Response to Reviewer #1 (D. Reusser)

The article investigates spatio-temporal patterns of parameter sensitivity on the example of the spatially explicit hydrological model HL-RDHM. The combined analysis of parameter sensitivity in space and time is novel to the hydrological community and an interesting extension of existing sensitivity analysis studies. The study is well designed and most of the article is clearly structured and well written.

We thank Dr. Reusser for his time and detailed, thoughtful comments. His work in the area of time-varying sensitivity analysis is one of the foundations of this manuscript, and we are fortunate to have his suggestions to improve the quality of the paper.

Major Comment #1: Provide detailed descriptions of model mechanisms, and use these to develop hypotheses regarding expected parameter sensitivities

There are a number of issues that would improve the work from a good article to an excellent article. (1) The authors present a large amount of data presenting sensitivities at different temporal scales in a well structured way. However, the authors could make a clearer statements about their expectations on how sensitivities should compare across the different temporal scales and for what reason. When presenting the results, they could then check if these expectations hold. At the moment, comparisons across different temporal scales are limited and some differences which were surprising to me are not mentioned and not discussed (see comments 23, 28, 29).

(2) The authors could do an even better job in connecting results from the sensitivity analysis to hydrological mechanisms (in the model). This involves (a) a more detailed explanation of the concepts and intended mechanisms behind SAC-SMA. The method section should make clear, what different mechanisms will mean in terms of (spatiotemporal) parameter sensitivities

We agree with the need to provide a more detailed explanation of model mechanisms and expectations for parameter sensitivity. (Reviewer #2 has requested a similar expansion of the methods section). To address this issue, we have expanded Section 2.1 as follows:

> Herman et al. (2013b) showed that time-varying parameter sensitivity can be linked to the underlying mechanisms of a model. Here, studying the formulation of the SAC-SMA model allows the development of hypotheses regarding the expected parameter sensitivities, and how these might change in space and time. At each timestep, evaporation first occurs from the additional impervious store, both upper zone stores, and the lower zone tension store. In all cases, evaporation is proportional to the saturation level of the storage element. Next, direct runoff occurs from the impervious area, specified by PCTIM, and the additional impervious area due to saturation, specified by ADIMP. Precipitation not assigned to direct runoff enters the upper zone free water store. Gravity drainage occurs from the upper and lower zones according to the rate constants UZK, LZPK, and LZSK, and is linearly proportional to the amount of water in each respective store. Finally, runoff is also generated when the storage capacity of the upper zone (UZFWM) is exceeded. The same process occurs when all of the lower zone storage

capacities are exceeded (LZTWM, LZFPM, LZFSM), but otherwise excess from any of the lower zones will spill into another.

After the runoff generation mechanisms have occurred, each timestep of the model concludes with a redistribution of water between stores according to their saturation levels. First, any deficiencies in the upper and lower tension stores are filled by the free water in their respective zones. Next, percolation occurs from the upper zone free water store to the lower zone based on the saturation level of the lower zone. It is important to note that the lower zone controls percolation in the SAC-SMA model, unlike many other water balance models where percolation is equivalent to spillover from the upper zone. The amount of percolation varies with the parameters Z_{Perc}, the maximum percolation rate under dry conditions, and R_{Exp}, the unitless exponent of the percolation equation (Koren et al., 2004). Finally, the parameter P_{Free} determines the fraction of percolation that enters the primary and secondary free water stores in the lower zone.

From this description of model mechanisms, we can hypothesize which parameters might be most sensitive in space and time. During and immediately after precipitation events, the parameters associated with quick responses should be most sensitive. This includes the impervious area parameters and the upper and lower zone storage maxima, which can cause direct runoff via overflow. We might expect these sensitive parameters to be spatially concentrated near the outlet of the watershed, since only this area will have sufficient time to contribute to streamflow while the event is occurring. Between precipitation events, the primary streamflow generation mechanism will be drainage from the storage zones, controlled by the rate constants UZK, LZPK, and LZSK; we would expect these to be most sensitive in the time following an event, and with a broader spatial distribution to reflect their slower response. As found in prior work (Herman et al., 2013b), the percolation parameters are unlikely to be highly sensitive at any time, for two reasons. First, the amount of percolation is controlled by the moisture deficiency in the lower zone, so the parameter LZTWM (for example) has more influence on the magnitude of percolation than do the percolation parameters themselves. Second, the percolation parameters do not contribute directly to streamflow, so their signature may be obscured by intermediate processes. In general, we expect the lower zone parameters to exhibit higher sensitivity over the course of the simulation than upper zone parameters, because the lower zone deficiencies are filled first during the redistribution routine. It is important to note that the spatiotemporal parameter sensitivities will depend on the metric chosen. For example, the sensitivity of the root mean squared error metric on a short timescale will emphasize transitions between quick-response processes, while a water balance error metric on a longer timescale will capture the integrated effects of interacting states and fluxes.

These expectations provide readers with more contextual support for our discussion of results in Sections 4.2 and 4.3. The results often match expectations – for example, the high sensitivity of upper zone and impervious area parameters during large events, compared to the more constant sensitivity levels of the lower zone parameters over time. However, there are a few surprises, such as the consistent (though

small) sensitivity of the percolation parameters, as highlighted in the last paragraph of Section 4.2. One especially interesting result is the near-zero sensitivity of grid cells far from the outlet during large events, suggesting that hydrograph peaks do not depend on a significant fraction of the model (as discussed in Paragraphs 2-3 of Section 4.1). Finally, the specific comments #23, #28, and #29 raise interesting points and will be discussed individually in the list of detailed comments.

Major Comment #2: Provide a specific example of the sequence of dominant parameters and compare to intended mechanisms, and connect this to the concept of hydrologic regimes

(b) Moreover, for a selected number of periods, the spatio-temporal sequence of most influential parameters could be described and compared to the intended mechanisms of the model. (See also comment 17, 24, 25)

(3) A central point of the study is "identifying key transitions between modeled hydrologic regimes". The authors should be explicit about their definition of a hydrologic regime, how a transition between hydrologic regimes are detected and how this is connected to parameter sensitivity analysis. This is currently not made sufficiently clear and somewhat disconnected. During the presentation and discussion of results, the authors could make clearer when we observe such a transition between hydrologic regimes.

