We would like to thank the reviewers for their competent and constructive reviews. Both point to some crucial issues regarding the manuscript. They clearly indicate where the paper needs to get improved such that the approach and the conclusions are sufficiently clear to the reader. In the following we would like to give some general answers to the major points of critique before we address each single topic individually.

General answers:

Both reviewers disagree with our statement that a more complex model is needed for better predictions. However, as this unanimous statement is somewhat superficial and the critique has several aspects we would like to differentiate:

i) Semantics about model complexity. Reviewer 1 argued that there was no need for a more complex model and suggested at the same time a model modification by a direct coupling of the saturated and the unsaturated zone in the model. The concept by (Seibert et al., 2003) was provided as a simple model for such a purpose. We agree very much with the proposition that a concept as described in the cited paper could be a useful strategy for further development of the model. However, we consider that a more complex model because the saturated and the unsaturated zone are dependent on each other. Hence, we agree with Reviewer 1 regarding the content on how the model should be developed in the future.

ii) Lack of evidence for more complex models: We agree with both reviewers that we did not provide strong and explicit arguments why we think a more complex model is required to simultaneously represent water table levels and fluctuations as well as discharge behavior in a satisfactory manner. Our approach was to use a model that is as parsimonious as possible (with the background of a possible practical application which necessitates fast computation and low data requirements). It turned out that model probably was too parsimonious. Our main arguments why we think a more complex model is needed for a satisfactory prediction of saturated areas are listed below. We will add them explicitly in the revised manuscript.

a) The current model neglects the effects of different antecedent soil moisture contents on the increase of the water table during different events. In reality however, the amount of water needed to reach saturation depends on the degree of saturation. This can only be achieved by some sort of water retention curve that is introduced into the model.

b) Because crops may differ strongly in their water requirements at any given moment accounting for antecedent soil moisture may require the inclusion of crop specific water abstraction from the unsaturated zone. This may be actually a reason why the relative responses of the water table in different piezometers relative to each other differed between events (see e.g. piezometer 5 and 8 in Fig. 7).

c) A dominant feature of the hydrology in the catchment are the tile drains. Because most of the area has no direct connection to stream, i.e. surface runoff cannot directly reach the stream (Doppler et al., 2012), we know that most of the discharge reaches the water course through tile drains (see below). At the same time, we know that the water table is often quite close to the soil surface despite the efficient drainage of the water through the soil. The current model version drains the soils too strongly. In order to get as much water through the soil while keeping the water table at a higher water level it is obvious that the water flux has to increase more strongly with the water table than described by Eq. [20]. Conceptually, this could be achieved by adding an additional preferential flow component that depends in a non-linear manner from the current water level.

d) Test of simple model predictions. The reviewers are right in that it is worth testing whether other simple models could perform better. Reviewer 2 referred to Topmodel as a logical candidate given the fact that we mentioned the (apparent) success of the wetness index (but see below). Because the spatial wetness distribution according to Topmodel is simply dependent on the distribution of the wetness index and the average moisture level in the catchment, one can test Topmodel predictions even without running the simulations. According to e.g. (Blazkova et al., 2002) one can calculate the depth to the water table as follows:

$$z_i = \frac{\overline{D} - m\left(ln\left(\frac{a}{tan\beta}\right) - \overline{ln\left(\frac{a}{tan\beta}\right)}\right)}{\Delta\theta} = \bar{z} - \frac{m}{\Delta\theta}\left(ln\left(\frac{a}{tan\beta}\right) - \overline{ln\left(\frac{a}{tan\beta}\right)}\right)$$

The second term is a constant which means, that at each location the there is a constant offset to the mean depth of the water table in the area that depends on the deviation between the local wetness index and its average value. Accordingly, one would expect the depth of the water table to be correlated with the local wetness index. Figure 1 depicts the data for the piezometers. It is obvious that for low index values there is a large scatter of the data and for high index values the water table hardly depends on the index. Based on these we conclude that a simple wetness index based approach does not outperform our model approach.

