

May 20, 2015

Dr. Ross Woods
Editor
Hydrology and Earth System Sciences

Dear Dr. Woods,

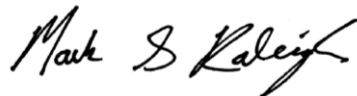
We have revised our *Hydrology and Earth System Sciences* manuscript “Exploring the impact of forcing error characteristics on physically based snow simulations within a global sensitivity analysis framework” (hess-2014-461) and are resubmitting it at this time. We believe that the thorough review has greatly benefited the manuscript and we hope that we have sufficiently addressed all concerns raised.

In the pages below, we include our point-by-point response to each of the six reviewers and a marked-up version of the manuscript to show the relevant modifications. We would like to draw your attention to the most substantial changes, which we summarize in the list below:

- Added a fifth error scenario that considers gauge undercatch levels of precipitation uncertainty (called NB_gauge).
- Corrected issue in NB+RE that occurred due to random errors with non-zero mean.
- Replaced the scatterplots with bar charts, and consolidated figures (only 9 figures now).
- Restructured and expanded both the results and discussions sections.
- Added a new figure that shows sensitivity indices for daily SWE and helps to illustrate key differences between error scenarios (Figure 8).
- Included a direct comparison to prior work on model structural uncertainty (Figure 9).
- Included more discussion on climate dependencies and how this might be better addressed in future work.
- Discussion and justification of the precipitation adjustments prior to the sensitivity analysis.
- Provided more justification and description of methodological choices made in the experimental setup (e.g., error ranges, specification of error distributions, etc.).

We appreciate your time and the reviewers’ time in the review process and look forward to hearing your decision. Thank you for considering our manuscript for publication.

Sincerely,



Mark S. Raleigh, PhD
Email: raleigh@ucar.edu
Phone: +1 (303) 497-8381

Response to Interactive comment by F. Pianosi (Referee)

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 disappearance 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 T_{air} 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., T_{air}) 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. 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 $\theta_{RE;i}$ 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 (see Fig. 5-7 in the revised manuscript). 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 6 (revised manuscript) 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, improve understanding of model sensitivity, and to address questions related to network design. For example, given budget constraints, 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 meteorological

data errors, as these errors are usually studied independently of each other (Flerchinger et al., 2009; Huwald et al., 2009), 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 because 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 cases of correlated errors (e.g., a positive measurement bias in T_{air} that causes a negative RH bias).”

Comment: Page 13755: “by creating k new parameters ($\theta_1, \theta_2, \dots, \theta_k$) 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

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

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 (ϕ) for the new forcing error parameters (section 3.3.1) for better discrimination from the native model parameters (θ). We added a new equation after equation 1 to help clarify, and changed all other references from θ to ϕ .

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 were 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 below (Figure R1-1) 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 Q_{si} , Q_{li} , and T_{air} (all of which lead to slower melt and reduce the chance of snow disappearing).

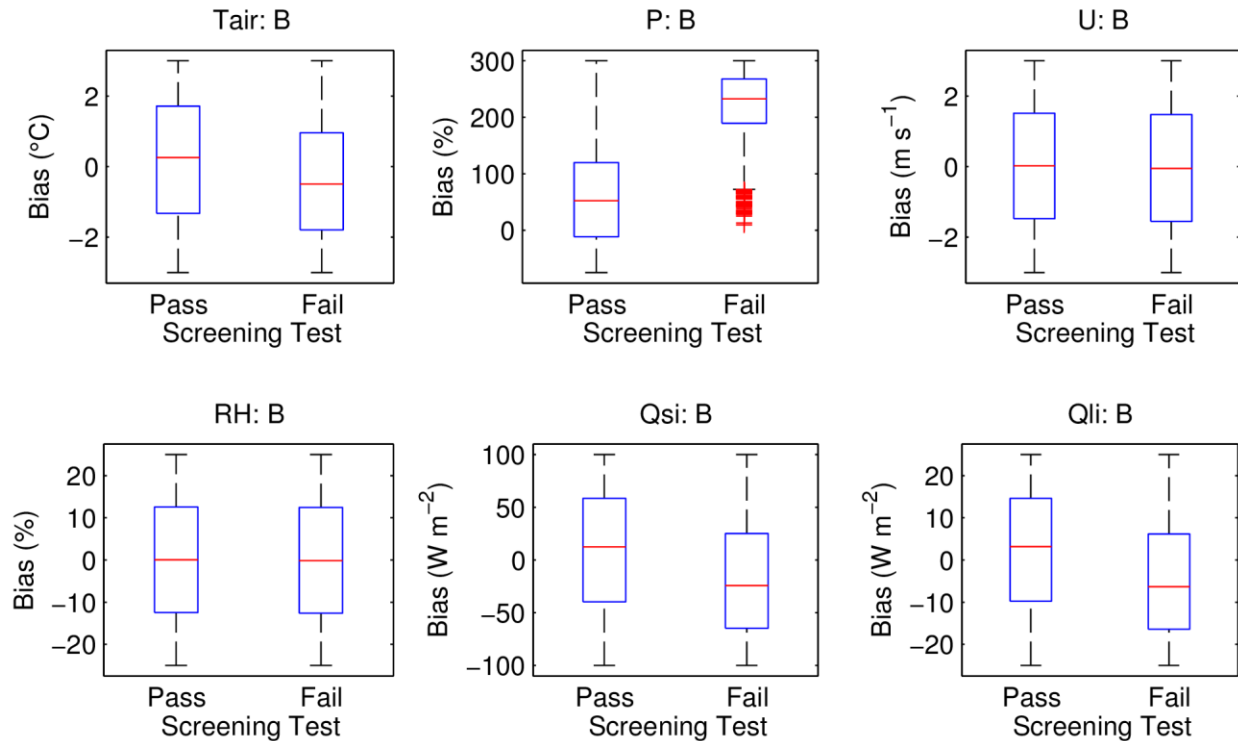


Figure R1-1 Categorical boxplots summarizing the relationship between imposed forcing biases and screening test results for the six forcings at SASP in scenario UB.

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 Q_{li} than for Q_{si} .” 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 4.2.3) 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 Q_{si} , thereby rendering Q_{si}

errors less important than Q_i errors. Additionally, the non-linear nature of the model may enhance the role of Q_i 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 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).”

REFERENCES

Peeters et al., 2014, Robust global sensitivity analysis of a river management model to assess nonlinear and interaction effects, HESS

Response to Interactive comment by J. Li (Referee)

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

***Comment:** This study applied Sobol' global sensitivity analysis for testing model sensitivity to coexisting errors in all forcings. Sensitivity analysis can reveal which forcing error characteristics matter most for hydrologic modelling. As there are fewer detailed studies focusing on forcing uncertainty, this work provides insights on the topic and provide a method that could be extended to more complex physically based models such as land surface models and climate models. It is a very interesting work, and the paper is clear and well structured. I think the paper should be considered for publication on HESS. Here, I have only some concerns about the Sobol' SA method used in this paper.*

Response: We thank you for your time in reviewing the paper.

***Comment:** (1) This study is too computational expensive. 1520000 Monte Carlo samples used here is too much, making that it will be impractical to be extended to other complex models. As I know, Sobol' method will cost a lot to estimate the interaction, such as second-order effect. But it can be less expensive to get the first-order effect and total effect. Did the study consider the SA results from fewer samples? In fact, I suggest either RS-HDMR or response surface based Sobol' can be used here to get similar results.*

Response: Computational expense is an important consideration of any SA study. We should note that while we evaluated the model over 1.8 million simulations, this was somewhat excessive because convergence was reached before all simulations were completed. Additionally, this number includes multiple error scenarios (5) and multiple sites (4), so it seems higher than in reality. Figure R2-1 (below) shows the time history of the total sensitivity indices (as a function of sample size) for Scenario NB (other scenarios exhibited similar levels of convergence). Examining this figure, it is evident that the same conclusions for the study (at least qualitatively) could have been drawn with fewer simulations. A dynamic system of calculating sensitivity indices as model completes simulations would optimize the analysis by stopping the process once convergence has been reached, but such a system was not implemented here. Such approaches might be needed when extended the error analysis framework to more complex model, such as land surface models. While we do not expect that this framework (and number of simulations) can be extended to all modeling endeavors, we note in our discussion the availability of more efficient sensitivity analysis methods and the need for improved efficiency.

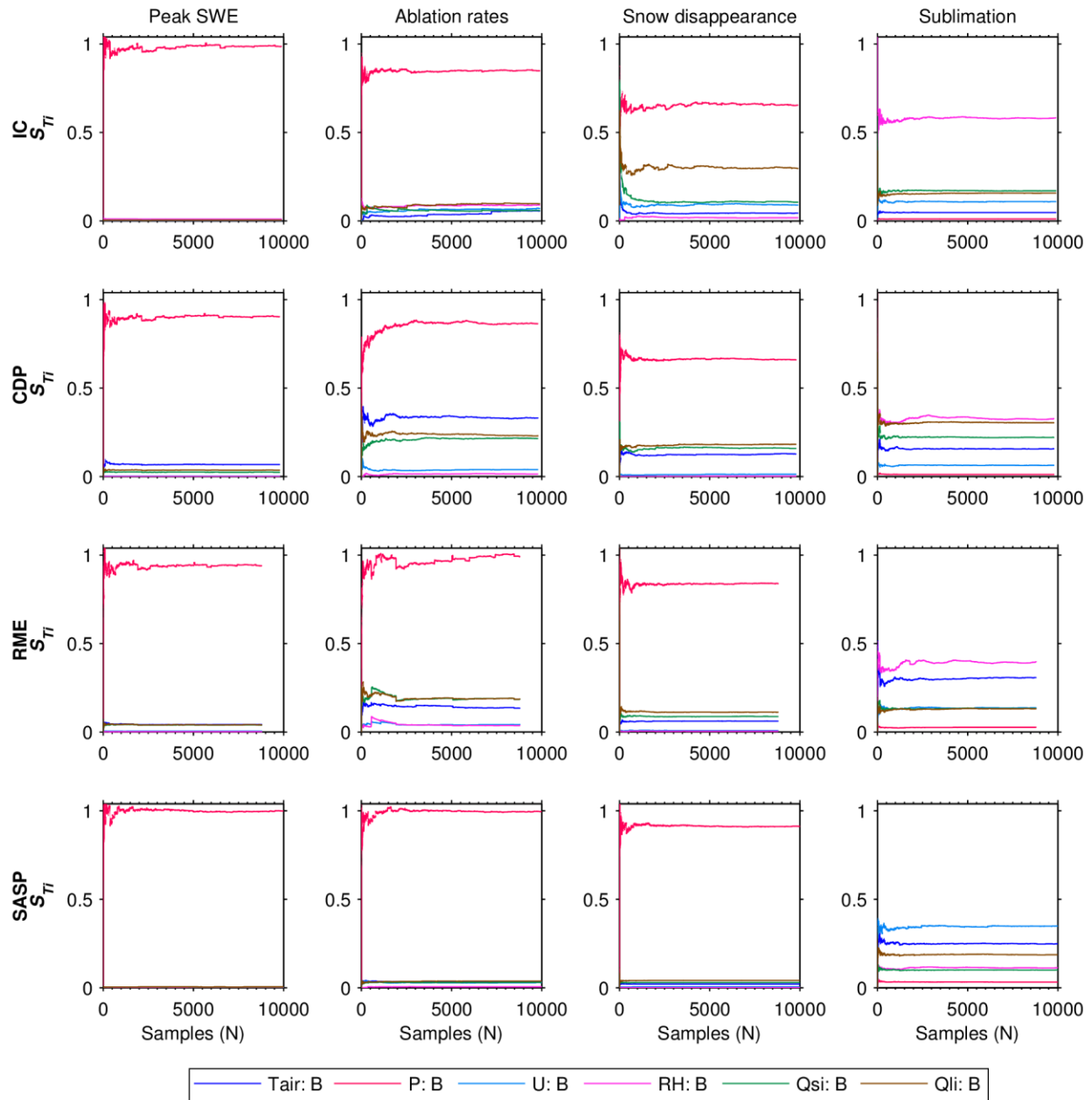


Figure R2-1 Convergence history of total-order sensitivity indices in scenario NB for the four model outputs at the four sites, as a function of sample size.

Manuscript Revisions: We now provide more context for the computational expenses at the end of the discussion section: “Finally, while the Sobol’ method is often considered the “baseline” method in global sensitivity analysis, we note that it comes at a relatively high computation cost (1 840 000 simulations across four sites and five error scenarios) and may be prohibitive for many modeling applications (e.g., for models of higher complexity and dimensionality). 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. Examination of the convergence rates indicated that most sensitivity indices stabilized after one-

third of the simulations completed, and hence the same results could have been found using significantly fewer simulations (no figures shown). 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), and is comparing how different methods classify sensitive factors differently (Pappenberger et al., 2008; Tang et al., 2007). We expect that detailed sensitivity analyses that concurrently consider uncertainty in forcings, parameters, and structure in a hydrologic model will be more feasible in the future with better computing resources and advances in sensitivity analysis methods." 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: (2) *This study used the total effect to quantify the sensitivity of different error type, different error distributions and error magnitudes. As the sum of total effect of each factor will be above 1, in order to quantify the contribution of each factor, I suggest to use the index $ST_i/\text{sum}(ST)$, which is the total effect of one factor divided by the sum of total effect of all the factors.*

Response: While we thank the reviewer for this logical suggestion, we declined to make this change because we do not find a strong precedent for this practice in the sensitivity analysis literature. We prefer to report the total sensitivity indices according to common practice.

Manuscript Revisions: No changes made regarding this point.

Response to Interactive comment by Anonymous Referee #3

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

***Comment:** Overview This manuscript explores the relative effects of bias and error distributions on the Utah Energy Balance model's sensitivity across Peak SWE, Ablation Rates, Snow Disappearance, and Sublimation predictions. The work exploits detailed forcing observations at 4 seasonally snow covered sites: (1) the tundra Imnavait Creek in the Brooks Range in Alaska, (2) the Col de Porte site in the Chartreuse Range in France, (3) Reynolds Mountain in Idaho, and the Swamp Angle Study Plot in the San Juan mountains of Colorado. The core contention of the work is that forcing bias and errors could dominate structural and parametric uncertainties for snow-affected regions with strong observation limitations. Overall I found this hypothesis somewhat self-evident, although the overall study does highlight the importance of observation errors and uncertainties. I believe this manuscript requires revision to reach its full potential.*

Response: We thank you for your careful review of the manuscript.

Major Comments

***Comment:** 1. Limited Analysis: The core results in Figures 5-11 are discussed with extreme brevity and little analysis. The authors have made the chose to provide a more detailed exposition in their Discussion but at present the Results do not even orient the reader very well across individual plots. Figures 5-8 are summarized in text that mixes results across figures and severely limited in its value. The question that emerges when reading this is that either the authors could compress their results into fewer and better designed figures or they could tease more model related insights in their analysis text.*

Response: This is a reasonable comment. Given the number of dimensions that we are examining (4 sites, 6-12 error parameters, 4 model outputs, and now 5 scenarios), we have opted to focus on providing more context and explanation of the results in the text. However, we were able to compress Figures 5-11 into three figures and thank you for the suggestion.

Manuscript Revisions: We now provide expanded description of the core results and compressed the figures, but reserve discussion of the results in section 5.

***Comment:** 2. Discussion Disconnected from Results: The most interesting portions of the discussion relate to the contention of the relative importance of structural uncertainty to forcing errors. Unfortunately, this text references other published work strongly and does a very poor job connecting to directly to the Results/Figures of this paper. Transitioning from Section 4 to Section 5 almost feels like your reading a different paper. Overall the structure and writing of the work varies significantly from the well written Introduction, the detailed Methods, and more detailed Discussion versus the extremely cursory Results.*

Response: We can understand how this is problematic and agree that the exposition of these sections can be improved.

Manuscript Revisions: We have rewritten and reorganized some parts of the discussion to provide better correspondence with the results and better connection to other published works. As an example of the latter, we have acquired the model results of Essery et al. (2013), which is referenced heavily in the early part of the discussion and now create a new figure (see Figure 9 in the revised manuscript) that directly compares our results (due to forcing uncertainty) and Essery's results (due to structural uncertainty).

Comment: 3. It is unclear how generalizable the results are beyond this study: Many of the results are not very insightful and seem to convey a very place-based specificity for deviating cases. The reporting of sensitivities in the Results are not well articulated in terms of their dependency on site location, the nuances of the Utah Energy Balance model, and scenarios. In its present form, I am not convinced that manuscript provides insights and it may be conflating several factors that could influence the differences in sensitivity (model choice, site selection, scenarios). Explanation of the stronger results, such as distribution choice minimally impacts computed sensitivities, is limited and not compelling. The core of the Discussion section is the best overall text of the paper. It would have been far better to lead with your core hypotheses in the Results section and test them explicitly through the analysis of your results. The Discussion would then emphasize key caveats, insights, and implications.

Response: While recognizing the importance of generalizing the results, we are hesitant to generalize relationships between site geo-characteristics/climate and sensitivities indices because of the relatively low number of sites represented (n=4 sites, 1 year each) and the confounding number of differences between our sites (e.g., snow climate, latitude, elevation, wind exposure/sheltering, etc.). We would require a much larger population of snow measurement sites in order to more robustly test relationships between sensitivity indices and site characteristics such as elevation and latitude. A successful example of relating climate characteristics to sensitivity can be found in van Werkhoven et al. (2008), which had 12 sites and 39 years each, making it possible to explore inter-site and inter-annual variations in climate and linkages to model sensitivity.

Manuscript Revisions: We now emphasize in Section 2 that we selected the four sites to check for climate dependencies, but are unable to generalize the results due to the low sample size. We note in the discussion however, that there are common results that emerge across all sites, such as the dominance of precipitation bias on SWE, ablation rates and snow disappearance (NB scenario) and longwave bias on all four outputs (NB_lab scenario). This suggests that there may be common features in model sensitivity to forcing errors across distinct climates.

Minor Comments

Comment: 1. It would have been interesting to explore 2nd order and 1st order differences from the total indices in the results.

Response: While we agree this would be interesting, we argue that this could make the study less focused and therefore elect to present only the total sensitivity indices. The total sensitivity indices provide a summative measure of both first-order and interaction effects and therefore convey the overall importance in a straightforward manner. Calculation of the 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.

While we do not present them in the manuscript, we can calculate the first-order indices with the existing model simulations. The comparison of the first- and total-order indices provides insights into how much of the variance is due to direct effects vs. interactions, and broad justification for only reporting one type of sensitivity indices. Figures R3-1 and R3-2 (below) show the first- (S_i) and total-order (S_{Ti}) indices for the NB and NB+RE scenarios. From these figures, it is evident that in many cases, the sensitivity is dominated by first-order effects, as suggested by the close alignment of S_i and S_{Ti} values. There are cases however when the interactions have greater importance (e.g., factors of secondary importance for the ablation rates). The general correspondence between the first- and total-order indices suggests to us that most of the story is captured with just a single index; hence, we focus on just total-order sensitivity indices for simplicity/clarity.

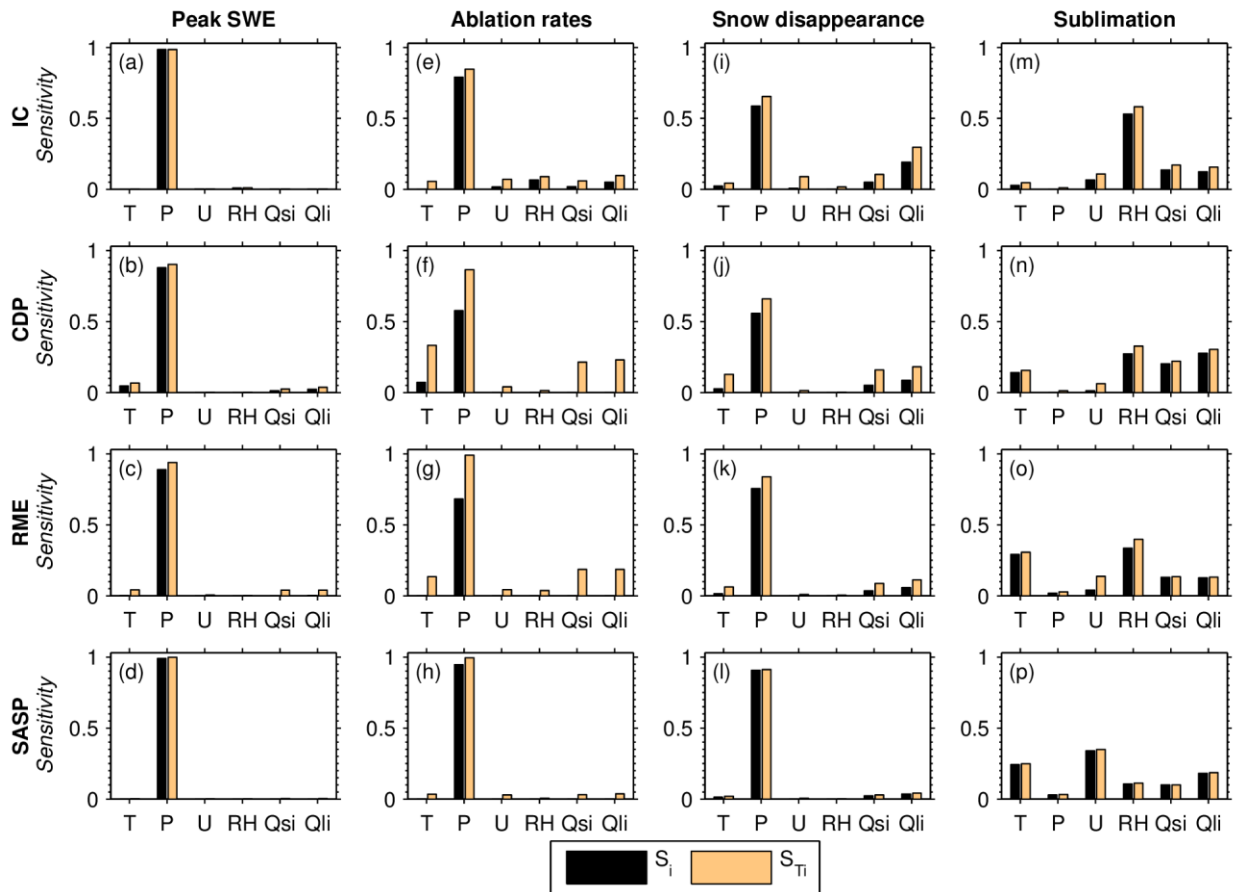


Figure R3-1 First order (S_i) and total-order (S_{Ti}) sensitivity indices for bias factors in the NB scenario.

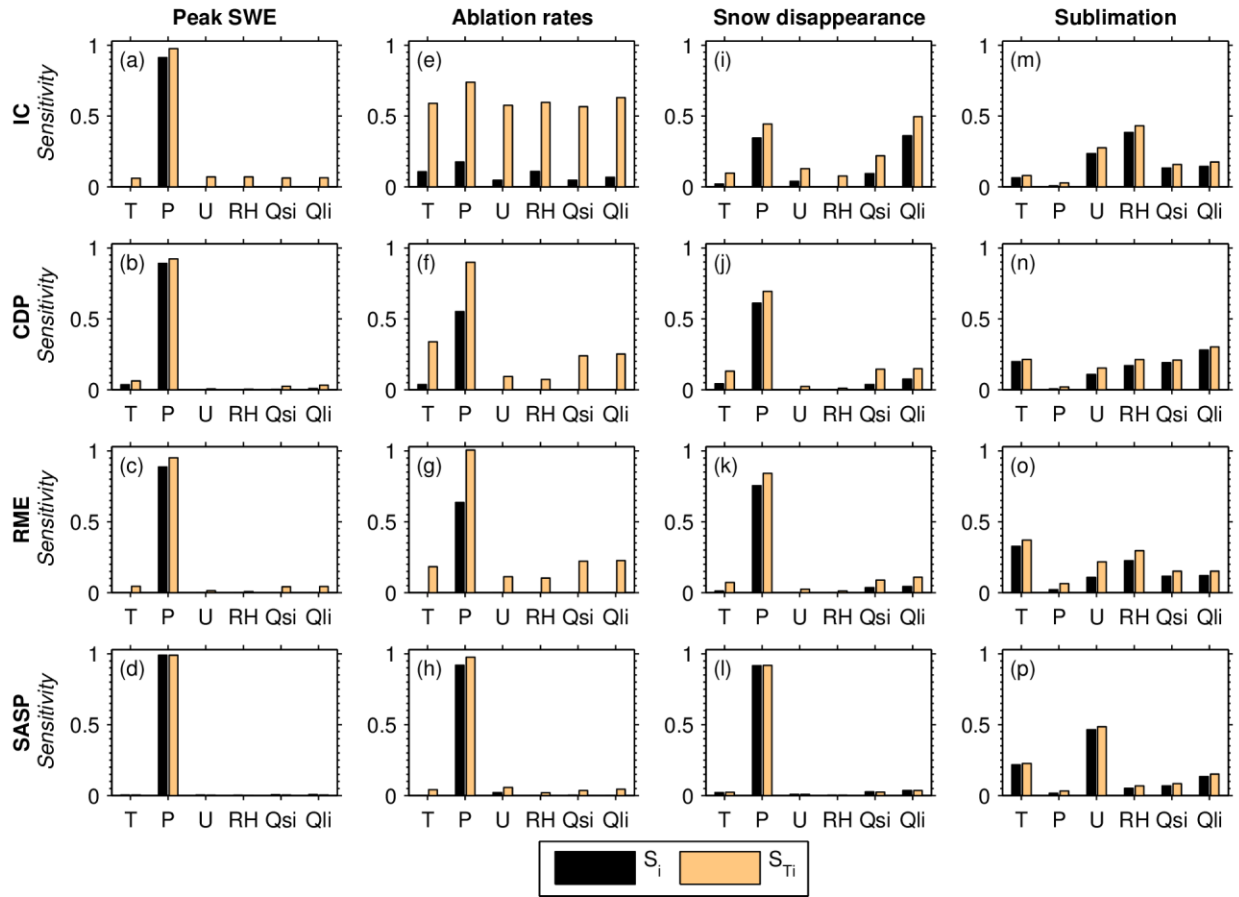


Figure R3-2 First order (S_i) and total-order (S_{Ti}) sensitivity indices for bias factors in the NB+RE scenario.

