Hydrol. Earth Syst. Sci. Discuss., 10, C1396–C1415, 2013
www.hydrol-earth-syst-sci-discuss.net/10/C1396/2013/

Hydrology and
Earth System
Sciences
Discussions

Interactive comment on "Resolving structural errors in a spatially distributed hydrologic model" *by* J. H. Spaaks and W. Bouten

J. H. Spaaks and W. Bouten

jspaaks@uva.nl

Received and published: 3 May 2013

1 Reply to general comments from Referees #1 and #2

We are grateful that the Referees consider our paper 'innovative', 'interesting', 'novel', 'of high relevance to the readers of HESS', 'compelling' and 'very well written'. The Referees' main criticism is related to whether the approach will work on real-world cases, in particular with regard to data abundance, data quality, unknowable boundary fluxes, and soil heterogeneities. Both Referees consider this their main criticism of the paper. Below, we will discuss our motives for designing the paper as we did.

We have tried to be very conscious about how we communicate the paper's main message to the reader—which elements of our 'story' should be included and what was better left out. We feel that the reader is best served by having one main message ('SODA's state updating patterns are more helpful in diagnosing model structure errors than are SCEM-UA's simulation-observation residuals'). To avoid obscuring this main message, less important messages should be avoided, if at all possible. This was in fact the primary reason for our use of artificial data in this study: had we used real data instead of artificial data, we could not have conclusively shown the potential of analyzing state updating patterns, because the truth would not be known and our message would be obscured by issues relating to data quality (including incommensurability, measurement error, etc.), data abundance relative to process heterogeneity, and uncertainty associated with for instance the estimate of water balance. This is because the results would still require *interpretation* on our part, as opposed to less subjective *clarification* that is needed when using artificial data. Since our interpretation would likely be different from someone else's, there would be great opportunity for the discussion to get bogged down in the specifics of one particular hillslope, wheras we were hoping to stimulate a discussion about the relative merits of SCEM-UA and SODA.

Because of these considerations, we do not intend to show any results based on real world data in the revised version of the manuscript. However, the Referees' comments have made it clear to us that we have painted too positive a picture with regard to the immediate possibilities of analyzing state updates. For the revised version of the manuscript, we will emphasize that this method is not the end-all and be-all of model diagnostics, and that much work remains to be done before we know how to usefully apply the method within the context of real-world experiments. We also propose to better explain our choice for artificial measurements in the discussion and to include additional simulation work, in which we decrease the number of observation locations. The results of this additional simulation will then serve to illustrate a discussion about the Referees' questions, specifically:

1. what happens if there are less measurements?
2. what happens if the measurements of pressure head are not taken at exactly the location where model structural error is introduced?

Referee #1 further expresses interest in the effect of multiple, interacting model deficiencies on the feasibility of the proposed approach, and suggests that this may be pursued in a future paper. We agree with the Referee that this is an important subject, and unfortunately also a complicated one. Having multiple interacting contributors to state updating will feature in the additional simulation we do for the revised version of the manuscript. This is because when the measurements of pressure head are not taken at the location where model structural error is introduced, state updates can not be performed at the locations of the hotspots anymore, but may occur some distance downslope. This means that the state updating pattern that is performed at the location away from the hotspot is in fact different from the state updating that would be performed had pressure head been observed at the hotspot. State updates can thus no longer be interpreted directly as additional water loss to the bedrock, but must instead be interpreted as the local effect on pressure head of failing to remove additional water some distance upslope. When pressure heads are not observed at the hotspots, state updates thus become a mixture of two signals: (1) the structural error introduced at the (unobserved) hotspots, which previously could be interpreted directly as additional water loss to bedrock; and (2) the altered dynamics of flow in the part of the domain between the current (observed) location and the (unobserved) hotspot. In the revised version of the manuscript, we will present and discuss simulation results that illustrate the above.