This conclusion is supported by the comparison of the dynamics of the different piezometers. According to the topmodel approach, the water table dynamics at different topographical positions should simply be shifted in height of the water table (if one assumes a constant drainable porosity with depth). Figure 7 in the manuscript however, reveals that the dynamic varies substantially between the piezometers. Our approach mostly failed to reproduce these differences; conceptually topmodel will do so as well.

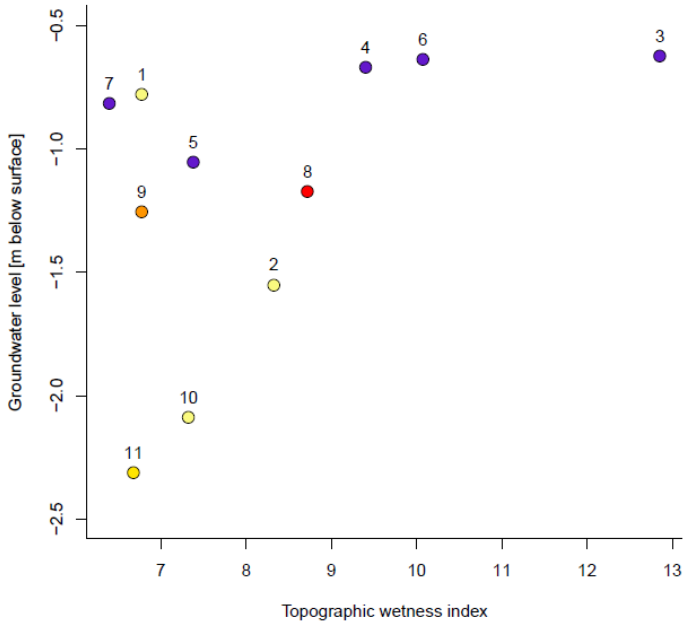


Fig. 1: Average groundwater depth in the piezometers as a function of the local wetness index. Numbers refer to the piezometer number, colors to the corresponding soil water regime class.

iii) In addition to these arguments we will reword the conclusion regarding model complexity in a more prudent form. We cannot prove that no simple model can reconcile the simulation of the discharge and water table dynamics. Hence, we will state that based on our model results and the arguments listed above a more complex model is strongly suggested.

A second major critique (put forward by reviewer 2) relates to the quality of the calibration procedure in the context of multi-objective calibration. Again, we fully understand that the reviewer criticized this aspect in the paper. We have been very short on some aspects and definitely need to explain the chosen approach in a better way and properly discuss the implications. As we understand the critique raised by Reviewer 2, the issue of calibration consists of two parts. The first deals with the question whether the algorithms used are appropriate for the complex objective functions we have used (see (Efstratiadis and Koutsoyiannis, 2010)).

We are confident that the calibration approach (a sequential algorithm composed of particle swarm optimization as a global search and a common simplex search as a local refinement method) was able to find the global optimum in the constrained parameter space. Certainly, as this algorithm includes a random component, there is no guarantee that the true optimum has been found. This is common to all global optimization methods. However, the optimization was carried out several times (upon slight modification of the deterministic model) prior to the final calibration. There was no instance when the calibration converged to a parameter set that described water table fluctuations more successfully. In the final calibration about 28’000 different parameter combinations have been evaluated during the optimization process, so our confidence has solid probabilistic foundations.

The second part of this issue is the relative weighting of the different variables (discharge and water table levels) considered in the optimization procedure. It could be argued that the poor model performance regarding water table levels does not reflect a principle limitation of the model structure but a wrong subjective attribution of weights to the water table data in the aggregating likelihood function (Efstratiadis and Koutsoyiannis, 2010). Because we used such a single function we did not explicitly quantify the trade-offs between discharge and water table simulations. Such an imbalance could be caused

i) By an unbalanced number of data points for the different variables
ii) Or a disparity between the standard deviations attributed to the two variables in the likelihood function (Eq. 22 in the manuscript).

