uncertain contributions to components in the representation. Hence, not only the Mualem-Brooks-Corey 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 measured positionspositions, 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 small–scale heterogeneities and eventually a non-represented 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. Being key for the identification of representation errors, a

2.3.3 Residual analysis

A visual analysis of the standardized residual increases the intuitive understanding of the model-data model–data mismatch (e.g., Legates and McCabe, 1999; Ritter and Muñoz-Carpena, 2013). Therefore, We analyze the standardized residual is visualized over time and over the 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 . The former visualization highlights the structural model-data mismatch and the latter

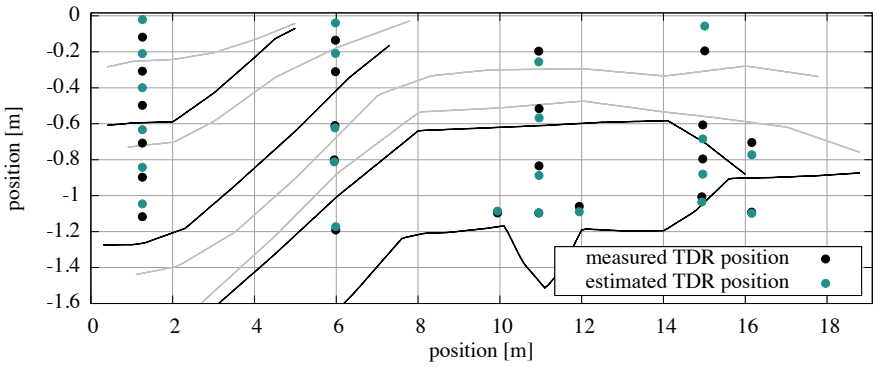

Figure 5. The subsurface architecture of ASSESS (Fig. 2) is shown with a comparison of measured and estimated TDR sensor positions based on a first evaluation of the hydraulic measurement data. The consistent deviation of the estimated TDR sensor positions reveal an unrepresented model error: The orientation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

permits easy facilitates the comparison of the standardized residual distribution to the expected Gaussian distribution . Additionallyof 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 model–data mismatch. As single measures a single measure might be misleading (Legates and McCabe, 1999), we apply (i) the calculate the root mean square error (e_{RMS}), (ii) the e_{RMS}) and the mean absolute error (e_{MA}), and (iii) the Nash–Sutcliffe model efficiency coefficient (e_{NS}) (**?**).

Comparison of measured and estimated TDR sensor positions based on an exemplary preliminary evaluation of the measurement data. The consistent deviation of the estimated TDR sensor positions reveal an unrepresented model error: The orientation of ASSESS (Sect. **??**). The black lines indicate material interfaces, whereas the grey lines indicate compaction interfaces, which were introduced during the construction of ASSESS. Additionally to those shown, GPR evidence indicates additional compaction interfaces. Note the different scales on the horizontal and the vertical axis.

2.4 Setup

In this section, we analyze the estimation of effective material properties for ASSESS based on 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 . For each of these studies, with four different setupswere implemented: : (i) ***naivebasic***: 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 *naivebasic* setup, we also estimate the sensor positions. (iii) ***miller***: In addition to the parameters estimated in the *naivebasic* setup, we estimate one Miller scaling factor for each TDR sensor. (iv) ***miller and position***: In addition to the parameters estimated in the *naivebasic* setup, we estimate both the sensor positions and one Miller scaling factor for each TDR sensor.

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 the cases are analyzed with four setups ***naive***, ***position***, ***miller***, and ***miller and position***. The more sensors per material are used in the inversion, the worse the representation of the *naive* 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.

The 1D study consists of three cases (***case I***, ***case II***, and ***case III***). Together with the material functions resulting from the ***initial state estimation*** (Sect. **??**), we visualize the resulting material functions of the best ensemble member for each setup (***naive***, ***position***, ***miller***, or ***miller and position*** – denoted by the color close to saturation). For all inversion results, the plot range is adjusted to the available water content range. 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. The main message of this figure is, that unrepresented model errors may lead to biased parameters.

2.4.1 1D study

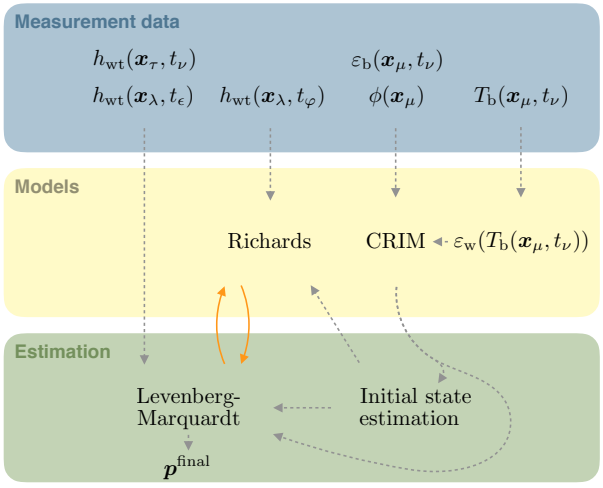

Figure 6. In this sketch, we visualize The available hydraulic potential h_{wt} is measured at the uncertainties with respect to (i) bottom of the material properties, (ii) groundwater well $\boldsymbol{x}_\lambda$ and at the TDR sensor position , (iii) of the small-scale heterogeneitytensiometer $\boldsymbol{x}_\tau$. The data set, and (iv) which is measured in the groundwater table position. During static phaseswell, these uncertainties can lead is split according to correlated estimated parametersthe measurement times: The data measured during the equilibration phases t_ϵ enter the Levenberg–Marquardt algorithm (Sect. 2.3.1) directly, e.g., whereas the data measured during the forcing phases t_φ are only used as an incorrect 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 groundwater table can be compensated by changing h_0 *and* λTDR sensors $\boldsymbol{x}_\mu$ at times t_ν. During transient phasesAdditionally using the porosity $\phi(\boldsymbol{x}_\mu)$, however, the addressed uncertain model components have distinct effects on bulk permittivity is transferred to water content (Sect. 2.2.4). The water content data enter the model, einitial state estimation (Sect. A1.6) yielding an initial water content distribution and optional initial parameter values for the Levenberg–Marquardt algorithm. g., as λ The water content data are also changes directly used in the conductivity functionLevenberg–Marquardt algorithm. HenceDashed grey arrows represent one–time preparation steps, whereas solid orange arrows represent the ability iterative steps of the parameter estimation Levenberg–Marquardt algorithm to separate these uncertain model components depends on yielding the available measurement datafinal material parameters $\boldsymbol{p}^{\mathrm{final}}$.