Manuscript Revisions: We have made no changes to the analysis, but we now comment in section 3.3.2, “First-order and higher-order sensitivities can be resolved; here, only the total-order sensitivities are examined (see below) for clarity and because the first-order sensitivity indices were typically comparable to the total-order sensitivity indices.”

Comment: 2. *A better explanation of the scales assumed in the measures used to report sensitivities and caveats as to what cannot capture would be helpful.*

Response: We assume you are referring to numerical scales in this comment, and can comment on this in the text.

Manuscript Revisions: In section 3.3.3, we explain that interpretation of the total sensitivity indices is straightforward because they represent the fraction of output variance due to a specific factor, and state that these indices scale from 0 to 1. We now include a caveat that the Sobol’ total sensitivity indices cannot account for the case of correlated errors (section 3.3.2), which may occur in the real-world.

Comment: 3. Very little treatment is provided for the convergence rates of the total order indices and their associated bootstrap intervals as a function of your sampling.

Response: The reviewer is correct that we did not provide much information on convergence rates. Figure R3-3 (below) shows the time history and convergence of the total sensitivity indices (as a function of sample size) for Scenario NB (other scenarios exhibited similar levels of convergence). Examining the figure, it is evident that the same conclusions for the study (at least qualitatively) could have been drawn with fewer simulations. A dynamic system of calculating sensitivity indices as model completes simulations would optimize the analysis by stopping the process once convergence has been reached, but such a system was not implemented here.

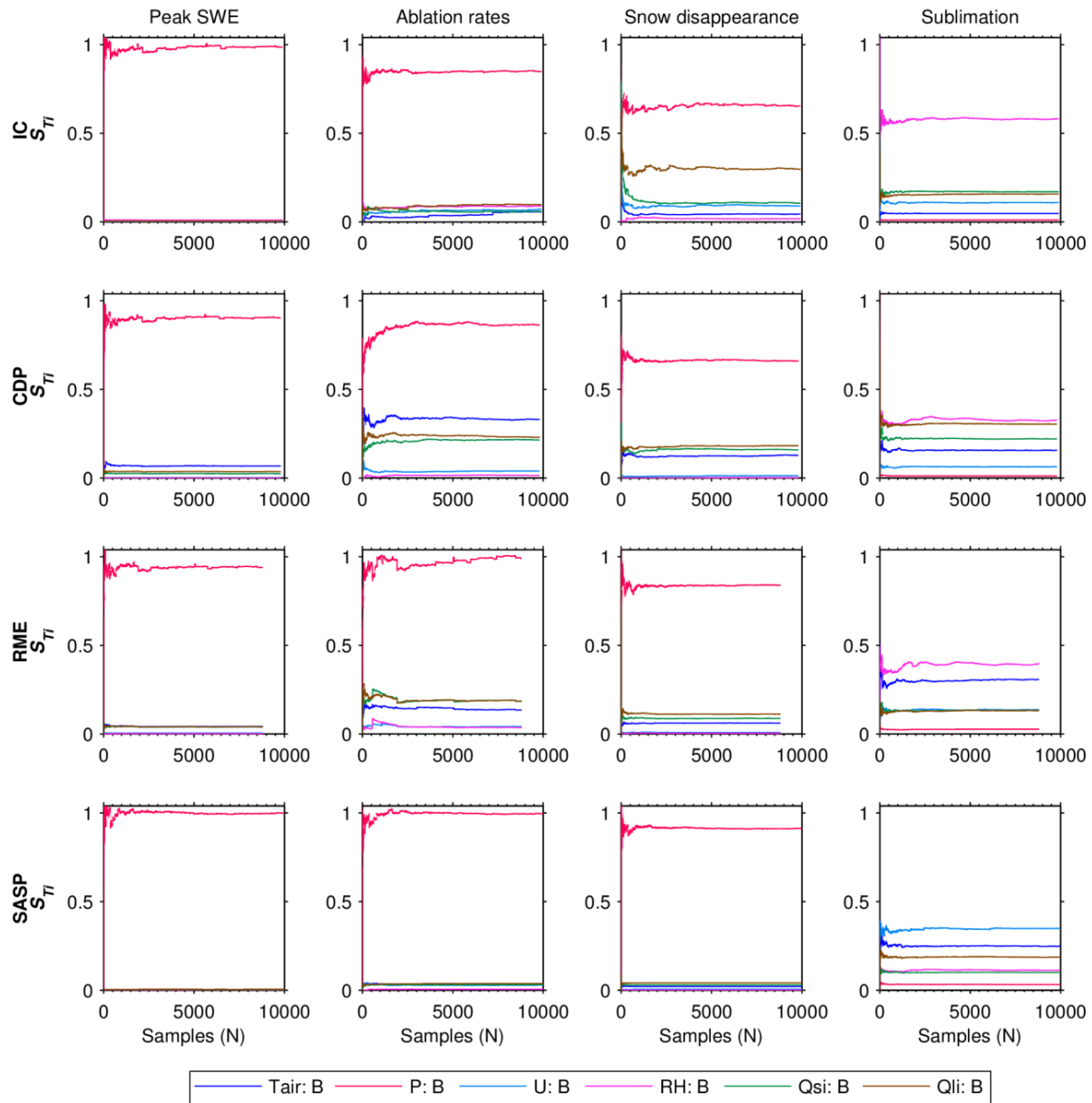


Figure R3-3 Convergence of total-order sensitivity indices in scenario NB for the four model outputs at the four sites, as a function of sample size.

We can quantitatively assess the level of convergence by examining the ratio of the 95% confidence interval (from the bootstrapping procedure) to the mean S_{Ti} values. Figure R3-4 (below) shows this ratio (as a percentage) for the error parameter with the highest S_{Ti} for each model output, scenario, and site. If we assume convergence has been reached when the ratio is less than 10% (based on Herman et al., 2013), then we can see that the majority cases in our study reached convergence, and only three out of 64 cases had a ratio greater than 15%. Even for these three cases, it is evident that the general order of importance of errors is established. For example, ablation rates at RME in Scenario NB had a CI that was 16% of the mean, but Figure R3-3 shows that the relative hierarchy of importance in biases is established in this case.

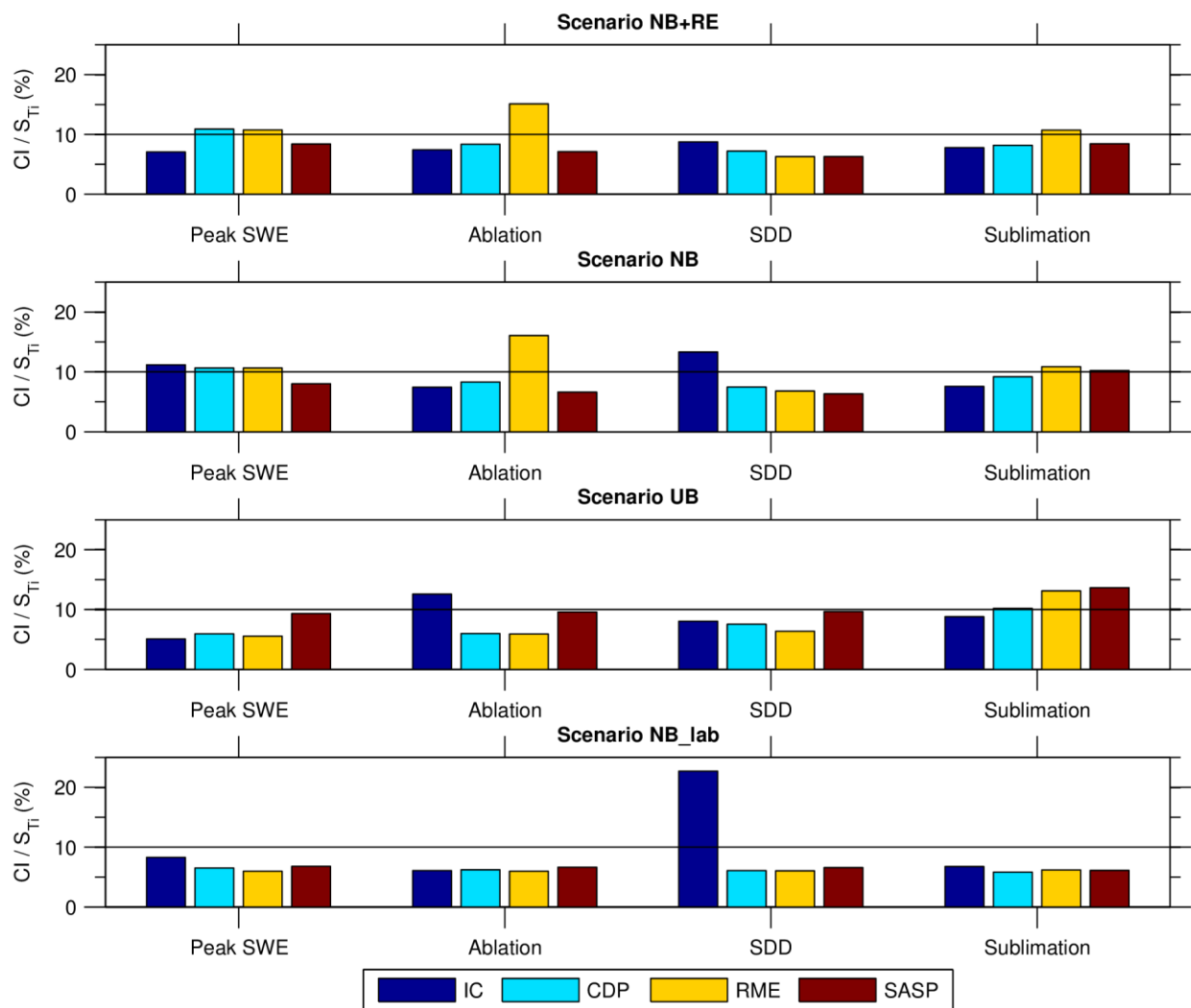


Figure R3-4 Ratio of 95% confidence intervals to the bootstrapped mean total-order sensitivity indices (%) for the most important factor for each scenario, site, and model output.

Manuscript Revisions: We now provide some description of the convergence rates and on the bootstrap confidence intervals (sections 4.2 and 5), but do not provide any additional figures in the manuscript.

Comment: 4. How stable and/or separable are the factor prioritization rankings? What results have higher confidence?

Response: The 95% confidence intervals (from the bootstrapping procedure) are presented in Figures 5-7 in the revised manuscript, and these provide a measure of our confidence in the rankings. The difference between the bootstrap mean and the final mean S_{Ti} values also provides a measure of stability.

Manuscript Revisions: We note in section 3.3.4: “For all cases, final S_{Ti} values were close to the mean bootstrapped values (i.e., 99% had a difference less than 0.001 and no difference was greater than 0.003), suggesting convergence.”

Comment: 5. It would improve the manuscript to better understand the justification of the ranges tested in the Sobol analysis. Would a slight change in your a priori ranges change factor rankings?

Response: The original manuscript outlines the justification for the ranges, but we now provide more information in the methods section. While we did not test for “slight changes” in the a priori ranges, we know that more substantial changes in these ranges can change the hierarchy of factors. Our original results already suggest that a change in the error ranges will change the rankings of factors (compare NB to NB_lab, where the only difference is field vs. laboratory levels of uncertainty). We also now include the new scenario (identical to NB but with lower precipitation error ranges to reflect gauge undercatch), and find again that the factor ranges do change with the a priori ranges in the forcing uncertainty.

Manuscript Revisions: We now expand on our justification of the error ranges (section 3.2.3). Additional treatment of this topic is included in the discussion.

REFERENCES

- Essery, R., S. Morin, Y. Lejeune, and C. B. Ménard, 2013: A comparison of 1701 snow models using observations from an alpine site. *Adv. Water Resour.*, **55**, 131–148, doi:10.1016/j.advwatres.2012.07.013.
- Herman, J. D., J. B. Kollat, P. M. Reed, and T. Wagener, 2013: Technical Note: Method of Morris effectively reduces the computational demands of global sensitivity analysis for distributed watershed models. *Hydrol. Earth Syst. Sci.*, **17**, 2893–2903, doi:10.5194/hess-17-2893-2013.

Saltelli, A., 2002: Making best use of model evaluations to compute sensitivity indices. *Comput. Phys. Commun.*, **145**, 280–297, doi:10.1016/S0010-4655(02)00280-1.

Van Werkhoven, K., T. Wagener, P. Reed, and Y. Tang, 2008: Characterization of watershed model behavior across a hydroclimatic gradient. *Water Resour. Res.*, **44**, W01429, doi:10.1029/2007WR006271.

Response to Interactive comment by R.L.H. Essery (Referee)

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

***Comment:** Raleigh et al. present an interesting attempt to systematically determine how uncertainty in forcing data influences uncertainty in snow simulations. Some of their conclusions seem quite obvious (biases are more significant than random errors, and uncertainty in measured precipitation is the most important factor), but a large effort is often required to demonstrate things that appear obvious with hindsight in hydrological modelling. The increased sensitivity to biases when random errors are introduced is a striking result, however, and there should be more exploration of how this arises.*

Response: Thank you for your interest and your review of the manuscript. We have revisited the striking result of enhanced sensitivity to biases when random errors are introduced, and found this to actually be related to a deficiency in the introduction of random errors into the system (eq. 4). 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.

Manuscript Revisions: We have corrected this coding issue, reran NB+RE and have found that this minimized the problem you have found. We find that the sensitivity to biases (after introducing random errors) is less pronounced in this case for most outputs/sites considered. The most obvious outlier is for ablation rates at IC, where there is heightened sensitivity to biases after random errors are introduced. In this case, the total-sensitivity indices are amplified because of more interactions in the system (e.g., first-order sensitivities were small compared to the total-order indices). We surmise that the relatively short ablation season at IC (on order of 10-20 days) is a critical reason why there is enhanced sensitivity across all error types; errors in a variety of factors can yield large impacts on ablation rates during the brief melt period.

***Comment:** I am slightly concerned about how the error distributions have been assigned. It is variances in model outputs that are examined but ranges in model inputs that are specified. The variance of a uniform distribution is larger than a normal distribution over the same range, so these scenarios are not really comparable.*

Response: These concerns are reasonable, and it is true that the variance of a uniform distribution is larger than the variance of a normal distribution over the same range. That is part of the purpose of this particular experiment, namely, to examine how the assumed probability distribution of errors influences model sensitivity. It is by design that the ranges are made the same for the uniform and normal distributions for a given forcing; this allows us to test in a controlled fashion whether/how more frequently occurring extreme errors (in the uniform distribution) change model sensitivity. If we had not matched the ranges in the two distributions,

then there would be two confounding reasons why the distributions were different (probability distribution shape and range), and we wished to isolate the differences due to shape only.

You are correct that we could have alternatively constructed the experiment such that we specified variance (or standard deviation) instead of the range. However, we constructed the experiment with range instead of variance because it was more straightforward and provided a more direct approach to encompass all magnitudes of errors found in our literature review for different forcing observation/estimation approaches. We also note that most sensitivity analyses use uniform distributions (e.g., Nossent et al., 2011; Peeters et al., 2014), which are specified by the range and not the variance. In considering normal distributions, we are extending these methods to other types of distributions.

Regardless of whether we base our distributions on range or variance, we note that there is “uncertainty in uncertainty.” In other words, we are not always certain about what the spread of uncertainty should be. Our understanding of the spread of uncertainty is poor due to the relatively low sample size of papers that report error statistics for different forcings. It can be shown experimentally (with a Monte Carlo sampling experiment) that for low sample sizes ($n < 150$), we have higher confidence in the range of a given normal distribution than in the variance. Figure R4-1 (below) shows the relative uncertainty in range and variance derived from such a Monte Carlo experiment. Given that there are few papers that systematically assess forcing errors in mountainous areas, we argue that it is not necessarily a bad idea to work in terms of range because we have comparatively higher confidence in the range than in the variance. For example, if we only have 10 papers that specify the mean bias of shortwave radiation, the confidence interval (CI) for our variance estimate (of the probability distribution) is about 80% greater (in a relative sense) than the CI for the range of the distribution (see figure below).

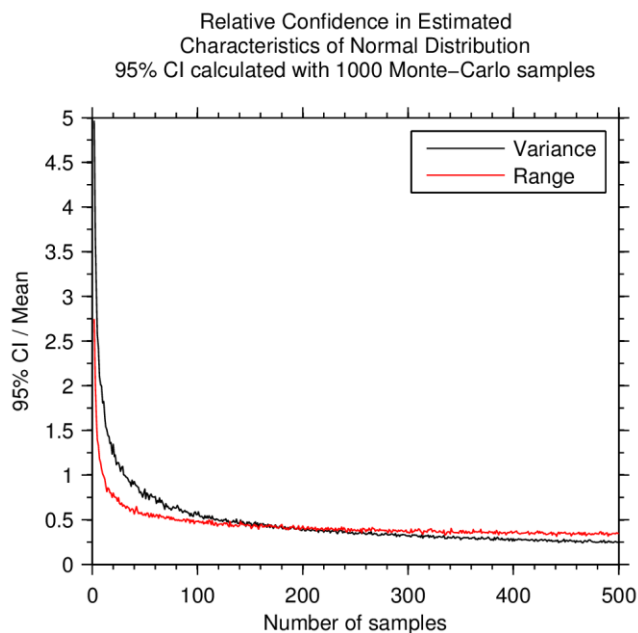


Figure R4-1 Confidence in the variance and range of a normal distribution determined with Monte-Carlo sampling ($n=1000$) with a random dataset (106 samples, mean=0, variance=1) as a function of sample size.

Manuscript Revisions: We clarify in section 3 why we prescribed the probability distributions in this manner.

Comment: It is not clear, in any case, how the ranges given in Table 3 determine the means and variances that would characterize the normal and lognormal distributions. Can this be clarified?

Response: Thank you for pointing this out. Yes, this can be clarified.

Manuscript Revisions: We now clarify in detail in section 3.3.5 how the normal and lognormal distributions are constructed based on specified characteristics.

Comment: Uncertainty in measurements of snowfall is certainly a major concern, but the upper bound chosen for precipitation biases in the model forcing (+300%) is enormous – much bigger than the stated error of less than 20% of peak SWE for most SNOTEL sites. The reason given for choosing this large uncertainty is to represent areas with drifting snow, but I would argue that the neglect of drifting snow is a missing process in the model, not an uncertainty in its forcing.

Response: We partially agree with you on this point. We think that the model scale and the process scale are important considerations in how we categorize the uncertainty due to drifting snow. For the case when the 1-d model is applied at a model element resolution that is greater than the process scale, we would classify the wind drift uncertainty as structural uncertainty in terms of sub-grid variability accounting. However, when the model element resolution is less than the process scale of drifting snow, it is impossible to account for snowdrift processes within the structural uncertainty because the model is applied independently of neighboring locations (i.e., no lateral snow mass transfer, by definition of a 1-d model). In this case, we argue that the drift uncertainty is somewhat ambiguous for the 1-d UEB model but still must be accounted for in either the parametric or the forcing uncertainty. We argue that drift uncertainty is analogous to the precipitation undercatch uncertainty (both are cases of wind-affecting precipitation), and therefore we treat the drift uncertainty as a source of forcing uncertainty for our 1-d model.

For models of higher dimensionality (e.g., 3-d), then we agree with your point. A 3-d snow model should account for lateral mass transfer via snow drifting. In this case, it is clear that large uncertainties in snow accumulation arise due to omission of the snowdrift processes in the model, and this is a case where the uncertainty is attributed to structural (and parametric) uncertainty.

While we make this point, we share your interest in how the study would have been different if we had “standard SNOTEL precipitation errors” as the upper limit of uncertainty in precipitation. To that end, we introduced a new scenario (NB_gauge) that repeated Scenario NB with all factors the same except we changed the ranges of precipitation bias of [-10% to +10%]. When we consider this lower range in precipitation uncertainty, we find that precipitation bias is never a major factor for these four outputs at the four sites, and other dominant factors emerge. At IC,

longwave bias emerges as the most important factor for ablation rates and snow disappearance while humidity bias matters most for peak SWE and sublimation. At the other sites, biases in shortwave and longwave radiation and air temperature are most important for peak SWE, ablation rates, and snow disappearance. Humidity bias is an important factor for sublimation at IC, CDP, and RME, while wind bias is important to sublimation at SASP.

Manuscript Revisions: We now introduce the new error scenario “NB_gauge” in the analysis and have updated the manuscript text to introduce this scenario and report/discuss the results.

Specific comments:

Comment: page 13749, line 25 SWE is measured at Col de Porte using a cosmic ray detector, not a snow pillow.

Response: Thank you for catching this mistake.

Manuscript Revisions: We have corrected the sentence.

Comment: 13750, 5 How was reasonable representation of observed SWE judged?

Manuscript Revisions: We now state in section 2, “We considered adjustment multipliers ranging from 0.5 to 2.5 (increments of 0.05) and selected the multiplier that yielded the lowest root mean squared error between observed and modeled SWE.”

Comment: 13751, 18 It could be made clear at this stage that normal distributions are used for additive errors and lognormal distribution for multiplicative errors.

Manuscript Revisions: Done.

Comment: 13752, 3 In contrast, data assimilation techniques often address random errors that are assumed to be unbiased.

Manuscript Revisions: Thank you for pointing this out – we now note this in the sentence.

Comment: 13755, 1 Overview of what?

Manuscript Revisions: We now change the title of this subsection to “Overview: model conceptualization and sensitivity”.

Comment: 13756, 12 “due to bias in forcing”

Manuscript Revisions: Done.

Comment: 13757, 15 “For all cases, final STi values were generally close ...” sounds a little contradictory; were they all close, or generally close?

Manuscript Revisions: We now rephrase this sentence to be more quantitative: “For all cases, final STi values were generally close to the mean bootstrapped values (i.e., 99% had a difference less than 0.001 and no difference was greater than 0.003), suggesting convergence.”

Comment: 13758, 20 Non-physical values would be less common if multiplicative perturbations were applied to all forcing variables that cannot be negative, not just precipitation.

Response: This is a valid point, but we are attempting to follow typical error reporting conventions and to provide easy interpretation of errors. For example, it is often the case that radiation errors are reported in the literature in an additive context (e.g., $+35 \text{ W m}^{-2}$) and not in a multiplicative context (e.g., $+10\%$). In the case of radiation, a multiplicative error (e.g., $\pm 10\%$) is not straightforward to interpret because the magnitude of the error will change with seasonality (e.g., 10% error in winter shortwave radiation is much less than 10% error in summer shortwave). Additionally, some errors only make sense in an additive context (e.g., temperature errors). Our treatment of errors reflects common practices in the literature to make it more easily understood by the community.

Manuscript Revisions: We clarify in section 3.3.5 why we prescribed multiplicative vs. additive errors.

Comment: 13761, 8 Can differences in which variables are of secondary importance be linked to differences in climate at the sites?

Response: The links with climate in these secondary variables are not always clear to us. At the warm maritime CDP site in scenario NB, it makes sense that T_{air} bias is important to peak SWE, as it helps control the partitioning of rain and snow. In contrast, it is not clear why Q_{si} and Q_{li} biases are of secondary importance for sublimation at IC and CDP but not at RME and SASP (where T_{air} bias is the second most importance factor).

While there may be interesting climate linkages, we note that we are hesitant to generalize relationships between site geo-characteristics/climates and sensitivities indices because of the relatively low number of sites represented ($n=4$ sites, 1 year each) and the confounding number of differences between our sites (e.g., snow climate, latitude, elevation, wind exposure/sheltering, etc.). We would require a much larger population of snow measurement sites in order to more robustly test relationships between sensitivity indices and site characteristics such as elevation and latitude. A successful example of relating climate characteristics to sensitivity can be found in van Werkhoven et al. (2008), which had 12 sites and

39 years each, making it possible to explore inter-site and inter-annual variations in climate and linkages to model sensitivity. We now emphasize in Section 2 that we selected the four sites to check for climate dependencies, but are unable to generalize the results due to the low sample size.

Manuscript Revisions: We now emphasize in Section 2 that we selected the four sites to check for climate dependencies, but are unable to generalize the results due to the low sample size. We note in the discussion however, that there are common results that emerge across all sites, such as the dominance of precipitation bias on SWE, ablation rates and snow disappearance (NB scenario) and longwave bias on all four outputs (NB_lab scenario). This suggests that there may be common features in model sensitivity to forcing errors across distinct climates.

Comment: 13762, 3 It is not so surprising that Q_{li} biases are more important than Q_{si} biases because of the high albedo of snow.

Response: We agree with you here. We also note that given how the literature often emphasizes the importance of net shortwave over all other terms for snowmelt (e.g., Bales et al., 2006), this could be considered a surprising result.

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 Q_{si} , thereby rendering Q_{si} errors less important than Q_{li} errors. Additionally, the non-linear nature of the model may enhance the role of Q_{li} through interactions with other factors.”

Comment: 13765, 11 Please consider doi:10.1029/2010EO450004

Response: Thank you for making us aware of this article and for helping us to see the problem with using parentheses to indicate the opposite meaning.

Manuscript Revisions: We have reworded the sentence to avoid this issue and have ensured that there are no other instances of this convention in the manuscript.

Comment: 13766, 2 Note that “constraining P uncertainty in snow-affected catchments” is the aim of WMO-SPICE <http://www.rap.ucar.edu/projects/SPICE/>

Manuscript Revisions: We now state: “Progress is being achieved with advanced pathways for quantifying snowfall precipitation, such as NWP models (Rasmussen et al., 2011, 2014) and through systematic intercomparisons of precipitation and snow gauges (e.g., Solid Precipitation Intercomparison Experiment, <http://www.rap.ucar.edu/projects/SPICE/>).”

Comment: 13766, 10 Probabilistic forcing is a common and long-standing approach in data assimilation

Manuscript Revisions: We now note: “We suggest that probabilistic model forcings (e.g., Clark and Slater, 2006), which have a legacy in data assimilation methods (e.g., precipitation uncertainty, Durand and Margulis, 2007), present one potential path forward where measures of forcing uncertainty can be explicitly included in the forcing datasets.”