The first possibility is excluded because the actual number of piezometric data exceed the discharge data by a factor of eleven (p. 12925, L. 7 – 9). The second possibility was avoided by the joint calibration of the standard deviations. Hence, if there was a parameter set that performed well on the water table levels – causing the standard deviation to be small - it had outperformed data sets that performed only well on discharge. Based on this argument we don't think it makes sense to run additional calibrations with this model structure (as requested by Reviewer 2).

Overall, we think that the results of this study are of interest to the hydrological community since they show how one can make more use of the spatial information contained in soil maps for model evaluation and what can be learned from it. Below we answer the reviewer comments in more detail.

**Reviewer 1**

*The paper to me (and I invite the authors to consider this) should be all about the model development and testing and the usefulness of additional soil data (with the procedure described to derive soil saturation frequencies) for model calibration with the potential to be applied for practical delineation of CSAs in the context of herbicide transport. However, already in the introduction quite a lot of material is presented about the practical issues of CSAs. I think it is important that the model was developed for this purpose, but it is an application that follows from what is presented in the paper. Therefore, it would be possible to shorten the introduction (section from p. 12907 to 12908) quite a bit which also results in more focus on the main objectives of the paper.*

Answer: We agree that this part can be considerably shortened to put more focus on the model, which is the main objective of the paper. It is however (as the reviewer stated), important to mention the practical purpose of the model development. For purely scientific use, some practical restrictions like computational time and data availability are less important. For practical applications, these aspects can decide on the usefulness of the model and we developed the model with this background.

*I also recommend to clearly stating two or three main objectives.*

Agree

*Reading the detailed model description I was immediately struck by the separate and decoupled (or unidirectional) treatment of the unsaturated and saturated zone. Despite that one goal for model development is stated as the representation of groundwater levels and the dynamic behaviour of the unsaturated and saturated zone in soils (p. 12911, l. 28), the decoupling is only mentioned as a limitation much later in the discussion (p. 12932, l. 6 citing work by Seibert et al., 2003 who used a conceptual model to achieve coupling). My concern here is that this in fact is one of the reasons the model failed to predict the groundwater levels with a higher degree of precision.*

We agree with the reviewer's assessment that our rather simple representation of the processes at the boundary between the unsaturated and the saturated zone is probably one of the major reasons for the relative model failure especially for the groundwater level dynamics (see above).

*From this model failure it is then concluded that a more complex approach would be needed to improve simulations (abstract, l. 28). I don't agree with this conclusion as a slightly different model structure with coupled saturated and unsaturated zones might already do a much better job at simulating groundwater levels and the frequency of soil saturation. Did you try this or is this planned for future work?*

We agree that a moderate modification on the interactions between saturated and unsaturated zone could result in an improvement (see comment above). However, because of the mutual dependency of the two zones we consider this as a more complex model. We intend testing different model structures (Fenicia et al., 2011) in the future. But this goes beyond the scope of this study.

*Furthermore, from the result section it becomes evident that the drainage system is over-emphasized (p. 12929, l. 4) in the model structure resulting in fluxes reaching the drains too quickly. You rightly identify this issue, but was it somehow addressed in the model building process or do you plan to do this in the future?*

