Hydrol. Earth Syst. Sci. Discuss., 11, C6748–C6760, 2015
www.hydrol-earth-syst-sci-discuss.net/11/C6748/2015/

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Exploring the impact of forcing error characteristics on physically based snow simulations within a global sensitivity analysis framework" *by* M. S. Raleigh et al.

M. S. Raleigh et al.

raleigh@ucar.edu

Received and published: 15 April 2015

Note: reviewer comments are in italics and the authors' responses and manuscript revisions are in normal face.

Comment: *This paper investigates how errors in meteorological observations affect the simulations of a physically based one-dimensional snow model (the Utah Energy Balance). Global sensitivity analysis (GSA) is used to quantify the relative contribution of different error characteristics (bias, magnitude, presence of random errors, error distribution) to the uncertainty in four snow variables (SWE, ablation rates, snow dis-*

appearance and sublimation). GSA results are presented for four study sites in distinct snow climate. Detailed studies focusing on forcing uncertainty are relatively few, and they are needed particularly in snow-affected watersheds where meteorological measurements are scarce and forcing uncertainty can significantly impact model simulation. This work provides useful insights on the topic and establish a methodology that could be extended to other physically based models or error types. I think the analysis here described is interesting and solid, the paper is clear and well structured, and its contribution is well placed in the literature. I have some concerns about the reliability and interpretation of some of the GSA results, and a number of specific comments that the authors may consider in revising their manuscript. I think the paper should be considered for publication on HESS after such revisions.

Response: Thank you for your encouraging and careful review.

Comment: *1) Some of the results in Figure 6 and 7 are a bit surprising and need clarification. For instance, in the cases of Fig. 5.a and 5.e, bias in P is the only influential parameter. However, when including random errors (Fig. 6.a and 6.e), all parameters become (almost equally) influential. In the text, this is explained as being due to interactions between parameters. I agree in principle but I think a more detailed analysis is needed. For instance, do bias parameters $\theta B;i$ become influential through interactions with parameter $\theta RE;i$ of the same meteorological variable? Or does this happen through interactions with $\theta RE;i$ of different forcings (for instance, bias $\theta B;i$ of Tair interacting with random error magnitude $\theta RE;i$ of P)? I guess the physical interpretation of the result and its implications would be very different in the two cases. For instance, if the interactions occur within the same forcing error equation, it would mean that the bias in the observations is not influential per se, but it becomes influential if there are also random errors. Does this make sense from the physical point of view? Or is it a result of some inadequacy in the error structure of Eq. (4)?*

Response: Thank you for making this excellent comment. After double-checking our code, it appears that there are some inadequacies in our implementation of the error

structures (eq 4) in scenario NB+RE. Specifically, we discovered that the random number generator (randn.m in Matlab) used to create the "noise" (i.e. random errors) did not always have a mean of 0 (though it was a value close to 0). This is because it is a discrete array with samples drawn from a population of mean 0; hence the sample mean is not guaranteed to be 0. Because of a non-zero mean in the noise, the "random error" term also introduced additional systematic errors that were not accounted for in the bias terms.

While our results support the role of random errors in introducing error interactions (pg. 13763, line 16), our focus on the total sensitivity indices (for a more focused analysis) prevented us from exploring specific interactions in the original manuscript. A more quantitative link would be interesting to pursue and would require examination of the second-order sensitivity indices to illuminate the relationship between biases in a specific forcing (e.g., Tair) and random errors in another (e.g., P). Calculation of these second-order terms would require nearly double the number of simulations (compare n(2k+2) vs. n(k+2) in the current analysis) (Saltelli, 2002), and hence we have not pursued this extended analysis due to the additional computational expenses required.

Manuscript Revisions: We have corrected this coding issue, reran NB+RE and have found that this minimized the problem you have found (see Figure 1 in this reply, below). The figures and text have been updated to reflect these corrections. We find two improvements with this fix: (1) there is now better discrimination between the bias and random error factors, and (2) the "nugget" effect (i.e., a minimum level of sensitivity across all factors) is substantially reduced across all scenarios, except for ablation rates at IC. We think that there exists a physical explanation for this one exception, namely that the short ablation season at IC accentuates the sensitivity of ablation rates to a variety of error types.

Comment: *Also, in all sites and for all outputs, the sensitivity indices of θRE;i are almost the same for all i. This is strange. Does it make sense that errors in all meteorological variables have the same importance, or is there a purely numerical explanation*

for this?