In addition to the questions raised above, Referee #2 also asks: How does analyzing the state updates help with real-world field experiments, in which boundary fluxes are

unknown? Will [SODA] still be able to improve the model structure if the error-free perfect "truth" is not known? These are questions that are very relevant for practical application of the method. With regard to unknowable boundary conditions, our results (Fig. 8–10) show that the correct parameter values were identified, and that the associated state updates provide insight into where sink hotspots are located. It is important to note that these locations were indeed 'unknown', in the sense that we did not explicitly specify them during calibration of the model with SODA. The fact that the hotspot locations were correctly identified is just due to the fact that any process associated with a noticable (as in 'measurable') effect on the value of the system's state (in this case: pressure head) will leave its signature in the space-time pattern of state updating (for example, Fig. 9). In the manuscript we showcase this for a sink, but it does work fundamentally the same for any process.

With regard to the 'error-free' aspect: whether or not an analysis of state updates can help to improve the model structure depends on whether the effect of any particular missing process is noticable in the observations of the system's state values. For example, if the measurement uncertainty is very large, or the process has only a small effect on the observed value of the system's state, an analysis of state updates may not yield any useful insights (but note that this not so much a shortcoming of the method; other methods would be equally hard-pressed to make sense out of so little useful information). A similar argument can be made for the measurement interval (time) or measurement spacing (space) relative to the scale of the process of interest: if the system's state is affected by a process that is not represented in the model, but that only works over short ranges (in time or space, respectively), the observations of the system's state may not show any evidence of the existence of this process if the measurement density is low. Again, this is not so much a shortcoming of the SODA method, but simply results from the fact that even powerful methods of analysis cannot *create* their own information—they can only *use* information contained in measurements. If the measurements do not contain any such information, then analysis is hopeless any-way. In the revised version of the manuscript, we will elaborate on these issues.

The remainder of this reply covers Referee #1's specific comments in detail.

2 Specific comments

1. (comment) The title is too general and should be a bit more specific.

   (reply) We agree with the Referee that the title may be too general. In the revised manuscript, we will include some more specific terms in the title in order to clarify what the paper is about.

2. (comment) Some references related to diagnostic approaches are missing (see at the end)

   (reply) We thank you for the useful references you suggested. We will incorporate them into the manuscript.

3. (comment) P 1826, L10: "forward model" and "inverse model" - the rational to choose these terms is not very clear. Why not using "reference model" and "test model" instead?

   (reply) The terminology is indeed a bit confusing. In the revised manuscript we will combine your suggestion with that of Referee #2; instead of "forward model" we will use "reference model", and instead of "inverse model" we will use "simplified model".

4. (comment) Fig 2 and Fig 4: as this is a virtual setup, presenting spatial distributions as smoothly distributed variables is somewhat misleading. I suggest to

   present rectangles for which a node is representative with the corresponding values.

   (reply) We agree with the referee on this, but the smoothness was unintentional and has an origin in the presentation software. In fact, the smoothness appears to be related to the PDF viewer/printer driver, which interprets the raster of soil depths (Fig. 2) or sink locations (Fig. 4) as a low resolution image/photo, and then tries to improve it by applying a smoother. In order to avoid this problem, we will change the way the image is constructed (each raster element will be a square polygon/patch object), so PDF viewers and printer drivers do not attempt smoothing.

5. (comment) P1827 L10-25: Do you use homogeneous soils and K values? Does this affect your main findings?

   (reply) We do indeed use homogeneous soil properties and K values. In the revised manuscript, we will add some further explanation to the text of the Methods section, and we will discuss the effect of soil heterogeneity on the applicability of the method in the Discussion section.

6. (comment) Fig 4: Please also refer to the symbol rsink.

   (reply) In the revised manuscript, Figs. 2 and 4 will be merged as per Referee #2's suggestion. In the new, combined figure, we make explicit reference to the $r_{sink(low)}$ and $r_{sink(high)}$ parameters in relation to where they are applied. Parameter r_{sink} is only used in the inverse model (which will be renamed to 'simplified model'), so we thought it better to introduce r_{sink} at a later point in the manuscript.

7. (comment) P 1828 L20: Explicitly state what the reference level for pressure head is. The method to obtain the initial state is not described sufficiently clear. Where

do spatial heterogeneities in the initial state in Fig 5 come from? Is this due to spatially heterogeneous soil depths and rsink values only?