Comments (2b) and (3) are connected, because the spatio-temporal sequence of the most influential parameters is what we intended to convey with the term “modeled hydrologic regime”. However, we agree that this terminology may be confusing for readers, since the unqualified term “hydrologic regime” suggests a connection to true watershed processes separate from our modeling efforts. We have replaced the term “modeled hydrologic regime” with “dominant parameters and processes” throughout the paper, including in the abstract, introduction, and discussion sections. This clarification has been added to the last paragraph of the introduction section:

> This study proposes high-resolution time-varying sensitivity analysis for a spatially distributed rainfall-runoff model, avoiding the biases introduced by representative event selection by identifying key transitions between dominant parameters and processes *a posteriori*. These parameters dominate the performance of the model at a particular time, distinct from the true dominant watershed processes independent of our modeling efforts.

Assessing the sequence of influential parameters at different times is the intent of Figure 9, which shows a qualitative summary of dominant parameters at increasing temporal resolution. In this work, the transitions between sets of dominant parameters and processes must be detected visually. There is, of course, a subjective component to the summarization of results in Figure 9, but it is important to note that this summary was compiled *a posteriori*, once the time-varying sensitivity results had been analyzed. This represents an improvement over a traditional *a priori* event selection, which may be biased by assumptions regarding the similarity between events without exploring the full dynamic variability of parameter sensitivity throughout the simulation.

As Reviewer #2 has noted, “The moving time window enables a clear identification of shifts in processes”. This is the concept that we intend to convey with our discussion of transitions between sets of

dominant parameters and processes. In order to clarify this point, we have augmented Section 4.3 with a discussion of the sequence of the most influential parameters in Figure 9 and how this can be used to identify transitions between dominant processes in the model:

> As Figure 9 shows, the dominant controls for the full aggregated period are a combination of lower zone parameters in the headwaters of the basin, and upper zone parameters near both the headwaters and outlet. The full period sensitivities are clearly influenced by the wet periods at the event scale, which exhibit the same responses, indicating that the aggregate period is biased toward these large events (a result consistent with the focus of the RMSE metric). By contrast, dry periods at the event scale exhibit very different sensitivity patterns, centered around slow drainage from the lower zone supplemental store. The summarized high-resolution sensitivity results in the bottom row of Figure 9 provide a more detailed understanding of model behavior than the full period or the event scale. In general, the parameters that appear most sensitive at the event scale are also the most active for the high-resolution moving window. These primarily include the upper zone parameters UZFWM and UZK and the lower zone parameters LZFPM and LZPK. This finding aligns with our initial hypotheses, since gravity drainage and overflow from exceeding storage maxima represent two of the primary runoff generation mechanisms in the model. The most sensitive cells during the rising and falling limbs of large events represent a decomposition of the event scale sensitivity during wet period, which may be particularly valuable depending on the part of the hydrograph being analyzed. As anticipated, the upper zone and impervious area parameters dominate model performance during and immediately following large events, since these create the quick response required to reproduce observed streamflow. The high-resolution dry period exhibits largely the same sensitivities as the event scale, which would be expected considering the lack of dynamic behavior during these dry periods. Finally, the small response reflects the common scenario in which quick runoff must be avoided to achieve good performance, a behavior which remains invisible at the event scale unless a small response event is explicitly chosen for analysis *a priori*.
>
> The high-resolution results in the bottom panel of Fig. 9 can also be interpreted to identify transitions between dominant parameters and processes in the model. During the rising limb of streamflow events, the dominant processes in the model are typically direct runoff from impervious area, and overflow/drainage from the upper zone free water store. As might be expected, these processes are most dominant near the outlet of the watershed, reflecting the need for a quick response to match the observed hydrograph. During the falling limb, the model transitions to a dominant process comprising slower drainage responses from the upper and lower zone. These processes are dominant in the headwaters as well in addition to the cells near the outlet, since the longer time lag allows cells further from the outlet to contribute to streamflow. During small responses, the dominant process consists of direct runoff from impervious area and overflow from upper zone tension water, both of which must be properly attenuated to avoid overshooting the observed peak. Finally, during dry periods, a dominant process consisting of slow release from the lower zone often dominates model performance. These types of insights

regarding transitions between modeled processes are not attainable from *a priori* selection of events assumed to be broadly representative. The coarser event scale sensitivities are typically obscured, and are not necessarily consistent even for seemingly similar events (as highlighted in Figures 4 and 5).

It should be emphasized that even though Fig. 9 represents a qualitative aggregation of the high-resolution sensitivity patterns, this aggregation is drawn *a posteriori* from the full range of dynamic parameter activation characterized using the three-hour moving window. The value of the high-resolution approach, as shown in Figs. 6-8, is its ability to isolate parameter activation in space and time while avoiding the potential biases introduced by *a priori* event selection and aggregation.

This clarifies the definition and interpretation of transitions between sets of dominant parameters and processes, which as Reviewer #2 notes is one of the primary strengths of the time-varying approach. The specific comments #17, #24, and #25 also raise important issues and will be discussed in the list of individual comments.

Detailed comments

• *Comment 1, P10776 L15 I would suggest to drop the word "surprisingly". Otherwise, the authors should make clear throughout the manuscript, why they expect different performance controls for events with different forcing and initial states.*

We have followed this suggestion.

• *Comment 2, P10777 L8-9: add Reusser et al. (2009) as a study on "tracing the causes of desirable or undesirable model performance"*
• *Comment 3, P10777 L18-25: Another example of global sensitivity analysis for a spatially distributed watershed model is Guse et al. 2013*

Citations added.

• *Comment 4, Section 2.1: A description of the concepts of HL-RDHM is lacking, except for a reference to Figure 1c. However a detailed description of concepts and mechanisms is necessary to understand detailed results about the timing of different parameters.*

This is a major component of our revision, please refer to the response to Major Comment #2 above.

