Responses to Editor's and Reviewers' comments

Editor

Dear Authors,

Your submission is close to the final acceptance. The revised paper has received additional, but useful (I guess), comments from reviewers. Please, take a look at these comments and see if they can still be included so as to improve or make clearer a few parts of your study.

Send me your review and a document explaining the changes done. Should you disagree with some comments, or think a change is unfeasible, please explain why.

I look forward to receiving your documents.

Response: We thank the Editor for appreciating our revisions and the suggestion to accept the paper after minor revisions. We thank all the Reviewers once more for their comments and suggestions, which further improved our manuscript.

Our revisions in response to Reviewers' comments are described and justified below. We indicate the line numbers of the marked manuscript.

Ty Ferre (Reviewer #1)

I want to start by saying that I really like what the authors have done. They have taken an unusually comprehensive and well-characterized data set and subjected it to improved inverse analysis. I have no objection to this paper being published as is ... except that it was not at all what I was expecting based on the title.

The fundamental problem here is that there is no good way to measure recharge flux. As a result, all that the authors can do is to look at the statistics of their predictions. How can a reader trust that the estimated fluxes are accurate based on the finding that your inverse method gave small uncertainties (in some cases) compared to the mean predicted values? Again, this is the way that it is right now ... we don't have a valid ground truth. But I would have been much more comfortable with the paper if the authors had presented it as a field study of recharge and then commented on the strengths and weaknesses of their selected analyses.

Not to overstate it, but what is the purpose of highlighting 'reverse hydrology' in the title. The word 'reverse' only shows up one more time in the paper and in an entirely different context! If the paper were refocused very slightly to better represent what (I think) it is, I would suggest accepting as is!

Best

Ty Ferre

Response: We thank Ty Ferre for appreciating our revisions and the helpful feedback concerning the title! We have adapted it to “Estimating vadose zone water fluxes from soil water monitoring data: a comprehensive field study in Austria”. This leaves out the term “reverse” and makes the scope of the study clearer.

Jasper Vrugt (Reviewer #4):

Review of "From soil water monitoring data to vadose zone water fluxes: a comprehensive example of reverse hydrology"

I have been asked by the Editor to provide a re-review of this paper. I looked at the first round of comments of the other reviewers and the document with track changes. The paper is generally well written and addresses an important topic in hydrology, namely the quantification of groundwater recharge rates and their associated uncertainty. The paper makes a useful contribution. I recommend a major revision. I list my comments - not in any particular order.

Response: We thank Jasper Vrugt for reviewing our manuscript and sharing his expertise on Bayesian inference. We have considered all comments which have been very helpful for improving the paper. In the following, we address our changes based on the comments and/or justify our choices.

0. Reverse hydrology? We have hydrology backward; inverse. Reverse hydrology is catchy but I personally would stick to jargon of inverse. Also, I am not so sure that the example is very comprehensive; comprehensive in analyzing different sites, but not comprehensive in numerical modeling, inverse estimation, and uncertainty quantification. I'll discuss this further below.

Response: We have changed the title to “Estimating vadose zone water fluxes from soil water monitoring data: a comprehensive field study in Austria” to refer the term “comprehensive” to the multiple sites representing different hydrological conditions and to leave out the term “reverse hydrology”.

1. Line 125: Units of S are missing

Response: The units have been included (Line 126).

2. Line 148: "The vast majority of the soil profiles indicated a distinct topsoil overlying deeper soil layers that had low to mild degrees of inhomogeneity" How did you determine this? Soils that may appear homogeneous visually, can be highly heterogeneous.

Response: We determined this from the available soil water measurements and profile information (texture data and soil horizons) established by/for the Austrian ministry who operate the soil water monitoring network. We did not rely on a visual assessment. In Lines 149-151 we clarified this: **“The available soil water measurements and profile information (texture data and soil horizons) indicated a distinct topsoil overlying deeper soil layers with low to mild degrees of inhomogeneity at the vast majority of the soil profiles.”**

3. Line 167-169: "For inverse parameter estimation during the half-year calibration periods, as well as for the model validation periods, we chose boundary conditions with respect to the conditions at the measurement plots, i.e. seepage face for the lysimeter sites and free drainage for sites with natural field conditions."