Response: This partially relates to the numerical implementation problem described in our previous response. As we indicated above, we have fixed the issue with non-zero mean for the random error assignment and found improved discrimination between sensitivity indices. In general, we think it is realistic to have similar sensitivity indices for random errors in different forcings because the nature of random errors is that they tend to cancel out (due to the requirement for bias=0). Additionally, in the revised results for NB+RE, most sensitivity indices for RE are close to zero, and in this case it is reasonable for them to all have the same level of non-importance.

Manuscript Revisions: See previous comment.

Comment: *2) I am not sure that Figure 9, 10, 11 are the most effective way to compare GSA results. The main conclusion drawn in the text is that overall GSA results are similar across scenarios NB, NB+RE and UB. Scatter plot visually confirm this. However, they do not facilitate one-to-one comparison of sensitivity indices (bar plots with two coloured bars would be better), which in my opinion would provide more interesting information. For instance, comparing Fig. 5.o with 7.o I can see a big increase in the influence of U bias when moving from scenario NB to UB; comparing Fig. 5.e with 7.e shows that in the NB scenario only P bias is important, while in the UB scenario the bias of other meteorological variables also matter. Can you explain these behaviours? Maybe an interpretation effort of these results might lead to learning important aspects of the model behaviour.*

Response: We have considered your comment here and have produced new figures with dual color bars instead of scatterplots (for one example, see Figure 2 in this reply document). We agree with you that this is a more effective way to show the data and thank you for the suggestion.

The example you have highlighted (Fig 5e vs 7e, ablation rates at IC) is a bit of an outlier in terms of the sites and outputs considered. Figure 2 (this document, below)

illustrates that while the values of the total-order indices change somewhat between NB and UB, the relative importance of the forcing errors does not usually change. The case you highlighted is the only one where there is a drastic shift in total-order indices between NB and UB. Nevertheless, we hypothesize in section 4.1 that the ablation rates at IC is a different case because the melt season is so short relative to the other sites, and thus the site may be comparatively less stable in terms of what types of errors dominate the melt rates. Additionally, under the UB scenario, the wind (U) bias is an important factor to ablation rates, and this might have a physical basis in that this site is the most exposed site and has the highest wind speeds. In UB, the uniform distribution makes extreme wind biases more common, and these considerably reduce or enhance the sensible heat contribution toward the ablation rates at IC.

Manuscript Revisions: All scatterplots have been changed to bar plots and text has been updated to reflect these new figures.

Comment: *3) Motivation of the study (in both the abstract and the introduction). I would add some comments on how the authors think that GSA results (which error characteristic matter most) could be used in practice. What are the implications of these results? How would you expect to use this piece of information? I think one way to use GSA results is to spot unexpected behaviours and thus have directions for further investigation of simulation results. However, I feel that this is somehow missing in the paper (see also my previous comment).*

Response: This is a reasonable observation and we thank you for making this suggestion. We now elaborate how we expect knowledge of specific error characteristics might be beneficial to practical applications.

Manuscript Revisions: We now state in the introduction, “In our view, it is important to clarify the relative impact of specific error characteristics on modeling applications, so as to prioritize future research directions and to inform network design. For example, given a constrained budget, is it better to invest in a heating apparatus for a radiometer

(to minimize bias due to frost formation on the radiometer dome) or in a higher quality radiometer (to minimize random errors associated with measurement precision)? Additionally, it is important to contextualize different types of meteorological data errors for snow modeling applications, as these errors are usually studied independently of each other and it is unclear how they compare in terms of model sensitivity.”

SPECIFIC COMMENTS **Comment:** *page 13755: "The goal of sensitivity analysis is to quantify how variance in specific input factors (...) influences variance in specific outputs". This sentence is inaccurate. First, the use of output variance as a proxy of output uncertainty is a specific assumption of variance-based SA (Sobol') and it is not a general assumption of GSA. Many other GSA methods are available that do not rely on this assumption, either because they simply do not look at output distribution (e.g. the Morris method) or because they consider other properties of the output distribution (e.g. density-based methods, see for instance Peeters et al. 2014). Second, also within the variance-based approach, the output variance is related to generic variability of input factors (reproduced by random sampling or Sobol' sampling) and not their variance only.*

Response: Thank you for catching this inaccurate statement. You are correct that this statement only applies to variance-based SA methods and excludes other SA methods.