• *Comment 5, Fig. 2: The figure is quite confusing at a first glance. The y-axis represents on the one hand different grid cells (abstracted spatial information) and on the other hand the level of streamflow (information about magnitude of certain quantities). This needs careful introduction. The figure caption is completely silent about this double meaning of the y axis and is misleading.*

We agree that this is an important clarification. It was previously explained in the text of Section 2.2 ("The vertical axis of Fig. 2 contains the 78 HRAP grid cells of the watershed, arranged according to distance from the outlet cell.") and also in the caption of Figure 2 ("The y-axis of each plot correponds to the 78 grid cells of the basin model, sorted from the outlet cell (1) to the cell furthest from the outlet (78)"). To further clarify, we have revised the relevant sentences in the caption of Figure 2 as follows to capture the dual meaning of the y-axis in these figures.

> The y-axis on the left side of each plot represents abstracted spatial information, where the 78 grid cells of the basin model are sorted from the outlet cell (1) to the cell furthest from the outlet (78). The y-axis on the right side of each plot shows to magnitude of streamflow. […]

• *Comment 6, Fig. 2: A short comment on the different visual impression related to rainfall intensity from the top panel compared to the lower three panel would be helpful (The top panel appears to represent much lower rainfall intensities due to the thinner line). The same issue is also somewhat relevant for Figures 6 - 8, because some details may get lost due to the thin lines and should be mentioned.*

This is another important point. We have appended the following clarification to the Figure 2 caption:

> Note that hours with high precipitation are more visible in time periods 1-3 than in the 6 month simulation period due to the reduced width of hourly intervals when plotting over the full period.

• *Comment 7, P10780, L16-20: At this point, it is unclear to the reader that this describes only a part of the experimental setup and will provide the basis to compare high-resolution results to event-scale results. The experimental setup related to different time scales should be clarified at this point.*

We agree that the experiment should be better framed in this paragraph (P2 of Section 2.2). We have revised this paragraph to explain the different temporal resolutions at which sensitivity analysis is performed in this study, as shown below.

> In order to explore the potential consequences of event scale diagnostics, we select *a priori* three sub-periods to represent watershed dynamics. These are highlighted in Fig. 2 for further analysis: (1) a large rainfall event with the highest intensity precipitation focused in the headwaters; (2) a large rainfall event with similar cumulative precipitation but uniform intensity throughout the basin, and (3) a prolonged dry period with low flow. Figure 3 shows the spatial distribution of forcing for each of the three selected sub-periods. We first perform sensitivity analysis over the full 6-month period and these three sub-periods to determine the relationship between parameter sensitivities over the full period and those derived for smaller, representative intervals. We then advance this comparison by computing spatially distributed parameter sensitivities at a high-resolution moving window with a 3-hour timestep. In summary, the experiment consists of sensitivity analysis at three temporal resolutions: the full 6-month period, three representative sub-periods, and the high-resolution moving window. We seek to

understand the similarities and differences in dominant model behavior at each of these temporal resolutions. […]

• *Comment 8, P10781, L7: Suggestion to rephrase: Each of N trajectories yields one estimate...*

We have followed this suggestion.

• *Comment 9, P10781, L10-11: It is not sufficiently clear what f(x1,...,xi+di,...,xp) is. Appears to have multiple function arguments, but f(x) is only defined for one argument.*

This is a typesetting error. Since there are multiple parameter inputs, the variable **x** should be a vector (bold, not italic) consisting of *p* components. The notation in the first statement simply means that the i^{th} component has been perturbed by a distance d_i. This will be corrected during proofing.

• *Comment 10, P10781, L24: Be more specific about whether Herman et al. 2013a benchmarked Morris method against first order or total order variance.*

The comparison was performed between the μ* values from Morris and the total-order indices from Sobol. This has been clarified in the manuscript.

• *Comment 11, P10782, L1-3: Suggestion to drop the introduction of the standard deviation, because it interrupts presentation of the averaging procedure. It is confusing to the reader that the averaging is discussed on P101781, L25-26 and also P10782, L3-5*

Agreed, we have removed this sentence.

• *Comment 12, Section 3: Did you take care to avoid conceptual inconsistencies? For example, we would expect the withdrawal from the upper zone to be quicker compared to the lower zone. However, parameter ranges are overlapping, such that withdrawal from the lower zone could actually be higher then withdrawal from the upper zone for some runs. How does the model deal with such conceptual inconsistencies? This needs to be briefly touched and discussed if this could affect results in a critical way.*

The sampling bounds for each parameter (Table 1) are based on the *a priori* gridded parameter values derived by the US National Weather Service (NWS) (Koren et al., 2004) and extended for the event scale sensitivity analysis performed by Van Werkhoven et al. (2008b). While the original NWS ranges did not contain overlapping values, the extended ranges allow us to explore a wider range of variability around the *a priori* values and thus provide a broader picture of model behavior. There are two issues that could arise due to overlapping values. First, the performance of the model could be degraded, which would affect the sensitivity results. We have not noticed any such effects—even in the absence of overlapping, the upper and lower zone drainage rates are similar to one another, and model performance is not harmed if the values inadvertently overlap. The second potential problem, as Dr. Reusser notes, is that overlapping parameters may cause a conceptual inconsistency (i.e. the performance is not degraded, but the overlap violates the intended formulation of the model). While the original NWS parameter ranges did

not overlap, this does not necessarily create a conceptual inconsistency. The lower zone of the model is not intended to represent deep groundwater, so it is quite possible that a watershed would experience comparable drainage rates from the upper and lower soil zones. We understand the reviewer's concern for conceptual inconsistencies in the parameter sampling, and we are always careful to use feasible samples (in terms of both performance and conceptual formulation) when performing a sensitivity analysis.

• *Comment 13, P10784, L5: Does this implementation include the improvement of Campolongo 2007?*

Yes, the absolute μ* values from Campolongo 2007 are included in this package. This sentence has been updated in the manuscript to reflect this.

• *Comment 14, P10784, L17-18: Please provide a table where you report the original range sensitivities for each experiment.*

This comment refers to the normalization of Morris μ* indices to the range [0,1]. For the event-scale analysis shown in Figures 4 and 5, the original range was [0, 0.08]. For the time-varying indices in Figures 6, 7, and 8, the original range was [0, 0.2]. In both cases the ranges were chosen to display the range of variability of the sensitivity indices as clearly as possible.

This clarification has been added: at the beginning of Section 4.1, the beginning of Section 4.2, and in the captions of Figures 4, 5, 6, 7, and 8.