Comment: The forcing error scenarios are described in Figure 1, Table 3 and section 3.2. Is the figure really necessary?

Response: We considered removing Figure 1, but other reviewers thought this figure was helpful in summarizing the scenarios, and so we have left it in the manuscript. We have also expanded Figure 1 to include the new NB_gauge scenario, which was added to address your concerns about the level of precipitation uncertainty.

Manuscript Revisions: None taken here.

REFERENCES

- Bales, R. C., N. P. Molotch, T. H. Painter, M. D. Dettinger, R. Rice, and J. Dozier, 2006: Mountain hydrology of the western United States. *Water Resour. Res.*, **42**, W08432, doi:10.1029/2005WR004387.
- Nossent, J., P. Elsen, and W. Bauwens, 2011: Sobol’ sensitivity analysis of a complex environmental model. *Environ. Model. Softw.*, **26**, 1515–1525, doi:10.1016/j.envsoft.2011.08.010.
- Peeters, L. J. M., G. M. Podger, T. Smith, T. Pickett, R. H. Bark, and S. M. Cuddy, 2014: Robust global sensitivity analysis of a river management model to assess nonlinear and interaction effects. *Hydrol. Earth Syst. Sci.*, **18**, 3777–3785, doi:10.5194/hess-18-3777-2014.
- Van Werkhoven, K., T. Wagener, P. Reed, and Y. Tang, 2008: Characterization of watershed model behavior across a hydroclimatic gradient. *Water Resour. Res.*, **44**, W01429, doi:10.1029/2007WR006271.

Response to Interactive comment by RR Rosolem (Referee)

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

***Comment:** The work presented by Raleigh et al. investigates the impact of uncertainty in individual meteorological forcing variables on simulation of snow processes at selected sites using the Utah Energy Balance (UEB) model. The manuscript investigates how different error distributions and magnitudes can impact quality of simulations of key snow variables by using the Sobol' sensitivity analysis methodology. The number of model simulations needed for individual sites/experiments varies approximately between 70,000 and 130,000. The authors found that model outputs were generally more sensitive to systematic biases in forcing in comparison to random error. In addition, simulations indicated that model was more sensitive to the magnitude of forcing rather than the distribution of errors.*

I particularly like the manuscript and I think it should be accepted for publication after minor revisions (see my comments below). This is a good example of model diagnostics employed in a relevant context (understanding impacts of forcing uncertainty). We usually focus on uncertainty in parameters, but forcing can play a significant role (especially with such models where both local in-situ and global gridded forcing data are commonly available). The large number of model simulation does not concern me because (1) evaluating the total number of simulations without actual simulation time is somewhat meaningless (how long does it take to run a single year simulation in this model?), and (2) the authors are clearly using such approach to diagnose model uncertainty in detail and recognize that there are more simple approaches that can be used but the emphasis here is on the benefits of using Sobol'. Finally, the manuscript is well written, it explains the strategy very well and includes very good tables and figures.

Response: We thank you for your positive and constructive feedback.

General Comments:

***Comment:** [1] Section 2: If the goal was to understand impact of forcing uncertainty on simulations, I do not understand why precipitation adjustments (due to wind conditions) were employed prior to the simulation? It would have been interesting to see the overall results related to precipitation. I suspect that would increase uncertainty even more.*

Response: The underlying assumption made here is that the original precipitation data had an unresolved bias prior to the simulations. We wished to begin the sensitivity analysis with reasonably realistic simulations of the observed snowpack, and hence made these precipitation adjustments. We argue that this is not problematic because we do not compare the sensitivity analysis SWE simulations to the observed SWE.

***Comment:** [2] Section 3.1: Very good explanation of why such metrics were used. Other studies should follow this example when listing metrics used in their experiments.*

Response: Thank you.

Comment: [3] Section 3.3.2: The Sobol' method assumes factors are independent from each other. Can you safely assume that for each forcing data analyzed (e.g., T_{air} and RH)?

Response: Thank you for making this excellent point. You are correct that in reality a bias in T_{air} will induce a bias (of the opposite sign) in RH . To avoid this issue, we could have constructed the analysis such that we considered errors in T_{air} and the vapor pressure, but did not do this for simplicity and for general applicability (given that many datasets report RH and not vapor pressure).

Manuscript Revisions: We now state 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 T_{air} that propagates a negative bias to RH).”

Comment: [4] Section 4.2: Could the fact that Q_{li} bias was found to be the most important factor (given its low error magnitudes compared to Q_{si}) indicate some structural limitation in radiation partitioning parameterization in the model (longwave versus shortwave radiation)?

Response: We think that the relative importance of Q_{si} errors is less than that of Q_{li} errors because the high albedo of snow minimizes how much energy Q_{si} transfers to the snowpack.

Manuscript Revisions: We now note this in Section 4.2: “In one sense, this was surprising, given that the bias magnitudes were lower for Q_{li} than for Q_{si} (Table 3). However, the albedo of snow minimizes the amount of energy transmitted to the snowpack from Q_{si} , thereby rendering Q_{si} errors less important than Q_{li} errors. Additionally, the non-linear nature of the model may enhance the role of Q_{li} through interactions with other factors.”

Comment: [5] Section 5: I particularly like the discussion on limitation of the analyses described by the authors.

Response: We appreciate that you liked this discussion.

Comment: [6] Table 2: What is the limitation of fixed ground heat flux? Isn't it calculated in the model? In addition, I imagine that setting it to zero all the time could potentially be problematic.

Response: The snow model provides an option for turning off the ground heat flux. Because ground heat flux typically has a small contribution to the energy balance, it is assumed negligible in some snow modeling applications (e.g., Essery, 1997; Jepsen et al., 2012; Letsinger and Olyphant, 2007), and we chose to mimic those approaches. This indeed would be problematic for calculating the energy balance during snow-free periods and in areas with intermittent

snowpacks, however, the focus of the study was on the snow-covered periods (minimum continuous duration of 15 days, as stated in section 3.3.5).

Comment: [7] *Figures 1 and 2: Excellent figures explaining/summarizing the methodology employed in the study.*

Response: Thank you.

Comment: [8] *Figure 5: Have the authors looked at relationships between certain site characteristics and the magnitude of sensitivity from each factor. For instance, Figures 5 and 7 show an interesting relationship between site elevation/latitude with precipitation forcing for snow disappearance (third column in both figures). Given the site arrangements in the figure, both cases show an increase in sensitivity with elevation (and consequently decrease with latitude). With respect to precipitation and elevation, this can show the difficulties of measuring precipitation according to elevation (especially given the fact that most continuous weather monitoring networks are placed in low/mid-elevation locations). I wonder if there could be other relationships the authors can investigate to see more of those relationships. I see this as a good additional exercise to understand forcing uncertainty and model diagnostics.*

Response: We had not considered this relationship before and thank you for making this suggestion. While this is worthy of further attention, we are hesitant to generalize relationships between site geo-characteristics and sensitivities indices because of the relatively low number of sites represented (n=4 sites, 1 year each) and the confounding number of differences between our sites (e.g., snow climate, latitude, elevation, wind exposure/sheltering, etc.). We would require a much larger population of snow measurement sites in order to more robustly test relationships between sensitivity indices and site characteristics such as elevation and latitude. A successful example of relating climate characteristics to sensitivity can be found in van Werkhoven et al. (2008), which had 12 sites and 39 years each, making it possible to explore inter-site and inter-annual variations in climate and linkages to model sensitivity. We now emphasize in Section 2 that we selected the four sites to check for climate dependencies, but are unable to generalize the results due to the low sample size.

Manuscript Revisions: We now emphasize in Section 2 that we selected the four sites to check for climate dependencies, but are unable to generalize the results due to the low sample size. We note in the discussion however, that there are common results that emerge across all sites, such as the dominance of precipitation bias on SWE, ablation rates and snow disappearance (NB scenario) and longwave bias on all four outputs (NB_lab scenario). This suggests that there may be common features in model sensitivity to forcing errors across distinct climates.

REFERENCES

Essery, R. L. H., 1997: Modelling fluxes of momentum, sensible heat and latent heat over heterogeneous snow cover. *Quart. J. Roy. Meteor. Soc.*, **123**, 1867–1883.

Jepsen, S. M., N. P. Molotch, M. W. Williams, K. E. Rittger, and J. O. Sickman, 2012: Interannual variability of snowmelt in the Sierra Nevada and Rocky Mountains, United States: Examples from two alpine watersheds. *Water Resour. Res.*, **48**, 1–15, doi:10.1029/2011WR011006.

Letsinger, S., and G. Olyphant, 2007: Distributed energy-balance modeling of snow-cover evolution and melt in rugged terrain: Tobacco Root Mountains, Montana, USA. *J. Hydrol.*, **336**, 48– 60, doi:10.1016/j.jhydrol.2006.12.012.

Van Werkhoven, K., T. Wagener, P. Reed, and Y. Tang, 2008: Characterization of watershed model behavior across a hydroclimatic gradient. *Water Resour. Res.*, **44**, W01429, doi:10.1029/2007WR006271.

Response to Interactive comment by A Winstral (Referee)

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

***Comment:** In this open discussion forum/review, the other reviewers have amply summarized the contents of this manuscript, so I don't find the need to restate the contents of this work. The other reviewers have also made some excellent suggestions. The paper is well-written and provides a potentially-extensive analysis of errors that haven't been previously assessed.*

Response: Thank you.

***Comment:** My major concerns are largely in line with the three major comments provided by Referee #3. Once addressed this work would move from a cursory analysis to an extensive one. As you can tell already, I would also like to see a better discussion of the results. Particularly, a more extensive analysis of how some of the site-specific results may/may not relate to site-specific climatology. This type of analysis could be initiated by providing a summary of conditions at each site during the years of analysis. Meteorological summary statistics with a brief description in the Study Site section should be included. This would give the readers (and the authors) guidance as to how the snow regimes differ at each site and how that might be influencing findings/results. These observed differences might be correlated with the modeling results providing greater context and transferability of the presented findings.*

Response: This is a reasonable point, and we now include more in-depth reporting of the results. While there may be interesting linkages between climate and model sensitivity, we note that we are hesitant to generalize relationships between site geo-characteristics and sensitivities indices because of the relatively low number of sites represented (n=4 sites, 1 year each) and the confounding number of differences between our sites (e.g., snow climate, latitude, elevation, wind exposure/sheltering, etc.). We would require a much larger population of snow measurement sites in order to more robustly test relationships between sensitivity indices and site characteristics such as elevation and latitude. A successful example of relating climate characteristics to sensitivity can be found in van Werkhoven et al. (2008), which had 12 sites and 39 years each, making it possible to explore inter-site and inter-annual variations in climate and linkages to model sensitivity. We now emphasize in Section 2 that we selected the four sites to check for climate dependencies, but are unable to generalize the results due to the low sample size.

Manuscript Revisions: We have reorganized and expanded the results and discussion sections to include more in-depth analysis of the site-specific sensitivities and our views on the generalizability of the results, and we now expand Table 1 to include summary statistics of site meteorology for context.

We now emphasize in Section 2 that we selected the four sites to check for climate dependencies, but are unable to generalize the results due to the low sample size. We note in the discussion however, that there are common results that emerge across all sites, such as the

dominance of precipitation bias on SWE, ablation rates and snow disappearance (NB scenario) and longwave bias on all four outputs (NB_lab scenario). This suggests that there may be common features in model sensitivity to forcing errors across distinct climates.

Further suggestions follow:

Comment: *Study Sites: as mentioned above, please summarize the observations at each site. This should be included in Table 1.*

Manuscript Revisions: We have expanded Table 1 to include summary statistics of the meteorology at each site (temperature, precipitation, and wind only).

Comment: *Lines 98-100 (the precipitation corrections): Nowhere in this paragraph is the term “undercatch” referenced. All prior works on these types of adjustments have been based on the theory of wind-induced undercatch. Schmucki et al. is certainly not the only work that should be referenced here. Given that I think the authors are trying to adjust for this process, a 60% adjustment at IC is a very large number (Schmucki et al. applied increases of 5-17% to account for undercatch)! Is there something else going on at this site (e.g. the SWE measurement is located in an enhanced deposition zone, wind speeds are extreme, etc.). Something needs to be stated to justify this large an adjustment.*

Manuscript Revisions: We now include the term “undercatch” in this paragraph and provide more references. The 60% correction at IC is consistent with analyses of undercatch errors at Wyoming-type gauges in wind-blown areas in the Alaska tundra (Yang et al., 2000), and we now make a note of this.

Comment: *On the other side of the coin however, the question of why was there a need to decrease the precipitation measurements at CDP and RME begs for an explanation. Perhaps this is reflective of a modeling deficiency or errors in other observations? A large amount of prior modeling has been conducted at these two sites. I am particularly familiar with the work done at RME where in order to properly model snow evolution at that site it was necessary to adjust the shielded-gauge precipitation catch for undercatch. The “corrected” published data, which generally increased solid precipitation by 10-12%, reflects the undercatch correction which has been applied in every study I know of that has been conducted at this site. This includes the 25-year analysis presented in Reba et al. (2011), which had a Nash-Sutcliffe efficiency coefficient of 0.90 for modeled SWE over the entire period. So I ask, why the need to decrease the data in order to properly model SWE in the current work? As the authors note, accurate precipitation data is vitally important to simulating SWE evolution. A more detailed explanation is needed to explain these eye-catching adjustments that were necessary to properly model SWE.*

Response: This is an excellent point; we can understand how this is eye-catching, as the pervasiveness of undercatch errors makes it a rare necessity to decrease precipitation data. As we initiated our analysis, we found that running an “off-the-shelf model” (i.e., no parameter adjustments) with “off-the-shelf forcing datasets” (most with precipitation undercatch

adjustments already made) rarely resulted in close agreement (i.e., within 10%) of modeled and observed SWE. We can point to multiple sources of uncertainty here, including: (1) model forcing, (2) model parameters, (3) model structure, and (4) model evaluation (e.g., SWE) data. Because you are most familiar with RME, we will focus on that site (WY 2007) as an example to explain why adjusting the initial precipitation data was the most straightforward approach to arrive at reasonable SWE simulations (relative to the observations).

In Figure R6-1 (below), we compare SWE and accumulated precipitation and snowfall datasets at RME, and contrast uncertainties due to evaluation data, and model structure (rain-snow partitioning as an example), parameters, and forcings. We make the following observations:

- **Evaluation data uncertainty:** Snow pillow SWE generally agrees with snow course SWE, though the pillow SWE ablates more rapidly than snow course SWE in April (Figure R6-1a, below). The consistency between these datasets does not provide evidence that the evaluation uncertainty is causing the discrepancy between modeled and observed SWE.
- **Structural uncertainty (rain-vs-snow):** Using four different methods for delineating snowfall results in a range of about 180 mm of accumulated snowfall by season's end (Figure R6-1a, below). Snowfall delineated with dewpoint temperature (from Reba et al. 2011) underestimates SWE whereas snowfall delineated with a linear temperature threshold (UEB) overestimates SWE (Figure R6-1a, below). Because we are looking at accumulated snowfall and not SWE, this does not take into account the three distinct mid-winter melt events, so the simulations with the dewpoint-based approach will have more SWE underestimation than what is suggested in Figure R6-1a (below).
- **Parameters (rain-vs-snow):** Perturbation of the UEB rain-snow threshold temperatures results in a range of about 70 mm of accumulated snowfall by season's end (Figure R6-1b, below). For the selected parameter values, this range is smaller than the range encompassed by the four methods of delineating rain and snow (Figure R6-1a, below).
- **Forcing (precipitation):** Assuming there is still a bias due to under- or over-correction in the original data, we examine snow accumulation under the case of -30% to +30% biases (Figure R6-1c, below). A range in snowfall accumulation of 125 mm exists when considering +/-10% bias and 250 mm when considering +/-20% bias.

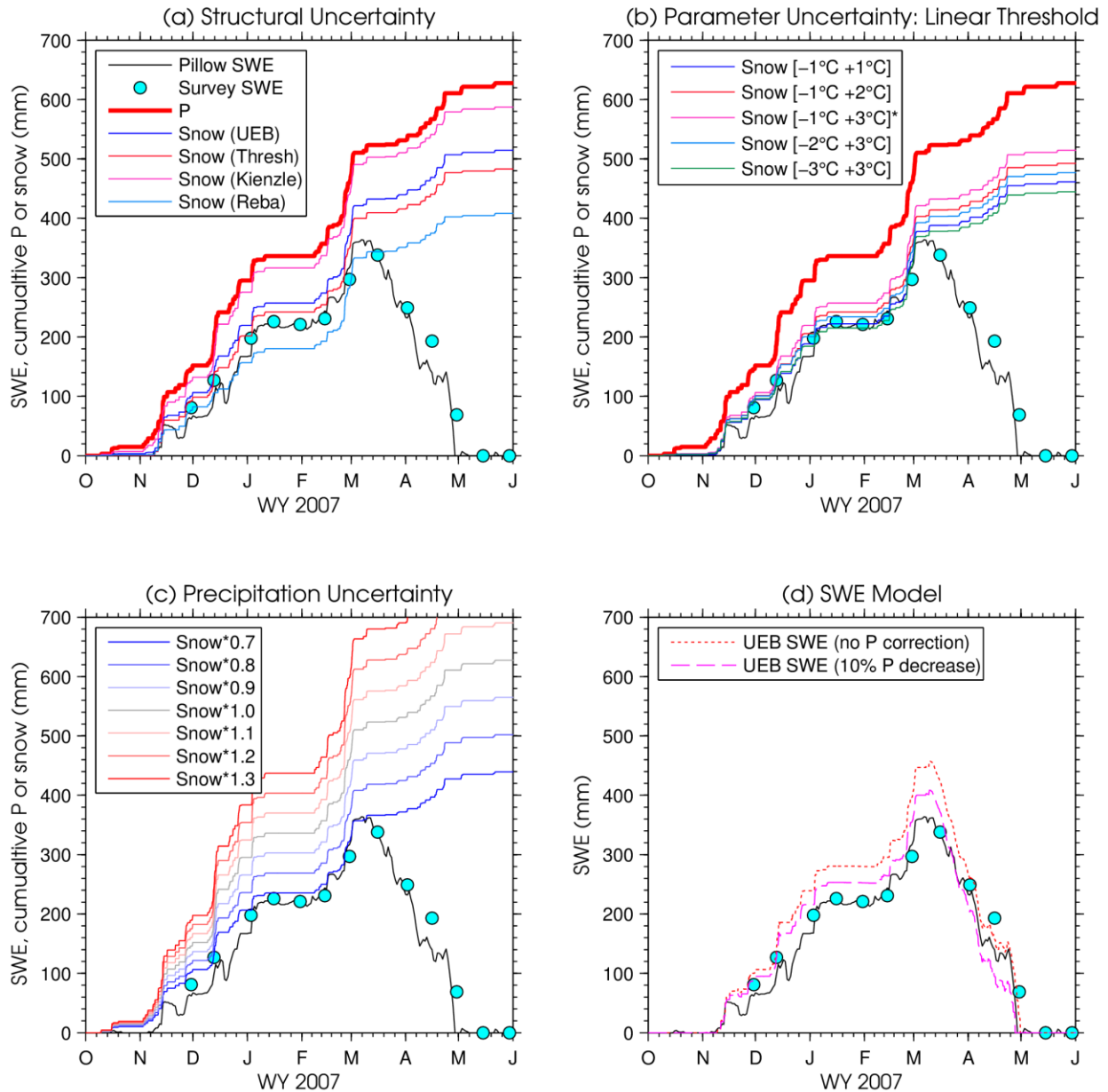


Figure R6-1 SWE, accumulated precipitation and snowfall at RME (WY 2007) as a function of uncertainties in rain-snow (a) structure and (b) parameters, and (c) precipitation. (d) Modeled SWE with adjusted P.

Based on the ranges in snowfall accumulation in these comparisons (and neglecting other processes such as snowmelt), it appears that the most likely cause of the mismatch between modeled and observed SWE is either (1) structural uncertainty (selected rain-snow delineation parameterization) or (2) precipitation bias (on the order of 10-15%). Addressing (1) would require modifying the source code of UEB to incorporate a different parameterization, but this might be somewhat arbitrary because no independent dataset exists (to our knowledge) that can provide clues which rain-snow delineation method is most realistic at each site and should be

selected. Therefore, we concluded that the more straightforward approach would be to address (2) by making some adjustments to the precipitation data.

We note that when forced with the precipitation data (no new adjustments), UEB consistently overestimates SWE throughout most of the season. In contrast, decreasing the precipitation by 10% yields closer agreement with the snow pillow SWE. The UEB simulations of SWE without new precipitation adjustments exhibit a Nash-Sutcliffe (NS) of 0.88 and RMSE of 40 mm, relative to snow pillow SWE. When the 10% decrease in precipitation is applied, UEB yields a Nash-Sutcliffe (NS) of 0.95 and RMSE of 25 mm SWE. These NS values are in fact comparable to the performance of Isnobal that you have referenced (from Reba et al., 2011).

Finally, we note that calibration of model parameters is a step that usually occurs after the sensitivity analysis has determined the most sensitive factors, and this is a reason why we did not calibrate the model prior to the analysis. However, if we consider the interplay between optimal rain-snow threshold parameters in UEB and a potential precipitation adjustment, we find that it is essential to adjust the precipitation in order to find an optimal parameter set (Figure R6-2, below). Leaving the precipitation unchanged would require potentially unrealistic snow and rain threshold temperatures (-4 C and 0 C, respectively) to arrive at the most optimal SWE simulations (figure below), and these parameters are at the edge of the parameter space (suggesting they are not really the optimal parameters). By decreasing the precipitation by 10%, it becomes possible to find a parameter set that is both optimal and realistic. While we are neglecting other processes, this brief analysis provides support for adjusting the precipitation data.

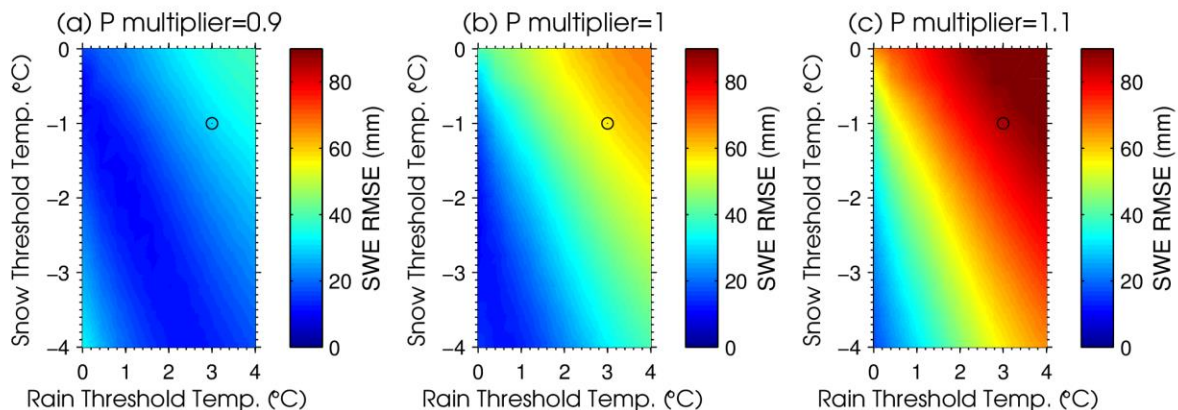


Figure R6-2 RMSE in modeled SWE at RME (WY2007), as a function of rain-snow thresholds and precipitation multipliers of (a) 0.9, (b) 1.0, and (c) 1.1. The black circle are the default UEB rain-snow parameters.

Manuscript Revisions: We briefly expand our discussion at the end of section 2 of why we adjusted the precipitation data.

Comment: Lines 236-238. I think this sentence would sound better if it was re-written in a manner that stated you provide a “brief (or other adjective)” description while further

analysis/details/information can be found in Saltelli and Amnami. (Just a personal opinion there).

Manuscript Revisions: We have changed the sentence to say “Below, we provide a brief summary of the Sobol’ sensitivity analysis methodology but note that further details can be found in Saltelli et al. (2010).”

Comment: Section 3.3.3. As mentioned in F. Pianosi’s comment, the transition from θ (parameterizations) in (1) to θ (forcings) in (2 and 4) should be cleared up.

Manuscript Revisions: We have clarified this point by introducing a distinct variable (ϕ , ϕ) for the forcing errors.

Comment: Lines 415-420. Could you please provide some direct quotes of the structural uncertainties found in Essery et al. (2013) so that the readers of this manuscript can directly see these comparisons rather than having to dig up the Essery work?

Response: There are no direct quotations in the Essery et al. work that are relevant to our discussion. In order to provide a more direct comparison, we have obtained the modeled SWE ensemble from Richard Essery and have created a new figure comparing the forcing uncertainty to structural uncertainty (see Figure 9 in revised manuscript). This illustrates our point that structural uncertainty is only marginally larger than uncertainty due to measurement precision for peak SWE and snow disappearance, and that field uncertainties (due to wind drift and gauge undercatch) are larger than the structural uncertainty. The uncertainty due to structure for ablation rates however is notably higher than the gauge and lab levels of uncertainty.

Manuscript Revisions: We now include the figure comparing the forcing uncertainty to the Essery et al. (2013) structural uncertainty and focus the discussion around that figure.

Comment: Lines 446-448. The Zuzel and Cox findings are being presented out of context. Zuzel and Cox assessed the most important factors for snowmelt for a given snowpack; precipitation (or accumulation amounts) was never a consideration in their analysis. The current findings are really not so "surprising" as the entire winter is analyzed including both accumulation and ablation phases. Great care should be taken when comparing the current findings to research findings solely focused on the ablation phase. If you choose to continue to use this reference, please review the work fully and put it in it’s proper context.

Response: Thank you for catching this problematic comparison.

Manuscript Revisions: We have rephrased this to say: “Prior investigations into the relative importance of forcings to ablation were typically framed for a snowpack at the end of winter, such that P uncertainty was not considered (e.g., Zuzel and Cox, 1975).”