Manuscript Revisions: We have now modified the sentence (based on Matott et al., 2009) to be more broadly encompassing: “The goal of sensitivity analysis is to determine which input factors are most important to specific outputs.”

Comment: *One assumption of the Sobol' method (at least in the implementation used in this work) is that input factors are uncorrelated. In this case, this means that: in the NB+NR scenario, bias and magnitude of random errors are independent; and in all scenarios, bias (and random errors) of different meteorological observations are independent. Are these reasonable assumptions?*

Response: For the error types, we argue that these are reasonable assumptions be-

cause by definition, bias and random errors are independent. Random errors introduce noise/variance without changing the mean value (i.e., the bias), whereas bias describes the systematic errors. As we note in section 3.2.1, there are no widely used metrics to report random errors separately from bias, as root mean square error and mean absolute error encapsulate both systematic and random errors. Hence, the random errors specified in our study are hypothetical in nature, and do not exactly conform to these widely used metrics.

For the same type of error but for different variables, it is possible that there will be error-linkages in real-world conditions. As one example with measured forcings in a sunny environment, an air temperature sensor (no mechanical ventilation) may be subject to a positive bias, which then can induce a negative bias in the RH data. As an example with estimated forcings, a positive bias in the maximum daily air temperature will bias the diurnal temperature range, which in turn would bias estimates of atmospheric transmissivity and hence bias the calculated shortwave and longwave radiation.

Manuscript Revisions: We now note in section 3.3.2, "A key assumption to the Sobol' approach is that the factors are independent; hence, our analysis does not consider the case of when specific error types are correlated (e.g., a positive measurement bias in Tair that propagates a negative bias to RH)."

Comment: *Page 13755: "by creating k new parameters (θ1, θ2,..., θk) that specify forcing uncertainty characteristics". This is a bit confusing, mainly because up to this point the symbol θ was used to refer to model parameters in contrast to forcing inputs F. The same confusion may arise in the following section, when the symbol θ and the term "parameters" may be interpreted as referring to model parameters (and Eq. (1) reinforce this misinterpretation). I would suggest to use a different symbol for the model parameters in Eq. (1) (for instance, p), and maybe insert a second equation like*

$$\mathbf{Y} = M(\mathbf{F}, \theta, p) \tag{1}$$

as a companion to Eq. (1) to clarify the point (and also to link to the error model of Eq.

(4)).

Response: We can see how this convention would be confusing, and thank you for pointing this out.

Manuscript Revisions: We followed your recommendation and introduced a new symbol (phi, ϕ) for the new forcing error parameters (section 3.3.1) for better discrimination from the native model parameters (theta, θ). We added a new equation after equation 1 to help clarify, and changed all other references from theta to phi.

Comment: *Page 13759: "The number of rejected samples varied with site and scenario...". I think the step of screening out meaningless simulations before estimating sensitivity indices is a very good practice, unfortunately not always applied in SA applications - the authors may want to stress the relevance, also referencing other works where this was done (for instance the already cited Pappenberger 2008). Also, it would be interesting to know if this screening provided further insights about the model response surface. For instance, did you find that discarded simulations where generated by input samples falling in a specific range or were they scattered across the input space? In the former case, can you give a physical interpretation to this result? Also, it is reported that the UB scenario at SASP had a very high number of meaningless simulations: can you give an interpretation for this? Does this relate to any specific property of the SASP site?*

Response: We have examined the characteristics of the discarded simulations and are able to provide a physical interpretation. We found that simulations were more often rejected because too much snow was simulated (and hence the snow never fully disappeared) instead of too little. SASP had the most rejected simulations in UB because it had the highest peak SWE and hence was more prone to have too much snow simulated. The boxplot in Figure 3 (this document, below) summarizes the characteristics of the passed and failed simulations for SASP in the UB scenario. The most distinct characteristics of the failed simulations was a high precipitation bias,

which lead to high peak SWE and no snow disappearance. This is not surprising given how the error ranges were assigned to precipitation (with a larger range on the positive bias end to mimic snow drift errors). Other contributing characteristics were cases with a negative bias in Qsi, Qli, and Tair (all of which lead to slower melt and reduce the chance of snow disappearing).