• *Comment 15, Section 4: Be explicit about when you consider a parameter to be sensitive. Is it a value of 0.5 throughout the entire manuscript?*

Yes, we consider a normalized sensitivity value greater than 0.5 to be sensitive. This is clarified in the second paragraph of the discussion section when introducing Figure 9. There is always an element of subjectivity in separating "sensitive" from "insensitive"—since the Morris indices do not have a direct interpretation as a percent variance (like Sobol), the indices can only be interpreted relative to one another. We believe that this threshold of 0.5 provides sufficient inclusion of interesting sensitivity patterns without including so many parameters that the signal becomes meaningless.

• *Comment 16, Figure 3-5: For each panel, please also provide an alternative representation using the approach arranging the cells along the y-axis. This will allow better comparison between results from the various figures (e.g. precipitation patterns show in Fig. 2 and 4)*

We agree that placing a vertical arrangement of grid cells adjacent to each map would clarify the interpretation of the high-resolution figures later in the paper. However, this would add a significant amount of complexity to the figures, and is not readily achievable. For example, Figures 4 and 5 each contain 56 individual map panels, and are already quite complex figures.

The key issue here is whether the high-resolution figures (with grid cells arranged along the y-axis) are easily interpretable for readers. We have taken precautions to explain this figure format each time it appears, in both the text and the captions, so that readers can quickly interpret our results.

• *Comment 17, P10785 L21-25: This is one possible explanation. An alternative explanation would be the following: Figure 2 shows multiple short rainfall events for Period2. River discharge reacts only after the last rainfall event which is concentrated in the lower catchment. We may hypothesise that only this last rainfall event was actually causing the discharge reaction and that earlier rainfall is insignificant. This would be an alternative explanation for the observed concentration of parameter sensitivity in the lower catchment. The way results are currently presented does not allow to make a clear judgment about which explanation is more consistent with the processes in the model. This is one possibility how authors may better link parameter sensitivity patterns to hydrological processes (see my general comment (2)). Presenting the results in a way that would clearly indicate the ongoing processes (in the model) and would allow to discriminate the two potential explanation could improve the benefit from this study.*

The comment refers to the comparison between the spatial distribution of precipitation in Periods 1 and 2, and the corresponding model response in each case (Section 4.1). In order to investigate model processes over these shorter periods, we can zoom in to these events for some selected parameters:

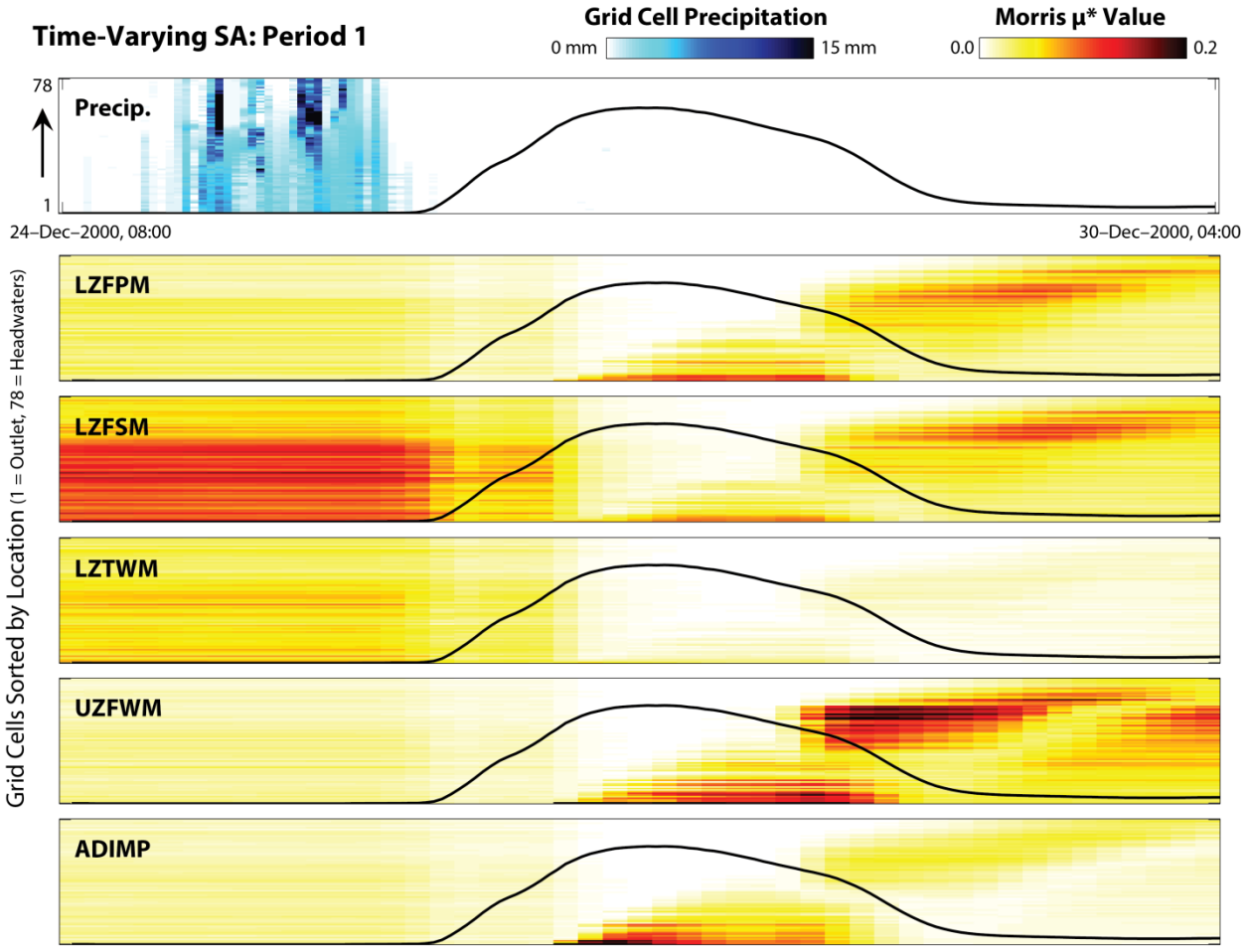


Figure R1: Time-varying parameter sensitivity during Period 1, for select parameters.