REFERENCES

- Essery, R., S. Morin, Y. Lejeune, and C. B. Ménard, 2013: A comparison of 1701 snow models using observations from an alpine site. *Adv. Water Resour.*, **55**, 131–148, doi:10.1016/j.advwatres.2012.07.013.
- Kienzle, S., 2008: A new temperature based method to separate rain and snow. *Hydrol. Process.*, **22**, 5067–5085, doi:10.1002/hyp.7131.
- Saltelli, A., P. Annoni, I. Azzini, F. Campolongo, M. Ratto, and S. Tarantola, 2010: Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Comput. Phys. Commun.*, **181**, 259–270, doi:10.1016/j.cpc.2009.09.018.
- Van Werkhoven, K., T. Wagener, P. Reed, and Y. Tang, 2008: Characterization of watershed model behavior across a hydroclimatic gradient. *Water Resour. Res.*, **44**, W01429, doi:10.1029/2007WR006271.
- Yang, D., and Coauthors, 2000: An evaluation of the Wyoming Gauge System for snowfall measurement. *Water Resour. Res.*, **36**, 2665–2677, doi:10.1029/2000WR900158.

1 Exploring the impact of forcing error characteristics on physically based snow 2 simulations within a global sensitivity analysis framework

3

4 **M.S. Raleigh¹, J.D. Lundquist² and M.P. Clark¹**

5 [1] National Center for Atmospheric Research, Boulder, Colorado, USA

6 [2] Civil and Environmental Engineering, University of Washington, Seattle, Washington, USA

7 Correspondence to: M.S. Raleigh (raleigh@ucar.edu)

8

9 Abstract

10 Physically based models provide insights into key hydrologic processes, but are associated with
11 uncertainties due to deficiencies in forcing data, model parameters, and model structure. Forcing
12 uncertainty is enhanced in snow-affected catchments, where weather stations are scarce and
13 prone to measurement errors, and meteorological variables exhibit high variability. Hence, there
14 is limited understanding of how forcing error characteristics affect simulations of cold region
15 hydrology [and which error characteristics are most important](#). Here we employ global sensitivity
16 analysis to explore how (1) different error types (i.e., bias, random errors), (2) different error
17 [probability](#) distributions, and (3) different error magnitudes influence physically based
18 simulations of four snow variables (snow water equivalent, ablation rates, snow disappearance,
19 and sublimation). We use Sobol' global sensitivity analysis, which is typically used for model
20 parameters, but adapted here for testing model sensitivity to co-existing errors in all forcings.
21 We quantify the Utah Energy Balance model's sensitivity to forcing errors with [1 840 000](#) Monte
22 Carlo simulations across four sites and [five](#) different scenarios. Model outputs were (1)
23 [consistently](#) more sensitive to forcing biases than random errors, (2) [generally](#) less sensitive to
24 forcing error distributions, and (3) [critically](#) sensitive to different forcings depending on the
25 relative magnitude of errors. For typical error magnitudes [found in areas with drifting snow](#),
26 precipitation bias was the most important factor for snow water equivalent, ablation rates, and
27 snow disappearance timing, but other forcings had a [more dominant](#) impact [when precipitation](#)
28 [uncertainty was due solely to gauge undercatch](#). Additionally, the relative importance of forcing
29 errors depended on the model output of interest. Sensitivity analysis can reveal which forcing
30 error characteristics matter most for hydrologic modeling.

31

32 **1. Introduction**

33 Physically based models allow researchers to test hypotheses about the role of specific processes
34 in hydrologic systems and how changes in environment (e.g., climate, land cover) may impact
35 key hydrologic fluxes and states (Barnett et al., 2008; Clark et al., 2011b; Deems et al., 2013;
36 Leavesley, 1994). Due to the complexity of processes represented, these models usually require
37 numerous meteorological forcing [inputs](#) and model parameters. Most inputs are not measured at
38 the locations of interest and require estimation; hence, large uncertainties may propagate from
39 hydrologic model inputs to outputs. Despite ongoing efforts to quantify forcing uncertainties
40 (e.g., Bohn et al., 2013; Clark and Slater, 2006; Flerchinger et al., 2009) and to develop
41 methodologies for incorporating uncertainty into modeling efforts (e.g., He et al., 2011b;
42 Kavetski et al., 2006a; Kuczera et al., 2010; Slater and Clark, 2006), many analyses continue to
43 ignore uncertainty. These often assume either that all forcings, parameters, and structure are
44 correct (Pappenberger and Beven, 2006) or that only parametric uncertainty is important (Vrugt
45 et al., 2008b). Neglecting uncertainty in hydrologic modeling reduces confidence in hypothesis
46 tests (Clark et al., 2011b), thereby limiting the usefulness of physically based models.

47

48 There are fewer detailed studies focusing on forcing uncertainty relative to the number of
49 parametric and structural uncertainty studies (Bastola et al., 2011; Benke et al., 2008; Beven and
50 Binley, 1992; Butts et al., 2004; Clark et al., 2008, 2011b, 2015a, 2015b; Essery et al., 2013;
51 Georgakakos et al., 2004; Jackson et al., 2003; Kuczera and Parent, 1998; Liu and Gupta, 2007;
52 Refsgaard et al., 2006; Slater et al., 2001; Smith et al., 2008; Vrugt et al., 2003a, 2003b, 2005;
53 Yilmaz et al., 2008). Di Baldassarre and Montanari (2009) suggest that forcing uncertainty has
54 attracted less attention because it is “often considered negligible” relative to parametric and
55 structural uncertainties. Nevertheless, forcing uncertainty merits more attention in some cases,
56 such as in snow-affected watersheds where meteorological and energy balance measurements are
57 scarce (Bales et al., 2006; Raleigh, 2013; Schmucki et al., 2014) and prone to errors [due to](#)
58 [environmental or instrumental factors](#) (Huwald et al., 2009; Lundquist et al., 2015; Rasmussen et
59 al., 2012). Forcing uncertainty is enhanced in complex terrain where meteorological variables
60 exhibit high spatial variability (Feld et al., 2013; Flint and Childs, 1987; Herrero and Polo, 2012;

61 Lundquist and Cayan, 2007). As a result, the choice of forcing data can yield substantial
62 differences in calibrated model parameters (Elsner et al., 2014) and in modeled hydrologic
63 processes, such as snowmelt and evapotranspiration (Mizukami et al., 2014; Wayand et al.,
64 2013). Thus, forcing uncertainty demands more attention in snow-affected watersheds.

65

66 Previous work on forcing uncertainty in snow-affected regions has yielded basic insights into
67 how forcing errors propagate to model outputs and which forcings introduce the most uncertainty
68 in specific outputs. However, these studies have typically been limited to: (1)
69 empirical/conceptual models (He et al., 2011a, 2011b; Raleigh and Lundquist, 2012; Shamir and
70 Georgakakos, 2006; Slater and Clark, 2006), (2) errors for a subset of forcings (e.g., precipitation
71 or temperature only) (Burles and Boon, 2011; Dadic et al., 2013; Durand and Margulis, 2008;
72 Lapo et al., 2015; Xia et al., 2005), (3) model sensitivity to choice of forcing parameterization
73 (e.g., longwave) without considering uncertainty in parameterization inputs (e.g., temperature
74 and humidity) (Guan et al., 2013), and (4) simple representations of forcing errors (e.g., Kavetski
75 et al., 2006a, 2006b). The last is evident in studies that only consider single types of forcing
76 errors (e.g., bias) and single distributions (e.g., uniform), and examines errors separately (Burles
77 and Boon, 2011; Koivusalo and Heikinheimo, 1999; Raleigh and Lundquist, 2012; Xia et al.,
78 2005). [Lapo et al. \(2015\) show that biases have a greater impact than random errors on modeled
79 snow water equivalent and surface temperature, but this analysis only considers longwave and
80 shortwave forcings and considers errors separately.](#) Examining uncertainty in one factor at a
81 time remains popular but fails to explore the uncertainty space adequately, ignoring potential
82 interactions between forcing errors (Saltelli and Annoni, 2010; Saltelli, 1999). [In contrast,
83 global sensitivity analysis explores the uncertainty space more comprehensively by considering
84 uncertainty in multiple factors at the same time.](#)

85

86 The purpose of this paper is to [use global sensitivity analysis to](#) assess how specific forcing error
87 characteristics influence outputs of a physically based snow model. To our knowledge, no
88 [previously published](#) study has investigated this topic in snow-affected regions. It is unclear how
89 (1) different error types (bias vs. random errors), (2) different error distributions, and (3)
90 different error magnitudes across all forcings affect model output. [The impact of forcing errors](#)

91 [on models can be tested by corrupting forcings with specified characteristics \(e.g., artificial](#)
92 [biases and random errors\) and quantifying the impact on model outputs \(e.g., Oudin et al., 2006;](#)
93 [Spank et al., 2013\), but we are unaware of any detailed studies that have done this type of](#)
94 [experiment for all meteorological forcings commonly required for physically based snow](#)
95 [models. We hypothesize that \(1\) model outputs are more sensitive to biases than random errors](#)
96 [in forcing variables, \(2\) the assumed probability distribution for biases will alter the relative](#)
97 [ranking of importance in forcing errors, and \(3\) the magnitude of forcing biases will have a](#)
98 [strong influence on which forcing errors are most important.](#)

99
100 [In our view, it is important to clarify the relative impact of specific error characteristics on](#)
101 [modeling applications, so as to prioritize future research directions, improve understanding of](#)
102 [model sensitivity, and to address questions related to network design. For example, given budget](#)
103 [constraints, is it better to invest in a heating apparatus for a radiometer \(to minimize bias due to](#)
104 [frost formation on the radiometer dome\) or in a higher quality radiometer \(to minimize random](#)
105 [errors associated with measurement precision\)? Additionally, it is important to contextualize](#)
106 [different meteorological data errors, as these errors are usually studied independently of each](#)
107 [other \(e.g., longwave radiation, Flerchinger et al., 2009; air temperature, Huwald et al., 2009\),](#)
108 [and it is unclear how they compare in terms of model sensitivity.](#)

109
110 The [overarching](#) research question is “how do assumptions regarding forcing error characteristics
111 impact our understanding of uncertainty in physically based model output?” Using the Sobol’
112 (1990) global sensitivity analysis framework, we investigate how artificial errors introduced into
113 high-quality observed forcings (temperature, precipitation, wind speed, humidity, shortwave
114 radiation, and longwave radiation) at four sites in contrasting snow climates propagate to four
115 snow model outputs (peak snow water equivalent, ablation rates, snow disappearance timing, and
116 sublimation) that are important to cold region hydrology. We select a single model structure and
117 set of parameters to clarify the impact of forcing uncertainty on model outputs. Specifically, we
118 use the physically based Utah Energy Balance (UEB) snow model (Mahat and Tarboton, 2012;
119 Tarboton and Luce, 1996) because it is computationally efficient. The presented framework
120 could be extended to other models.

121

122 **2. Study sites and data**

123 We selected four seasonally snow covered study sites (Table 1) in distinct snow climates (Sturm
124 et al., 1995; Trujillo and Molotch, 2014). The sites included (1) the tundra Innvait Creek (IC,
125 930 m) site (Euskirchen et al., 2012; Kane et al., 1991; Sturm and Wagner, 2010), located north
126 of the Brooks Range in Alaska, USA, (2) the maritime Col de Porte (CDP, 1330 m) site (Morin
127 et al., 2012) in the Chartreuse Range in the Rhône-Alpes of France, (3) the intermountain
128 Reynolds Mountain East (RME, 2060 m) sheltered site (Reba et al., 2011) in the Owyhee Range
129 in Idaho, USA, and (4) the continental Swamp Angel Study Plot (SASP, 3370 m) site (Landry et
130 al., 2014) in the San Juan Mountains of Colorado, USA. [We selected these sites because of the
131 quality and completeness of the forcing data, and because they spanned contrasting climates
132 \(Table 1\), allowing us to check for potential climate-dependencies in sensitivity to forcing errors.
133 Generalization of the results with climate was not possible due to the low sample size of sites.](#)

134

135 The sites had high-quality observations of model forcings at hourly time steps. Serially complete
136 published datasets are available at CDP, RME, and SASP (see citations above). At IC, data were
137 available from multiple co-located stations (Bret-Harte et al., 2010a, 2010b, 2011a, 2011b,
138 2011c; Griffin et al., 2010; Sturm and Wagner, 2010). These data were quality controlled, and
139 gaps in the data were filled as described in Raleigh (2013).

140

141 We considered only one year for analysis at each site (Table 1) due to the high computational
142 costs of the experiment. Measured evaluation data (e.g., snow water equivalent, SWE) at daily
143 resolution were used [only](#) for qualitative assessment of model output. SWE was observed at
144 snow pillows at IC and RME. [At CDP, a cosmic ray detector collected SWE data.](#) At SASP,
145 acoustic snow depth data were converted to daily SWE using density [inferred](#) from nearby
146 [SNOW TELemetry \(SNOTEL\) \(Serreze et al., 1999\) sites](#) and local snow pit measurements
147 (Raleigh, 2013).

148

149 We adjusted the available precipitation data at each site with a multiplicative factor to [correct for](#)
150 [potential undercatch errors](#) (e.g., Goodison et al., 1998; Rasmussen et al., 2012; Yang et al.,
151 2000) [and to](#) ensure the base model simulation with all observed forcings reasonably represented
152 observed SWE before conducting the sensitivity analysis. [Several studies have demonstrated the](#)
153 [necessity of](#) precipitation adjustments for realistic SWE simulations, even at well-instrumented
154 sites (e.g., Hiemstra et al., 2006; Reba et al., 2011; Schmucki et al., 2014). Precipitation
155 adjustments were most necessary at IC, where windy conditions preclude effective
156 measurements (Yang et al., 2000). In contrast, only modest adjustments were necessary at the
157 other three sites because they were located in sheltered clearings and because the data already
158 had some corrections applied in the published data. [We considered adjustment multipliers](#)
159 [ranging from 0.5 to 2.5 \(increments of 0.05\) and selected the multiplier that yielded the lowest](#)
160 [root mean squared error between observed and modeled SWE](#). Precipitation [multipliers were 1.6](#)
161 [at IC and 1.15 at SASP, and 0.9 at CDP and RME](#). [The undercatch errors at IC were consistent](#)
162 [with the 61-68% undercatch errors found by](#) Yang et al. (2000) [for Wyoming-type gauges in](#)
163 [wind-blown regions](#).

164

165 The initial discrepancies between modeled and observed SWE ([prior to applying the above](#)
166 [precipitation multipliers](#)) may have resulted from deficiencies in the measured forcings, model
167 parameters, model structure, and measured verification data, [and justification of our decision to](#)
168 [apply precipitation multipliers was warranted](#). [Manual observations of SWE \(e.g., snow surveys,](#)
169 [snow pits\) generally supported the automatically collected SWE observations \(no figures](#)
170 [shown\), and thus differences between observed and modeled SWE did not likely stem from](#)
171 [issues in the verification data](#). [Sites where we decreased the precipitation data \(CDP and RME\)](#)
172 [were also the warmer sites and experienced more mixed rain-snow events in the winter](#). Hence,
173 [we considered multiple hypotheses to explain the SWE differences at these sites: \(1\) the choice](#)
174 [of rain-snow parameterization, \(2\) the choice of parameters \(e.g., threshold temperatures\) for the](#)
175 [rain-snow parameterization, and \(3\) the quality of the forcing data \(e.g., precipitation\)](#). For these
176 [warmer sites, an exploratory analysis revealed that either \(1\) or \(3\) could explain the SWE](#)
177 [differences, but auxiliary data \(e.g., precipitation phase data\) were not available to discriminate](#)
178 [these hypotheses](#). [Choosing a different rain-snow parameterization might minimize the SWE](#)
179 [differences at the warmer sites but would not rectify the SWE differences at the colder sites \(IC](#)

180 [and SASP\) where most winter precipitation falls as snow. Therefore, the most straightforward](#)
181 [and consistent approach was to adjust the precipitation data and to leave the native UEB](#)
182 [parameterizations intact.](#) It was beyond the scope of this study to optimize model parameters and
183 unravel the relative contributions of uncertainty for factors other than the meteorological
184 forcings. [Nevertheless, we suggest these precipitation adjustments minimally affected the](#)
185 [sensitivity analysis, as we did not quantitatively compare the model outputs to the observed](#)
186 [response variables \(e.g., SWE\).](#)

187

188 3. Methods

189 3.1. Model and output metrics

190 The Utah Energy Balance (UEB) is a physically based, one-dimensional snow model (Mahat and
191 Tarboton, 2012; Tarboton and Luce, 1996; You et al., 2013). UEB represents processes such as
192 snow accumulation, snowmelt, albedo decay, surface temperature variation, liquid water
193 retention and refreezing, and sublimation. [Due to the one-dimensional structure of the model,](#)
194 [UEB does not account for lateral mass transfer of snow \(e.g., wind-induced snow drifting\), and](#)
195 [therefore these processes must be represented in other model components \(e.g., precipitation](#)
196 [uncertainty, see Sec. 3.2.3\).](#) UEB has a single bulk snow layer and an infinitesimally thin
197 surface layer for energy balance computations at the snow-atmosphere interface. UEB tracks
198 state variables for snowpack energy content, SWE, and a dimensionless snow surface age (for
199 albedo computations). We ran UEB at hourly time steps with six forcings: air temperature (T_{air}),
200 precipitation (P), wind speed (U), relative humidity (RH), incoming shortwave radiation (Q_{si}),
201 and incoming longwave radiation (Q_{li}). We used fixed parameters across all scenarios (Table 2).
202 We initialized UEB during the snow-free period; thus, model spin-up was unnecessary.

203

204 With each UEB simulation, we calculated four summary output metrics: (1) peak (i.e.,
205 maximum) SWE, (2) mean ablation rate, (3) snow disappearance date, and (4) total annual snow
206 sublimation. The first three metrics are important for the timing and magnitude of water
207 availability and identification of snowpack regime (Trujillo and Molotch, 2014), while the fourth
208 impacts the partitioning of annual P into runoff and evapotranspiration. We calculated the snow

209 disappearance date as the first date when 90% of peak SWE had ablated, similar to other studies
210 that use a minimum SWE threshold for defining snow disappearance (e.g., Schmucki et al.,
211 2014). The mean ablation rate was calculated in the period between peak SWE and snow
212 disappearance, and was taken as the absolute value of the mean of all SWE decreases.

213

214 3.2. Forcing error scenarios

215 To test how error characteristics in forcings affect model outputs, we [examined five](#) scenarios
216 (Fig. 1 and Table 3) with different assumptions regarding error types, distributions, and
217 magnitudes (i.e., error ranges). In the first scenario, only bias (normally [distributed for additive](#)
218 [errors](#) or lognormally distributed [for multiplicative precipitation errors](#)) was introduced into all
219 forcings at a level of high uncertainty (based on values observed in the field, see Sec. 3.2.3
220 below). This scenario was named “NB,” where N denotes normal (or lognormal) error
221 distributions and B denotes bias only. The remaining scenarios were identical to NB except one
222 aspect was changed: scenario NB+RE considered both bias and random errors (RE) [in all](#)
223 [forcings](#), scenario UB considered uniformly distributed biases [in all forcings](#), [scenario NB_gauge](#)
224 [considered precipitation error magnitudes associated with gauge undercatch](#), and scenario
225 NB_lab considered error magnitudes [for all forcings](#) at minimal values (i.e., specified instrument
226 accuracy as found in a laboratory). Constructed in this way (Fig. 1), we could test model
227 sensitivity to (1) bias vs. random errors by comparing NB and NB+RE, (2) error distributions by
228 comparing NB and UB, and (3) error magnitudes by comparing NB [\(high forcing uncertainty\) to](#)
229 [both NB_gauge \(moderate uncertainty in precipitation but high uncertainty for all other forcings\)](#)
230 and NB_lab [\(low forcing uncertainty\)](#).

231

232 3.2.1. Error types

233 Forcing data inevitably have some (unknown) combination of bias and random errors. However,
234 hydrologic sensitivity analyses have tended to focus more on bias with little or no attention to
235 random errors (Raleigh and Lundquist, 2012), [whereas data assimilation methods often focus on](#)
236 [random errors but assume bias does not exist](#) (e.g., Dee, 2005). [Rarely is there](#) any consideration
237 of interactions between [these](#) error types. [As a recent example](#), Lapo et al. (2015) tested biases

238 and random errors in Q_{si} and Q_{li} forcings, finding that biases generally introduced more variance
239 in modeled SWE than random errors. Their experiment considered biases and random errors
240 separately (i.e., no error interactions allowed), and examined only a subset of the required
241 forcings (i.e., radiation). Here, we examined co-existing biases in all forcings in NB, UB,
242 [NB_gauge](#), and NB_lab, and co-existing biases and random errors in all forcings in NB+RE.

243

244 Table 3 shows the assignment of error types for the [five](#) scenarios. We relied on studies that
245 assess errors in measurements or estimated forcings to identify typical characteristics of biases
246 and random errors. Published bias values were more straightforward to interpret than random
247 errors because common metrics, such as root mean squared error (RMSE) and mean absolute
248 error (MAE), encapsulate both systematic and random errors. Hence, when defining random
249 errors, published RMSE and MAE served as qualitative guidelines.

250

251 3.2.2. Error distributions

252 [In their recent review of global sensitivity analysis applications in hydrological modeling](#), Song
253 et al. (2015) [identified the selection of probability distributions \(this section\) and ranges \(Sec.](#)
254 [3.2.3\) as among the most important considerations. While it is common practice in sensitivity](#)
255 [analysis to assume a uniform distribution when sampling model parameters](#) (e.g., Campolongo et
256 al., 2011; Rosero et al., 2010), [this may fail to represent the real distribution of errors in](#)
257 [meteorological forcing data, as the uniform distribution implies that extreme and small biases are](#)
258 [equally probable. It is more likely that real error distributions more closely resemble non-](#)
259 [uniform distributions, with higher probability of smaller biases and lower probability of more](#)
260 [extreme biases \(e.g., normal distributions\). Investigators in other fields](#) (e.g., Foscarini et al.,
261 2010; Touhami et al., 2013) [have tested how distribution assumptions \(uniform vs. normal\)](#)
262 [change their computed measures of model sensitivity. These studies broadly suggest that the](#)
263 [grouping of most important factors may be similar under different distribution assumptions,](#)
264 [particularly in cases when interactions are minimal, but the relative ranking of factors within](#)
265 [those groups may vary depending on the distribution. Here we test how the assumed probability](#)
266 [distribution influences the sensitivity of a snow model to forcing errors.](#)

267
268 [We designed](#) the UB scenario [with the](#) naive hypothesis that the probability distribution of biases
269 was uniform [for all six meteorological variables](#). [In contrast](#), error distributions (Table 3) were
270 [assumed non-uniform \(described below\) in](#) scenarios NB, NB+RE, [NB_gauge](#), and NB_lab.
271 Unfortunately, error distributions are reported less frequently than error statistics (e.g., bias,
272 RMSE) in the literature. [We assumed that](#) T_{air} and RH errors follow normal distributions
273 (Mardikis et al., 2005; Phillips and Marks, 1996), as do Q_{si} and Q_{li} errors. Conflicting reports
274 over the distribution of U indicated that errors may be approximated with a normal (Phillips and
275 Marks, 1996), a lognormal (Mardikis et al., 2005), or a Weibull distribution (Jiménez et al.,
276 2011). For simplicity, we assumed that U errors were normally distributed. Finally, we assumed
277 P errors followed a lognormal distribution to account for snow redistribution due to wind
278 drift/scour (Liston, 2004) [or to account for precipitation gauge undercatch](#) (Durand and Margulis,
279 2007). Error distributions were truncated in cases when the introduced errors violated physical
280 limits (e.g., negative U ; see Sec. 3.3.5).

281

282 3.2.3. Error magnitudes

283 We considered [three](#) magnitudes of forcing uncertainty ([Table 3](#)): levels of uncertainty found (1)
284 [in the field for all forcings \(i.e., NB\)](#), (2) [in the field for all forcings except precipitation \(which](#)
285 [has uncertainty due to precipitation gauge undercatch, i.e., NB_gauge\)](#), and (3) [in](#) a controlled
286 laboratory setting ([i.e., NB_lab](#)). [These](#) cases were considered because they sampled realistic
287 errors ([NB and NB_gauge](#)) and minimum errors ([NB_lab](#)). We expected that the error ranges
288 exerted a major control on model uncertainty and sensitivity, [as demonstrated in several prior](#)
289 [sensitivity analyses \(see review of Song et al., 2015\)](#).

290

291 [Consideration of error magnitudes was achieved in each scenario by assigning a range to each](#)
292 [error probability distribution \(see Sec. 3.2.2 and Table 3\)](#). While non-uniform distributions (e.g.,
293 [normal](#)) are typically described by measures other than the range (e.g., mean and variance), we
294 [scaled these distributions \(see Sec. 3.3.5 for details\) such that they were bounded within a](#)
295 [specified range](#). This convention was necessary to ensure that differences between scenarios NB

296 and UB were due solely to the shape of the error probability distributions, and not due to
297 differences in both distribution shape and the domain. Additionally, this followed the typical
298 practice of sensitivity analysis where the range specifies the domain of the distribution.

299

300 We considered field uncertainties in all forcings in NB, NB+RE, and UB, and in all forcings
301 except precipitation in NB_gauge. Field uncertainties depend on the source of forcing data and
302 on local conditions (e.g., Flerchinger et al., 2009; Lundquist et al., 2015). To generalize the
303 analysis, we chose error ranges for the field uncertainty that enveloped the reported uncertainty
304 of different methods for acquiring forcing data. T_{air} error ranges spanned errors in measurements
305 (Huwald et al., 2009) and commonly used models, such as lapse rates and statistical methods,
306 (Bolstad et al., 1998; Chuanyan et al., 2005; Fridley, 2009; Hasenauer et al., 2003; Phillips and
307 Marks, 1996). U error ranges spanned errors in topographic drift models (Liston and Elder,
308 2006; Winstral et al., 2009) and numerical weather prediction (NWP) models (Cheng and
309 Georgakakos, 2011). RH error ranges spanned errors in observations (Déry and Stieglitz, 2002)
310 and empirical methods (e.g., Bohn et al., 2013; Feld et al., 2013). Q_{si} error ranges spanned errors
311 in empirical methods (Bohn et al., 2013), radiative transfer models (Jing and Cess, 1998),
312 satellite-derived products (Jepsen et al., 2012), and NWP models (Niemelä et al., 2001b). Q_{li}
313 error ranges spanned errors in empirical methods (Bohn et al., 2013; Flerchinger et al., 2009;
314 Herrero and Polo, 2012) and NWP models (Niemelä et al., 2001a).