From the perspective of flow, the drainage system is not too much over-emphasized. The tile-drains deliver the largest part of discharge during low and high flow and they react very quickly to rainfall, which is correctly predicted by the model. We had two flow measuring sites in the drainage system. Together they delivered 58'330 $m^3$ within the calibration period and they cover around 75% of the drained area. Flow from the whole drained area can therefore be expected to be around 77'770 $m^3$. This is about 62% of the total measured flow in this period (124'470 $m^3$). The model predicts a tile drain flow of 98'670 $m^3$ which is 82% of the predicted total flow (120'910 $m^3$). The model therefore does overestimate the proportion of flow coming from tile drains but the overestimation is not very strong. However, the modelled drainage system is too efficient in terms of lowering the groundwater levels. In our opinion this is at least partially due to the spatial resolution. With our resolution of 16 x 16 m we are in the range of the spacing of the individual drains (about 15 m spacing). This means that the water level in the whole drained area is homogeneously lowered. However, in reality there are steep water level gradients between neighboring drains. With the grid cell approach we don't see any possibility to reproduce this behavior without substantially increasing the spatial resolution (which then has a huge influence on the computational time). A reasonable representation of the tile drains would require a new model approach, which is beyond the possibilities of this study.

*I therefore think that it is premature to conclude about the need for a more complex model if different perfectly plausible concepts were not exhaustively tested. Recognising that this is an interesting and difficult scientific problem specifically considering the low-relief and tile-drained catchment, I would urge the authors to more clearly state the potential impact of a different model structure with reference to e.g. Gupta et al. (2012) and/or incorporate the testing of competing model concepts of similar complexity.*

We agree that we have no prove for our statement that a more complex is needed for better predictions (but see general comment above). We will soften this statement in the revised version of the manuscript. The comparison of different model structures would be very interesting, it is however beyond the goal of this study.

*Specific comments:*
*Abstract: I think the abstract could be substantially shortened. As it stands it contains a lot of methodological information which is not really needed at this point.*

Agree

*Abstract, Line 28: I don't agree with the conclusion that you necessarily need to increase model complexity to adequately simulate groundwater levels if other competing model structures were not tested (see general comments).*

See comments above

*Page 12908, Line 28: In the context of hydrologically active you could refer to work by Ambroise (2005).*

Agree

*Page 12910: In the context of using HOST as additional information for model calibration and evaluation I would suggest including work by e.g. Dunn and Lilly (2001). See reference below.*

This paper indeed fits the discussion on how soil information has already been used in hydrological modeling. We will include this reference in the revised version of the manuscript.

*Introduction: I would ask the authors to consider formulating clear study objectives.*

Agree, see above

*Introduction and Discussion: Consider including physics-based Integrated Surface-Subsurface Hydrological Models as the more complex counterpart to the model presented (see e.g. recent work by Partington et al., 2013).*

We will mention examples of this type of models to show what we mean by complex models.

*Page 12915, Line 15: Please, spell out Dinf and/or explain what type of algorithm this is.*

Agree, it is the algorithm proposed by Tarboton (1997). It's a multiple-flow-direction algorithm hence allowing for flow divergence. We will add the reference and explanation in the revised version of the manuscript.

*Page 12925, Line 10: I think the information on calibrated parameters, initial ranges and best-fit parameters is best presented in the paper rather than as supplementary material.*

The table is very long and we are not sure if it is relevant for the majority of the readers. However, if the editor considers it useful to have it in the main text we can add the table there.

*Page 12926, Line 4: Delete one "of".*

Agree

*Page 12927, Line 21: Please, consider including a more quantitative assessment of fitted groundwater levels.*

We will add the following table showing the rmse at the piezometers to the manuscript.

| Piezometer Number | RMSE [m] |
|---|---|
| P 1 | 0.792 |
| P 2 | 0.310 |
| P 3 | 0.218 |
| P 4 | 0.281 |
| P 5 | 0.832 |
| P 6 | 0.857 |
| P 7 | 1.001 |
| P 8 | 0.328 |
| P 9 | 0.441 |
| P 10 | 0.411 |
| P 11 | 1.003 |

*Page 12927, Line 24-26: Please, give the reason why it's not shown or just don't mention it.*

We added some data on this (see answer on drainage system above) and will also show this data in the revised version of the manuscript.