The precipitation in Period 1 is concentrated in the headwaters of the basin. The initial response from the upper zone (here characterized by the sensitivity of UZFWM) begins near the outlet, but eventually becomes very strong in the headwaters as time progresses.

We can compare this to a zoomed-in view of Period 2, again for select parameters:

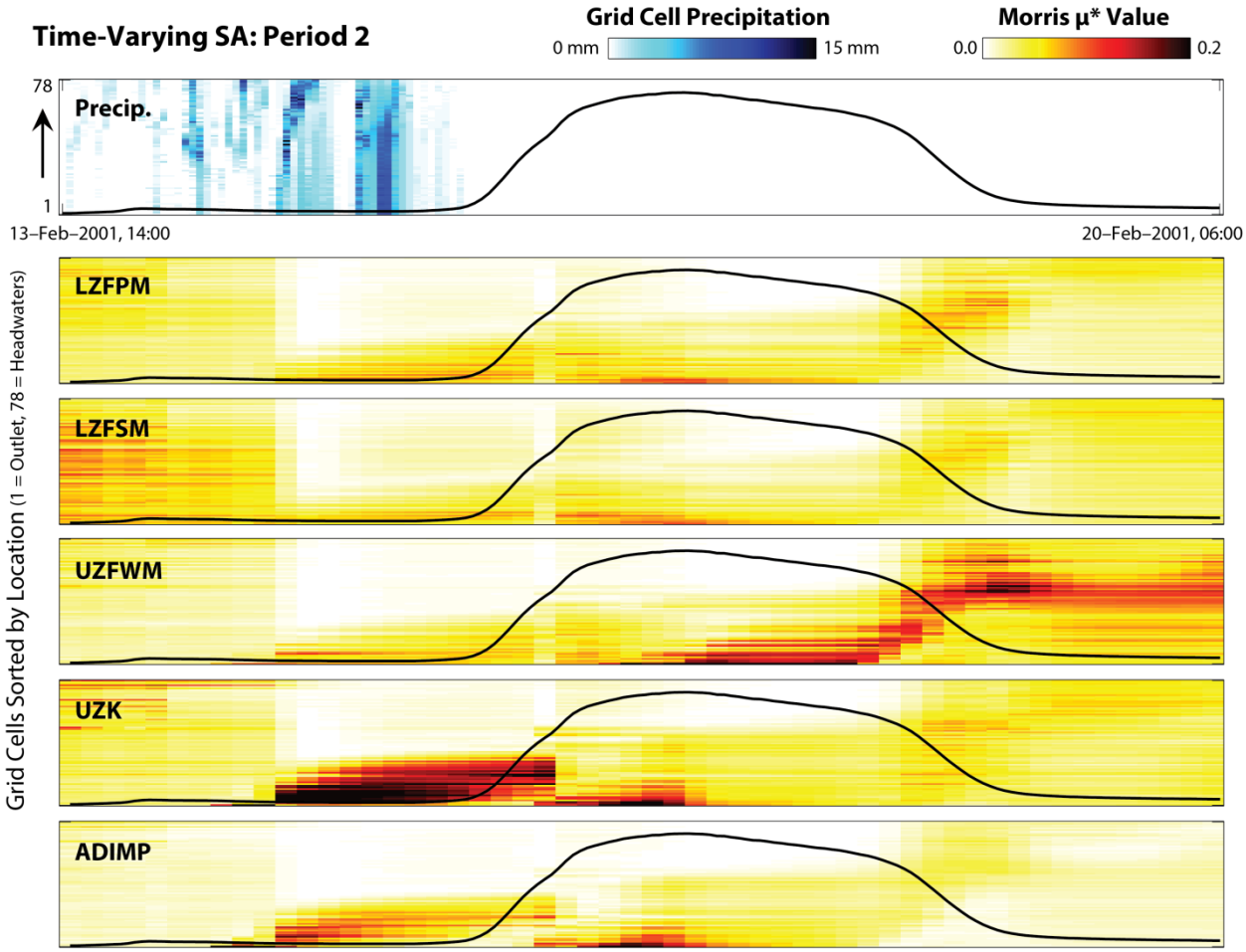


Figure R2: Time-varying parameter sensitivity during Period 2, for select parameters.

The precipitation in Period 2 is more distributed over the watershed. As Dr. Reusser has noted, this period really contains two separate precipitation events, one in the headwaters and a second more toward the outlet. The question is whether the initial event provokes a model response, or if the response does not occur until after the second event. From this zoomed-in view, we can clearly see that the first event does cause a strong response – not in terms of streamflow, but in terms of the sensitivity of the upper zone parameters UZK and ADIMP. This sensitivity is concentrated near the outlet and never truly propagates to the headwaters (with the exception of UZFWM, much later). From the comparison of Figures R1 and R2, we note that the headwater cells are only likely to become sensitive when precipitation is concentrated there. Otherwise, the streamflow response will be biased toward the outlet cells, as we might expect for a peak-focused metric like RMSE.

As Dr. Reusser correctly suggests, there may be internal information which would complicate our interpretation of these events, such as antecedent moisture conditions. This underscores one of the key points of the paper, the difficulty (or impossibility) of identifying "representative" events. The shape and magnitude of the streamflow response in Events #1 and #2 are extremely similar, yet their sensitivity signatures are quite different in space and time. We have added Figures R1 and R2 as supplemental material, along with the following explanation in Section 4.1 Paragraph 2:

> These differences in the responses of Periods 1 and 2 are potentially complicated by internal model states, such as antecedent moisture conditions, which could alter the response signatures. However, as shown in the supplemental material, zooming in to the time-varying sensitivity of Periods 1 and 2 clearly reveals the differences in the spatial distribution of precipitation between the two events. The sensitivity responses to each event begin almost immediately following the precipitation, and thus their differences may be traced primarily to the precipitation distribution over the watershed.

• *Comment 18, P10787, L9-12: The previous comment needs to be reflected here as well.*

The comment refers to the following sentence:

> These findings align with previous work: spatially concentrated precipitation will cause a similar concentration of sensitivity, whereas distributed precipitation will cause sensitivity in cells near the outlet (Tang et al., 2007; Van Werkhoven et al., 2008b).

Based on Figures R1 and R2 above, we do not believe the sentence requires revision at this time.