315

316 P error ranges spanned both undercatch (e.g., Rasmussen et al., 2012) and wind drift/scour errors
317 in NB, NB+RE, and UB, but only undercatch errors in NB_gauge. We assumed that P biases
318 due to gauge undercatch in NB_gauge ranged from -10% to +10% because Meyer et al. (2012)
319 found 95% of SNOTEL sites (often in forest clearings) had observations of accumulated P
320 within 20% of peak SWE. Results of NB, NB+RE, and UB were thus most relevant to areas
321 with prominent snow redistribution (e.g., alpine zone), whereas NB_gauge results were more
322 relevant to areas with minimal wind drift errors. It could be argued that uncertainty due to snow
323 drift processes is a structural issue and not a source of forcing error; however, this distinction
324 depends strongly on what type of model is considered. This process is clearly a structural
325 component for snow models with explicit (e.g., three dimensional models with dynamic wind

326 [transport](#), Lehning et al., 2006) [or implicit \(one dimensional models with probabilistic subgrid](#)
327 [snow variability routines, e.g., Clark et al., 2011a\) treatment of snow redistribution. However,](#)
328 [when a one dimensional snow model is applied at length scales shorter than drift process length](#)
329 [scales \(as assumed here with UEB\), then it is not possible to account for snow drift in a structural](#)
330 [sense. Therefore, we treat drifting snow as a form of precipitation error in NB, NB+RE, and UB.](#)
331 [Because UEB lacks dynamic wind redistribution, accumulation uncertainty was not linked to \$U\$](#)
332 [errors but instead to \$P\$ errors \(e.g., drift factor, Luce et al., 1998\).](#)

333

334 In contrast, scenario NB_lab assumed laboratory levels of uncertainty ([i.e., measurement](#)
335 [accuracy](#)) for each forcing. [Skiles et al. \(2012\) considered a similar scenario in their sensitivity](#)
336 [analysis of the SNOBAL model \(Marks and Dozier, 1992; Marks et al., 1992\) to instrument](#)
337 [accuracy at SASP, finding a 5 day range in uncertainty in modeled snow disappearance, with](#)
338 [longwave uncertainty having the greatest impact. An emerging sensitivity analysis \(Sauter and](#)
339 [Obleitner, 2015\) with the CROCUS model \(Brun et al., 1992\) applied on the Kongsvegen](#)
340 [Glacier \(Svalbard\) indicates that longwave measurement uncertainty has an approximately](#)
341 [comparable effect on modeled snow depth as \$\pm 25\%\$ precipitation uncertainty, but is the most](#)
342 [dominant influence on the modeled energy balance and turbulent heat flux \(relative to the](#)
343 [measurement uncertainty of other forcings\). Here we build on these efforts to examine how](#)
344 [instrument accuracy impacts modeled snow variables in a variety of seasonal snow climates. In](#)
345 [general, laboratory](#) uncertainty levels vary with the type and quality of sensors, as well as related
346 accessories (e.g., [radiation shield for the temperature sensor](#)), which we did not explicitly
347 consider. [Because the actual sensors available varied between sites \(Table 1\) and we needed](#)
348 [consistent errors across sites within scenario NB lab, we](#) assumed that the manufacturers'
349 specified accuracy of meteorological sensors at a typical SNOTEL site were representative of
350 minimum uncertainties in forcings because of the widespread use of SNOTEL data in snow
351 studies. While we used the specified accuracy for [idealized](#) P measurements in NB_lab, we note
352 that the instrument uncertainty of $\pm 3\%$ was likely unrepresentative of errors likely to be
353 encountered. For example, corrections applied to the P data (see Sec. 2) exceeded this
354 uncertainty by factors of 3 to 20.

355

3.3. Sensitivity analysis

Numerous approaches that explore uncertainty in numerical models have been developed in the literature of statistics (Christopher Frey and Patil, 2002), environmental modeling (Matott et al., 2009), and optimization/calibration [of hydrology and earth systems models](#) (Beven and Binley, 1992; Duan et al., 1992; Kavetski et al., 2002, 2006a, 2006b; Kuczera et al., 2010; Razavi and Gupta, 2015; Song et al., 2015; Vrugt et al., 2008a, 2008b). Among these, global sensitivity analysis is an elegant platform for testing the impact of input uncertainty on model outputs and for ranking the relative importance of inputs while considering co-existing sources of uncertainty. Global methods are ideal for non-linear models (e.g., snow models). The Sobol' (1990, hereafter Sobol') method is a robust global method based on the decomposition of variance (see below). We investigate Sobol', as it is often the baseline for testing sensitivity analysis methods (Herman et al., 2013; Li et al., 2013; Rakovec et al., 2014; Tang et al., 2007).

368

3.3.1. Overview: [model conceptualization and sensitivity](#)

One can visualize any hydrology or snow model (e.g., UEB) as:

$$\mathbf{Y} = M(\mathbf{F}, \boldsymbol{\theta}) \quad (1)$$

where \mathbf{Y} is a matrix of model outputs (e.g., SWE), $M(\)$ is the model operator, \mathbf{F} is a matrix of forcings (e.g., T_{air} , P , U , etc.), and $\boldsymbol{\theta}$ is an array of model parameters (e.g., [Table 2](#)). The goal of sensitivity analysis is to [determine which input factors](#) (\mathbf{F} and $\boldsymbol{\theta}$) [are most important to](#) specific outputs (\mathbf{Y}) (Matott et al., 2009). Sensitivity analyses tend to focus more on the model parameter array ($\boldsymbol{\theta}$) than on the forcing matrix (Foglia et al., 2009; Herman et al., 2013; Li et al., 2013; Nossent et al., 2011; Rakovec et al., 2014; Rosero et al., 2010; Rosolem et al., 2012; Tang et al., 2007; van Werkhoven et al., 2008). Here, we extend the sensitivity analysis framework to forcing uncertainty by creating k new parameters ($\phi_1, \phi_2, \dots, \phi_k$) that specify forcing uncertainty characteristics (Vrugt et al., 2008b) [and reformulate equation 1 as:](#)

$$\mathbf{Y} = M(\mathbf{F}, \boldsymbol{\theta}, \boldsymbol{\phi}) \quad (2)$$

382 By fixing the original model parameters (Table 2), we focus solely on the influence of forcing
 383 errors on model output (Y). Note it is possible to consider uncertainty in both forcings and
 384 parameters in this framework.

385

386 3.3.2. Sobol' sensitivity analysis

387 Sobol' sensitivity analysis uses variance decomposition to attribute output variance to input
 388 variance. First-order and higher-order sensitivities can be resolved; here, only the total-order
 389 sensitivities [were examined \(see below\) for clarity and because the resulting first-order](#)
 390 [sensitivity indices were typically comparable to the total-order sensitivity indices \(e.g., 83% of](#)
 391 [all cases had total-order and first-order indices within 10% of each other\), suggesting minimal](#)
 392 [error interactions](#). The Sobol' method is advantageous in that it is model independent, can
 393 handle non-linear systems, and is among the most robust sensitivity methods (Saltelli and
 394 Annoni, 2010; Saltelli, 1999). The primary limitation of Sobol' is that it is computationally
 395 intensive, requiring a large number of samples to account for variance across the full parameter
 396 space. [A key assumption to the Sobol' approach is that the factors are independent; hence, our](#)
 397 [analysis does not consider cases of correlated errors \(e.g., a positive measurement bias in \$T_{\text{air}}\$ that](#)
 398 [causes a negative \$RH\$ bias\). Below, we provide a brief summary of the Sobol' sensitivity](#)
 399 [analysis methodology but note that further details can be found in](#) Saltelli et al. (2010).

400

401 3.3.3. Sensitivity indices and sampling

402 Within the Sobol' global sensitivity analysis framework, the total-order sensitivity index (S_{Ti})
 403 describes the variance in model outputs (Y) due to a specific [forcing error](#) (ϕ_i), including both
 404 unique (i.e., first-order) effects and all interactions with all other parameters:

$$405 S_{Ti} = \frac{E[V(Y | \phi_{-i})]}{V(Y)} = 1 - \frac{V[E(Y | \phi_{-i})]}{V(Y)} \quad (3)$$

406 where E is the expectation (i.e., average) operator, V is the variance operator, and ϕ_{-i} signifies all
 407 parameters except ϕ_i . The latter expression defines S_{Ti} as the variance remaining in Y after
 408 accounting for variance due to all other parameters (ϕ_{-i}). S_{Ti} values have a range of [0, 1].

409 Interpretation of S_{Ti} values was straightforward because they explicitly quantified the variance

410 introduced to model output by each parameter (i.e., forcing errors). As an example, an S_{Ti} value
 411 of 0.7 for bias parameter ϕ_i on output Y_j indicates 70% of the output variance was due to bias in
 412 forcing i (including unique effects and interactions).

413

414 A number of numerical methods are available for evaluating sensitivity indices, and most adopt a
 415 Monte-Carlo approach (Saltelli et al., 2010). Evaluation of Eq. (3) requires two sampling
 416 matrices, which we refer to as matrices \mathbf{A} and \mathbf{B} (Fig. 2a). To construct \mathbf{A} and \mathbf{B} , we first
 417 specified the number of samples (N) in the parameter space and the number of parameters (k),
 418 depending on the error scenario (Table 3). Selecting sampling points for these two matrices was
 419 achieved using the quasi-random Sobol' sequence (Saltelli and Annoni, 2010). The sequence
 420 can be approximated as a uniform distribution in the range [0, 1]. Figure 2a shows an example
 421 Sobol' sequence in two dimensions. For each scenario and site, we generated a ($N \times 2k$) Sobol'
 422 sequence matrix with quasi-random numbers in the [0, 1] range, and then divided it in two parts
 423 such that matrices \mathbf{A} and \mathbf{B} were each distinct ($N \times k$) matrices. Calculation of S_{Ti} required
 424 perturbing factors; therefore, a third Sobol' matrix (\mathbf{A}_B) was constructed from \mathbf{A} and \mathbf{B} . In matrix
 425 \mathbf{A}_B , all columns were from \mathbf{A} , except the i th column, which was from the i th column of \mathbf{B} ,
 426 resulting in a ($kN \times k$) matrix (Fig. 2a). Sec. 3.3.5 provides specific examples of this
 427 implementation. From Eq. (3), we compute S_{Ti} as (Jansen, 1999; Saltelli et al., 2010):

$$428 \quad S_{Ti} = \frac{\frac{1}{2N} \sum_{j=1}^N (f(\mathbf{A})_j - f(\mathbf{A}_B^{(i)})_j)^2}{V(Y)} \quad (4)$$

429 where $f(\mathbf{A})$ is the model output evaluated on the \mathbf{A} matrix, $f(\mathbf{A}_B^{(i)})$ is the model output evaluated
 430 on the \mathbf{A}_B matrix where the i th column is from the \mathbf{B} matrix, and i designates the parameter of
 431 interest. Evaluation of S_{Ti} required $N(k+2)$ simulations at each site and scenario.

432

433 3.3.4. Bootstrapping of sensitivity indices

434 To test the reliability of S_{Ti} , we used bootstrapping with replacement across the $N(k+2)$ outputs,
 435 similar to Nossent et al. (2011). The mean and 95% confidence interval were calculated using
 436 the Archer et al. (1997) percentile method and 10 000 samples. For all cases, final S_{Ti} values
 437 were close to the mean bootstrapped values (i.e., 99% had a difference less than 0.001 and no

438 [difference was greater than 0.003](#)), suggesting convergence. Thus, we report only the mean and
439 95% confidence intervals of the bootstrapped S_{Ti} values.

440

441 3.3.5. Workflow and error introduction

442 Figure 2 shows the workflow for creating the Sobol' A , B , and A_B matrices, [mapping](#) Sobol'
443 values to errors, applying errors to the original forcing data, executing the model and saving
444 outputs, and calculating S_{Ti} values. The workflow was repeated at all sites and scenarios. Each
445 step is described in more detail below:

446

447 Step 1) Generate an initial ($N \times 2k$) Sobol' matrix (with N and k values for each scenario, Table
448 3), separate into A and B , and construct A_B (Fig. 2a). NB+RE had $k=12$ (six bias and six random
449 error parameters). All other scenarios had $k=6$ (all bias parameters).

450

451 Step 2) In each simulation, map the Sobol' value of each forcing error parameter (ϕ) to the
452 specified error distribution and range (Fig. 2b, Table 3). [Here we treat the Sobol' values as](#)
453 [quantiles, which allows us to map the Sobol' values to errors via different probability](#)
454 [distributions. For a uniform distribution, the quantile value scales linearly between the specified](#)
455 [lower and upper error ranges \(Fig. 2b\). This linear scaling is not possible for normal \(or](#)
456 [lognormal\) distributions \(due to differences in distribution shape\) and we therefore map the](#)
457 [quantile values to normal \(or lognormal\) distributions scaled within the specified range. We](#)
458 [begin by generating a probability distribution of random numbers with specified mean=0 and](#)
459 [standard deviation of 1 for the case of a normal distribution, and with specified mean=20 and](#)
460 [standard deviation of 0.5 for the case of a lognormal distribution. The random numbers of the](#)
461 [distribution are normalized in the \[0, 1\] range by subtracting the minimum value and dividing by](#)
462 [the maximum value, and then quantiles of these normalized values are computed. The final step](#)
463 [of the mapping is to multiply the normalized quantile by the specified range of uncertainty and](#)
464 [adding the lower bound value.](#) For example, [a \$Q_{si}\$ bias parameter of \$\phi=0.75\$ \(quantile value\) in](#)
465 [the \$\[-100 \text{ W m}^{-2}, +100 \text{ W m}^{-2}\]\$ range would map to a \$Q_{si}\$ bias of \$+50 \text{ W m}^{-2}\$ when assuming a](#)
466 [uniform probability distribution but only \$+14 \text{ W m}^{-2}\$ when assuming a normal distribution. For](#)

467 | [context, a bias parameter of +50 W m⁻² or higher has about a 25% probability of occurring in the](#)
 468 | [uniform distribution but only 2% in the normal distribution.](#)

469

470 Step 3) In each simulation, perturb (i.e., introduce artificial errors) the observed time series of the
 471 *i*th forcing (F_i) with bias (all scenarios), or both bias and random errors (NB+RE only) (Fig. 2c):

$$472 \quad F'_i = F_i \phi_{B,i} b_i + (F_i + \phi_{B,i})(1 - b_i) + \phi_{RE,i} R c_i \quad (5)$$

473 where F'_i is the perturbed forcing time series, $\phi_{B,i}$ is the bias parameter for forcing *i*, b_i is a binary
 474 switch indicating multiplicative bias ($b_i=1$) or additive bias ($b_i=0$), $\phi_{RE,i}$ is the random error
 475 parameter for forcing *i*, R is a time series of randomly distributed noise (normal distribution,
 476 mean=0) scaled in the [-1, 1] range, and c_i is a binary switch indicating whether random errors
 477 are introduced ($c_i=1$ in scenario NB+RE and $c_i=0$ in all other scenarios). For T_{air} , U , RH , Q_{si} , and
 478 Q_{li} , $b_i=0$; for P , $b_i=1$. [The decision to treat biases as multiplicative for \$P\$ but additive for all](#)
 479 [other forcings was made based on practical considerations \(e.g., multiplicative bias in \$T_{air}\$ are](#)
 480 [difficult to interpret\) and on convention of past studies that report forcing errors. However, we](#)
 481 [note this is somewhat subjective, as errors in some forcings \(e.g. radiation\) have been reported in](#)
 482 [both conventions.](#) For P , U , and Q_{si} , we restricted random errors to periods with positive values.
 483 We checked F'_i for non-physical values (e.g., negative Q_{si}) and set these to physical limits. This
 484 was most common when perturbing U , RH , and Q_{si} ; negative values of perturbed P only
 485 occurred when random errors were considered (Eq. 5). Due to this resetting of non-physical
 486 errors, the error distribution was truncated (i.e., it was not always possible to impose extreme
 487 errors). Additional tests (not shown) suggested that distribution truncation changed sensitivity
 488 indices minimally (i.e., <5%), and [thus we assumed this truncation](#) did not alter the relative
 489 ranking of forcing errors.

490

491 Step 4) Input the $N(k+2)$ perturbed forcing datasets into UEB (Fig. 2d). At each site, NB+RE
 492 required 140 000 simulations, whereas the other [four](#) scenarios each required 80 000 simulations,
 493 for a total of [1 840 000](#) simulations in the analysis. The doubling of k in NB+RE did not result
 494 in twice as many simulations because the number of simulations scaled as $N(k+2)$.

495

496 Step 5) Save the model outputs for each simulation (Fig. 2e). [The outputs included daily time](#)
497 [series of SWE, and four summary outputs including peak SWE, mean ablation rate, snow](#)
498 [disappearance date, and total snow sublimation.](#)

499

500 Step 6) Calculate S_{Ti} for each forcing error parameter and model output (Fig. 2f) based on Sect.
501 3.3.3-3.3.4. Prior to calculating S_{Ti} , we screened the model outputs for cases where UEB
502 simulated too little or too much snow (which can occur with perturbed forcings); [this was an](#)
503 [essential step to ensure meaningful results.](#) [Other studies](#) (e.g., Pappenberger et al., 2008) [have](#)
504 [also applied screening methods to model output prior to calculating sensitivity indices.](#) For a
505 valid simulation, we required a minimum peak SWE of 50 mm, a minimum continuous snow
506 duration of 15 days, and identifiable snow disappearance. We rejected samples that did not meet
507 these criteria to avoid meaningless or undefined metrics (e.g., peak SWE in ephemeral snow or
508 snow disappearance for a simulation that did not melt out). The number of rejected samples
509 varied with site and scenario (Table 4). On average, 94% passed the requirements. All cases had
510 at least 86% satisfactory samples, except in UB at SASP, where only ~34% met the
511 requirements. [In this case, the most common reason for rejecting a simulation was that too much](#)
512 [snow was simulated, such that it never disappeared by the end of the model run. The rejected](#)
513 [runs were characterized by high \(positive\) precipitation biases and low \(negative\) biases in \$T_{air}\$,](#)
514 [\$Q_{sj}\$, and \$Q_{lj}\$.](#) Despite this attrition, S_{Ti} values still converged in all cases.

515

516 4. Results

517 4.1. [Propagation of forcing uncertainty to model outputs](#)

518 Figure 3 shows density plots of daily SWE from UEB at the four sites and [five](#) forcing error
519 scenarios (Fig. 1, Table 3), while Fig. 4 summarizes the model outputs. As a reminder, NB
520 assumed normal (or lognormal) biases at field level uncertainty. The other scenarios were the
521 same as NB, except NB+RE considered both biases and random errors, UB considered uniform
522 distributions, [NB gauge considered gauge undercatch biases in precipitation,](#) and NB_lab
523 considered lower error magnitudes [in all forcings](#) (i.e., laboratory level uncertainty).

524

525 Large uncertainties in SWE were evident, particularly in NB, NB+RE, and UB (Fig 3.a-l). The
526 large range in modeled SWE within these three scenarios often translated to large ranges in mean
527 ablation rates (Fig 4.e-h), snow disappearance dates (Fig 4.i-l) and total sublimation (Fig 4.m-p).
528 In contrast, SWE and output uncertainties in NB_gauge and NB_lab were comparatively small
529 (Fig. 3m-t and Fig. 4). Model output ranges were generally larger in NB_gauge than NB_lab.
530 The envelope of SWE simulations in NB_lab more tightly encompassed observed SWE at all
531 sites, except during early winter at IC (Fig. 3m), which was possibly due to initial *P* data quality
532 and redistribution of snow to the snow pillow site.

533

534 NB and NB+RE generally yielded similar SWE density plots (Fig. 3a-h), but NB+RE yielded a
535 slightly higher frequency of extreme SWE simulations. NB and NB+RE also had very similar
536 (but not equivalent) mean outputs values and ensemble spreads at all sites except IC (Fig. 4).
537 This initial observation suggested that random errors in the forcings had minimal impact on
538 model behavior at CDP, RME, and SASP. NB+RE and NB model outputs were slightly
539 different at IC (particularly for the ablation rates), indicating that random errors had some
540 influence there, and this was possibly due to the low snow accumulation (~200 mm peak SWE
541 observed) at that site and brief snowmelt season (less than 10 days in the observations).

542

543 NB and UB yielded generally very different model outputs (Fig. 3 and Fig. 4). The only
544 difference in these two scenarios was the assumption regarding error distribution (Table 3).
545 Uniformly distributed forcing biases (scenario UB) yielded a relatively uniform ensemble of
546 SWE simulations (Fig. 3i-l), larger mean values of peak SWE and ablation rates, and later snow
547 disappearance, as well as larger uncertainty ranges in all outputs. At some sites, UB also had a
548 higher frequency of simulations where seasonal sublimation was negative (i.e., condensation).

549

550 Contrasting NB and NB_gauge, NB_gauge had a lower uncertainty range in SWE and slightly
551 higher mean peak SWE at all sites (Fig. 3 and Fig. 4). With the exception of RME, the ranges in
552 ablation rates in NB_gauge were at least 50% smaller than in NB (Fig. 4 e-h). Snow

553 disappearance ranges were marginally smaller in NB_gauge relative to NB (Fig. 4i-l). Finally,
554 sublimation ranges were very similar between NB and NB_gauge (Fig. 4m-p).

555

556 Relative to NB, NB_lab had smaller uncertainty ranges in all model outputs (Fig. 3 and Fig. 4),
557 an expected result given the lower magnitudes in forcing errors in NB_lab (Table 3). Likewise,
558 NB_lab SWE simulations were generally less biased than NB, relative to observations (Fig. 3).
559 NB_lab generally had higher mean peak SWE and ablation rates, and later mean snow
560 disappearance timing than NB (Fig 4).

561

562 **4.2. Model sensitivity to forcing error characteristics**

563 Total-order sensitivity indices (S_{Ti}) were calculated for four summary variables of model output
564 (peak SWE, mean ablation rates, snow disappearance dates, and total sublimation) and for daily
565 SWE output at all sites and error scenarios. Examination of the total-order indices with sample
566 size indicated that most indices stabilized after evaluating the model at 3 000 to 5 000 samples
567 (no figures shown). Below we sequentially compare sensitivity indices from different scenarios
568 to scenario NB to test the impact of differences in error characteristics (type, probability
569 distribution, and magnitudes).

570

571 **4.2.1. Impact of error types**

572 We first focus on sensitivity to forcing bias, as this error type was common to scenarios NB and
573 NB+RE. Figure 5 shows the computed total-order sensitivity indices from the two scenarios
574 (with sensitivities to biases and random errors shown separately in NB+RE). Both NB and
575 NB+RE showed that UEB peak SWE was most sensitive to P bias at all sites (Fig.5a-d). In both
576 scenarios, P bias was also the most important factor for ablation rates and snow disappearance at
577 all sites (Fig. 5e-l). For ablation rates in NB, T_{air} bias was the next most important factor (after P
578 bias) at CDP₂ while biases in Q_{si} and Q_{li} were secondarily important at RME (Fig.5f-g). For
579 ablation rates at IC in NB+RE, most types of errors had some baseline influence (i.e., $S_{Ti} \geq 0.5$) on
580 model sensitivity (Fig. 5e). In both NB and NB+RE, biases in the radiation terms were of
581 secondary importance to snow disappearance timing (Fig. 5i-k). In contrast to the other three

582 model outputs, sublimation in NB and NB+RE was insensitive to P bias and the most important
583 factors varied somewhat between sites and scenarios (Fig. 5m-p). In both scenarios, sublimation
584 was most sensitive to RH bias at IC and U bias at SASP. At CDP and RME, sublimation was
585 most sensitive to RH bias in NB; however, in NB+RE, sublimation was most sensitive to Q_{li} bias
586 at CDP and to T_{air} bias at RME (Fig. 5n-o). In both scenarios, biases in T_{air} , Q_{sj} , or Q_{li} were
587 generally of secondary importance for sublimation.

588

589 We hypothesized that the snow model outputs would have higher sensitivity to biases than to
590 random errors in the forcings. The results of our analysis generally supported this hypothesis.
591 Across all outputs and sites, S_{Ti} values for random errors were always less than or comparable to
592 the smallest S_{Ti} bias values, and the most important factor was always a bias term (Figure 5).
593 Furthermore, there was typically high correspondence between NB and NB+RE (bias terms
594 only) in terms of identifying the most important forcing error (e.g., P bias in peak SWE and
595 ablation rates at all sites, Fig. 5a-h). The main exceptions were snow disappearance at IC (Fig.
596 5i), and sublimation at CDP and RME (Fig. 5n-o), where the two scenarios identified different
597 errors as the most important factor. However, even in these exceptional cases, the two scenarios
598 yielded similar groupings of more important vs. least important errors. For example, biases in
599 T_{air} and RH were important to sublimation at RME in both scenarios (Fig. 5o), though they
600 distinguished these sensitivities differently (i.e., NB found RH bias was more important whereas
601 NB+RE found T_{air} bias was more important).

602

603 While there was general correspondence between NB and NB+RE (bias terms), sensitivity
604 indices were not identical across cases, due to interactions between biases and random errors in
605 NB+RE. Random errors changed model sensitivity to biases, and the change in sensitivity was
606 more notable (i.e., absolute change exceeding 0.10) for ablation rates and snow disappearance at
607 IC (Fig. 5e,i) and sublimation at all sites (Fig. 5m-p). Random errors amplified model sensitivity
608 to biases in some cases (e.g., U bias in all sublimation scenarios) but diminished model
609 sensitivity to biases in other cases (e.g., RH bias in all sublimation scenarios). Because
610 consideration of second-order sensitivity indices was beyond the scope of the study, we were

611 unable to determine which specific interactions were important in terms of error types, and leave
612 this topic for future work.