First of all, you can remove the word "inverse" in front of parameter estimation. Secondly, I would argue that a six-month calibration period may be too short to get a) an accurate characterization of the soil hydraulic properties - let alone their uncertainty, and b) to remove the dependence of the initial soil moisture state (=wetness of profile) and the resulting parameter estimates. This is a serious issue and authors need to demonstrate that their parameter estimates are not too dependent on the initial wetness; otherwise the inference and uncertainty estimates depend on the choice of the initial state. Not desirable.

Response: We removed “inverse” from the sentence (Line 169).

- We restricted the calibration to half-year since most of the sites are influenced by snow during winter. Snow simulation and parameterization of the snow routine introduce additional numerical burdens with more frequent non-converging model runs as well as additional complexity and potential biases in the calibration. Also, the use of spring-summer months, which have an alternation of wet-dry periods, is expected to increase the informativeness of soil water measurements.

- We used a model spin-up period of two months to relax the effect of initial conditions on the estimation procedure. We apologize for not including this detail in earlier versions of the manuscript. We have now mentioned it in the text and made the justification of our choice of calibration periods clearer, see Lines 105-111: **“The length of calibration periods was chosen to be similar for all sites, long enough to be informative for a range of soil water conditions. We excluded the winter season requiring the simulation of snow accumulation and melt processes as it increases the computational cost and numerical sensitivity of the simulations and introduces additional complexity and potential biases in the calibration. The use of spring-summer months, which have an alternation of wet-dry periods, is expected to increase the informativeness of soil water measurements.”**

4. Line 182: Bayes theorem ... (remove "The")

Response: We removed “The” before “Bayes theorem” in Lines 70 and 184.

5. Why does Equation 7 not appear after Line 182-183? Unusual

Response: We shifted the equation to the indicated lines.

6. Some call P(D | M,Ω) the data likelihood but this is really the conditional probability as the parameters are assumed given (appear on right hand side of "|"), likelihood you write as L(Ω|D,M).

Response: We changed the description in Line 189: **“… P(D | M,Ω) is the conditional probability of the data given the model and parameters…”** and changed the notation of the Likelihood to L (Ω|D,M).

7. Measurements errors are assumed to be IID and lead to the standard normal likelihood. a) These assumptions are questionable at best and should be verified a-posteriori using diagnostic checks of the residuals (histogram of residuals, ACF and QQ plots); b) the posterior parameter distribution is strongly dependent on the choice of likelihood function - and, thus, the uncertainty estimates of the recharge rates are suspect. I would strongly recommend using a distribution-free likelihood function instead. This will adapt to the residual properties at hand; hence, satisfy residual

assumptions made. For example, check: https://www.sciencedirect.com/science/article/pii/S002216942201112X

The universal and generalized likelihood functions are your best bet to getting the most accurate estimates of recharge uncertainty.

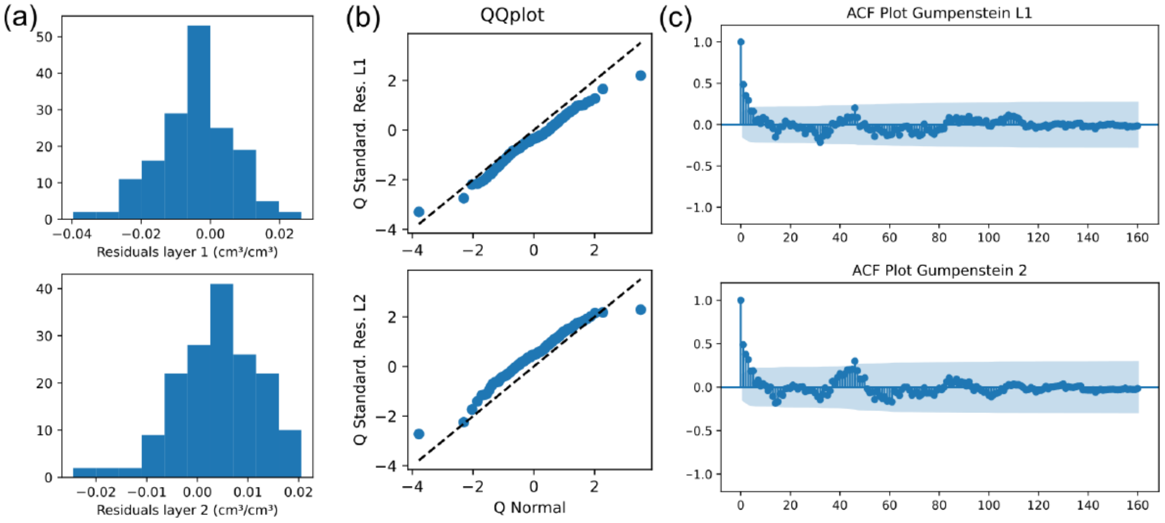
Response: We thank the Reviewer for suggesting these new likelihood functions. Here are the main justifications for our choice:

- The calibration procedure includes only volumetric water content measurements from TDR sensors. While the deterministic part of the measurement signal is correlated, the stochastic part (i.e., measurement error) is not as it is based on an electromagnetic instantaneous pulse. The correlated part of the signal is implicitly described when numerically solving the Richards equation, as the variables (theta and h) at time step t+dt are solved using their counterparts at time t. Therefore, we decided to use a more physically realistic likelihood to carry out a process-based probabilistic inference. Significant discrepancies between model predictions and observations are used as indicators that the model structure needs to be improved. On one hand, we agree that this approach might be restrictive, however, on the other, we think it might better target model inadequacies, which can be masked when using other likelihoods. Nevertheless, we thank the Reviewer for suggesting the universal likelihood. This is certainly something we want to explore in future studies!

- We have added a plot showing the diagnostic checks of the model residuals as example for Gumpenstein (layer 1 upper graphs; layer 2 lower graphs) where we do not see severe violations of our assumptions.

In Lines 197-199 we added: **"We used volumetric water content measurements from TDR sensors in the calibration where the measurement error is based on electromagnetic instantaneous pulses and can be assumed to be independent, homoscedastic, and normally distributed. This leads to a Gaussian likelihood function […]"**.

In Lines 204-210 we added: **"The choice of likelihood function is critical to the outcome of Bayesian inference and is the subject of ongoing debate. A recent promising approach that should be explored in future studies is the universal likelihood proposed by Vrugt et al. (2022). Instead of making prior assumptions about the distribution of model residuals in the likelihood function, this approach is distribution-adaptive to the actual residual properties. However, in the present study, we used the Gaussian likelihood function as described above for process-based probabilistic inference, where we use significant, systematic discrepancies between model predictions and observations that violate our assumptions as indicators that the model structure needs improvement. We show the residual checks as example for the location Gumpenstein in the Appendix (Fig. A1)."**



8. Subscripts that are acronyms should not be italic. "s" in theta_s, should be upright, otherwise it is considered a variable. This comment applies to all super/subscripts in the paper.

Response: We changed the subscripts in question to non-italic.

9. The authors use the MULTINEST sampling algorithm - I do not want to be difficult, but I would recommend the authors to have at least a quick look at the DREAM algorithm. This has been developed within the context of hydrologic problems - and is much better benchmarked than the MULTINEST algorithm. In fact, I am not sure if this algorithm has ever been used for hydrologic problems ; if so then it is important to show that it actually infers the correct parameter distributions - it should, but this is not a guarantee. Note that DREAM can also handle multimodal surfaces. This is demonstrated in theory and practice in related manuals and papers. The DREAM algorithm also provides estimates of the evidence - see Volpi et al. https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2016WR020167

Response:

- The MULTINEST algorithm has been tested and benchmarked with hydrological models in previous studies (https://doi.org/10.1016/j.watres.2020.115973, https://doi.org/10.1016/j.jhydrol.2020.124681), (https://doi.org/10.1016/j.jhydrol.2018.06.055). In Schübl et al. 2022 (https://doi.org/10.1016/j.jhydrol.2022.128429) we tested MULTINEST with artificially generated data and similar HYDRUS models, where the algorithm reliably inferred the true parameter values as well as standard deviations of the artificial errors in the calibration data. This is mentioned in Lines 216-218 (**"The Nested Sampling algorithm as proposed by Skilling (2006) has been used successfully for parameter estimation and uncertainty quantification in studies with non-linear hydrological or biogeochemical models (Brunetti et al., 2020a; Elsheikh et al., 2013)."**) We further added in Lines 218-220: **"It has been tested in Schübl et al. (2022) with synthetic data scenarios for SHP estimation**

with similar HYDRUS models where it reliably inferred the true parameter values as well as standard deviations of the artificial errors in the calibration data."