• *Comment 19, P10788, L4-6: How critical is the choice of the size of the moving window for the results presented? How was the choice made? How do the 24 hours compare to a typical event? How does potential noisiness in rainfall and discharge data affect the results for such a short time window?*

The moving window size is quite critical, for two reasons. First, the window size must be large enough to smooth out noise in forcing and discharge data, particularly in the simulated discharge. Second, even in the absence of noise, different parameters will be sensitive for different window sizes. A coarse window will fail to identify fast processes, and a finely resolved window will miss slower processes.

The 24-hour window represents a balance between these two issues. It is large enough to smooth out any noise in the RMSE calculation (the model runs on an hourly timestep). Since most events during the simulation period are approximately 48-72 hours in length, a 24-hour window is also small enough to capture processes within each event, which is useful from a flood forecasting standpoint where fast-response processes are important.

The choice of a 3-hour window timestep also helps to smooth out any noise in the sensitivity indices, since a large portion of the moving window will overlap with the previous one. The following text has been added to the beginning of Section 4.2 to address these issues:

The choice of window size contains two competing considerations: it must be large enough to smooth out any noise in the performance metric calculations, yet also small enough to capture dominant model processes that occur on fast timescales. The moving window size and timestep used in this study reflect a balance between these two issues. Since the model runs on an hourly timestep, a 24-hour window size (with significant overlap due to the 3-hour timestep) will smooth noise in the calculation of the RMSE metric. Additionally, since most large events during the simulation period are approximately 48-72 hours in length, the 24-hour window is also sufficiently small to capture quick responses, which is critical from a flood forecasting standpoint.

• *Comment 20, P10788, L13-14: Please also provide movies for single parameters and combinations of parameter groups if there are no size limits on the multimedia supplements.*

There is a 50 MB limit for multimedia supplements, which is plenty of space to include more movies. We have included animations of individual parameter sensitivity and upper/lower zone, as suggested. These are divided into select events of interest during the simulation period in order to limit the size of each individual file.

• *Comment 21, P10788, L15-21: This is a repetition and can be replaced by a reference to Figure 2*

The comment refers to the explanation of the figure format where grid cells are arranged on the y-axis according to their distance from the outlet. This is a repetition of the explanation in Figure 2, but an important one—as discussed in Comment #16, we believe it is necessary to provide multiple explanations of this figure format since it is crucial to readers' interpretation of the results.

• *Comment 22, Figures 6-8: Also provide Fig 6-8 for selected events as supplementary material (issue with the visual impression apparent in figure 2).*

We have produced these figures to show time-varying sensitivity indices zoomed in to the individual periods #1-3. They are included in the supplement in PDF form. These also correspond to Figures R1 and R2 in this document, which show Period 1 and 2, respectively.

• *Comment 23, P10789 L2-4: Comparing fig 4 and fig 6 we find important differences: In Fig 6 Parameter LZSK shows high sensitivities (brown) in the lower part of the catchment. In Figure 4 however, sensitivities are high in the middle of the catchment (brown) but low in the lower catchment (yellow) - see also my general comment (1). The authors need to analyse, present and discuss such differences. The statement on p10789, L 3 (insighs largely align with those found ... in Figs 4) is not sufficient. Also, This is another reason to provide the representation with spatial information on the y axis in Figure 4.*

We thank the reviewer for bringing this to our attention, because it highlights an interesting and important issue when comparing sensitivity indices across timescales.

Figure 4 shows the event-scale sensitivity indices, while Figure 6 shows the time-varying indices. We assume the reviewer is referring to Period #1, when the sensitivity of LZSK is concentrated near the middle of the watershed (in Figure 4). Sensitivity indices are always a relative measure of influence, so the sensitivity of LZSK depends on the magnitude of other parameters' sensitivity. At the event scale, LZSK is insensitive in the lower portion of the watershed because other parameters are so strongly dominant (for example, UZFWM), whereas in the middle of the watershed its effects are sufficient to warrant a moderate level of sensitivity. In Figure 6, however, sensitivity is calculated for many individual timesteps, so there is less "competition" among parameters to achieve high levels of sensitivity. Notice that during Period 1, LZSK is sensitive until the event begins, at which point the upper zone parameters take control. We have clarified this issue at the location in the text pointed out by the reviewer:

> These insights largely align with those found at the event scale in Figures 4 and 5. However, when comparing sensitivity indices across temporal resolutions, it is important to note that parameter sensitivity is measured relative to other parameters. Thus, a larger time window may cause some parameters to appear insensitive at certain locations due to the dominance of others. This phenomenon is visible, for example, for LZSK parameter near the outlet of the watershed, where it is sensitive in Figure 6 but not in Figure 4. Compared to the event scale analysis, the high temporal resolution in Figure 6 has the advantage of clarifying the timing of parameter activation.

• *Comment 24, P10789 L16-18: Is this expectation due to the water routing only or are there other mechanisms leading to this expectation. Be explicit. This might already be presented in section 2.1.*

Yes, this is due to routing. We have revised as follows:

> As expected, there is a lag between the time at which the event begins and the time at which the headwater cells begin to affect the model performance due to routing.

• *Comment 25, P10790 L1-6: I like this part already a lot. It could be further improved if you provide more explanation about the conceptual bases of the two parameters UZTWM and PCTIM in Section 2.1. Moreover, can we go a step further in understanding the mechanisms? Are impervious areas and tension storage capacity more likely to refer to independent processes or are they interconnected?*

The comment refers to the phenomenon in which these two parameters appear sensitive when the streamflow response cannot be too large (i.e., the absence of these processes is required to achieve good performance). The conceptual bases of these and other parameters have been provided in Section 2.1 (see response to Major Comment #1). The tension storage capacity UZTWM causes runoff via overflow, and the impervious area PCTIM converts precipitation directly to runoff. These processes are unlikely to be interconnected in the model, because direct runoff from PCTIM will occur immediately and independently of any other process. (In a real-world watershed, however, the two processes may be indistinguishable from one another). We have added clarification in this sentence:

> If impervious area is too high (causing high direct runoff), or tension storage capacity too low (causing runoff via overflow), the model may overestimate streamflow and create significant errors in the RMSE metric.