In order to investigate whether the 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): The (Table 4): ***case ICase I*** includes the measurement data of sensor 1 in material C and sensor 2 in material Asensor 1 in material C and sensor 2 in material A, and thus comprises one sensor per material. The ***case IICase II*** includes two sensors per material, namely the sensors 10
5 sensors 10 and 11 in material C and sensors 12 material C and sensors 12 and 13 in material B. Finally, the material B. ***case IIICase III*** includes three sensors per material, namely the sensors 25sensors 25, 26, 27 in material A and sensors 28material A and sensors 28, 29, 30 in material B. Note that material B. Note (i) that the cases are located at different positions in ASSESS (Fig.**??**).(Fig. 2) and (ii) that since the hydraulic potential is not measured in the domain covered with these 1D studies, the respective inversions are only based on the TDR water content measurements.

As described above, the analysis is organized in four different setups (*naivebasic*, *position*, *miller*, and *miller and position*). The *naivebasic* 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 [1] (material A in *case II* and material C in *case III*) are estimated for *case II* and *case IIIcase II* and *case III*. The other setups remain accordingly. We use the manually measured groundwater table data as Dirichlet boundary condition. Uncertainties concerning the position of the sensors and the subsurface material interfaces (Sect. **??**) directly translate to uncertainties in the boundary condition. Accounting for the orientation of ASSESS (Sect. **??**), we add a constant offset to the Dirichlet boundary condition for each case (*case I*: -0.02 , *case II*: -0.05 , *case III*: -0.12). 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. The forward simulations were calculated on a grid with 1×400 cells on $1.9\ \mathrm{m}$ and 10^{-8} as limit of the Newton solver (Sect. **??**). Following Jaumann (2012), the standard deviation of the TDR measurements is assumed as 0.007.Since the hydraulic potential is not measured within the domain of these Further details concerning the implementation of the 1D studies, the inversions are only based on the TDR water content measurements.study are given in Sect. A2.1.

For each of the different setups, we ran an ensemble of 20 inversions starting from Latin-Hypercube Latin–Hypercube sampled initial parameter sets . In 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 evaluation. The according statistical measures (e_{RMS}, e_{MA}, and e_{NS}) are given in Table 5. Here, we only refer to the e_{MA}, because the e_{RMS} and the e_{NS} are behaving accordingly if not stated differently.

Combining the data of all applied TDR sensors,

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

3 Results and discussion

3.1 1D study

3.1.1 Objectivity of the measurement data

The standardized residual for each case is presented in Fig. **??**.Investigating the resulting standardized residuals of Fig. 7 combining the resulting data of all applied TDR sensors. Investigating them for *case I*, it is striking that all setups describe the

[1] Material A in *case II* and material C in *case III*

measurement data qualitatively equally well. Since the estimation of the material properties is only based on one sensor per material in this case, the material parameterization offers enough freedom to describe the measurement data. Hence, it also describes accommodates unrepresented model errors, such as the sensor position and small-scale 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 measurement data, the simulated imbibition phase is too slow for sensor 1 sensor 1 and too fast for sensor 2. sensor 2. Also the simulated drainage phase is too slow for sensor 1 sensor 1 and drainage behavior of sensor 2 sensor 2 is consistently wrong. This structural model-data model–data mismatch hints at unrepresented model errors due to the restriction to a 1D domain. Yet1D domain, which is further discussed in Sect. 3.1.3. Still, the residuals of all setups are smaller than 5 5 standard deviations, which translates to 3.5 % a volumetric water content of 0.035.

We noted in section 2.3 that by applying the χ^2 cost function, we implicitly assume that the model can describe the measurement data up to a white Gaussian noise. However, this is generally not the case, because the measurement error may include a bias (accuracy and precision) and the representation might neglect processes in the dynamics, for example. Inspecting the probability plots of the three cases, we spot a characteristic *S*-shape: The slope < 1 for small residuals indicates that the precision of the simulation is smaller than the standard deviation of the Gaussian distribution with the standard deviation of the TDR measurements. The slope > 1 for large residuals shows that these residuals are larger than the presumed Gaussian distribution. Generally, the *S*-shape indicates non-Gaussian distributions. Since the large residuals are of structural instead of random nature and because the large residuals The large residuals are not random and preferably occur in transient phases, we attribute them mainly . 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 *naivebasic* 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 small–scale heterogeneity at the position of both sensors . Consistentlysimultaneously. Consequently, representing the small-scale small–scale heterogeneity improves the description of the measurement data. As before, the largest residuals occur during the highly transient phases, especially during the drainage phase. Except for two outliers, the residuals stay smaller than 5 5 standard deviations here as well. Considering three sensors per material in *case III*, the e_{MA} increases even further in the *naivebasic* setup. Consequently, representing small-scale 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 are visualized for the respective materials.Comparing the results of the for the different cases and setups (Fig. 8), we notice a vertical shift in the that the resulting soil water characteristic for functions are shifted within each material. It seems reasonable to attribute this vertical shift to