Manuscript Revisions: We now stress the relevance of screening out meaningless simulations and cite the Pappenberger paper as an example where this was also done (section 3.3.5). We also generalize the characteristics of the rejected simulations (at the end of "Step 6" in section 3.3.5).

Comment: *Page 13762: "This was surprising given that bias magnitudes are lower for Qli than for Qsi." Misleading. It seems to suggest that the input with the larger variability range is expected to have the larger influence on the model output, which is not true unless the model is linear (and which motivates the use of complex SA methods to obtain input ranking).*

Response: You are correct that the non-linear nature of the model does not guarantee this is true. However, we note that albedo also plays a role in minimizing the effect of errors in shortwave radiation.

Manuscript Revisions: We have rephrased this sentence (section 4.2) to provide a more physically based explanation of what is happening here: "However, the albedo of snow minimizes the amount of energy transmitted to the snowpack from Qsi, thereby rendering Qsi errors less important that Qli errors. Additionally, the non-linear nature of the model may enhance the role of Qli through interactions with other factors."

Comment: *Page 13766: "1 520 000 simulations for examining only a single year at four sites across four error scenarios." Misleading: the number of simulated years influences the computing time of each simulation but not the number of simulations. See also next comment on the issue of number of simulations vs computing time.*

Response: We understand your argument and agree.

Manuscript Revisions: We have removed the reference to the number of years and rephrased this to say "1 840 000 simulations across four sites and five error scenarios". Note that we have now include a fifth scenario to address concerns raised by another reviewer about precipitation uncertainty, and this brings the total number of simulations to 1 840 000.

Comment: *Page 13767: "will be more feasible in the future with better computing resources and advances in sensitivity analysis methods". The computing issue here is not completely clear. Over one million model evaluations is a big number but what is the actual computing time? Given that the model is one-dimensional I would expect every model evaluation to be rather fast, and therefore even 1 million evaluations to be a reasonable target. Also, before Rakovec et al. (2014), there exist other well established GSA methods (for instance Morris method or FAST) requiring much less model evaluations than Sobol'. This is not a criticism of the choice of using Sobol', just a comment about the fact that computational complexity in this case is also due to the fact that you chose the GSA method that requires by far the highest number of model evaluations.*

Response: This is a valid point and we thank you for pointing this out.

Manuscript Revisions: We now note at the end of the discussion section: "For context, the typical time required for a single simulation was 1.4 seconds, resulting in a total computational expense of 720 hours (30 days) across all scenarios. Ongoing research is developing new sensitivity analysis methods that compare well to Sobol' but with reduced computational demands (e.g., FAST, Cukier, 1973; method of Morris, 1991; DELSA, Rakovec et al., 2014)."

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 11, 13745, 2014.

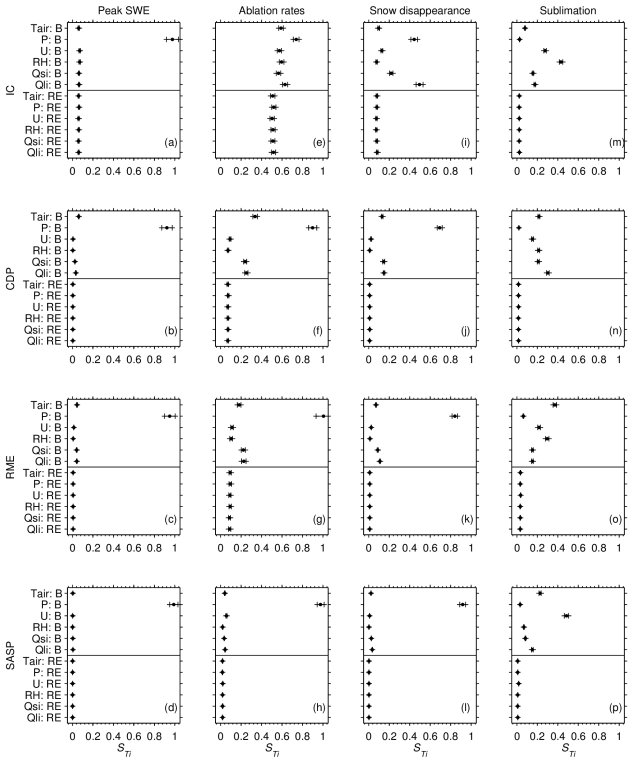

Fig. 1. Total sensitivity indices for scenario NB+RE with code fixed to have mean=0 for the random errors (revised Figure 6 of the original manuscript).