• *Comment 26, P10790 L6: This phenomenon is not visible at event scale because you did not analyse such periods. I'm quite confident that if you define an additional event (for example just before Period1) you would be able to see this. I suggest that you add this to the discussion about the impossibility to find representative events.*

This is a point well taken. The event scale analysis did not include any "small response" events, where this would have occurred. We have revised as follows:

> This phenomenon is not visible for the events analyzed in Figure 4 because it would be difficult to predict at the time of *a priori* event selection.

Also, as suggested, we have added a brief sentence to the discussion section using this discovery as an example of the weakness of representative event selection:

> For example, this difficulty is demonstrated by our inability to foresee the phenomenon in which parameters UZTWM and PCTIM are most sensitive when modeled responses require attenuation to match observations.

• *Comment 27, There is a second exception to the statement from P10789L27-29: UZK shows high sensitivity quite a long time after an event during low flow periods. This should also be briefly presented.*

We have added this second exception as suggested:

> As Figure 7 indicates, the upper zone parameters typically do not control model performance during low-flow periods and small events. There are two interesting exceptions to this, however. First, the drainage rate UZK remains sensitive for several hours after each event as the upper zone drains its storage. That is, gravity drainage from the upper zone typically occurs slowly enough such that it continues to release water well into the low-flow periods. Second, the parameters UZTWM (upper zone tension storage) and PCTIM (percent impervious area) are most sensitive following rainfall events which do not lead to large streamflow events. […]

• *Comment 28, P10790 L8: Please present and discuss other inconsistencies between Figure 4 and Figure 7. For example: Event1, UZK Figure 4 shows high sensitivity in the middle of the catchment, while Figure 7 shows sensitivity only in the lower part of the catchment.*

This is a helpful addition. We have added the following text at the location suggested:

> An additional difference between Figures 4 and 7 is that, while the UZK parameter is sensitive throughout the watershed at the event scale (e.g., during Period 1), it only

appears sensitive near the outlet at the high-resolution timescale. This could be due to the dominance of other parameters in the upper region of the watershed at the high-resolution timescale, or simply a difference in the sensitivity of the upper zone drainage response at different timescales.

Beyond that, another noteworthy difference concerns parameter UZFWM during Period 2. At the event scale, its sensitivity is clearly restricted to the lower portion of the watershed near the outlet. However, at the high-resolution timescale, its sensitivity propagates to the headwater region during the falling limb of the hydrograph. Due to the nature of the RMSE metric, this portion of the hydrograph contains less influence in the calculations. We have appended this additional explanation:

Finally, the UZFWM parameter during Period 2 is only sensitive near the watershed outlet when considered at the event scale, but the high-resolution results in Figure 7 show that the sensitivity of UZFWM propagates to the headwaters during the falling limb of the event. Since the RMSE metric focuses on peak flows, the falling limb does not play a significant role in the calculation of aggregated sensitivity for each period, even though these are clearly visible at the high-resolution timescale.

• *Comment 29, P10790 L19: Please present and discuss inconsistencies between Figure 4 and Figure 8. For example REXP, Event 1: Figure 4 high sensitivity in the middle of the catchment while absent in Figure 8*

Yes, this is likely an issue of the timescale of response. REXP will contribute to model performance on a slower timescale, so it appears occasionally at the event scale, but not so much for the high-resolution timescale. We have added the following clarification at the location suggested by the reviewer:

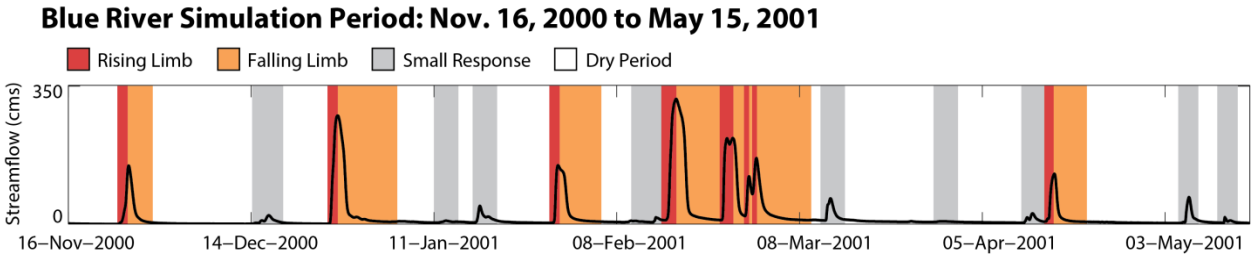
Comparing to the event scale results in Figure 4, these parameters are generally much less sensitive at the high-resolution timescale. For example, the REXP parameter modifies the rate of percolation from the upper to lower zone but does not cause runoff directly, which makes it less likely to influence model performance during the 24-hour moving window than over the course of an aggregated event.

• *Comment 30, Figure 9: Such a graphical summary is very nice. However, the distinction between events/periods appears somewhat arbitrary. For the event time scale, I would avoid combining events 1 and 2. There are important differences in parameter sensitivity. For the high-resolution case, the authors need to provide reason, why the four cases presented are sufficient to represent all periods. For example, by relating each point in time to one of the four cases.*

We intend for Figure 9 to serve as a qualitative summary of the results found in Figs. 4-8. It is true that some subjectivity exists in separating the events/periods. At the event scale, we combine Events 1 and 2 into a generalized “Wet Period”, using insights taken from both events. We believe that dividing the event scale analysis into wet and dry periods is the clearest way to present results, since the format of Figure 9 does not incorporate any information about characteristics that may cause differences in sensitivity between wet periods of similar magnitude. These characteristics, such as the spatial distribution of

precipitation and antecedent moisture conditions, have been discussed throughout the text in terms of the differences they may cause in parameter sensitivity. We believe it will be clear to readers that the "wet period" event scale results in Figure 9 contains general insights from Events 1 and 2, while the specifics of those individual events can be seen previously in Figures 4 and 5.

For the high-resolution case, the four cases shown (rising limb, falling limb, small response, and dry period) are again intended to reflect general insights from Figures 4-8. We do not necessarily intend for these to collectively represent all timesteps in the simulation period, but rather the most interesting classifications that were observed in the results. However, because this particular watershed is driven primarily by infrequent large events, without significant temperature or elevation effects (e.g. snow), it is possible to broadly separate the hydrograph into these event types as shown below.