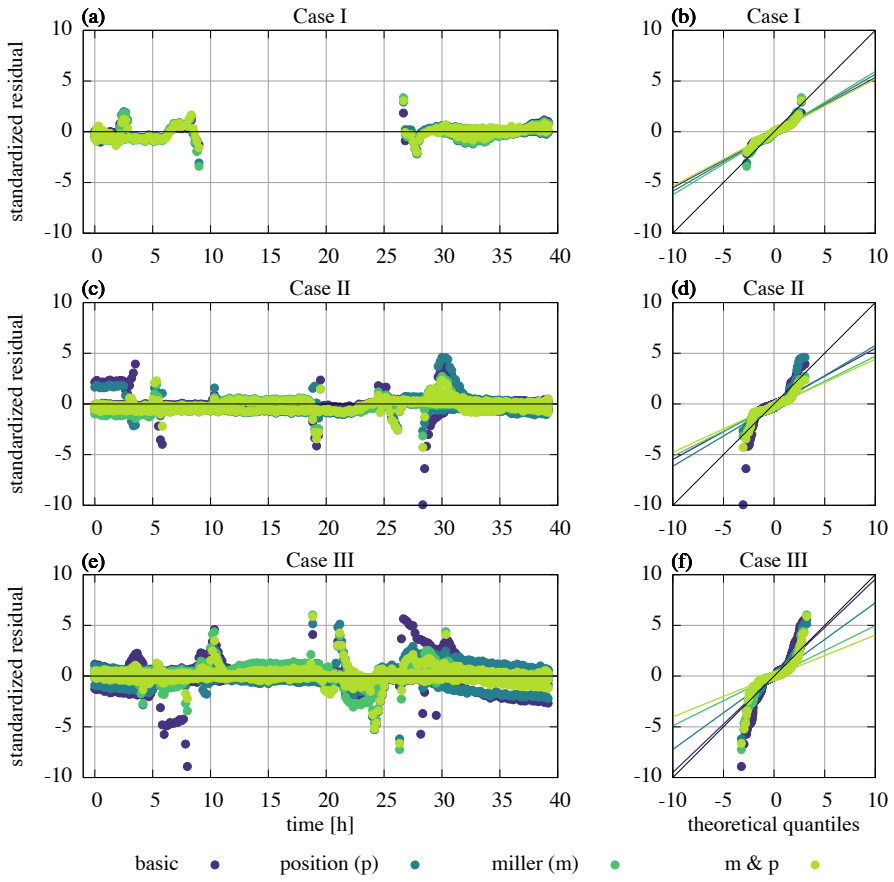

Figure 7. For the 1D study, the standardized residuals of the best ensemble member are visualized over time (left) and over the theoretical quantiles of a Gaussian with the estimated standard deviation of the TDR measurements (0.007) (right). The cases are analyzed with four setups ***basic***, ***position***, ***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.

the high number of estimated uncertain model components (Sect. 2.4), because during During static phases and if only few measurement sensors are available, these the parameters for the estimated uncertain model components (Sect. 2.4) can be correlated(Fig. **??**). . 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 , for example as the Brooks-Corey parameter λ and the Miller scaling factors also influence the hydraulic conductivity.If monitored close enough with TDR sensors and hydraulic potential measurements, the parameter estimation algorithm can separate the effects better leading to a more unique solution (Sect. 3.2). In order to further analyze this vertical shift, we (Fig. 9 and Sect. 3.2.3).

We also ran the inversions without estimating the offset to the Dirichlet boundary condition (Sect. A1.5), which are not shown here. Besides destabilizing the convergence of the Levenberg-Marquardt algorithmdue to the increased input errorLevenberg–Marquardt algorithm, this fully transfers the uncertainty in the boundary condition to the sensor position. Hence, those setups , which setups that estimate the sensor position , clearly outperform the others. It is worth noting that not estimating the offset to the Dirichlet boundary condition Additionally, this does not remove the vertical shift of the soil water characteristics.

Hence, as the given measurement data are merely sensitive to the curvature of the soil water characteristic, we will mainly focus on its curvature, e.g., when comparing the inversion results with the initial state material functions in the subsequent evaluation.

3.1.3 Lateral flow

The three cases cover the three materials at different locations in ASSESS and are based on distinct measurement data with respect to both quantity and measurement 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. **??**). Thus, also Fig. 2). To illustrate that this leads to a different sensitivity on the unrepresented model errorshave different effects. Subsequently, 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*, the at the position of TDR sensor 2, the water content increases as the sensor monitors the relaxation of the capillary fringeleading to an increasing water content. Due to lateral flow, this data includes the relaxation of the whole test site. The different characteristic behavior of the measurement data during the equilibration phase is shown in Fig.**??**. 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 the model–data mismatch during this equilibration phase, the parameter estimation algorithm increases the hydraulic conductivity to compensate for the non-represented non–represented lateral flow with additional vertical flow from above the sensor. This interpretation is supported by the fact that the Hence, the hydraulic conductivity of *case I* is larger than the hydraulic conductivity for both the *case III* and for the 2D study, which is discussed in subsequent section.Material B is in the middle of ASSESS and thus the Sect. 3.2.4.

The measurement data of material B used in the inversions of *case II* and *case III* are based on comparable measurement data. Thereforedo 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 the two setups in those setups of *case II*, in which no Miller scaling factor was estimated. These setups show a deviating larger curvature of the soil water characteristic and of the hydraulic conductivity function . This effect which is explained in more detail in the subsequent section. Regarding material CSect. 3.2.4 in more detail. Additionally, we can identify both effects – the vertical shift and the deviating curvature the previously discussed shift of the soil water characteristic . (Sect. 3.1.2).