- Allison et al. (2014) (https://doi.org/10.1093/mnras/stt2190) compared Nested Sampling, classic MCMC Metropolis-Hastings, and the affine invariant MCMC ensemble sampler, which shares multiple similarities with DREAM and DE-MC. They found that Nested Sampling delivers high-fidelity estimates for posterior statistics at low computational cost, and has comparable accuracy to MCMC techniques.

Therefore, based on our experience and existing studies, we trust that MULTINEST can provide reliable estimates of the posterior distribution.

10. Per my earlier point; The uncertainty estimates of Table 1 are strongly dependent on the likelihood function and, possibly, the choice of initial conditions. I believe this paper would be substantially stronger if this were investigated in more detail - and the likelihood function traded for a distribution-free formulation.

Response: Please refer to the previous responses.

11. Fig. 2b: I am a bit surprised that the posterior distribution of theta_r is relatively well defined. The soil moisture observations in the left plot show that the profile is quite wet during the calibration period; soil moisture values do not go lower than about 0.22. This is much larger than the residual moisture content, meaning that there will be hardly any information in the soil moisture observations for estimating theta_r. Hence, I would expect a much larger posterior uncertainty of this parameter - extending over almost its entire prior parameter range; unless the range of alpha and n are chosen so that high values of theta_r are discouraged. Certainly, theta_r plays a role in characterizing soil moisture flow at low moisture contents.

Response: In the old version of the manuscript, Figure 2(b) was a summary of the resulting relative uncertainty ranges for the parameters at all 14 sites while 2(a) on the left showed an example for the calibration only for the site Gumpenstein. To avoid confusion, we separated these two figures in the newly revised manuscript (Fig. 2 and Fig. 3). We agree that given the wet calibration period at Gumpenstein we expect theta_r to show a large posterior uncertainty. As given in Table 1, the uncertainty ranges for theta_r in Gumpenstein (according to the 95% confidence intervals) were 1.3 – 7.8% in the top soil layer and 1.7 – 10.1% in the bottom soil layer. The uncertainty ranges were thus quite large. The respective prior ranges for alpha and n are given in the manuscript (0.0001-0.5 and 1.01-2.70, respectively); they were not chosen to discourage high values of theta_r.

12. On a related note, my personal experience suggests that MCMC-HYDRUS sampling is difficult due to the numerical errors of HYDRUS - this results in very low acceptance rates; requiring an efficient MCMC method to traverse the many pits in the response surface introduced by the numerical errors of HYDRUS. What is the acceptance rate of MULTINEST? How many posterior samples do you have? And how do we know that the algorithm has formally converged? The advantages of multi-chain methods such as DREAM is that you can much better assess convergence of the chains by looking at the within and between-variance of the parameters (univariate scale reduction factor). The multivariate scale-reduction factor compares the covariances as well.

Response:

- HYDRUS errors: Numerical errors from HYDRUS that can lead to difficult posterior sampling are: 1) numerical diffusion, 2) non-convergence due to improper settings, 3) mass balance. The numerical diffusion was limited by adopting a relatively fine mesh, refined at the top to accommodate pressure gradients induces by atmospheric conditions. A crucial value to reduce the number of non-convergent HYDRUS runs is hCritA. If set too high, it will lead to floating precision error in the solver and non-convergence. We implemented a subroutine that set this value based on the soil hydraulic parameters proposed by the Bayesian sampler. In particular, hCritA is set equal a pressure that leads to a volumetric water content slightly higher than the residual water content. This drastically reduced the number of non-convergent runs. Finally, a large negative log-likelihood value was attributed to simulations affected by high mass balance error (>5%).