We will include a PDF version of this figure in the supplement so that readers can see which specific time periods we are referring to in the general summary of results in Figure 9. This explanation has been added to Paragraph 2 of the discussion section:

> For the high-resolution timescale, the four cases shown (rising limb, falling limb, small response, and dry period) are intended to reflect general insights from Figures 4-8. These cases do not necessarily reflect all timesteps in the simulation period, but rather the most interesting classifications that were observed in the results. However, since the Blue River watershed is driven primarily by infrequent large events without significant temperature or elevation effects, it is possible to broadly separate the hydrograph into these four classifications, as shown in the supplemental material.

• *Comment 31, P10791 L20: How much is this determined by your selection of performance metric? RMSE is known to focus on large events.*

The reviewer is correct, this result does depend on the choice of performance metric. As shown in Figure 5, the choice of a metric other than RMSE (in this case, ROCE, the water balance error) produces very different results. We have amended this statement as follows:

> The full period sensitivities are clearly influenced by the wet periods at the event scale, which exhibit the same responses, indicating that the aggregate period is biased toward these large events (a result consistent with the focus of the RMSE metric).

• *Comment 32, Section 4.3: A critical reflection and discussion of the limitations of your approach is lacking.*

We agree that this discussion was missing from the initial manuscript. We have added the paragraph below to Section 4.3, which includes discussion of several of the limitations that have been correctly identified in the reviewer's prior comments.

> The high-resolution sensitivity approach presented here requires several important considerations. First, as with any sensitivity analysis, the results strongly depend on the choice of performance metric. Figures 4 and 5 show the substantial differences in sensitivity indices that occur when changing from the RMSE metric to the ROCE metric. We focus on the RMSE metric in this study because its emphasis on large events is consistent with our goal of understanding model behavior in the context of quick-response flood forecasting. Interestingly, as the window size of the analysis decreases, different performance metrics begin to behave similarly (i.e., in the limit as window size approaches zero, most metrics reduce to a percent error at a single point). This leads to the second important consideration, the choice of window size. Modeled processes which dominate performance at one timescale may be invisible at another, so it is crucial to choose a window size commensurate with the purpose of the analysis. Our moving window size of 24 hours (with a 3-hour timestep) reflects the need to capture dominant processes on a fast timescale while also containing a sufficient number of timesteps to smooth out any noise in the performance metric. Finally, the visualization approach presented in Figures 6-8 (in which the spatial dimensions are compressed into a single distance measure on the y-axis) is readily applicable to the narrow Blue River watershed, but may be difficult to extend to other watersheds. This is particularly true if significant spatial heterogeneity exists in land cover or soil storage properties. In such instances, it may be preferable to represent the y-axis as, for example, the soil storage capacity of each grid cell, or whichever characteristic is expected to govern grid cell sensitivity. For this case study, the primary characteristic of interest is simply the distance from the watershed outlet, but this may not hold true for all applications.

• *Comment 33, P10793 L19: This opportunity to identify location and timing for data collection is very interesting. Could you elaborate a little more and explain how you would extract this information from the large amount of data presented?*

The goal would be to identify where and when to perform data collection in order to improve the modeled representation of hydrologic processes. Of course, since this is a conceptual soil moisture model, we cannot collect data to truly measure the parameters. However, we can use the spatial and temporal distributions of sensitivity to inform measurements of runoff fluxes in a particular set of grid cells to achieve a certain goal. For example, if our objective is to improve model representation of flood forecasts, we would notice that peak flows are typically controlled by a few grid cells near the outlet of the watershed, immediately following a precipitation event. By measuring fluxes at these points during an event, we would obtain additional observations against which the model could be calibrated. Such observations could also be used to falsify the model, if the modeled internal fluxes between grid cell boundaries prove inaccurate despite the accuracy of the overall output. Conversely, if our objective is to improve the modeled representation of the long-term water balance, we would need to spread

measurements across the watershed, as evidenced by Figure 5. We have addressed this issue in the conclusion section as follows:

> However, it also presents a valuable opportunity to overcome the complexity of distributed parameter identification by restricting search to only those parameters which are active at a specific time and location. It also suggests an opportunity to identify locations and timing for optimal data collection to improve the modeled representation of hydrogic processes, particularly under nonstationary conditions in which dominant watershed processes fall outside observed ranges. For example, the results of this study indicate that large streamflow events in the model are controlled primarily by upper zone fluxes quite close to the watershed outlet; collecting flux data in only this area during a large event could provide justification to falsify the internal processes of the model, and to improve them by calibrating against the new observations.

• *Comment 34, P10793 L21-22 How far is your visualization approach limited to this special case of a long and narrow catchment? I could imagine that results are more confusing for other catchment shapes if arranged according to distance from the catchment outlet. Imagine a catchment with bare bedrock on the one side of the river, while the other side has soils with large storage capacity. How could your visualization approach be extended to cover such a case?*

This is a valid point. While our visualization approach applies easily to the Blue River, watersheds of different shapes or soil characteristics will pose more difficulty. We have addressed this issue in our discussion of potential limitations in Comment #32:

> Finally, the visualization approach presented in Figures 6-8 (in which the spatial dimensions are compressed into a single distance measure on the y-axis) is readily applicable to the narrow Blue River watershed, but may be difficult to extend to other watersheds. This is particularly true if significant spatial heterogeneity exists in land cover or soil storage properties. In such instances, it may be preferable to represent the y-axis as, for example, the soil storage capacity of each grid cell, or whichever characteristic is expected to govern grid cell sensitivity. For this case study, the primary characteristic of interest is simply the distance from the watershed outlet, but this may not hold true for all applications.

To achieve a similar visualization in such a case would require some creativity, but there is typically some variable expected to have a large effect on the sensitivity of each grid cell that could be plotted on the y-axis, using the same plotting approach presented here. It would be certainly be more difficult, but not impossible, to identify a plotting format that conveys the maximum possible insight in this scenario.

We thank Dr. Reusser again for his highly thoughtful contributions to our work.

Technical issue

P10778 L7-8 The statement "This approach has been limited to lumped models" needs reformulation since WaSiM-ETH used in Reusser et al. (2011) is spatially explicit.

We have corrected this error in the text.