613

614 **4.2.2. Impact of probability distribution of errors**

615 We hypothesized that the assumed probability distribution of errors would alter the relative
616 hierarchy of forcing biases. However, the results did not consistently support this hypothesis
617 (Fig. 6). In all cases, scenarios NB and UB identified the same factor as the most important and
618 similar factors as the least important at all sites. Specifically, P bias was most important for peak
619 SWE, ablation rates, and snow disappearance at all sites in both scenarios (Fig. 6a-l). The only
620 exception was in scenario UB at IC, where ablation rates had similar sensitivity to P bias and U
621 bias. In both scenarios, T_{air} bias was the second most important factor for peak SWE and
622 ablation rates at the warmest site, CDP. Both scenarios showed that RH bias was the least
623 important factor to snow disappearance at all four sites (Fig. 6i-l). Finally, both NB and UB
624 showed that P bias was least important for sublimation (in contrast to the other model outputs)
625 and that RH and U biases were among the most sensitive factors for sublimation (Fig. 6m-p).
626 More specifically, sublimation was most sensitive to RH bias at IC, CDP, and RME, and U bias
627 as SASP (Fig. 6m-p).

628

629 For a few specific forcings and outputs, the selected probability distribution played a role in
630 model sensitivity to that type of forcing bias. For example, assumption of a uniform probability
631 distribution (UB) for forcing errors enhanced the sensitivity of sublimation to U and RH biases
632 but reduced sublimation sensitivity to Q_{si} and Q_{li} biases at all sites (Fig. 6m-p). In contrast,
633 assuming a normal distribution (NB) of biases yielded the opposite results. Additionally,
634 modeled ablation rates at IC were notably more sensitive to forcing biases (precipitation
635 excluded) in scenario UB than in NB.

636

637 **4.2.3. Impact of error magnitude**

638 We hypothesized that the relative magnitude of forcing errors would exert a strong control on
639 model sensitivity. Comparing NB to NB_gauge and to NB_lab generally supported this

640 hypothesis (Fig. 7). The contrast in S_{Ti} values between scenarios NB, NB_gauge, and NB_lab
641 implied that the specified ranges of forcing errors was a critical determinant of model sensitivity.

642
643 While P bias was the most important factor at all sites in NB for peak SWE, ablation rates, and
644 snow disappearance, P bias was never the most important factor for these model outputs in
645 NB_gauge, and in many cases was among the least important errors (Fig. 7a-l). In NB_gauge,
646 peak SWE was most sensitive to RH bias at IC, T_{air} bias at CDP and RME, and Q_{li} bias at SASP
647 (Fig. 7a-d). Ablation rates in NB_gauge were most sensitive to T_{air} bias at CDP and to Q_{li} bias at
648 IC, RME, and SASP (Fig. 7e-h). Snow disappearance was also most sensitive to Q_{li} bias at all
649 four sites in NB_gauge (Fig. 7i-l). However, for sublimation at all sites, NB and NB_gauge
650 yielded very similar sensitivities to forcing biases (Fig. 7m-p). Specifically, in both NB and
651 NB_gauge, modeled sublimation was most sensitive to RH bias at IC, CDP, and RME and to U
652 bias at SASP (Fig. 7m-p). The similarity in sublimation sensitivity indices between NB and
653 NB_gauge emerged because these scenarios only differed in terms of P uncertainty (Table 3) and
654 because P bias was not important to modeled sublimation. The contrast between sensitivity
655 indices in these two scenarios and for these four outputs illustrated that model sensitivity may
656 depend on both the magnitudes of uncertainty for specific forcings and on the output of interest.

657
658 Whereas NB_gauge demonstrated that reducing the magnitude of forcing uncertainty in one
659 factor (i.e., precipitation) was sufficient to change which factors were most and least important,
660 NB_lab showed that changing the magnitude of forcing uncertainty in all terms could yield a
661 substantially different pattern of model sensitivity (Fig. 7). As a primary example, scenarios NB
662 and NB_lab did not agree whether P bias or Q_{li} bias was the most important factor for peak
663 SWE, ablation rates, and snow disappearance dates at all four sites (Fig. 7a-l). For sublimation,
664 NB_lab sensitivity indices indicated that Q_{li} bias was most important, whereas RH bias (IC,
665 CDP, and RME) and U bias (SASP) were most important in NB (Fig. 7m-p). Across all sites
666 and outputs in NB_lab, Q_{li} bias was consistently the most important factor (Fig. 7). In one sense,
667 this was surprising, given that the bias magnitudes were lower for Q_{li} than for Q_{si} (Table 3).
668 However, the albedo of snow minimizes the amount of energy transmitted to the snowpack from
669 Q_{si} , thereby rendering Q_{si} errors less important than Q_{li} errors. Additionally, the non-linear

670 nature of the model may enhance the role of Q_{li} through interactions with other factors. The
671 general lack of importance in P bias in NB_lab (main exception was peak SWE at IC, Fig. 7a)
672 was due to the discrepancy between the laboratory specified accuracy for P gauges and typical
673 errors encountered in the field.

674

675 **4.2.4. Relative controls of forcing error characteristics on SWE sensitivity**

676 The above results sequentially compared sensitivity indices from different error scenarios to NB
677 in order to ascertain how different assumptions regarding error types, distributions, and
678 magnitudes translated to changes in model sensitivity. To summarize the relative controls of
679 these three forcing error characteristics on model sensitivity, we calculated daily sensitivity
680 indices of modeled SWE to forcing biases at each site and scenario (Fig. 8). We also examined
681 the correspondence between changes in S_{Ti} values and the timing within the snow season.

682

683 Comparing the broad patterns in the time varying S_{Ti} values across the five scenarios, it was
684 evident that error magnitudes were the greatest determinant in model sensitivity to forcing errors
685 through the snow season (compare Fig. 8a-l with Fig. 8m-t). NB, NB+RE, and UB exhibited
686 similar patterns, with high S_{Ti} in P bias throughout the year and with the other forcing biases
687 yielding low S_{Ti} values in the winter and increasing S_{Ti} values in the spring and early summer for
688 some forcings (Fig. 8a-l). In contrast, NB_gauge and NB_lab (Fig. 8m-t) had lower S_{Ti} values
689 for P bias, and more coherent changes in S_{Ti} values that were more synchronized with the
690 specific part of the snow season.

691

692 After error magnitudes, the next most important determinant to model sensitivity was the
693 probabilistic distribution of forcing errors (compare Fig. 8a-d and Fig. 8i-l). Relative to NB, UB
694 tended to yield lower S_{Ti} values for P bias. UB also had higher S_{Ti} values for biases in T_{air} , Q_{li} ,
695 and Q_{si} as time progressed at IC, CDP, and RME (Fig. 8i-k). Finally, the addition of random
696 errors was least important to model sensitivity, as the evolution of S_{Ti} bias values was very
697 similar between NB and NB+RE at most sites (compare Fig. 8a-d and Fig. 8e-h). Random errors

698 [mattered the most to modeled SWE at IC, but random errors only changed \$S_{Ti}\$ values \(on](#)
699 [average\) by less than 10%.](#)

700

701 **5. Discussion**

702 Here we examined the sensitivity of physically-based snow simulations to forcing error
703 characteristics (i.e., types, [probability](#) distributions, and magnitudes) using Sobol' global
704 sensitivity analysis. [A key result is that among these three](#) characteristics, the magnitude of
705 biases had the most significant impact on UEB simulations (Figs. 3-4) and on model sensitivity
706 (Figs. 7-8). The assumed [probability](#) distribution of biases was important in that it increased the
707 range of model outputs (compare NB and UB in Fig. 4), but [surprisingly, this usually translated](#)
708 [to only modest changes in](#) model sensitivity to forcing errors (Figs. 6 and 8). [Random errors](#)
709 [were usually less important than biases. Although random errors changed model sensitivity to](#)
710 [biases through error interactions, this effect was only large in specific conditions \(e.g., ablation](#)
711 [rates at IC, Fig. 5e\), and the snow model was never more sensitive to random errors than to](#)
712 [biases \(Fig. 5\). Below we discuss these three error characteristics \(in order of importance, as](#)
713 [suggested by the results\), place forcing errors in the context of structural uncertainty, and](#)
714 [identify limitations of the analysis and future research directions.](#)

715

716 **5.1. [Ranges of error magnitudes](#)**

717 [The results supported our hypothesis that the magnitude of biases strongly influences the relative](#)
718 [importance of forcing errors. The three magnitudes of uncertainty considered \(NB, NB_gauge,](#)
719 [and NB_lab\) all resulted in different patterns in model sensitivity to forcing biases, and these](#)
720 [patterns also varied with the output of interest \(Fig. 7\). Modeled peak SWE, ablation rates, and](#)
721 [snow disappearance were consistently sensitive to \$P\$ bias in scenario NB and to \$Q_{fi}\$ bias in](#)
722 [scenario NB_lab, but there was less consistency in the dominant forcing errors across these three](#)
723 [outputs in scenario NB_gauge. While peak SWE, ablation rates, and snow disappearance dates](#)
724 [had similar sensitivities to forcing errors \(particularly to \$P\$ biases\), sublimation exhibited notably](#)
725 [different sensitivity to forcing errors. \$P\$ bias was frequently the least important factor for](#)
726 [sublimation, in contrast to the other model outputs. Biases in \$RH\$, \$U\$, and \$T_{air}\$ were often the](#)

727 major controls on modeled sublimation in NB, NB+RE, UB, and NB_gauge, while Q_{li} bias
728 controlled modeled sublimation in NB_lab. These [field](#) results partially agree with the
729 sensitivity analysis of Lapp et al. (2005), who showed the most important forcings for
730 sublimation in the Canadian Rockies were U and Q_{si} . [However, they did not consider \$Q_{li}\$ in their](#)
731 [sensitivity analysis and so the experiments are not exactly comparable.](#) These results suggest
732 that no single forcing is important across all modeled variables, and model sensitivity strongly
733 depends on the output of interest.

734

735 [The dominant effect of \$P\$ bias on modeled peak SWE, ablation rates, and snow disappearance in](#)
736 [the field scenarios \(e.g., NB\) confirmed previous reports that \$P\$ uncertainty is a major control on](#)
737 [snowpack dynamics \(Durand and Margulis, 2008; He et al., 2011b; Schmucki et al., 2014\). It](#)
738 [was surprising that \$P\$ bias was often the most critical forcing error for ablation rates in these](#)
739 [scenarios \(Fig. 5-6\). Prior investigations into the relative importance of forcings to ablation were](#)
740 [typically framed for a snowpack at the end of winter, such that \$P\$ uncertainty was not considered](#)
741 [\(e.g., Zuzel and Cox, 1975\). The results here showed that ablation rates were highly sensitive to](#)
742 [\$P\$ bias and this is likely](#) because it controlled the timing and length of the ablation season.
743 Positive P bias extends the fraction of the ablation season in the warmest summer months when
744 ablation rates and radiative energy approach maximum values, [whereas negative \$P\$ bias truncates](#)
745 [the fraction of ablation in the warm season.](#) Trujillo and Molotch (2014) reported a similar result
746 based on SNOTEL observations.

747

748 [The contrast between scenarios NB, NB_gauge, and NB_lab highlights that selection of the error](#)
749 [ranges is a critical step in sensitivity analysis. However, we recognize that there is some](#)
750 [subjectivity in the specification of these ranges. Quantification of errors in forcing estimation](#)
751 [methods is best achieved through comparisons with surface observations](#) (e.g., Bohn et al., 2013;
752 Flerchinger et al., 2009), [but it remains challenging to specify error ranges with confidence](#)
753 (Song et al., 2015). [Key considerations controlling the ranges and impacts of forcing errors](#)
754 [include the representativeness of the forcing data \(e.g., reanalysis, numerical weather model](#)
755 [output, extrapolated surface measurements, etc.\) in the study area, the length scale of dominant](#)
756 [processes \(e.g., snow drifting\), and the configuration of the snow model \(e.g., spatial scale,](#)

757 complexity). Here we selected ranges in the field scenarios to encompass errors encountered
758 across a variety of possible forcing data sources (Table 3), but ultimately the appropriate ranges
759 must be tailored to the specific application. This supports the need for continual evaluation of
760 forcing datasets across a variety of climates and environmental conditions.

761

762 **5.2. Probability distribution of errors**

763 The results did not universally support our hypothesis that the assumed probability distribution
764 of biases was important to the relative ranking of forcing errors. The relative consistency in the
765 dominant forcing errors between NB and UB may have emerged because the probability
766 distributions of all six forcing biases varied together between these two scenarios (i.e., all forcing
767 biases were uniform in UB and either normal or lognormal in NB). While we did not conduct
768 additional tests, we suspect that changing the probability distribution of just a single forcing error
769 (e.g., T_{air} bias) from normal to uniform would have uniquely enhanced model sensitivity to that
770 particular forcing error (Touhami et al., 2013).

771

772 The similarity of results between scenarios NB and UB conform to findings in previous studies
773 (e.g., Foscarini et al., 2010; Touhami et al., 2013) where uniform and normal distributions
774 identified similar factors as the most important. These previous studies imply that greater
775 differences in sensitivity indices (as a function of distribution) will emerge when factor
776 interactions are more prominent. The case with the strongest error interactions here (i.e.,
777 ablation rates at IC) also yielded the largest differences in sensitivity indices between scenarios
778 NB and UB, which is consistent with the prevailing logic.

779

780 **5.3. Error types**

781 The results were consistent with our hypothesis that the snow model is more sensitive to biases
782 than to random errors in the forcings. While previous investigations supported this idea for
783 shortwave and longwave forcings in physically based snow models (i.e., Lapo et al., 2015), the
784 current study showed that biases are more important than random errors for all commonly
785 required meteorological forcings (and not just irradiances). The model was more sensitive to

786 [biases and less sensitive to random errors due to the systematic nature of biases. In contrast, the](#)
787 [effect of random errors tended to cancel out when integrating model outputs over long periods.](#)
788 [Our selected model outputs were generally a function of several months of mass and energy](#)
789 [exchange in the snowpack, thereby ensuring minimization of effects from random errors.](#)
790 [Random errors only had a greater impact on ablation rates at IC \(Fig. 5e\), and this was because](#)
791 [the relatively brief snowmelt period presented an opportunity for the random errors to not cancel](#)
792 [out. Hence, the model may have greater sensitivity to random errors for other model outputs not](#)
793 [considered here that integrate over relatively short time scales \(e.g., snowmelt over a single day\).](#)

794

795 **5.4. Contextualizing forcing and structural uncertainties**

796 Our central argument at the onset was that forcing uncertainty may be comparable to parametric
797 and structural uncertainty in snow-affected catchments. To support our argument [and to place](#)
798 [our results in context](#), we compare our results at CDP in 2005-2006 to Essery et al. (2013), who
799 assessed the impact of structural uncertainty [in a suite of local snowpack processes \(i.e., snow](#)
800 [compaction, fresh snow density, snow albedo evolution, surface heat and moisture fluxes, snow](#)
801 [cover fraction, snow hydrology, and thermal conductivity\)](#) on SWE simulations from 1701
802 physically based snow models at the same site/year. [Figure 9 compares the 95% uncertainty](#)
803 [ranges in peak SWE, ablation rates, and snow disappearance in NB, NB gauge, and NB lab to](#)
804 [the ranges found across the 1701 snow models of Essery et al. \(2013\).](#) [From the comparisons at](#)
805 [this site, it is clear that the uncertainty associated with drifting snow \(i.e., scenario NB\)](#)
806 [overwhelms the structural uncertainty in local snowpack processes for all three model outputs.](#)
807 [As discussed previously, it could be argued that the uncertainty due to drifting snow is a](#)
808 [structural issue \(not a forcing issue\) and that this does not represent the uncertainty of sheltered](#)
809 [areas where drifting snow less important. Hence, NB gauge may be a better determinant of the](#)
810 [level of uncertainty that can be attributed unambiguously to errors in forcing data. In that case,](#)
811 [the output uncertainty range due to model forcing is still larger than that due to the structural](#)
812 [uncertainty \(as considered by Essery et al., 2013\) in the cases of peak SWE and snow](#)
813 [disappearance but is smaller for ablation rates \(Fig. 9\). As expected, the case of forcing](#)
814 [uncertainty in NB lab yields the lowest range in model outputs at CDP \(Fig. 9\), though it is](#)
815 [interesting to note that the uncertainty in peak SWE due to structural uncertainty \(90 mm\) is only](#)

816 [marginally larger than that due to the specified instrument accuracy \(60 mm\). These](#)
817 [comparisons illustrate that](#) forcing uncertainty cannot be discounted, and the magnitude of
818 forcing uncertainty is a critical factor in how forcing uncertainty compares to [other sources of](#)
819 [uncertainty \(e.g., structural\). This resonates with the recent work of](#) Magnusson et al. (2015)
820 [who found that uncertainty in the \$P\$ forcing was a greater determinant of model performance than](#)
821 [structural considerations.](#)

822

823 **5.5. Caveats and future research**

824 Limitations of the analysis are [that the impact of forcing error characteristics on model behavior](#)
825 [is evaluated through the lens of a single sensitivity analysis method and a single snow model. It](#)
826 [is possible that alternative sensitivity analysis methods might yield different results than the](#)
827 [Sobol' method, as suggested in previous studies](#) (e.g., Pappenberger et al., 2008). [Likewise, we](#)
828 [recognize it is possible that](#) different snow models may yield different sensitivities to forcing
829 uncertainty. [As one](#) example, both Koivusalo and Heikinheimo (1999) and Lapo et al. (2015)
830 found UEB (Tarboton and Luce, 1996) and the SNTHERM model (Jordan, 1991) exhibited
831 significant differences in radiative and turbulent heat exchange. [As another example, the role of](#)
832 [U bias on snowpack formation may vary strongly depending on the snow model configuration.](#)
833 [Because of the lack of wind transport in UEB, we lumped snow drift uncertainty into P](#)
834 [uncertainty via a “drift factor” formulation \(Luce et al., 1998\) and this could not account for the](#)
835 [role of wind in snow drift/scour processes \(Mott and Lehning, 2010; Winstral et al., 2013\). This](#)
836 [convention would be unnecessary for a model that explicitly models this process \(e.g., the](#)
837 [SNOWPACK model, Lehning et al., 2006\), and for this type of model we would expect the role](#)
838 [of U bias to be enhanced \(relative to UEB\) for outputs such as peak SWE and snow](#)
839 [disappearance timing. While sensitivity may vary with model selection in these examples, there](#)
840 [is also evidence suggesting that similar results may emerge when using different snow models](#)
841 [for a similar type of error scenario. Despite using different models, a somewhat different suite of](#)
842 [forcing variables, and slightly different error ranges, our NB lab experiment corroborated](#)
843 [independent reports that \$Q_{li}\$ measurement uncertainty was most important to both modeled snow](#)
844 [disappearance \(Skiles et al., 2012\) and sublimation/latent heat exchange \(Sauter and Obleitner,](#)
845 [2015\). Our analysis demonstrated this result was consistent across four snow climates and this](#)

846 result was apparent in four different model outputs (Fig. 7). The implication here is that more
847 work is needed to better understand how different snow models respond to forcing uncertainty.

848

849 Generalizing the relationship between model sensitivity and site climate is a research topic of
850 high interest. Although we found similarities in model sensitivity to specific forcing errors
851 across sites (e.g., high sensitivity to P bias in peak SWE, ablation rates, and snow disappearance
852 in NB, NB+RE, and UB, Fig. 8a-l), we note that the sites exhibited some differences in
853 sensitivity when P uncertainty was reduced to gauge levels (Fig. 8m-p). Additionally, the sites
854 exhibited differences in the relative importance of secondary forcing errors (Fig. 6-7). There
855 may be interesting linkages between climate and model sensitivity, but we were unable to
856 generalize relationships between site geo-characteristics and sensitivity indices because of the
857 relatively low number of sites represented here (n=4 sites, 1 year each) and the confounding
858 number of differences between sites. A much larger population of snow measurement sites is
859 required in order to test relationships between sensitivity indices and site characteristics, and this
860 is an important avenue of future research. A successful example of relating climate
861 characteristics to sensitivity indices when many study sites and years are available can be found
862 in van Werkhoven et al. (2008).

863

864 While the Sobol' method is often considered the "baseline" method in global sensitivity analysis,
865 we note the limitation is that it comes at a relatively high computation cost (1 840 000
866 simulations across four sites and five error scenarios) and it may be prohibitive for many
867 modeling applications (e.g., for models of higher complexity and dimensionality). For context,
868 the typical time required for a single simulation was 1.4 seconds, resulting in a total
869 computational expense of 720 hours (30 days) across all scenarios. Examination of the
870 convergence rates indicated that most sensitivity indices stabilized after one-third of the
871 simulations completed, and hence the same results could have been found using significantly
872 fewer simulations (no figures shown). Ongoing research is developing new sensitivity analysis
873 methods that compare well to Sobol' but with reduced computational demands (e.g., FAST,
874 Cukier, 1973; method of Morris, 1991; DELSA, Rakovec et al., 2014), and is comparing how
875 different methods classify sensitive factors differently (Pappenberger et al., 2008; Tang et al.,

876 | 2007). We expect that detailed sensitivity analyses that concurrently consider uncertainty in
877 | forcings, parameters, and structure in a hydrologic model will be more feasible in the future with
878 | better computing resources and advances in sensitivity analysis methods.

879

880 | The question remains: “what can be done about forcing errors in hydrologic modeling?” First,
881 | the results suggest model-based hypothesis testing must account for uncertainties in forcing data.
882 | The results also [highlight](#) the need for continued research in constraining P uncertainty in snow-
883 | affected catchments. [Progress is being](#) achieved [with](#) advanced pathways for quantifying
884 | snowfall precipitation, such as NWP models (Rasmussen et al., 2011, 2014) [and through](#)
885 | [systematic intercomparisons of precipitation and snow gauges \(e.g., Solid Precipitation](#)
886 | [Intercomparison Experiment, <http://www.rap.ucar.edu/projects/SPICE/>\)](#). However, in a broader
887 | sense, the hydrologic community should [also](#) consider whether deterministic forcings (i.e., single
888 | time series for each forcing) are a reasonable practice for physically-based models, given the
889 | large uncertainties in both future (e.g., climate change) and historical data (especially in poorly
890 | monitored catchments) and the complexities of hydrologic systems (Gupta et al., 2008). We
891 | suggest that probabilistic model forcings (e.g., Clark and Slater, 2006), [which have a legacy in](#)
892 | [data assimilation methods](#) (e.g., precipitation uncertainty, Durand and Margulis, 2007), present
893 | one potential path forward where measures of forcing uncertainty can be explicitly included in
894 | the forcing datasets. The challenges are (1) to ensure statistical reliability in our understanding
895 | of forcing errors and (2) to assess how best to input probabilistic forcings into current model
896 | architectures.

897

898 | **6. Conclusions**

899 | Application of the Sobol’ sensitivity analysis framework across sites in contrasting snow
900 | climates reveals that forcing uncertainty can significantly impact model behavior in snow-
901 | affected catchments. Model output uncertainty due to forcings can be comparable to or larger
902 | than model uncertainty due to model structure. [Furthermore, this work demonstrates that](#)
903 | [sensitivity analysis can be applied to understand the role of specific error characteristics in model](#)
904 | [behavior](#). Key considerations in model sensitivity to forcing errors are the magnitudes of forcing
905 | errors and the outputs of interest. For the [physically-based snow](#) model tested, random errors in

906 forcings are generally less important than biases, and the [probability](#) distribution of biases is
907 relatively less important [to model sensitivity](#) than the magnitude of biases.

908

909 The analysis shows how forcing uncertainty might be included in a formal sensitivity analysis
910 framework through the introduction of new parameters that specify the characteristics of forcing
911 uncertainty. The framework could be extended to other physically based models and sensitivity
912 analysis methodologies, and could be used to quantify how uncertainties in model forcings and
913 parameters interact. [Based on this framework, it](#) would be interesting to assess the interplay
914 between co-existing uncertainties in forcing errors, model parameters, and model structure, and
915 to test how model sensitivity changes relative to all three sources of uncertainty.

916

917 **Acknowledgements**

918 M. Raleigh was supported by a post-doctoral fellowship in the Advanced Study Program at the
919 National Center for Atmospheric Research (NCAR). NCAR is sponsored by the National
920 Science Foundation. J. Lundquist was supported by NSF (EAR-838166 and EAR-1215771).
921 [The manuscript was improved thanks to thoughtful comments from F. Pianosi, J. Li, R. Essery,](#)
922 [R. Rosolem, A. Winstral, and one anonymous reviewer.](#) Thanks to M. Sturm, G. Shaver, S.
923 Bret-Harte, and E. Euskirchen for assistance with Innavaik Creek data, S. Morin for assistance
924 with Col de Porte data, D. Marks for assistance with Reynolds Mountain data, C. Landry for
925 assistance with Swamp Angel data, and E. Gutmann and P. Mendoza for feedback. For Innavaik
926 Creek data, we acknowledge U.S. Army Cold Regions Research and Engineering Laboratory, the
927 NSF Arctic Observatory Network (AON) Carbon, Water, and Energy Flux monitoring project
928 and the Marine Biological Laboratory, Woods Hole, and the University of Alaska, Fairbanks.
929 The experiment was improved thanks to conversations with D. Slater.

930

931 **References**

932 Archer, G. E. B., Saltelli, A. and Sobol, I. M.: Sensitivity measures, anova-like Techniques and
933 the use of bootstrap, J. Stat. Comput. Simul., 58(2), 99–120, doi:10.1080/00949659708811825,
934 1997.