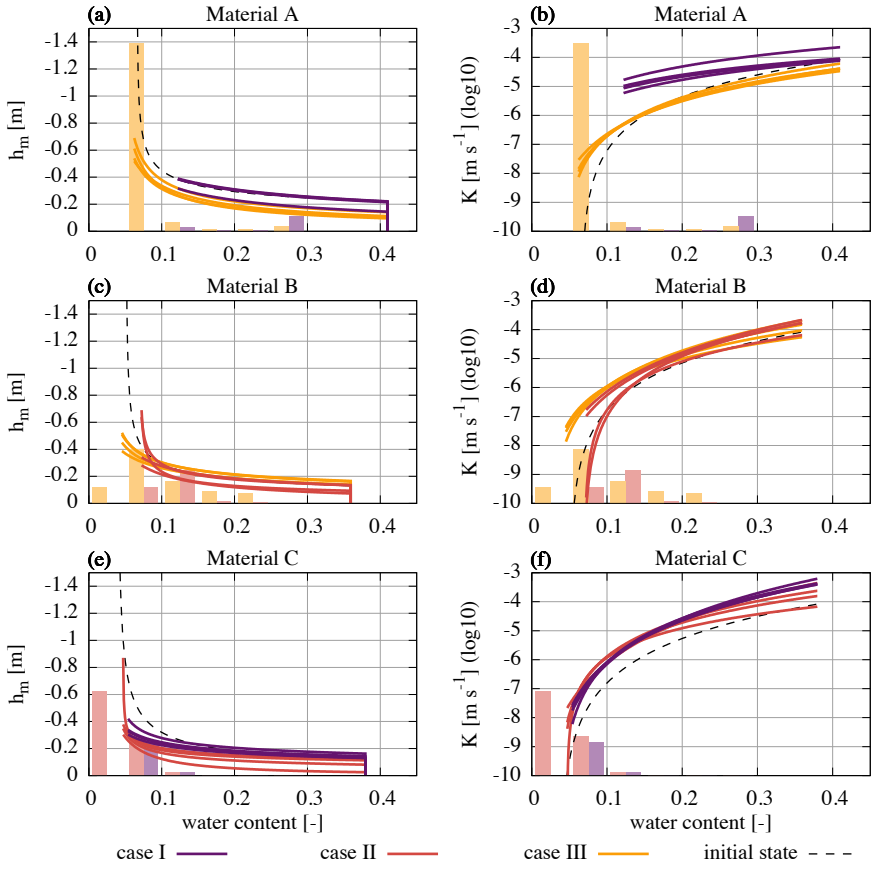

Figure 8. The estimated material functions of the best ensemble member are shown for each of the three cases (***case I***, ***case II***, and ***case III***) and the four setups of the 1D study. Additionally, we present the material functions resulting from the ***initial state estimation*** (Sect. 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.

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. Although the initial parameter sets for the 1D inversions were Latin Hypercube sampled, the

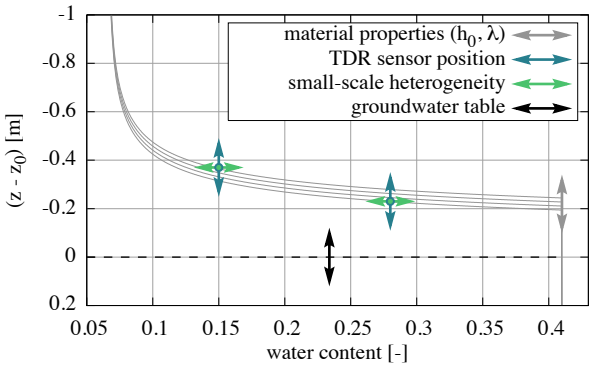

Figure 9. The estimation of uncertain model components can lead to correlated estimated parameters, e.g., as an incorrect position of the groundwater table (z_0) can be compensated by changing h_0 *and* λ during static phases. During transient phases, however, the components have distinct effects, e.g., 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.

3.1.4 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 which (Sect. A1.6), although the initial parameter sets for the 1D inversions were Latin Hypercube sampled. This allows to use the latter to initialize gradient-based the initial state material functions to initialize gradient–based inversion methods. The estimate of the initial state material function for material C material C deviates strongest from the inversion result results compared to the other two materials, since in material C 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(Sect. **??**). . Iteratively restarting the inversion using the previous inversion results as initial state material functions is likely to improve the representation. Since $K_{w,0}$ K_s and τ are prescribed a priori and are not estimated for 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.

In summary, we demonstrated that the more sensors per material are used in the inversion, the larger the probability gets to observe states (and model errors) which can not be described accurately and precisely enough with the *naive* setup. Naturally, this increases the size of the structural model-data mismatch. Hence, in order to avoid biased parameters, significant model errors have to be represented. The estimation of TDR sensor positions and Miller scaling factors constitutes a major step in this direction, as this decreased the e_{MA} by more than a factor of 2 in the case with most sensors per material. Due the low number of TDR sensors monitoring transient phases and the lack of hydraulic potential measurement data, the Levenberg-Marquardt algorithm was not able to completely separate the estimated uncertain model components. This effect becomes most

evident in the discussed vertical shift of the soil water characteristic. We found that the restriction to a 1D domain leads to an overestimation for the hydraulic conductivity function of material Adue to unrepresented lateral flow. Finally, we observed that the initial state material functions are reasonably close to the inversion results to use them as initial values for gradient-based inversion methods.