(reply) The reference level of pressure head is the soil-bedrock interface, i.e. the lower boundary of the domain. A pressure head of 0 was assigned to all nodes at the lower boundary of the domain at t=0. Nodes in the upper 4 layers were assigned a pressure head of $-z$, in which z is the vertical distance from a given node to the lower boundary of the domain at a particular X,Y location. With this initial state, SWMS_3D was run until t=96 h. During this period, soil water was redistributed due to hydraulic head differences. The slope of the domain, convergence of flow due to varying soil depth, as well as water removal from the domain at the sink hotspots were the driving factors in this redistribution. The pressure head pattern at t=96 h was then saved to file and served as a starting point for all further simulations. In the revised version of the manuscript, we explain in more detail how the pressure head was initialized.

8. (comment) P1829 L11: Related to the insufficient description of the initial state procedure: It is not sufficiently clear why 188 m3 of water is present in the soil at the initial state.

   (reply) The explanation which we will include in the revised version of the manuscript clarifies why there is 188 m3 of soil water in the domain at t=96.

9. (comment) P1829 L 20: This is a very rich set of observations that is hardly available in a real case catchment. The study will benefit much if a reduced set of pressure head observations is used in an alternative scenario.

   (reply) This relates to the Referee's 'main concern'. Please see our remarks in the previous section.

10. (comment) P1830 L 9: Be explicit about how you treated the other parameters.

Do you assume perfect knowledge about these? How does this influence results?

   (reply) The other parameters were indeed treated as perfectly known. We will add this to the revised version of the manuscript. We will also add some discussion about the effect of introducing more calibration parameters.

11. (comment) P1833 L14: Provide information about how to interface SCEM-UA and the model. Give exact information about implementation and version numbers for the model and SCEM-UA. Same for SODA further down.

   (reply) We will prepare a zip file with the software and data pertaining to this paper for uploading to HESS as a supplement. We do feel that the paper is not the best place to provide technical detail on the interface between SWMS_3D and the optimization framework (be it SCEM-UA or SODA). However, we are glad to elaborate on it here; we run SWMS_3D as a Windows binary (*.EXE). The binary expects a number of input files, and writes a few output files. In order to use such a binary within an optimization framework, we use so-called wrapper functions. These wrapper functions are written in MATLAB (as are SCEM-UA and SODA). The optimization framework samples a point in the parameter space (the parameter vector). This parameter vector is then passed along to the wrapper function, which (automatically) writes the input files. Next, MATLAB makes a system call to a *.BAT file, which in turn triggers the SWMS_3D binary. Once the binary starts, it looks for its input files (which were just written by the wrapper function). SWMS_3D runs for whatever simulated time vector was specified in the input files, and subsequently writes the output files. When the binary has finished writing the output files, focus is returned to MATLAB, which then calls another wrapper function which parses the output that was just written by SMWS_3D. From the many outputs generated, the output wrapper function selects whatever is needed (particularly, the state of the model, i.e., 3-D pressure heads) as well as whatever else is needed for calculating the objective score (e.g. discharge from the seepage face, sink volumes). It then continues with the next parameter vector, etc.

12. (comment) P1835 L6: Based on the values for rsink and the number of nodes, it should be possible to calculate a weighted average - this might be what we observe for rsink. What else do you expect than a value in between the two extremes for rsink?

    (reply) We have collected our replies to some of the Referee's comments about Section 3.1 'Interpretation of the SCEM-UA results' (specifically comments 12, 15, 16, 17, 18, 19, and 20 from this list) under item 29 in this list.

13. (comment) Figure 5,7 and 9: Mark hot spots with an asterisk or similar. Use linear legend as the square representation suggests a bivariate color schema, which you are not using.

    (reply) In the revised version of the manuscript, we will add a visual marker in order to more easily identify hotspot locations in Figures 5, 7, and 9 (these figure numbers refer to the old manuscript). We will further replace the bi-variate color scheme of figure 9 (old manuscript) and, where applicable, of any figures in the revised version of the manuscript with univariate color scales. Figure 7 (old manuscript) does indeed need a bi-variate colorscale (note, for instance, the clear color difference between X06Y03 and X09Y18).