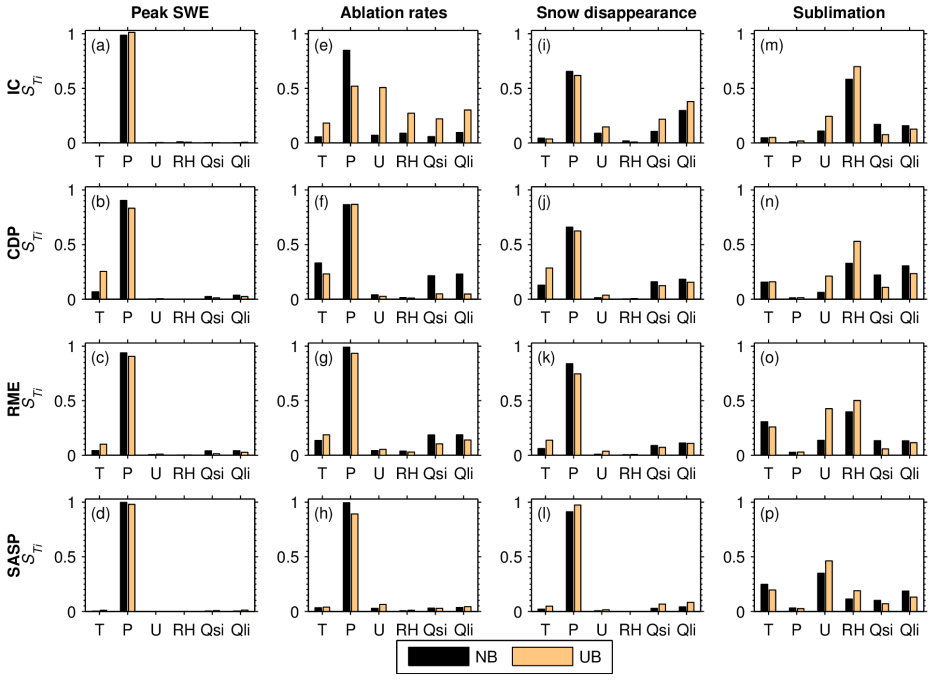

Fig. 2. Example of bar plots comparing total-order sensitivity indices from scenarios NB and UB for the four model outputs at the four sites.

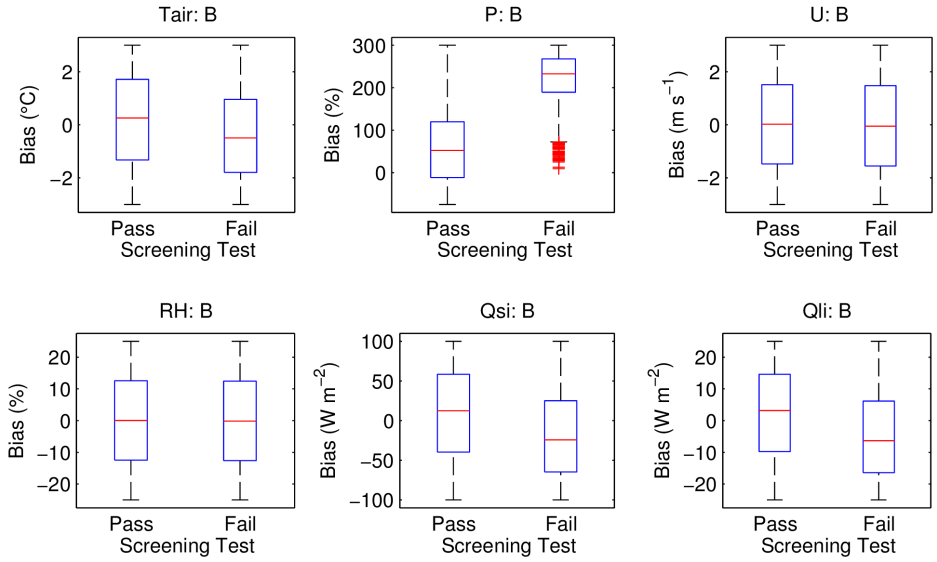

Fig. 3. Categorical boxplots summarizing the relationship between imposed forcing biases and screening test results for the six forcings at SASP in scenario UB.