3.2 2D study

In this section, we expand the investigated domain to

3.2.1 Objectivity of the measurement data

For the 2D and analyze the performance of study, the number of sensors is comparable to the improved representation. To this end, we set up a 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 studywhich includes the four different setups *naive*, *position*, *miller*, and *miller and position* as described above[1]. Since the position of both the tensiometer and the groundwater well is within in the modeled domain, we use the hydraulic potential measurement data as well as the TDR measurement data in this study. Thus, the 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 *positionbasic* setup is adjusted, such that both the positions of TDR sensors and the tensiometer are estimated. Considering that the ensemble members of the increase, and (iii) estimating sensor positions and Miller scaling factors improve the description of the TDR data significantly. The standardized residuals confirm the last two expectations (Fig. 10). However, similar to the 1D studyconverged reasonably close to the initial state material functions, the inversions , even the residuals of the *miller and position* setup still reach more than 5 standard deviations for the 2D study are directly initialized with these parameters.The 2D simulations in this work are calculated on a grid with 100×100 grid cells covering $19.1\ \mathrm{m} \times 1.9\ \mathrm{m}$. The limit of the Newton solver is set to 10^{-8} 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 $30 - 40\ \mathrm{h}$. The largest residuals occurring during the drainage phase around $30\ \mathrm{h}$ come from the TDR sensors 6, 9, 13, and 17. We identified that these sensors are located close to a compaction interface

[1] 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 all TDR sensors. No position was estimated for sensors 7, 8, 14, 15, 16, 23, 24, 31, and 32. No Miller scaling factor was estimated for sensors 8, 14, 15, 16, 23, 24, 31, and 32.

(Sect. **??**). 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 m) is assessed from A1.6). Hence, the accuracy ($\pm$ 5 hPa) as specified by the manufacturer[1]. Lacking an independent estimate for the accuracy of the manual groundwater table position measurement, we employ the accuracy of material interfaces in ASSESS large residuals indicate that this horizontal compaction layer is not correctly represented with a point–scale 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. **??**).Same as for the tensiometer, this leads to a standard deviation of 0.025 m. 3.2.2).

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: The 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 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* setuphas more than 3 times the number of parameters[1] of the *naive* setup. The total number of evaluated TDR sensors[1] increased to 25. Hence, the measurement data cover more complicated flow phenomena compared to the 1D studies. Therefore, we expect that (i), the fits yield a slope < 1, indicating that distribution of the resulting parameters are more reliable and (ii) the description 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 worse for 2D compared to 1D. The standardized residuals visualized in Fig. **??** confirm this expectation and demonstrate that – similar to the 1D study– representing sensor position and small-scale heterogeneity uncertainty improves the description of the TDR data. Still, the probability plot (Fig. **??**b) displays a characteristic *S*-shaped curve for the TDR data highlighting persisting large residuals during transient phases . 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 *naivebasic* setup merely increases by less than a factor of 2 compared to the 2 compared to *case III* of the 1D studyand that estimating . 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 0.0034.

[1] In order to transfer the given uniform distribution with range $\pm$ 5 hPa $\approx \pm$ 0.05 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 $(0.05\ \mathrm{m} \cdot 2)/4 = 0.025$ m.

[1] Number of estimated parameters for the different setups: *naive*: 17, *position*: 41, *miller*: 42, *miller and position*: 66.

[1] We evaluated 5, 12, 8 TDR sensors for the materials A, B, C.

3.2.2 Hydraulic potential

The description of the hydraulic potential measurement data , however, does exclusively improve data only improves in those setups, in which the sensor position is estimated and the general (Fig. 10 and Table 6). Also the temporal structure of the model-mismatch model–mismatch does not change significantly . Due to the large input flux during the experiment, a correct representation of the manually measured groundwater table data is impossible in 2D. As soon as the quasi-equilibrium between the well and the site is exceededwith 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 instantiates leads to a 3D water flux . The pressure difference between the tensiometer and the groundwater table in the well (Fig. **??**) shows that the site is not in quasi-equilibrium during the forcing. Thus, we expect that the simulation predicts 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. **??**. The structural model-data 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. Since each

3.2.3 Separation of uncertain model components

3.2.4 Effect of unrepresented model errors

Each setup is started from the same initial material functionfunctions (Sect. A1.6). Therefore, the difference between the resulting material properties of the setups (Fig. **??**) (Fig. 11) is a direct consequence of the representation of uncertainties in the sensor position and small-scale heterogeneities.A more intuitive understanding can be gained for example by closely investigating small–scale heterogeneities.

To investigate this, consider the initial state estimation for material B shown in Fig. **??**. Fig. A2. The measurement data of the sensors 5, 12, and 29 which are approximately $0.6\ \mathrm{m}$ above groundwater table considerably deviate from the estimated function considerably. In order to cope with this deviation, the least squares 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 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). As soon as If the uncertainty in sensor position and small-scale small–scale heterogeneities are represented in the model, the outlying measurement data can be described without altering the effective material properties.

The 2D study is based on an increased number of water content measurements, additional hydraulic potential measurements, and a 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 the estimated uncertain model components. Solely for material A, the setups

show a vertical shift in the soil water characteristic. This can be explained with the relatively low number of water content measurements monitoring transient phases. Although the number of measurements in the dynamical water content range of material A is comparable to that of material C, Fig. **??** shows that fewer sensors monitor the transient phases in material A compared to material C.Although the It is worth noting, that although the uncertainty of the measured grain size distribution (Table 1) is large, the resulting material properties confirm these measurements to the extent, the measurements in that material A is the finest of all materials and that and the properties of materials B and C are similar.

4 Summary and Conclusions

We presented a fluctuating groundwater table experiment in a complicated and We applied a structural error analysis on a representation of the effectively 2D architecture (ASSESS), which was monitored with TDR, GPR, and hydraulic potential measurement devices. This kind of experiment provides high variability of the measured water content up to and beyond the validity limits of the Richards equation.Using inversion methods for parameter estimation, it is key to start the simulations close to the measurement data. Hence, we employed the Brooks-Corey parameterization to estimate the water content between the TDR sensors. Therefore, we assumed hydraulic equilibrium and approximated the hydraulic potential with the negative distance to the groundwater table ASSESS. Subsequently, we associated the approximated hydraulic potential at the position of the TDR sensors with the measured water content and fitted Brooks-Corey parameters for each material. With the resulting parameters, we calculated an estimate for the initial water content distribution.We implemented a structural error analysis which is based on the insight that the structural model-data This representation includes TDR and hydraulic potential measurement data which were acquired during a fluctuating groundwater table experiment. Based on the assumption that structural model–data mismatch indicates incomplete quantitative understanding of reality. We demonstrated that the method can detect significant unrepresented model errors, such as the inclined architecture of ASSESS.However, as the sufficient complexity of the model and the measurement data for the required accuracy and precision are unknown a priori, we analyzed the effect of unrepresented model errors by implementing different setups of the , we implemented a 1D and a 2D studies with increasing model complexity. In these setups, the model complexity was gradually increased starting study organized in different setups with increasingly complex models. Starting with the estimation of effective hydraulic material properties and adding we added the estimation of sensor positions, small-scale 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.