14. (comment) P1835 L11: How did you determine the spatial auto-correlation? Please report the measure for the spatial auto-correlation.

    (reply) The term spatial auto-correlation was ill-chosen in that it suggests we performed some geostatistical analysis/kriging on the residuals. This is not the case—we simply meant to point out the spatial coherence/structure that is apparent in the error patterns. In the revised version of the manuscript, we will change

    the wording in order to avoid confusion about this.

15. (comment) P1835 L20-21: Results supporting this argument are not clearly presented. I have a hard time to see this in Fig. 7.

    (reply) We have collected our replies to some of the Referee's comments about Section 3.1 'Interpretation of the SCEM-UA results' (specifically comments 12, 15, 16, 17, 18, 19, and 20 from this list) under item 29 in this list.

16. (comment) Section 3.1: Your argumentation is not very convincing: If you want to improve understanding, you would make good effort to better understand what is going on in subsurface. A uniform leakage would not be a good assumption for this. Also rsink = rsink(low) may be a good assumption for 93 out of 98 nodes, but not for the overall model. I would suggest to leave out the entire part about SCEM-UA.

    (reply) We have collected our replies to some of the Referee's comments about Section 3.1 'Interpretation of the SCEM-UA results' (specifically comments 12, 15, 16, 17, 18, 19, and 20 from this list) under item 29 in this list.

17. (comment) P1836 L1-5: Present the results of leave completely away

    (reply) We have collected our replies to some of the Referee's comments about Section 3.1 'Interpretation of the SCEM-UA results' (specifically comments 12, 15, 16, 17, 18, 19, and 20 from this list) under item 29 in this list.

18. (comment) P1836 L 8: not very clear what you mean by "activation of response modes" – nowhere introduced.

    (reply) We have collected our replies to some of the Referee's comments about Section 3.1 'Interpretation of the SCEM-UA results' (specifically comments 12, 15, 16, 17, 18, 19, and 20 from this list) under item 29 in this list.

19. (comment) P1836 L11-15: Not very clear how this derives from the results shown.

    (reply) We have collected our replies to some of the Referee's comments about Section 3.1 'Interpretation of the SCEM-UA results' (specifically comments 12, 15, 16, 17, 18, 19, and 20 from this list) under item 29 in this list.

20. (comment) P1836 L 28-30: This thought, while interesting, could be made clearer by being clearer about some underlying conceptual ideas. For example introduce before, how and when patterns in residuals are related to physical processes.

    (reply) We have collected our replies to some of the Referee's comments about Section 3.1 'Interpretation of the SCEM-UA results' (specifically comments 12, 15, 16, 17, 18, 19, and 20 from this list) under item 29 in this list.

21. (comment) P1837 L13-20 How were implicit sinks treated in the objective functions - where they neglected? What values do you get for OF1? Please present influence of the different OF on the selection of the parameters.

    (reply) Implicit sinks were not part of any objective function, they are simply the result of adjusting the pressure head: if you adjust it downward, that amounts to an extraction of water, whereas an upward adjustment represents adding water to the soil. In the revised version of the manuscript, we will describe this more clearly.

    Because of the state adjustment that we perform, any pressure head errors introduced at the hotspots are quickly canceled out (although not immediately, due to the model integration step being less than the interval with which we update the model states). However, it becomes therefore possible to match both objectives very well (almost zero misfit). As a result, we did not think it very interesting to go into detail about any influence of the different OF on the selection of the parameters.

22. (comment) P1838 L7 Briefly state that you will explain nodes that need updating but are not hotspots a bit later.

    In the revised version of the manuscript, we will mention this.

23. (comment) Section 3.2.1 Title is not well chosen. The first part of the section is not related to "experimental design". In general, I find your suggestions for experimental design not very helpful - mostly what you are saying is: "make as much and as reliable measurements as possible". Could you try to make your recommendations in view of limited budgets? Is it better to use only few reliable (small errors) or a larger number, not so reliable sensors (larger errors)? If logger space is limited, is it better to make more frequent measurements of should longer periods be measured? (you are contradiction yourself within the manuscript. P1839 L 18: measure more frequently; P1836 L10: measure longer). Check your recommendations with experimentalists for the revised version or leave them out.