- 935 Di Baldassarre, G. and Montanari, A.: Uncertainty in river discharge observations: a quantitative
936 analysis, *Hydrol. Earth Syst. Sci.*, 13(6), 913–921, doi:10.5194/hess-13-913-2009, 2009.
- 937 Bales, R. C., Molotch, N. P., Painter, T. H., Dettinger, M. D., Rice, R. and Dozier, J.: Mountain
938 hydrology of the western United States, *Water Resour. Res.*, 42, W08432,
939 doi:10.1029/2005WR004387, 2006.
- 940 Barnett, T. P., Pierce, D. W., Hidalgo, H. G., Bonfils, C., Santer, B. D., Das, T., Bala, G., Wood,
941 A. W., Nozawa, T., Mirin, A. A., Cayan, D. R. and Dettinger, M. D.: Human-induced changes in
942 the hydrology of the western United States, *Science (80-.)*, 319(5866), 1080–1083,
943 doi:10.1126/science.1152538, 2008.
- 944 Bastola, S., Murphy, C. and Sweeney, J.: The role of hydrological modelling uncertainties in
945 climate change impact assessments of Irish river catchments, *Adv. Water Resour.*, 34(5), 562–
946 576, doi:10.1016/j.advwatres.2011.01.008, 2011.
- 947 Benke, K. K., Lowell, K. E. and Hamilton, A. J.: Parameter uncertainty, sensitivity analysis and
948 prediction error in a water-balance hydrological model, *Math. Comput. Model.*, 47(11-12),
949 1134–1149, doi:10.1016/j.mcm.2007.05.017, 2008.
- 950 Beven, K. and Binley, A.: The future of distributed models: model calibration and uncertainty
951 prediction, *Hydrol. Process.*, 6(3), 279–298, doi:10.1002/hyp.3360060305, 1992.
- 952 Bohn, T. J., Livneh, B., Oyster, J. W., Running, S. W., Nijssen, B. and Lettenmaier, D. P.: Global
953 evaluation of MTCLIM and related algorithms for forcing of ecological and hydrological
954 models, *Agric. For. Meteorol.*, 176, 38–49, doi:10.1016/j.agrformet.2013.03.003, 2013.
- 955 Bolstad, P. V., Swift, L., Collins, F. and Régnière, J.: Measured and predicted air temperatures at
956 basin to regional scales in the southern Appalachian mountains, *Agric. For. Meteorol.*, 91(3-4),
957 161–176, doi:10.1016/S0168-1923(98)00076-8, 1998.
- 958 Bret-Harte, S., Euskirchen, E., Griffin, K. and Shaver, G.: Eddy Flux Measurements, Tussock
959 Station, Innvait Creek, Alaska - 2011, , Long Term Ecological Research Network,
960 doi:10.6073/pasta/44a62e0c6741b3bd93c0a33e7b677d90, 2011a.
- 961 Bret-Harte, S., Euskirchen, E. and Shaver, G.: Eddy Flux Measurements, Fen Station, Innvait
962 Creek, Alaska - 2011, , Long Term Ecological Research Network,
963 doi:10.6073/pasta/50e9676f29f44a8b6677f05f43268840, 2011b.
- 964 Bret-Harte, S., Euskirchen, E. and Shaver, G.: Eddy Flux Measurements, Ridge Station, Innvait
965 Creek, Alaska - 2011, , Long Term Ecological Research Network,
966 doi:10.6073/pasta/5d603c3628f53f494f08f895875765e8, 2011c.
- 967 Bret-Harte, S., Shaver, G. and Euskirchen, E.: Eddy Flux Measurements, Fen Station, Innvait
968 Creek, Alaska - 2010, , Long Term Ecological Research Network,
969 doi:10.6073/pasta/dde37e89dab096bea795f5b111786c8b, 2010a.

- 970 Bret-Harte, S., Shaver, G. and Euskirchen, E.: Eddy Flux Measurements, Ridge Station, Imnavait
971 Creek, Alaska - 2010, , Long Term Ecological Research Network,
972 doi:10.6073/pasta/fb047eaa2c78d4a3254bba8369e6cee5, 2010b.
- 973 Brun, E., David, P., Sudul, M. and Brunot, G.: A numerical model to simulate snow-cover
974 stratigraphy for operational avalanche forecasting, *J. Glaciol.*, 38(128), 13–22, 1992.
- 975 Burles, K. and Boon, S.: Snowmelt energy balance in a burned forest plot, Crowsnest Pass,
976 Alberta, Canada, *Hydrol. Process.*, doi:10.1002/hyp.8067, 2011.
- 977 Butts, M. B., Payne, J. T., Kristensen, M. and Madsen, H.: An evaluation of the impact of model
978 structure on hydrological modelling uncertainty for streamflow simulation, *J. Hydrol.*, 298(1-4),
979 242–266, doi:10.1016/j.jhydrol.2004.03.042, 2004.
- 980 Campolongo, F., Saltelli, A. and Cariboni, J.: From screening to quantitative sensitivity analysis.
981 A unified approach, *Comput. Phys. Commun.*, 182(4), 978–988, doi:10.1016/j.cpc.2010.12.039,
982 2011.
- 983 Cheng, F.-Y. and Georgakakos, K. P.: Statistical analysis of observed and simulated hourly
984 surface wind in the vicinity of the Panama Canal, *Int. J. Climatol.*, 31(5), 770–782,
985 doi:10.1002/joc.2123, 2011.
- 986 Christopher Frey, H. and Patil, S. R.: Identification and Review of Sensitivity Analysis Methods,
987 *Risk Anal.*, 22(3), 553–578, doi:10.1111/0272-4332.00039, 2002.
- 988 Chuanyan, Z., Zhongren, N. and Guodong, C.: Methods for modelling of temporal and spatial
989 distribution of air temperature at landscape scale in the southern Qilian mountains, China, *Ecol.*
990 *Modell.*, 189(1-2), 209–220, doi:10.1016/j.ecolmodel.2005.03.016, 2005.
- 991 Clark, M. P., Hendrikx, J., Slater, A. G., Kavetski, D., Anderson, B., Cullen, N. J., Kerr, T., Örn
992 Hreinsson, E. and Woods, R. A.: Representing spatial variability of snow water equivalent in
993 hydrologic and land-surface models: A review, *Water Resour. Res.*, 47(7),
994 doi:10.1029/2011WR010745, 2011a.
- 995 Clark, M. P., Kavetski, D. and Fenicia, F.: Pursuing the method of multiple working hypotheses
996 for hydrological modeling, *Water Resour. Res.*, 47(9), 1–16, doi:10.1029/2010WR009827,
997 2011b.
- 998 Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., Freer, J. E.,
999 Gutmann, E. D., Wood, A. W., Brekke, L. D., Arnold, J. R., Gochis, D. J. and Rasmussen, R. M.:
1000 A unified approach for process-based hydrologic modeling: 1. Modeling concept, *Water Resour.*
1001 *Res.*, 51, doi:10.1002/2015WR017198, 2015a.
- 1002 Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., Freer, J. E.,
1003 Gutmann, E. D., Wood, A. W., Gochis, D. J., Rasmussen, R. M., Tarboton, D. G., Mahat, V.,
1004 Flerchinger, G. N. and Marks, D. G.: A unified approach for process-based hydrologic modeling:

- 1005 2. Model implementation and case studies, *Water Resour. Res.*, 51,
1006 doi:10.1002/2015WR017200, 2015b.
- 1007 Clark, M. P. and Slater, A. G.: Probabilistic Quantitative Precipitation Estimation in Complex
1008 Terrain, *J. Hydrometeorol.*, 7(1), 3–22, doi:10.1175/JHM474.1, 2006.
- 1009 Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., Wagener, T.
1010 and Hay, L. E.: Framework for Understanding Structural Errors (FUSE): A modular framework
1011 to diagnose differences between hydrological models, *Water Resour. Res.*, 44(12),
1012 doi:10.1029/2007WR006735, 2008.
- 1013 Cukier, R. I.: Study of the sensitivity of coupled reaction systems to uncertainties in rate
1014 coefficients. I Theory, *J. Chem. Phys.*, 59(8), 3873, doi:10.1063/1.1680571, 1973.
- 1015 Dadic, R., Mott, R., Lehning, M., Carenzo, M., Anderson, B. and Mackintosh, A.: Sensitivity of
1016 turbulent fluxes to wind speed over snow surfaces in different climatic settings, *Adv. Water*
1017 *Resour.*, 55, 178–189, doi:10.1016/j.advwatres.2012.06.010, 2013.
- 1018 Dee, D. P.: Bias and data assimilation, *Q. J. R. Meteorol. Soc.*, 131(613), 3323–3343,
1019 doi:10.1256/qj.05.137, 2005.
- 1020 Deems, J. S., Painter, T. H., Barsugli, J. J., Belnap, J. and Udall, B.: Combined impacts of
1021 current and future dust deposition and regional warming on Colorado River Basin snow
1022 dynamics and hydrology, *Hydrol. Earth Syst. Sci.*, 17(11), 4401–4413, doi:10.5194/hess-17-
1023 4401-2013, 2013.
- 1024 Déry, S. and Stieglitz, M.: A note on surface humidity measurements in the cold Canadian
1025 environment, *Boundary-Layer Meteorol.*, 102, 491–497, doi:10.1023/A:1013890729982, 2002.
- 1026 Duan, Q., Sorooshian, S. and Gupta, V.: Effective and efficient global optimization for
1027 conceptual rainfall-runoff models, *Water Resour. Res.*, 28(4), 1015–1031,
1028 doi:10.1029/91WR02985, 1992.
- 1029 Durand, M. and Margulis, S. A.: Correcting first-order errors in snow water equivalent estimates
1030 using a multifrequency, multiscale radiometric data assimilation scheme, *J. Geophys. Res.*,
1031 112(D13), 1–15, doi:10.1029/2006JD008067, 2007.
- 1032 Durand, M. and Margulis, S. A.: Effects of uncertainty magnitude and accuracy on assimilation
1033 of multiscale measurements for snowpack characterization, *J. Geophys. Res.*, 113(D2), D02105,
1034 doi:10.1029/2007JD008662, 2008.
- 1035 Elsner, M. M., Gangopadhyay, S., Pruitt, T., Brekke, L. D., Mizukami, N. and Clark, M. P.: How
1036 Does the Choice of Distributed Meteorological Data Affect Hydrologic Model Calibration and
1037 Streamflow Simulations?, *J. Hydrometeorol.*, 15(4), 1384–1403, doi:10.1175/JHM-D-13-083.1,
1038 2014.

- 1039 Essery, R., Morin, S., Lejeune, Y. and B Ménard, C.: A comparison of 1701 snow models using
1040 observations from an alpine site, *Adv. Water Resour.*, 55, 131–148,
1041 doi:10.1016/j.advwatres.2012.07.013, 2013.
- 1042 Euskirchen, E. S., Bret-Harte, M. S., Scott, G. J., Edgar, C. and Shaver, G. R.: Seasonal patterns
1043 of carbon dioxide and water fluxes in three representative tundra ecosystems in northern Alaska,
1044 *Ecosphere*, 3(1), doi:10.1890/ES11-00202.1, 2012.
- 1045 Feld, S. I., Cristea, N. C. and Lundquist, J. D.: Representing atmospheric moisture content along
1046 mountain slopes: Examination using distributed sensors in the Sierra Nevada, California, *Water*
1047 *Resour. Res.*, 49, doi:10.1002/wrcr.20318, 2013.
- 1048 Flerchinger, G. N., Xaio, W., Marks, D., Sauer, T. J. and Yu, Q.: Comparison of algorithms for
1049 incoming atmospheric long-wave radiation, *Water Resour. Res.*, 45(3), 1–13,
1050 doi:10.1029/2008WR007394, 2009.
- 1051 Flint, A. L. and Childs, S. W.: Calculation of solar radiation in mountainous terrain, *Agric. For.*
1052 *Meteorol.*, 40(3), 233–249, doi:10.1016/0168-1923(87)90061-X, 1987.
- 1053 Foglia, L., Hill, M. C., Mehl, S. W. and Burlando, P.: Sensitivity analysis, calibration, and
1054 testing of a distributed hydrological model using error-based weighting and one objective
1055 function, *Water Resour. Res.*, 45(6), doi:10.1029/2008WR007255, 2009.
- 1056 Foscarini, F., Bellocchi, G., Confalonieri, R., Savini, C. and Van den Eede, G.: Sensitivity
1057 analysis in fuzzy systems: Integration of SimLab and DANA, *Environ. Model. Softw.*, 25(10),
1058 1256–1260, doi:10.1016/j.envsoft.2010.03.024, 2010.
- 1059 Fridley, J. D.: Downscaling Climate over Complex Terrain: High Finescale (<1000 m) Spatial
1060 Variation of Near-Ground Temperatures in a Montane Forested Landscape (Great Smoky
1061 Mountains)*, *J. Appl. Meteorol. Climatol.*, 48(5), 1033–1049, doi:10.1175/2008JAMC2084.1,
1062 2009.
- 1063 Georgakakos, K., Seo, D., Gupta, H., Schaake, J. and Butts, M.: Towards the characterization of
1064 streamflow simulation uncertainty through multimodel ensembles, *J. Hydrol.*, 298(1-4), 222–
1065 241, doi:10.1016/j.jhydrol.2004.03.037, 2004.
- 1066 Goodison, B., Louie, P. and Yang, D.: WMO solid precipitation measurement intercomparison:
1067 Final report, in *Instrum. Obs. Methods Rep. 67*, vol. 67, p. 211, World Meteorol. Organ.,
1068 Geneva, Switzerland., 1998.
- 1069 Griffin, K., Bret-Harte, S., Shaver, G. and Euskirchen, E.: Eddy Flux Measurements, Tussock
1070 Station, Innvait Creek, Alaska - 2010, , Long Term Ecological Research Network,
1071 doi:10.6073/pasta/7bba82256e0f5d9ec3d2bc9c25ab9bcf, 2010.

- 1072 Guan, B., Molotch, N. P., Waliser, D. E., Jepsen, S. M., Painter, T. H. and Dozier, J.: Snow
1073 water equivalent in the Sierra Nevada: Blending snow sensor observations with snowmelt model
1074 simulations, *Water Resour. Res.*, 49(8), 5029–5046, doi:10.1002/wrcr.20387, 2013.
- 1075 Guan, H., Wilson, J. L. and Makhnin, O.: Geostatistical Mapping of Mountain Precipitation
1076 Incorporating Autosearched Effects of Terrain and Climatic Characteristics, *J. Hydrometeorol.*,
1077 6(6), 1018–1031, doi:10.1175/JHM448.1, 2005.
- 1078 Gupta, H. V., Wagener, T. and Liu, Y.: Reconciling theory with observations: elements of a
1079 diagnostic approach to model evaluation, *Hydrol. Process.*, 22, 3802–3813,
1080 doi:10.1002/hyp.6989, 2008.
- 1081 Hasenauer, H., Merganicova, K., Petritsch, R., Pietsch, S. A. and Thornton, P. E.: Validating
1082 daily climate interpolations over complex terrain in Austria, *Agric. For. Meteorol.*, 119, 87–107,
1083 doi:10.1016/S0168-1923(03)00114-X, 2003.
- 1084 He, M., Hogue, T. S., Franz, K. J., Margulis, S. A. and Vrugt, J. A.: Characterizing parameter
1085 sensitivity and uncertainty for a snow model across hydroclimatic regimes, *Adv. Water Resour.*,
1086 34(1), 114–127, doi:10.1016/j.advwatres.2010.10.002, 2011a.
- 1087 He, M., Hogue, T. S., Franz, K. J., Margulis, S. A. and Vrugt, J. A.: Corruption of parameter
1088 behavior and regionalization by model and forcing data errors: A Bayesian example using the
1089 SNOW17 model, *Water Resour. Res.*, 47(7), 1–17, doi:10.1029/2010WR009753, 2011b.
- 1090 Herman, J. D., Kollat, J. B., Reed, P. M. and Wagener, T.: Technical Note: Method of Morris
1091 effectively reduces the computational demands of global sensitivity analysis for distributed
1092 watershed models, *Hydrol. Earth Syst. Sci.*, 17(7), 2893–2903, doi:10.5194/hess-17-2893-2013,
1093 2013.
- 1094 Herrero, J. and Polo, M. J.: Parameterization of atmospheric longwave emissivity in a
1095 mountainous site for all sky conditions, *Hydrol. Earth Syst. Sci.*, 16(9), 3139–3147,
1096 doi:10.5194/hess-16-3139-2012, 2012.
- 1097 Hiemstra, C. A., Liston, G. E. and Reiners, W. A.: Observing, modelling, and validating snow
1098 redistribution by wind in a Wyoming upper treeline landscape, *Ecol. Modell.*, 197(1-2), 35–51,
1099 doi:10.1016/j.ecolmodel.2006.03.005, 2006.
- 1100 Hutchinson, M. F., McKenney, D. W., Lawrence, K., Pedlar, J. H., Hopkinson, R. F., Milewska,
1101 E. and Papadopol, P.: Development and Testing of Canada-Wide Interpolated Spatial Models of
1102 Daily Minimum–Maximum Temperature and Precipitation for 1961–2003, *J. Appl. Meteorol.*
1103 *Climatol.*, 48(4), 725–741, doi:10.1175/2008JAMC1979.1, 2009.
- 1104 Huwald, H., Higgins, C. W., Boldi, M.-O., Bou-Zeid, E., Lehning, M. and Parlange, M. B.:
1105 Albedo effect on radiative errors in air temperature measurements, *Water Resour. Res.*, 45(8), 1–
1106 13, doi:10.1029/2008WR007600, 2009.

- 1107 Jackson, C., Xia, Y., Sen, M. K. and Stoffa, P. L.: Optimal parameter and uncertainty estimation
1108 of a land surface model: A case study using data from Cabauw, Netherlands, *J. Geophys. Res.*,
1109 108(D18), 4583, doi:10.1029/2002JD002991, 2003.
- 1110 Jansen, M. J. W.: Analysis of variance designs for model output, *Comput. Phys. Commun.*,
1111 117(1-2), 35–43, doi:10.1016/S0010-4655(98)00154-4, 1999.
- 1112 Jepsen, S. M., Molotch, N. P., Williams, M. W., Rittger, K. E. and Sickman, J. O.: Interannual
1113 variability of snowmelt in the Sierra Nevada and Rocky Mountains, United States: Examples
1114 from two alpine watersheds, *Water Resour. Res.*, 48(2), 1–15, doi:10.1029/2011WR011006,
1115 2012.
- 1116 Jiménez, P. A., Dudhia, J. and Navarro, J.: On the surface wind speed probability density
1117 function over complex terrain, *Geophys. Res. Lett.*, 38(22), doi:10.1029/2011GL049669, 2011.
- 1118 Jing, X. and Cess, R. D.: Comparison of atmospheric clear-sky shortwave radiation models to
1119 collocated satellite and surface measurements in Canada, *J. Geophys. Res.*, 103(D22), 28817,
1120 doi:10.1029/1998JD200012, 1998.
- 1121 Jordan, R.: A One-Dimensional Temperature Model for a Snow Cover: Technical
1122 Documentation for SNTHERM.89, p. 58, Special Report 91-16, US Army CRREL, Hanover,
1123 NH, USA., 1991.
- 1124 Kane, D. L., Hinzman, L. D., Benson, C. S. and Liston, G. E.: Snow hydrology of a headwater
1125 Arctic basin: 1. Physical measurements and process studies, *Water Resour. Res.*, 27(6), 1099–
1126 1109, doi:10.1029/91WR00262, 1991.
- 1127 Kavetski, D., Franks, S. W. and Kuczera, G.: Confronting input uncertainty in environmental
1128 modelling, in *Calibration of Watershed Models*, edited by Q. Duan, H. V. Gupta, S. Sorooshian,
1129 A. N. Rousseau, and R. Turcotte, pp. 49–68, American Geophysical Union, Washington, D.C.,
1130 2002.
- 1131 Kavetski, D., Kuczera, G. and Franks, S. W.: Bayesian analysis of input uncertainty in
1132 hydrological modeling: 1. Theory, *Water Resour. Res.*, 42(3), W03407,
1133 doi:10.1029/2005WR004368, 2006a.
- 1134 Kavetski, D., Kuczera, G. and Franks, S. W.: Bayesian analysis of input uncertainty in
1135 hydrological modeling: 2. Application, *Water Resour. Res.*, 42(3), W03408,
1136 doi:10.1029/2005WR004376, 2006b.
- 1137 Koivusalo, H. and Heikinheimo, M.: Surface energy exchange over a boreal snowpack:
1138 comparison of two snow energy balance models, *Hydrol. Process.*, 13(14-15), 2395–2408,
1139 doi:10.1002/(SICI)1099-1085(199910)13:14/15<2395::AID-HYP864>3.0.CO;2-G, 1999.

- 1140 Kuczera, G. and Parent, E.: Monte Carlo assessment of parameter uncertainty in conceptual
1141 catchment models: the Metropolis algorithm, *J. Hydrol.*, 211(1-4), 69–85, doi:10.1016/S0022-
1142 1694(98)00198-X, 1998.
- 1143 Kuczera, G., Renard, B., Thyer, M. and Kavetski, D.: There are no hydrological monsters, just
1144 models and observations with large uncertainties!, *Hydrol. Sci. J.*, 55(6), 980–991,
1145 doi:10.1080/02626667.2010.504677, 2010.
- 1146 Landry, C. C., Buck, K. A., Raleigh, M. S. and Clark, M. P.: Mountain system monitoring at
1147 Senator Beck Basin, San Juan Mountains, Colorado: A new integrative data source to develop
1148 and evaluate models of snow and hydrologic processes, *Water Resour. Res.*, 50,
1149 doi:10.1002/2013WR013711, 2014.
- 1150 Lapo, K. E., Hinkelman, L. M., Raleigh, M. S. and Lundquist, J. D.: Impact of errors in the
1151 downwelling irradiances on simulations of snow water equivalent, snow surface temperature,
1152 and the snow energy balance, *Water Resour. Res.*, 6(4), doi:10.1002/2014WR016259, 2015.
- 1153 Lapp, S., Byrne, J., Townshend, I. and Kienzle, S.: Climate warming impacts on snowpack
1154 accumulation in an alpine watershed, *Int. J. Climatol.*, 25(4), 521–536, doi:10.1002/joc.1140,
1155 2005.
- 1156 Leavesley, G. H.: Modeling the effects of climate change on water resources - a review, *Clim.*
1157 *Change*, 28(1-2), 159–177, doi:10.1007/BF01094105, 1994.
- 1158 Lehning, M., Völksch, I., Gustafsson, D., Nguyen, T. A., Stähli, M. and Zappa, M.: ALPINE3D:
1159 a detailed model of mountain surface processes and its application to snow hydrology, *Hydrol.*
1160 *Process.*, 20, 2111–2128, doi:10.1002/hyp.6204, 2006.
- 1161 Li, J., Duan, Q. Y., Gong, W., Ye, A., Dai, Y., Miao, C., Di, Z., Tong, C. and Sun, Y.: Assessing
1162 parameter importance of the Common Land Model based on qualitative and quantitative
1163 sensitivity analysis, *Hydrol. Earth Syst. Sci.*, 17(8), 3279–3293, doi:10.5194/hess-17-3279-2013,
1164 2013.
- 1165 Liston, G. E.: Representing Subgrid Snow Cover Heterogeneities in Regional and Global
1166 Models, *J. Clim.*, 17(6), 1381–1397, doi:10.1175/1520-
1167 0442(2004)017<1381:RSSCHI>2.0.CO;2, 2004.
- 1168 Liston, G. E. and Elder, K.: A Meteorological Distribution System for High-Resolution
1169 Terrestrial Modeling (MicroMet), *J. Hydrometeorol.*, 7(2), 217–234, doi:10.1175/JHM486.1,
1170 2006.
- 1171 Liu, Y. and Gupta, H. V.: Uncertainty in hydrologic modeling: Toward an integrated data
1172 assimilation framework, *Water Resour. Res.*, 43(7), doi:10.1029/2006WR005756, 2007.