In order to investigate, whether the soil water movement at ASSESS can be described with We showed that estimated material properties resulting from a 1D model, we created three cases with increasing number of sensors per material located at distinct positions in ASSESS. For each case, we generated an ensemble of Latin-Hypercube sampled initial parameters for the Levenberg-Marquardt algorithm. Since the resulting material properties of the best inversions are reasonably close to the

parameters estimated for the initial water content distribution, these may also be used as initial parameters for gradient-based optimization algorithms. We found that with an increasing number of sensors per material, 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 structural model-data mismatch increased for those setups, in which only the effective material propertieswere estimated. Representing 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 heterogeneities, however, small–scale heterogeneity improved the description of the measurement data significantly in the cases with more than one sensor per material. We showed that also due to unrepresented lateral flow, the resulting material properties of 1D cases are likely to be biased. Since all setups 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 dynamics, 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 were initialized with the parametersestimated for the initial water content distribution, the difference between the resulting material functions show are initialized with these parameters, the comparison of the results directly display the quantitative effect of the according unrepresented model errors . Representing sensor position uncertainty and small-scale heterogeneities improved the description of the water content data significantly, as this decreased the associated e_{MA} by more than a factor of 2 to 0.0034on the estimated material properties.

Since the three approaches (i) initial state estimation, (ii) 1D inversion, and (iii) 2D inversion yield similar 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. The method does not estimate the other parameters $K_{\mathrm{w},0}$ Lacking direct measurements of the unsaturated hydraulic conductivity, the method cannot estimate the remaining 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 small–scale heterogeneities. Yet, the method is fast (few seconds on a local machine) and suitable to provide initial parameters for gradient-based gradient–based inversion methods.

The 1D inversions are comparably fast (minutes up to hours on a local machine) and can represent transient states. They allow to estimate all necessary hydraulic material parameters In contrast to the initial state estimation, 1D inversions can estimate all

parameters of the material functions. However, due to unrepresentable lateral flow , the resulting parameters are likely to be biasedmore 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. Of Therefore, we expect that of the three approaches discussed, this the closest one to reality. Therefore, we expect one yields the most reliable material propertieshere. Still, unrepresented model errors , such as including 3D flow phenomena during strong forcing, may lead to biased resulting parameters.

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

A1.2 Orientation of ASSESS

A1.3 Evaluation of TDR traces

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 $\pm\, 0.05\ \mathrm{m}$ 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 $0.015\ \mathrm{m}$.

A1.5 Boundary condition

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 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

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 accu-

racy 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., Kühne, A., Antz, B., Roth, K., and Wollschläger, U.: Observation of volumetric water content and reflector depth with multichannel Ground-Penetrating Radar in an artificial sand volume, , 6th International Workshop on Advanced Ground Penetrating Radar (IWAGPR), 2011.

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.

Guennebaud, G., Jacob, B., et al.: Eigen v3, http://eigen.tuxfamily.org, 2010.

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.and , Šimůnek, J.: Review of inverse estimation of soil hydraulic properties, Nunzio, R., and Wolfgang, D.: Simultaneous determination of water transmission and retention properties. Inverse Methods., in: Proceedings of the International Workshop Characterization and Measurement of Hydraulic Properties of Unsaturated Porous MediaMethods of Soil Analysis. Part 4. Physical Methods, edited by van Genuchten, M. T., Leij, F. J., and Wu, L., pp. 643–659, University of California, Riverside, 1999. 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. 2015.

Klenk, P., Jaumann, S., and Roth, K.: Monitoring infiltration processes with high-resolution surface-based Ground-Penetrating Radar, Hydrology and Earth System Sciences Discussion, , 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.

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

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.

Roth, K., Schulin, R., Flühler

Scharnagl, B., Vrugt, J., Vereecken, H., and Attinger, W.: Calibration of Time Domain Reflectometry for water content measurement using a composite dielectric approach, Water Resources Research, 26, 2267–2273, , 1990. 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 SciencesDiscussions, 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. 2008a.

Vrugt, J. A., ter Braak, C. J. F., Clark, M. P., Hyman, J. M., and Robinson, B. A.: Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation, Water Resources Research, 44, doi:10.1029/2007WR006720, 2008b.

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.

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 (***naive***, ***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 significantly. 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.