    (reply) In hindsight we agree with the Referee's comment about the title of this section. For the revised version of the manuscript, we will change it, as well as parts of the section itself. We do feel that advise on experimental design in terms of budgets is beyond the scope of the current paper, although it certainly is an interesting opportunity for future research. Finally, there is one comment that we would like to make about the contradiction that the Referee refers to, between "measure longer" on the one hand, and "measure more frequently" on the other. We maintain that we are not necessarily contradicting ourselves, because the suggestion "measure longer" is made within the context of using SCEM-UA, whereas the suggestion "measure more frequent" is made within the context of using SODA. The purpose of the first should not so much be to improve one's understanding of a system (for reasons outlined under item 29 of this reply, as well as in the manuscript itself), but more to predict certain variables (notably discharge) *given* the model structure. The model's structural deficiency may not manifest itself for all events (for example, it may only become apparent for large events), so tuning the parameters using a long calibration set is likely to give the best predictions (on average).

In any case, the purpose of parameter tuning should be quite different from that of the second: analysis of state updates using SODA. As with SCEM-UA, here it can be also advantageous to have a long calibration set because it is more likely to include certain rare events. However, it is more important that the measurements of the state are not taken too far apart, otherwise the errors that are introduced on the states will spread to neighboring states, and it will become more difficult to tie state errors to a certain time and place, and by extension, to physically meaningful processes—this was the purpose of the analysis, after all.

24. (comment) P1839 L30: if X15Y39 never shows the behaviour $B = r_{sink(high)} * h$, is it correct to speak of a hotspot then? In my view, in case of h $<$ 0 for all times you can not distinguish the two kinds of nodes.

    (reply) In this case, we feel that it is correct to speak of a hotspot. The reason for this is that node X15Y39 in the forward model (which will be renamed to 'reference model') can indeed work according to the middle case of Eq. 5. Even though the mechanism at this location does allow for the quick removal of excess soil water (h$>$0), the soil does not become wet enough to enable this model behavior during the simulation we perform in the manuscript. This is on purpose, because it allows us to demonstrate that some aspects of a given model's behavior simply cannot be pinned down if these aspects are not represented in the data (no matter how good the method of analysis). Had we chosen to apply more (virtual) rain to the soil, node X15Y39 may have become wet enough to show the behavior $B = r_{sink(high)} * h$.

25. (comment) General: Provide virtual observations and results from the deficient

model as supplementary material, for others to test their method on your example.

    (reply) We will prepare a zip file with the relevant data as well as the MATLAB scripts and functions as supplementary material.

26. (comment) Additional references related to diagnostic approaches: Bastidas, L., T. Hogue, S. Sorooshian, H. Gupta, and W. Shuttleworth (2006), Parameter sensitivity analysis for different complexity land surface models using multicriteria methods, Journal of Geophysical Research, 111: D20101

    (reply) We will incorporate this reference into the revised version of the manuscript.

27. (comment) Reusser, D., and E. Zehe (2011), Inferring model structural deficits by analyzing temporal dynamics of model performance and parameter sensitivity, Water Resources Research, 47(7), W07,550.

    (reply) We will incorporate this reference into the revised version of the manuscript.

28. (comment) Sieber, A., and S. Uhlenbrook (2005), Sensitivity analyses of a distributed catchment 738 model to verify the model structure, Journal of Hydrology, 310(1-4), 216-235.

    (reply) We will incorporate this reference into the revised version of the manuscript.

29. (comments 12, 15, 16, 17, 18, 19, and 20)

    (reply) From the Referee's comments on Section 3.1 'Interpretation of the SCEM-UA results' of the manuscript, it appears that we have not formulated our points clearly enough. For the revised version of the manuscript we will make sure to change the text in such a way that the following points are made more clearly, and in more detail than before:

(a) what happens when soil water is not extracted at the hotspots due to model structure error;

(b) why the water that is not extracted at the hotspots makes it difficult to interpret the residuals in the lower part of the hillslope as 'new and local' errors;

(c) why that is detrimental to diagnosing how a given model structure could be improved;

(d) how the performance of the 3rd objective (pressure head) is affected by errors that are not new and not local;

(e) why the optimal parameter combination varies with the characteristics of the event, in particular with regard to rain intensity and duration;

In the remainder of this reply, we provide a more detailed discussion of these aspects.