**Reviewer 2**

*In my opinion, the causes for the relative failure of the modelling exercise should be attributed to an insufficient calibration approach, as the authors show that topography is a better predictor of the patterns of frequently saturated areas than the model results.*

There are two points to be mentioned here.

i) Based on the arguments described above we do not agree that the model failure should be attributed to a poor or improper calibration procedure. However, we agree that we did not present our arguments explicitly and clear enough. We will add text to deal with the issue of multi-objective optimization and will put our approach in this context.

ii) There is a misconception in the comment above in that it is stated topography was a better predictor for saturated areas. This misunderstanding is probably due to our wording in the paragraph on p. 12934, L. 15 – 28. This paragraph is indeed misleading because it talks about Critical Source Areas CSAs while we actually just compare the spatial patterns of topoindeces with spatial patterns of soil class distributions. Neither of them represent CSAs per se. We just wanted to express that the soil classes of the soil map resemble the classes based on the topoindex more closely than the patterns of average water table depth according to the model simulations. However, neither the soil map nor the topoindeces are directly compared to piezometer data here (this would be relevant for the CSAs!). As shown in the Figure above, the topoindex is actually a fairly poor predictor for water table levels.
Although we disagree with the opinion of the reviewer here, we learn from the comment very clearly that we need to be more precise and explicit in our wording. We will modify the text accordingly.

*Although the manuscript may deserve publication for several reasons, a substantial part of the model implementation should be revised before publishing in HESS a paper which claims the need for a more complex model whereas another calibration of the same model or the implementation of a simpler model could provide better results.*

Based on the arguments listed above we do not share the same conclusion. However, we realize that more careful wording is needed from our side.

*A key for understanding this unexpected result may be found in the model calibration process. If well understood, the calibration of the model was made searching a unique parameter set that maximized the likelihood of the groundwater level simulations at 11 piezometers, along with the simulation of a transformed function of water discharge at the outlet. The results of this exercise were good after the graph on Fig. 5 (observed and simulated discharges). But this choice of calibration approach hides a real multi-objective calibration; the poor simulations of the dynamics of water tables shown in Fig. 7 suggest that the model converged to an acceptable simulation for a few piezometers (3, 4 and 8) as well as discharge, but had difficulties to simulate the remaining piezometers because of the disparate values. As suggested by the authors, the model seems to optimise discharge by simulating an excessive role of tile drains, but this is not necessarily because of the model structure, but may be attributed to the results of the calibration procedure used. One may suspect that the degrees of freedom of the model mean that other parameter sets might provide flow simulations of similar quality with diverse simulations of water table depths and tile drains contribution.*

This was indeed a multi-objective calibration problem (see comment above, we will make this explicit in the revised version and discuss the implications). The Bayesian approach to this is to carry out inference the normal way with calculating likelihoods. The 'weighing' between the objectives is done by setting different priors for different error variances (Reichert and Schuwirth, 2012). By formulating the error variance priors for transformed discharge and piezometer levels we have expressed our expectation for the relative weight of discharge and groundwater components. As explained above, the priors were selected so wide and the number of data for water table levels exceeded discharge data 11-fold that we did not artificially force the optimization into a region where only discharge could be successfully simulated. In our opinion the calibration was successful, as both discharge and average groundwater levels could be simulated with good precision. The biggest problem was that the model was unable to simulate the transient peaks in groundwater levels during storms simultaneously at all sites, which interestingly did not systematically appear in all piezometers. We think that this was due to i) a process that was not represented in the model structure, ii) to spatio-temporal heterogeneity (see above). If this structural deficiency is true (arguments for that are mentioned above), there was no way for the model to provide a better fit to the piezometer data regardless of other calibration algorithms or different weighing between the different fit objectives.

*Irrespectively of the successfulness of the simulations, it is difficult to accept the publication of a hydrological modelling exercise that does not take into account the present awareness of multi-objective calibration (e.g. Efstratiadis and Koutsoyiannis, 2010) nor the uncertainties associated with model simulations (e.g. Beven, 2006).*

We agree (see above) and will extend the text accordingly.