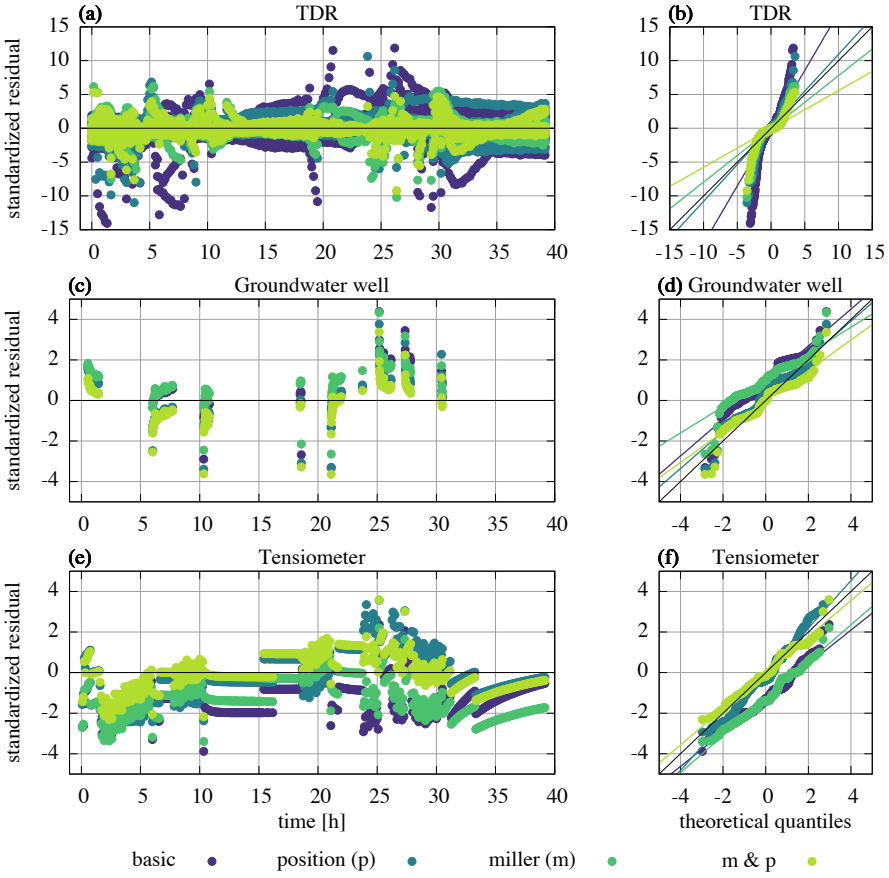

Figure 10. The standardized residuals of the 2D study are visualized over time (left) and in a probability plot (right) for all TDR and hydraulic potential sensors. The color associates the results with the four setups of the study (***basic***, ***position***, ***miller***, and ***miller and position***). 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.

We show the resulting material functions for all three materials involved in the 2D study which is analyzed with four setups (***naive***, ***position***, ***miller***, and ***miller and position***). The plot range is adjusted to the available water content range for each material. The line width of the 2D inversion results corresponds to two times the formal standard deviation of the hydraulic parameters. The height of the histogram bars denotes the number of available water content measurements and is normalized over all figures. Since the inversions for all setups are initialized with the material functions resulting from the ***initial state estimation*** (Sect. **??**), 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 visualized.

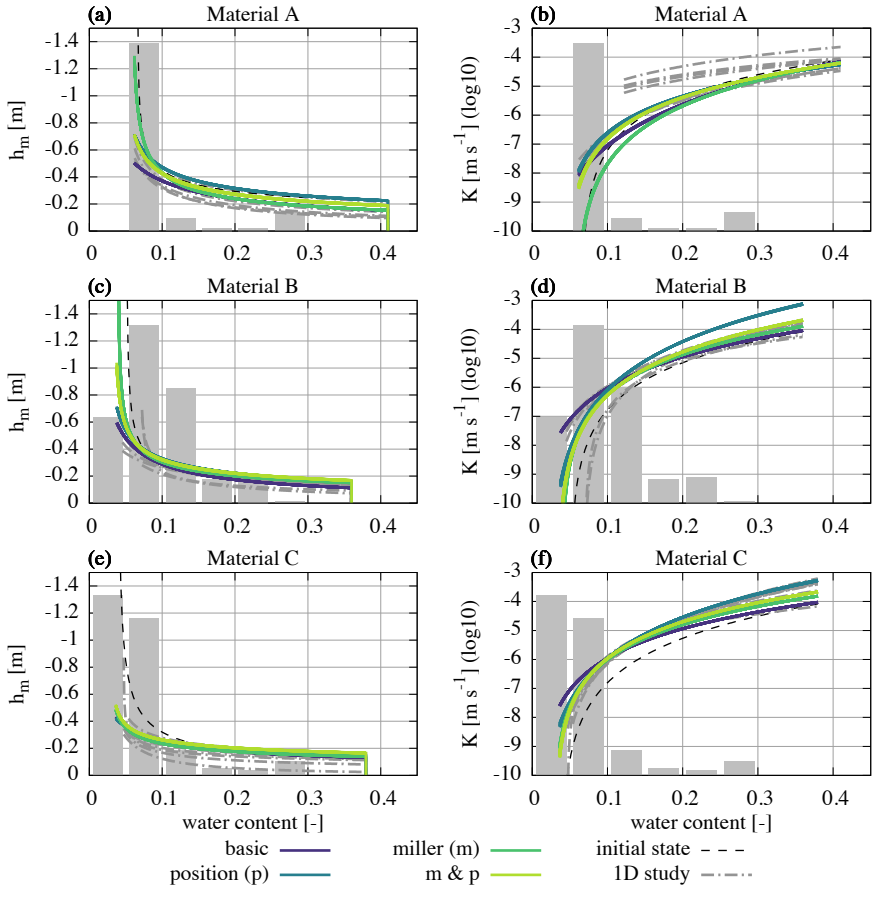

Figure 11. We show the resulting material functions for all three materials involved in the 2D study which is analyzed with the four setups ***basic***, ***position***, ***miller***, and ***miller and position***. The plot range is adjusted to the available water content range for each material. The 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.