29.1 Relatively small event

For the sake of argument let us assume that there exists a 'relatively small event' (however that may be defined), for which transient saturation does not coincide with any of the hotspot locations, and that there is therefore no water loss to the bedrock at these locations (Eq. 6; i.e. the equation that describes water loss to bedrock as used in the simplified model). In principle, this means that SCEM-UA can find the correct combination of K_s and r_sink, and with it, the simplified model is able to match the observed pressure head patterns with zero error, despite the difference in structure that exists between the reference model and the simplified model.

29.2 Thought experiment 1

We established earlier that for a relatively small event, the optimal value for r_sink will be $r_\mathsf{sink(low)}$, simply because the behavior that differentiates Eq. 6 from Eq. 5 never occurs for this event size. Now let's see what happens if we keep the parameter values thus identified for the small event, but increase the event size (with no further calibration). For the resulting simulation, the simplified model behaves exactly like the reference model up to the point where saturation occurs at one of the hotspots, let's say, at X09Y15. At X09Y15, the reference model removes quite a bit of water from the soil domain. In contrast, the simplified model removes only a little bit (because it applies $r_\mathsf{sink} = r_\mathsf{sink\,(low)}$ where the reference model used $r_\mathsf{sink\,(high)}$—a 30x difference). The implication of this is that the pressure head at X09Y15 will start to differ from the pressure heads in the artificial observations from this point forward.

29.3 Spreading of errors

The water that was *not* extracted at X09Y15 will affect the pressure heads in the direct vicinity of X09Y15; generally speaking the nodes downslope from X09Y15 are wetter than what was recorded in the artificial observations. Because the soil downslope from X09Y15 is wetter, the rate at which water is extracted from the soil in the area downslope from X09Y15 is increased relative to the artificial observations (note the h in Eq. 6). Depending on (1) how long saturated conditions last at X09Y15; (2) how much water is not extracted at X09Y15; (3) how much the sink rate downslope from X09Y15 is increased, the error signal downslope from X09Y15 becomes less pronounced. Note, however, that the damping effect at nodes downslope from X09Y15 only works while $h > 0$ there; for nodes that are too wet but do not have transient saturation, no water is removed from the soil at all, so there is no mitigation of any errors. Once saturated conditions cease, the error signal is just subject to diffusion (because of lateral redistribution according to the Richards equation), but is not dampened any more.

It is noteworthy that pressure head values downslope from X09Y15 are incorrect *despite* the r_{sink} parameter having the correct value $r_{\text{sink (low)}}$ for that part of the domain. The non-zero pressure head residuals in this part of the domain are purely the result of errors that occurred some time before, in other parts of the domain. They are thus not 'new' and not 'local'. In terms of diagnosing a model structure, this complicates matters greatly, because when errors cannot be considered new and local, it becomes very difficult to tie especially large errors to other circumstances ('physically meaningful processes' P1836 L30 of the manuscript) that could possibly explain what process is responsible for them.

29.4 Thought experiment 2

Now let's perform a second thought experiment in which we make a slight change to the value of the r_{sink} parameter. For instance, we set it to a slightly higher value than in the previous thought experiment. The new, slightly higher value will lead to more water being extracted from the soil domain at all locations where transient saturated conditions occur. For the upper part of the hillslope, slightly more water will be extracted than what is recorded in the artificial observations, and as a result the pressure heads in that part of the domain can no longer be matched perfectly until transient saturation occurs at X09Y15 as before; instead, (small) errors are introduced almost immediately, specifically when transient saturation occurs *anywhere*. However, if we look at how the lower part of the hillslope is affected by our increasing the value of the r_{sink} parameter, we see a different picture: there, the pressure heads are now matched better than what we had with the true (for that part of the hillslope) value of r_{sink}. This is because the water that was not extracted at X09Y15 can now be extracted more quickly at one of the nodes downslope from X09Y15.