*The authors may choose their way to offer a more convincing manuscript. If a more explicit multiobjective calibration taking into account the uncertainty associated with the simulations is not possible, the comparison with an alternative calibration of the model with other optimisation criteria or cancelling out the role of tile drains might provide more insight on the problems addressed.*
*Alternatively, a test with a simpler model based on topography (e.g. TOPMODEL) might provide comparative elements for a wider discussion on the needs (or not) for more complex models.*

We agree that the comparison with Topmodel gives valuable insight (see above). We will include the corresponding results (see above) in the paper.

*Specific comments*
*The list and values of used parameters should be preferably shown in the paper instead of in the supplementary material.*

See above

*Page 11, line 15: A more formal definition and citation of the "topographic wetness index" is needed. Is this the TOPMODEL topographic index ln(a/tanï¸Ac) index (Beven and Kirkby, 1979)?*

Agree, indeed we use the TOPMODEL topographic index ($\ln(A_{upslope}/\tan(\text{local slope}))$). We will add this equation and reference to the manuscript.

*Page 23, line 20: It should be noted that Fig. 6 provides a false impression of good results, because the large differences among elevations hide the errors in water table depths.*

We agree that the prediction goal is the depth of the water table below the surface and therefore Fig. 7 is the more honest way to show the data. Still in our opinion Fig. 6 does not give a false impression. It shows that our model is able to reproduce the general hydrological behavior of the catchment. We consider a maximum error of around 1m to be rather good when the elevation range is around 30m and the model domain is 10m thick.

*Page 29, line23: The topographic wetness index map may be used for modelling the spatial dynamics of the saturated area. TOPMODEL performs a lumped simulation of the average depth to the water table and the map is then used to distribute it, so the extent of the saturated area at every time step can be mapped.*

See comment above.

*Fig. 7: Some quality cross-check of records should be made in order to avoid errors (piezometer 2 shows an odd decrease by the end of the calibration period) as well as too local responses, before being used for model calibration.*

Piezometer measurements are always local responses and the piezometers do react very differently to rain events. It is not possible to decide which of the piezometers shows the more “general” behavior and which has a more “local” response. We checked what seems to be an odd water level decrease in piezometer 2. There is no hint that it is a measurement error.

Since we want to achieve local, small scale predictions, these local responses are exactly what the model would need to finally predict. So these measurements are what we have to judge our model on if we want to use it for real applications. We see that it was a very optimistic approach that we chose when we tried to reproduce these local responses with e.g. homogeneous subsoil properties. It turned out that the approach failed for the specific (and quite demanding) objective. Still we think that the results can be interesting for the hydrological community.

Blazkova, S., K. Beven, P. Tacheci, and A. Kulasova. 2002. Testing the distributed water table predictions of TOPMODEL (allowing for uncertainty in model calibration): The death of TOPMODEL? Water Resources Research 38:1257.

Doppler, T., L. Camenzuli, G. Hirzel., M. Krauss, A. Lück, and C. Stamm. 2012. The spatial variability of herbicide mobilization and transport: a field experiment at catchment scale. Hydrological and Earth System Sciences 16:1947 - 1967.

Efstratiadis, A., and D. Koutsoyiannis. 2010. One decade of multi-objective calibration approaches in hydrological modelling: a review. Hydrological Sciences Journal 55:58-78.

Fenicia, F., D. Kavetski, and H.H.G. Savenije. 2011. Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. Water Resources Research 47:W11511, doi:10.1029/2010WR010174.

Seibert, J., A. Rodhe, and K. Bishop. 2003. Simulating interactions between saturated and unsaturated storage in a conceptual runoff model. Hydrological Processes 17:379-390.

Tarboton, D. G. 1997. A new method for the determination of flow directions and upslope areas in grid digital elevation models. Water Resources Research, 33, 309–319.