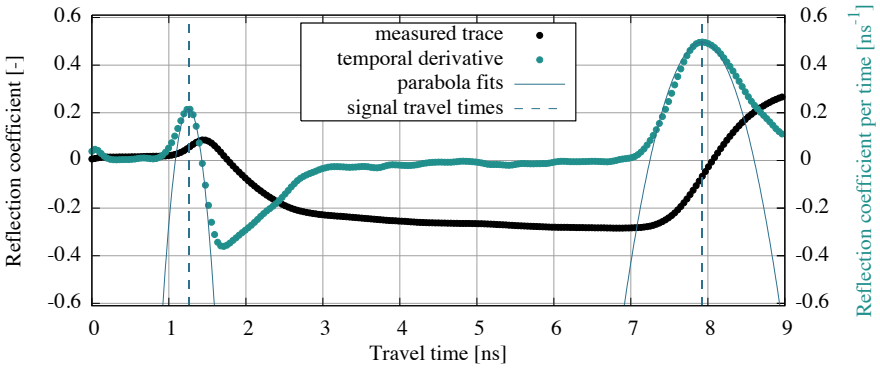

Figure A1. The evaluation of a TDR trace is based on the detection of the inflection points caused by the probe head and the end of the rod. This is done automatically after calculating of the first temporal derivative of the trace. Parabolas are fitted to the maxima of the temporal derivative to increase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

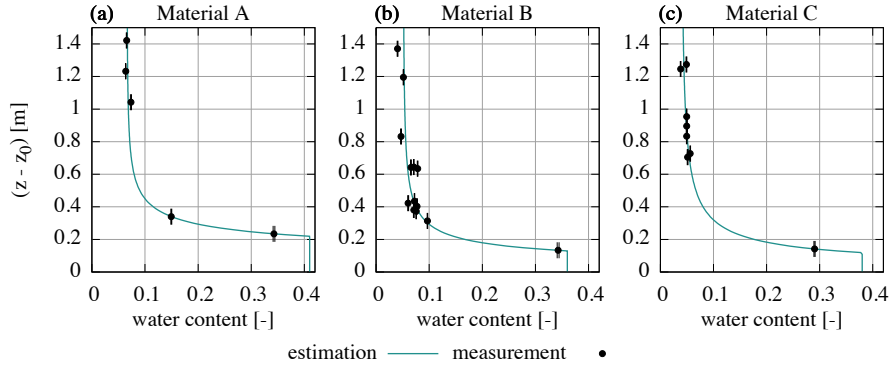

Figure A2. We use the Brooks–Corey parameterization to estimate the initial water content distribution between the TDR sensors. Assuming hydraulic equilibrium, we approximate the matric potential h_m with the negative distance to the groundwater table position z_0: $h_\mathrm{m} \approx -(z - z_0)$. 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 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.

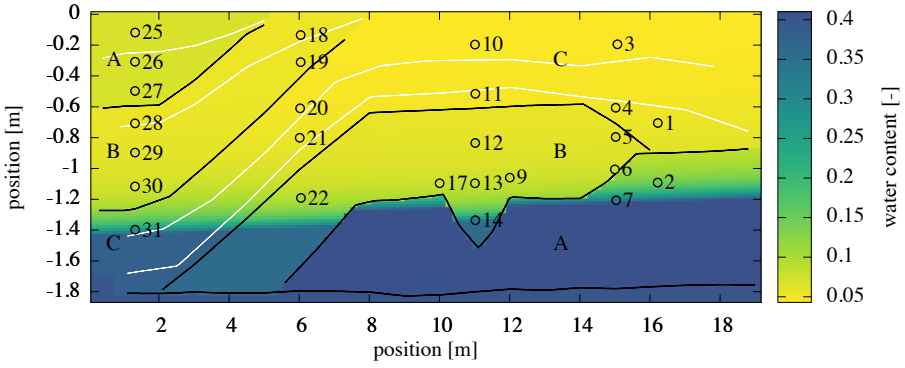

Figure A3. The estimated initial water content distribution is based on the TDR measurement data (Fig. A2, shown as face color of the circled dots). Since the saturated water content θ_s is fixed for each material a priori, only TDR sensors in unsaturated material are shown. Due to the orientation of 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.

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 g out of approximately 400 Mg were sampled. Due to rounding, the sum of the values is not always 100.100.

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

Table 2. During the experiment, ASSESS was forced with a fluctuating groundwater table. Therefore, 17.8 m^3 17.8 m^3 of water were pumped in and 14.7 m^3 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$80 m^2. All times are given in UTC.

phase	time start	time end	duration [min]	water volume [m^3]	flux [10^{-6} 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 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(Fig. 2). **??**). Note that each material is covered twice.

case	sensors	materials	position [m]
I	1 & 2	C, A	16.1616.16
II	10, 11 & 12, 13	C, B	10.9510.95
III	25, 25, 27 & 28, 29, 30	A, B	1.261.26

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 *naive 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} e_{NS}
I	naive basic		0.0043 0.004	0.0033 1.000.003
I	position	(p)	0.0037 0.004	0.0028 1.000.003
I	miller	(m)	0.0045 0.005	0.0035 1.000.004
I	m & p		0.0037 0.004	0.0028 1.000.003
II	naive basic		0.0067 0.007	0.0034 0.960.003
II	position	(p)	0.0053 0.005	0.0030 0.980.003
II	miller	(m)	0.0042 0.004	0.0027 0.990.003
II	m & p		0.0042 0.004	0.0029 0.990.003
III	naive basic		0.0090 0.009	0.0056 0.960.006
III	position	(p)	0.0062 0.006	0.0040 0.980.004
III	miller	(m)	0.0054 0.005	0.0031 0.980.003
III	m & p		0.0043 0.004	0.0023 0.990.002

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 hydraulic potential which describe 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_{NS} e_{RMS}	e_{MA} e_{NS}
naive basic		0.0156 0.017	0.0099 0.011	0.92 0.04	0.036 0.030 0.990.03
position	(p)	0.0098 0.011	0.0063 0.006	0.97 0.02	0.028 0.023 0.990.02
miller	(m)	0.0073 0.008	0.0047 0.005	0.98 0.03	0.036 0.031 0.990.03
m & p		0.0059 0.006	0.0034 0.004	0.99 0.02	0.022 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 $10^{-0.728\pm0.006}$ m s^{-1} and -0.034 ± 0.001 m, respectively.

material	$0.018\ h_0$ [m]	1.00λ [−]	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