Summarizing, after increasing the value of the r_{sink} parameter slightly, the performance on the upper part of the slope is slightly worse, but on the lower part of the slope is better. Moreover, saturated conditions occur for less time on the upper parts of the hillslope than on the lower part, so the errors on the upper

part of the hilslope are not only relatively small, but also, they do not last long. All in all then, the erroneous parameter value for the r_{sink} parameter that we use in this thought experiment is associated with a better score for objective function 3 (Eq. 11) compared to what we had in thought experiment 1. During optimization, the parameter tuning algorithm will thus prefer it over the more informative parameter combination that we used in thought experiment 1, because the algorithm has no use for 'bad' parameterizations. While 'better scores' do seem like a good thing to have, in fact they are detrimental to model diagnosis: recall that for thought experiment 1, we could not interpret the residual patterns as new and local due to the spreading of errors, and that most non-zero residuals could actually be attributed to errors that were introduced some time before as well as somewhere else (as much as a few nodes upslope perhaps). Now compare that to the situation we are facing in thought experiment 2, where non-zero pressure head residuals are not only not new and not local, they could be the result of errors that still have to occur in the simulated future (but within the calibration data, obviously). Furthermore, non-zero residuals in one part of the domain could result from parts of the domain that are (physically) completely removed from it; although two parts of the domain may not physically be linked, their dynamics *are* linked in that the parameters are applied to both. If we were having a difficult time relating the patterns in pressure head residuals resulting from thought experiment 1 to physically meaningful processes, that job is made much more difficult by parameter compensation in thought experiment 2.

29.5 Compensation varies with event size

With the additional detail provided herein, it should now be more easy to see that the optimal parameter value for r_{sink} is related to event size: larger events will generally mean that pressure heads at the hotspot are larger, so more water will flow past the hotspot locations. Also, hotspots that were not previously wet enough may now see transient saturation, and the structural difference between the reference model and the simplified model will become apparent at these locations as well. Both of these effects influence the optimal parameter values during calibration, but in a complex way. It is thus far from trivial to express the optimal r_{sink} value as a function of event size.

So, calibrating the simplified model yields different optimal parameter values depending om what size of event the model is calibrated to. But what if the simplified model is calibrated to a multi-year precipitation data set with, say, 100 events. Let's assume that a relatively small event is most common in this data set. Based on what we explained earlier, it is likely that the optimal parameter value for the 100-event calibration is dominated by the common, small events, as opposed to by the rare large event that is associated with different combination of parameter values. This is what we mean by "the compensation also reflects the frequency with which the [model] deficiency manifested itself during the calibration period" (P 1836 L12–14).

When the model is applied primarily to generate predictions (most commonly of streamflow), it can therefore be argued that the model should be calibrated to a long period of data, during which all types of events are seen (but with varying frequency, of course). In the absence of specific information about future events, calibrating the model to a data set that is as long as possible will give the best solution (on average!).

As a scientist though, one is interested in improving the current model (since models are essentially an explicit representation of our understanding of a system). The Referee points this out in his/her comment 15: if one suspects that the model is deficient in its representation of the subsurface, then one would do further experiments ("make good effort") to gain a better insight into how exactly the model representation deviates from what happens in reality. But we have an important note to make here: how do we know that something is not right in the model representation? How can we be even a little bit confident in our assertions

about the rightness or otherwise of certain model components if (virtually) all we have is calibration results such as those of Fig. 7, with all the parameter compensation and error propagation effects, as explained above? So, in principle we agree with the Referee, but we simply state that if a model does not perform well (for initially unknown reasons), better types of diagnostic analyses are needed exactly because "making a good effort" of measuring whatever process may be relevant requires knowledge (or at least some guidance) about what to measure next, and how to go about it. We are very much proponents of such an iterative approach to science and we will elaborate a little bit more on this aspect in the revised version of the manuscript.

We concede that Section 3.1 of the manuscript does not describe all of what is included here in sufficient detail. In the revised version of the manuscript, we will therefore offer a more elaborate explanation using some of the material developed here.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 10, 1819, 2013.