- Nested Sampling uses a different sampling approach than a classical MCMC scheme, therefore the acceptance rate does not have the same meaning here (it is supposed to decline with each iteration as the sampling approximates the bulk of the posterior). With Nested Sampling, the convergence is monitored via the accumulation of the evidence integral and the remaining prior volume. We describe this in lines 232-236 of the manuscript: **"At each iteration of the algorithm, the current maximum likelihood sample point is multiplied with the remaining prior volume to estimate the maximum remaining volume of the BME integral. Sampling is then terminated according to a tolerance (convergence) criterion, which defines when the remaining contribution from the current live points to the integral is considered to be small enough. At this point, it is expected, that the bulk of the posterior has been sampled sufficiently. The tolerance parameter in this study was set to 0.5"**. More detailed information is given in the papers by Feroz et al. (https://doi.org/10.1111/j.1365-2966.2007.12353.x, https://doi.org/10.1111/j.1365-2966.2009.14548.x).

- The number of posterior samples depends on the algorithm convergence, which was different for each monitoring station. On average, 4100 samples were used to characterize the posterior, which was randomly sampled 100 times to propagate the posterior uncertainty in model simulations. We have added further details, see Lines 237-240: **"The number of posterior samples provided by MULTINEST depends on the algorithm convergence with each model. On average, we obtained 4100 posterior samples and corresponding sample weights to characterize posterior parameter distributions. We used 100 random samples from the posterior to propagate parameter uncertainty in the model for long-term simulations to quantify the resulting uncertainty in recharge simulations."**

13. Line 306 - 309: "Overall, the validation of the models was acceptable with RMSE values ranging between 0.014-0.067 cm3 cm−3. Scatterplots including the coefficients of determination R2 (0.34– 0.98) for the validation period are shown in Fig. A3 in the Appendix." How did you determine that validation behavior was acceptable? I find the RMSE of 0.067 quite large; much larger than the measurement error of the data. As the authors discuss, this is a result in part of

measurement errors of rainfall / boundary conditions; This reiterates the importance to evaluate the likelihood assumptions using diagnostic tests of the residuals.

Response: We removed the word “acceptable” from this sentence and rephrased it to **“Overall, in the validation periods RMSE values ranged between…”** (Lines 323-324). We agree that 0.067 cm^3/cm^{-3} is a quite large error in the validation. It was found for the lysimeter site Pettenbach. We discussed the reasons for this specific case in the manuscript (Lines 318-322): **“At the Pettenbach lysimeter station, a crop rotation including fertilization was applied. It is possible, that this affected soil properties, which were assumed to be constant in the modeling. For example, Lu et al. (2020) showed in their review that root growth and decay can alter soil hydraulic properties; Whalley et al. (2005) found, that growing different plants had a significant effect on the porosity of the soil aggregates, and Schjønning et al. (2002) observed the development different pore systems in soils depending on crop rotation and fertilization.”**

14. Figure 5 and 6 document only the impact of parameter uncertainty on the bottom boundary flux. But what about model uncertainty? This will make the credible intervals much larger. I think it is worthwhile to consider model uncertainty as well.

Response: We agree that it would be very interesting to comprehensively assess parameter and model structural uncertainty, e.g. in the framework of a Bayesian Model Averaging analysis (BMA) with multiple soil hydrological models/model structures. It is true that in this work we address only the propagated parameter uncertainty originating from the inverse estimation with one model. For an analysis including multiple models/ model structures we would have to focus on one or few study sites. Our focus here was to evaluate and compare soil water fluxes and parameter uncertainties from 14 different sites with the same estimation technique. We discussed this and other limitations of our study in the manuscript in Lines 355-372.

15. Section 3.3 is a nice part of this paper - trying to relate what has been found to soil properties, etc.

Response: We thank the Reviewer for appreciating this part of the manuscript!

I hope these comments are useful to further improve this paper,

Jasper Vrugt

Irvine, Feb. 15, 2023

Response: We thank Jasper Vrugt again for appreciating our work and the useful comments which helped improve our paper!