- 1173 Luce, C. H., Tarboton, D. G. and Cooley, K. R.: The influence of the spatial distribution of snow
1174 on basin-averaged snowmelt, *Hydrol. Process.*, 12(10-11), 1671–1683, doi:10.1002/(SICI)1099-
1175 1085(199808/09)12:10/11<1671::AID-HYP688>3.0.CO;2-N, 1998.
- 1176 Lundquist, J. D. and Cayan, D. R.: Surface temperature patterns in complex terrain: Daily
1177 variations and long-term change in the central Sierra Nevada, California, *J. Geophys. Res.*, 112,
1178 D11124, doi:10.1029/2006JD007561, 2007.
- 1179 Lundquist, J. D., Wayand, N. E., Massmann, A., Clark, M. P., Lott, F. and Cristea, N. C.:
1180 Diagnosis of insidious data disasters, *Water Resour. Res.*, 51, doi:10.1002/2014WR016585,
1181 2015.
- 1182 Luo, W., Taylor, M. C. and Parker, S. R.: A comparison of spatial interpolation methods to
1183 estimate continuous wind speed surfaces using irregularly distributed data from England and
1184 Wales, *Int. J. Climatol.*, 28(7), 947–959, doi:10.1002/joc.1583, 2008.
- 1185 Magnusson, J., Wever, N., Essery, R., Helbig, N., Winstral, A. and Jonas, T.: Evaluating snow
1186 models with varying process representations for hydrological applications, *Water Resour. Res.*,
1187 51, doi:10.1002/2014WR016498, 2015.
- 1188 Mahat, V. and Tarboton, D. G.: Canopy radiation transmission for an energy balance snowmelt
1189 model, *Water Resour. Res.*, 48(1), 1–16, doi:10.1029/2011WR010438, 2012.
- 1190 Mardikis, M. G., Kalivas, D. P. and Kollias, V. J.: Comparison of Interpolation Methods for the
1191 Prediction of Reference Evapotranspiration—An Application in Greece, *Water Resour. Manag.*,
1192 19(3), 251–278, doi:10.1007/s11269-005-3179-2, 2005.
- 1193 Marks, D. and Dozier, J.: Climate and energy exchange at the snow surface in the Alpine Region
1194 of the Sierra Nevada: 2. Snow cover energy balance, *Water Resour. Res.*, 28(11), 3043–3054,
1195 doi:10.1029/92WR01483, 1992.
- 1196 Marks, D., Dozier, J. and Davis, R. E.: Climate and energy exchange at the snow surface in the
1197 Alpine Region of the Sierra Nevada: 1. Meteorological measurements and monitoring, *Water*
1198 *Resour. Res.*, 28(11), 3029–3042, doi:10.1029/92WR01482, 1992.
- 1199 Matott, L. S., Babendreier, J. E. and Purucker, S. T.: Evaluating uncertainty in integrated
1200 environmental models: A review of concepts and tools, *Water Resour. Res.*, 45(6),
1201 doi:10.1029/2008WR007301, 2009.
- 1202 Meyer, J. D. D., Jin, J. and Wang, S.-Y.: Systematic Patterns of the Inconsistency between Snow
1203 Water Equivalent and Accumulated Precipitation as Reported by the Snowpack Telemetry
1204 Network, *J. Hydrometeorol.*, 13(6), 1970–1976, doi:10.1175/JHM-D-12-066.1, 2012.
- 1205 Mizukami, N., Clark, M. P., Slater, A. G., Brekke, L. D., Elsner, M. M., Arnold, J. R. and
1206 Gangopadhyay, S.: Hydrologic Implications of Different Large-Scale Meteorological Model

- 1207 Forcing Datasets in Mountainous Regions, *J. Hydrometeorol.*, 15(1), 474–488,
1208 doi:10.1175/JHM-D-13-036.1, 2014.
- 1209 Morin, S., Lejeune, Y., Lesaffre, B., Panel, J.-M., Poncet, D., David, P. and Sudul, M.: An 18-yr
1210 long (1993–2011) snow and meteorological dataset from a mid-altitude mountain site (Col de
1211 Porte, France, 1325 m alt.) for driving and evaluating snowpack models, *Earth Syst. Sci. Data*,
1212 4(1), 13–21, doi:10.5194/essd-4-13-2012, 2012.
- 1213 Morris, M. D.: Factorial Sampling Plans for Preliminary Computational Experiments,
1214 *Technometrics*, 33(2), 161–174, doi:10.1080/00401706.1991.10484804, 1991.
- 1215 Mott, R. and Lehning, M.: Meteorological Modeling of Very High-Resolution Wind Fields and
1216 Snow Deposition for Mountains, *J. Hydrometeorol.*, 11(4), 934–949,
1217 doi:10.1175/2010JHM1216.1, 2010.
- 1218 Niemelä, S., Räisänen, P. and Savijärvi, H.: Comparison of surface radiative flux
1219 parameterizations: Part I. Longwave radiation, *Atmos. Res.*, 58(1), 1–18, doi:10.1016/S0169-
1220 8095(01)00084-9, 2001a.
- 1221 Niemelä, S., Räisänen, P. and Savijärvi, H.: Comparison of surface radiative flux
1222 parameterizations: Part II. Shortwave radiation, *Atmos. Res.*, 58(2), 141–154,
1223 doi:10.1016/S0169-8095(01)00085-0, 2001b.
- 1224 Nossent, J., Elsen, P. and Bauwens, W.: Sobol’ sensitivity analysis of a complex environmental
1225 model, *Environ. Model. Softw.*, 26(12), 1515–1525, doi:10.1016/j.envsoft.2011.08.010, 2011.
- 1226 Oudin, L., Perrin, C., Mathevet, T., Andréassian, V. and Michel, C.: Impact of biased and
1227 randomly corrupted inputs on the efficiency and the parameters of watershed models, *J. Hydrol.*,
1228 320(1-2), 62–83, doi:10.1016/j.jhydrol.2005.07.016, 2006.
- 1229 Pappenberger, F. and Beven, K. J.: Ignorance is bliss: Or seven reasons not to use uncertainty
1230 analysis, *Water Resour. Res.*, 42(W05302), doi:10.1029/2005WR004820, 2006.
- 1231 Pappenberger, F., Beven, K. J., Ratto, M. and Matgen, P.: Multi-method global sensitivity
1232 analysis of flood inundation models, *Adv. Water Resour.*, 31(1), 1–14,
1233 doi:10.1016/j.advwatres.2007.04.009, 2008.
- 1234 Phillips, D. and Marks, D.: Spatial uncertainty analysis: propagation of interpolation errors in
1235 spatially distributed models, *Ecol. Modell.*, 91(1-3), 213–229, doi:10.1016/0304-3800(95)00191-
1236 3, 1996.
- 1237 Rakovec, O., Hill, M. C., Clark, M. P., Weerts, A. H., Teuling, A. J. and Uijlenhoet, R.:
1238 Distributed Evaluation of Local Sensitivity Analysis (DELSA), with application to hydrologic
1239 models, *Water Resour. Res.*, 50, doi:10.1002/2013WR014063, 2014.

- 1240 Raleigh, M. S.: Quantification of uncertainties in snow accumulation, snowmelt, and snow
1241 disappearance dates, University of Washington., 2013.
- 1242 Raleigh, M. S. and Lundquist, J. D.: Comparing and combining SWE estimates from the SNOW-
1243 17 model using PRISM and SWE reconstruction, *Water Resour. Res.*, 48(1),
1244 doi:10.1029/2011WR010542, 2012.
- 1245 Rasmussen, R., Baker, B., Kochendorfer, J., Meyers, T., Landolt, S., Fischer, A. P., Black, J.,
1246 Thériault, J. M., Kucera, P., Gochis, D., Smith, C., Nitu, R., Hall, M., Ikeda, K. and Gutmann,
1247 E.: How Well Are We Measuring Snow: The NOAA/FAA/NCAR Winter Precipitation Test Bed,
1248 *Bull. Am. Meteorol. Soc.*, 93(6), 811–829, doi:10.1175/BAMS-D-11-00052.1, 2012.
- 1249 Rasmussen, R., Ikeda, K., Liu, C., Gochis, D., Clark, M., Dai, A., Gutmann, E., Dudhia, J.,
1250 Chen, F., Barlage, M., Yates, D. and Zhang, G.: Climate Change Impacts on the Water Balance
1251 of the Colorado Headwaters: High-Resolution Regional Climate Model Simulations, *J.*
1252 *Hydrometeorol.*, 15(3), 1091–1116, doi:10.1175/JHM-D-13-0118.1, 2014.
- 1253 Rasmussen, R., Liu, C., Ikeda, K., Gochis, D., Yates, D., Chen, F., Tewari, M., Barlage, M.,
1254 Dudhia, J., Yu, W., Miller, K., Arsenault, K., Grubišić, V., Thompson, G. and Gutmann, E.:
1255 High-Resolution Coupled Climate Runoff Simulations of Seasonal Snowfall over Colorado: A
1256 Process Study of Current and Warmer Climate, *J. Clim.*, 24(12), 3015–3048,
1257 doi:10.1175/2010JCLI3985.1, 2011.
- 1258 Razavi, S. and Gupta, H. V.: What do we mean by sensitivity analysis? The need for
1259 comprehensive characterization of “Global” sensitivity in Earth and Environmental Systems
1260 Models, *Water Resour. Res.*, 51, doi:10.1002/2014WR016527, 2015.
- 1261 Reba, M. L., Marks, D., Seyfried, M., Winstral, A., Kumar, M. and Flerchinger, G.: A long-term
1262 data set for hydrologic modeling in a snow-dominated mountain catchment, *Water Resour. Res.*,
1263 47(7), W07702, doi:10.1029/2010WR010030, 2011.
- 1264 Refsgaard, J. C., van der Sluijs, J. P., Brown, J. and van der Keur, P.: A framework for dealing
1265 with uncertainty due to model structure error, *Adv. Water Resour.*, 29(11), 1586–1597,
1266 doi:10.1016/j.advwatres.2005.11.013, 2006.
- 1267 Rosero, E., Yang, Z.-L., Wagener, T., Gulden, L. E., Yatheendradas, S. and Niu, G.-Y.:
1268 Quantifying parameter sensitivity, interaction, and transferability in hydrologically enhanced
1269 versions of the Noah land surface model over transition zones during the warm season, *J.*
1270 *Geophys. Res.*, 115, D03106, doi:10.1029/2009JD012035, 2010.
- 1271 Rosolem, R., Gupta, H. V., Shuttleworth, W. J., Zeng, X. and de Gonçalves, L. G. G.: A fully
1272 multiple-criteria implementation of the Sobol’ method for parameter sensitivity analysis, *J.*
1273 *Geophys. Res. Atmos.*, 117, D07103, doi:10.1029/2011JD016355, 2012.
- 1274 Saltelli, A.: Sensitivity analysis: Could better methods be used?, *J. Geophys. Res.*, 104(D3),
1275 3789, doi:10.1029/1998JD100042, 1999.

- 1276 Saltelli, A. and Annoni, P.: How to avoid a perfunctory sensitivity analysis, *Environ. Model.*
1277 *Softw.*, 25(12), 1508–1517, doi:10.1016/j.envsoft.2010.04.012, 2010.
- 1278 Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M. and Tarantola, S.: Variance based
1279 sensitivity analysis of model output. Design and estimator for the total sensitivity index, *Comput.*
1280 *Phys. Commun.*, 181(2), 259–270, doi:10.1016/j.cpc.2009.09.018, 2010.
- 1281 Sauter, T. and Obleitner, F.: Assessment of the uncertainty of snowpack simulations based on
1282 variance decomposition, *Geosci. Model Dev. Discuss.*, 8(3), 2807–2845, doi:10.5194/gmdd-8-
1283 2807-2015, 2015.
- 1284 Schmucki, E., Marty, C., Fierz, C. and Lehning, M.: Evaluation of modelled snow depth and
1285 snow water equivalent at three contrasting sites in Switzerland using SNOWPACK simulations
1286 driven by different meteorological data input, *Cold Reg. Sci. Technol.*, 99, 27–37,
1287 doi:10.1016/j.coldregions.2013.12.004, 2014.
- 1288 Serreze, M. C., Clark, M. P., Armstrong, R. L., McGinnis, D. A. and Pulwarty, R. S.:
1289 Characteristics of the western United States snowpack from snowpack telemetry (SNOTEL)
1290 data, *Water Resour. Res.*, 35(7), 2145–2160, doi:10.1029/1999WR900090, 1999.
- 1291 Shamir, E. and Georgakakos, K. P.: Distributed snow accumulation and ablation modeling in the
1292 American River basin, *Adv. Water Resour.*, 29, 558–570, doi:10.1016/j.advwatres.2005.06.010,
1293 2006.
- 1294 Skiles, S. M., Painter, T. H., Deems, J. S., Bryant, A. C. and Landry, C. C.: Dust radiative
1295 forcing in snow of the Upper Colorado River Basin: 2. Interannual variability in radiative forcing
1296 and snowmelt rates, *Water Resour. Res.*, 48(7), doi:10.1029/2012WR011986, 2012.
- 1297 Slater, A. G. and Clark, M. P.: Snow Data Assimilation via an Ensemble Kalman Filter, *J.*
1298 *Hydrometeorol.*, 7(3), 478–493, doi:10.1175/JHM505.1, 2006.
- 1299 Slater, A. G., Schlosser, C. A., Desborough, C. E., Pitman, A. J., Henderson-Sellers, A., Robock,
1300 A., Vinnikov, K. Y., Entin, J., Mitchell, K., Chen, F., Boone, A., Etchevers, P., Habets, F.,
1301 Noilhan, J., Braden, H., Cox, P. M., de Rosnay, P., Dickinson, R. E., Yang, Z.-L., Dai, Y.-J.,
1302 Zeng, Q., Duan, Q., Koren, V., Schaake, S., Gedney, N., Gusev, Y. M., Nasonova, O. N., Kim,
1303 J., Kowalczyk, E. A., Shmakin, A. B., Smirnova, T. G., Verseghy, D., Wetzol, P. and Xue, Y.:
1304 The Representation of Snow in Land Surface Schemes: Results from PILPS 2(d), *J.*
1305 *Hydrometeorol.*, 2(1), 7–25, doi:10.1175/1525-7541(2001)002<0007:TROSIL>2.0.CO;2, 2001.
- 1306 Smith, P. J., Beven, K. J. and Tawn, J. A.: Detection of structural inadequacy in process-based
1307 hydrological models: A particle-filtering approach, *Water Resour. Res.*, 44(1), W01410,
1308 doi:10.1029/2006WR005205, 2008.
- 1309 Sobol', I.: On sensitivity estimation for nonlinear mathematical models, *Mat. Model.*, 2(1), 112–
1310 118, 1990.

- 1311 Song, X., Zhang, J., Zhan, C., Xuan, Y., Ye, M. and Xu, C.: Global sensitivity analysis in
1312 hydrological modeling: Review of concepts, methods, theoretical framework, and applications, *J.*
1313 *Hydrol.*, 523(225), 739–757, doi:10.1016/j.jhydrol.2015.02.013, 2015.
- 1314 Spank, U., Schwärzel, K., Renner, M., Moderow, U. and Bernhofer, C.: Effects of measurement
1315 uncertainties of meteorological data on estimates of site water balance components, *J. Hydrol.*,
1316 492, 176–189, doi:10.1016/j.jhydrol.2013.03.047, 2013.
- 1317 Sturm, M., Holmgren, J. and Liston, G. E.: A Seasonal Snow Cover Classification System for
1318 Local to Global Applications, *J. Clim.*, 8(5), 1261–1283, doi:10.1175/1520-
1319 0442(1995)008<1261:ASSCCS>2.0.CO;2, 1995.
- 1320 Sturm, M. and Wagner, A. M.: Using repeated patterns in snow distribution modeling: An Arctic
1321 example, *Water Resour. Res.*, 46(12), 1–15, doi:10.1029/2010WR009434, 2010.
- 1322 Tang, Y., Reed, P., Wagener, T. and van Werkhoven, K.: Comparing sensitivity analysis
1323 methods to advance lumped watershed model identification and evaluation, *Hydrol. Earth Syst.*
1324 *Sci.*, 11(2), 793–817, doi:10.5194/hess-11-793-2007, 2007.
- 1325 Tarboton, D. and Luce, C.: Utah Energy Balance Snow Accumulation and Melt Model (UEB), in
1326 Computer model technical description users guide, Utah Water Res. Lab. and USDA For. Serv.
1327 Intermt. Res. Station, p. 64, Logan, UT., 1996.
- 1328 Thornton, P. E., Hasenauer, H. and White, M. A.: Simultaneous estimation of daily solar
1329 radiation and humidity from observed temperature and precipitation: an application over
1330 complex terrain in Austria, *Agric. For. Meteorol.*, 104(4), 255–271, doi:10.1016/S0168-
1331 1923(00)00170-2, 2000.
- 1332 Touhami, H. Ben, Lardy, R., Barra, V. and Bellocchi, G.: Screening parameters in the Pasture
1333 Simulation model using the Morris method, *Ecol. Modell.*, 266(1), 42–57,
1334 doi:10.1016/j.ecolmodel.2013.07.005, 2013.
- 1335 Trujillo, E. and Molotch, N. P.: Snowpack regimes of the Western United States, *Water Resour.*
1336 *Res.*, 50, doi:10.1002/2013WR014753, 2014.
- 1337 Vrugt, J. A., Braak, C. J. F., Gupta, H. V. and Robinson, B. A.: Equifinality of formal (DREAM)
1338 and informal (GLUE) Bayesian approaches in hydrologic modeling?, *Stoch. Environ. Res. Risk*
1339 *Assess.*, 23(7), 1011–1026, doi:10.1007/s00477-008-0274-y, 2008a.
- 1340 Vrugt, J. A., ter Braak, C. J. F., Clark, M. P., Hyman, J. M. and Robinson, B. A.: Treatment of
1341 input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte
1342 Carlo simulation, *Water Resour. Res.*, 44(12), doi:10.1029/2007WR006720, 2008b.
- 1343 Vrugt, J. A., Diks, C. G. H., Gupta, H. V., Bouten, W. and Verstraten, J. M.: Improved treatment
1344 of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data
1345 assimilation, *Water Resour. Res.*, 41(1), W01017, doi:10.1029/2004WR003059, 2005.

- 1346 Vrugt, J. A., Gupta, H. V., Bastidas, L. A., Bouten, W. and Sorooshian, S.: Effective and
1347 efficient algorithm for multiobjective optimization of hydrologic models, *Water Resour. Res.*,
1348 39(8), doi:10.1029/2002WR001746, 2003a.
- 1349 Vrugt, J. A., Gupta, H. V., Bouten, W. and Sorooshian, S.: A Shuffled Complex Evolution
1350 Metropolis algorithm for optimization and uncertainty assessment of hydrologic model
1351 parameters, *Water Resour. Res.*, 39(8), doi:10.1029/2002WR001642, 2003b.
- 1352 Wayand, N. E., Hamlet, A. F., Hughes, M., Feld, S. I. and Lundquist, J. D.: Intercomparison of
1353 Meteorological Forcing Data from Empirical and Mesoscale Model Sources in the N.F.
1354 American River Basin in northern Sierra Nevada, California, *J. Hydrometeorol.*, 14(3), 677–699,
1355 doi:10.1175/JHM-D-12-0102.1, 2013.
- 1356 Van Werkhoven, K., Wagener, T., Reed, P. and Tang, Y.: Characterization of watershed model
1357 behavior across a hydroclimatic gradient, *Water Resour. Res.*, 44(1), W01429,
1358 doi:10.1029/2007WR006271, 2008.
- 1359 Winstral, A. and Marks, D.: Simulating wind fields and snow redistribution using terrain-based
1360 parameters to model snow accumulation and melt over a semi-arid mountain catchment, *Hydrol.*
1361 *Process.*, 16(18), 3585–3603, doi:10.1002/hyp.1238, 2002.
- 1362 Winstral, A., Marks, D. and Gurney, R.: An efficient method for distributing wind speeds over
1363 heterogeneous terrain, *Hydrol. Process.*, 23, 2526–2535, doi:10.1002/hyp.7141, 2009.
- 1364 Winstral, A., Marks, D. and Gurney, R.: Simulating wind-affected snow accumulations at
1365 catchment to basin scales, *Adv. Water Resour.*, 55, 64–79, doi:10.1016/j.advwatres.2012.08.011,
1366 2013.
- 1367 Xia, Y., Yang, Z.-L., Stoffa, P. L. and Sen, M. K.: Using different hydrological variables to
1368 assess the impacts of atmospheric forcing errors on optimization and uncertainty analysis of the
1369 CHASM surface model at a cold catchment, *J. Geophys. Res.*, 110(D1), D01101,
1370 doi:10.1029/2004JD005130, 2005.
- 1371 Yang, D., Kane, D. L., Hinzman, L. D., Goodison, B. E., Metcalfe, J. R., Louie, P. Y. T.,
1372 Leavesley, G. H., Emerson, D. G. and Hanson, C. L.: An evaluation of the Wyoming Gauge
1373 System for snowfall measurement, *Water Resour. Res.*, 36(9), 2665–2677,
1374 doi:10.1029/2000WR900158, 2000.
- 1375 Yilmaz, K. K., Gupta, H. V. and Wagener, T.: A process-based diagnostic approach to model
1376 evaluation: Application to the NWS distributed hydrologic model, *Water Resour. Res.*, 44,
1377 W09417, doi:10.1029/2007WR006716, 2008.
- 1378 You, J., Tarboton, D. G. and Luce, C. H.: Modeling the snow surface temperature with a one-
1379 layer energy balance snowmelt model, *Hydrol. Earth Syst. Sci. Discuss.*, 10(12), 15071–15118,
1380 doi:10.5194/hessd-10-15071-2013, 2013.

1381 Zuzel, J. F. and Cox, L. M.: Relative importance of meteorological variables in snowmelt, Water
1382 Resour. Res., 11(1), 174–176, doi:10.1029/WR011i001p00174, 1975.

1383

1384 **7. Tables**1385 **Table 1** Basic characteristics of the snow study sites, ordered from left-to-right by increasing elevation.

<u>Site Name</u>	<u>Innavait Creek</u>	<u>Col de Porte</u>	<u>Reynolds Mountain East (sheltered site)</u>	<u>Swamp Angel Study Plot</u>
<u>Site ID</u>	<u>IC</u>	<u>CDP</u>	<u>RME</u>	<u>SASP</u>
<u>Location</u>	<u>Alaska, USA</u>	<u>Rhône-Alpes, France</u>	<u>Idaho, USA</u>	<u>Colorado, USA</u>
<u>Latitude (N)</u>	<u>68.62</u>	<u>45.30</u>	<u>43.07</u>	<u>37.91</u>
<u>Longitude (E)</u>	<u>-149.30</u>	<u>5.77</u>	<u>-116.75</u>	<u>-107.71</u>
<u>Elevation (m)</u>	<u>930</u>	<u>1330</u>	<u>2060</u>	<u>3370</u>
<u>Study Period (WY)</u>	<u>2011</u>	<u>2006</u>	<u>2007</u>	<u>2008</u>
<u>Snow Climate</u>	<u>Tundra</u>	<u>Mountain (maritime)</u>	<u>Mountain (intermountain)</u>	<u>Mountain (continental)</u>
<u>Sensors</u>	<u>T_{air}: Vaisala HMP45C P: Campbell Scientific TE 525 U: Met One 014A RH: Vaisala HMP45C Q_{si}: Kipp & Zonen CMA 6 Q_{lj}: none (taken as residual from measurements of all other radiation components^A)</u>	<u>T_{air}: PT 100/4 wires P: PG2000, GEONOR U: Chauvin Arnoux Tavid 87 – non-heated RH: Vaisala HMP 45D Q_{si}: Kipp & Zonen CM14 Q_{lj}: Eppley PIR</u>	<u>T_{air}: Vaisala HMP 45 P: Belfort Universal Gages U: Met One 013/023 RH: Vaisala HMP 45 Q_{si}: Eppley Precision Spectral Pyranometer Q_{lj}: Eppley PIR</u>	<u>T_{air}: Vaisala CS500 P: ETI Noah II U: RM Young Wind Monitor 05103-5 RH: Vaisala CS500 Q_{si}: Kipp & Zonen CM21 Q_{lj}: Kipp & Zonen CG-4</u>
<u>Operators</u>	<u>NRCS, CRREL, Ameriflux</u>	<u>Météo-France</u>	<u>Northwest Watershed Research Center, Agricultural Research Service</u>	<u>Center for Snow and Avalanche Studies</u>
<u>Oct-Dec T_{air} (°C)</u>	<u>-16.1</u>	<u>2.0</u>	<u>0.2</u>	<u>-3.7</u>
<u>Jan-Mar T_{air} (°C)</u>	<u>-14.7</u>	<u>-1.6</u>	<u>-2.0</u>	<u>-8.7</u>
<u>Apr-Jun T_{air} (°C)</u>	<u>-1.4</u>	<u>8.9</u>	<u>8.4</u>	<u>2.7</u>
<u>Oct-Mar P^B (mm)</u>	<u>200</u>	<u>690</u>	<u>480</u>	<u>1000</u>
<u>Mean annual U (m s⁻¹)</u>	<u>2.2</u>	<u>1.0</u>	<u>1.6</u>	<u>1.1</u>

1386 ^A At IC, Q_{lj} was taken as $Q_{lj} = Q_{net} - (Q_{si} - Q_{so}) + (5.67 \times 10^{-8}) T_{surf}^4$, where Q_{net} is measured net radiation ($W m^{-2}$), Q_{si} is measured incoming shortwave radiation
1387 ($W m^{-2}$), Q_{so} is measured reflected shortwave radiation ($W m^{-2}$), and T_{surf} is measured snow surface temperature (°C).

1388 ^B Note that precipitation data were adjusted with a multiplier (see Section 2) prior to conducting the sensitivity analysis.

1389

1390 **Table 2** UEB model parameters used across all simulations and sites

Description of parameter	Units	Value
Rain threshold temperature	°C	+3.0
Snow threshold temperature	°C	-1.0
Snow emissivity	--	0.99
Bulk snow density	kg m ⁻³	300
Liquid water holding capacity	fraction	0.05
Snow saturated hydraulic conductivity	m hr ⁻¹	20
Visual new snow albedo	--	0.85
Near infrared new snow albedo	--	0.65
New snow threshold depth to reset albedo	m	0.01
Snow surface roughness	m	0.005
Forest canopy fraction	fraction	0
Ground heat flux	W m ⁻²	0

1391 **Table 3** Details of error types, distributions, and uncertainty ranges for the [five](#) scenarios. Bold
 1392 face in the error type, distribution, and uncertainty range indicates defining characteristics,
 1393 relative to scenario NB.

Forcing	Error Type ^A	Distribution ^B	Range	Units	Citations and Notes
Scenario NB ($k=6, N=10000$)					
T_{air}	B	Normal	[-3.0, +3.0]	°C	Bolstad et al. (1998); Chuanyan et al. (2005); Fridley (2009); Hasenauer et al. (2003)
P	B	Lognormal	[-75, +300] ^C	%	Goodison et al. (1998); Luce et al. (1998); Rasmussen et al. (2012); Winstral and Marks (2002)
U	B	Normal	[-3.0, +3.0]	m s ⁻¹	Winstral et al. (2009)
RH	B	Normal	[-25, +25]	%	Bohn et al. (2013); Déry and Stieglitz (2002); Feld et al. (2013)
Q_{si}	B	Normal	[-100, +100]	W m ⁻²	Bohn et al. (2013); Jepsen et al. (2012); Jing and Cess (1998); Niemelä et al. (2001b)
Q_{li}	B	Normal	[-25, +25]	W m ⁻²	Bohn et al. (2013); Flerchinger et al. (2009); Herrero and Polo (2012); Niemelä et al. (2001a)
Scenario NB+RE ($k=12, N=10000$)					
This scenario has six bias parameters (identical to NB above), plus the following six random error parameters					
T_{air}	RE	Normal	[0.0, 7.5]	°C	Chuanyan et al. (2005); Fridley (2009); Hasenauer et al. (2003); Huwald et al. (2009); Phillips and Marks (1996)
P	RE	Lognormal	[0.0, 25]	%	Guan et al. (2005); Hasenauer et al. (2003); Hutchinson et al. (2009)
U	RE	Normal	[0.0, 5]	m s ⁻¹	Cheng and Georgakakos (2011); Liston and Elder (2006); Luo et al. (2008); Winstral et al. (2009)
RH	RE	Normal	[0.0, 15]	%	Bohn et al. (2013); Liston and Elder (2006); Phillips and Marks (1996)
Q_{si}	RE	Normal	[0.0, 160]	W m ⁻²	Hasenauer et al. (2003); Jepsen et al. (2012); Liston and Elder (2006); Thornton et al. (2000)
Q_{li}	RE	Normal	[0.0, 80]	W m ⁻²	Bohn et al. (2013); Flerchinger et al. (2009); Liston and Elder (2006)
Scenario UB ($k=6, N=10000$)					
Identical to NB, except all probability distributions are uniform					
Scenario NB_gauge ($k=6, N=10000$)					
Identical to NB, except P uncertainty mimics documented differences between P and SWE at SNOTEL sites					
P	B	Lognormal	[-10, +10]	%	Meyer et al. (2012)
Scenario NB_lab^D ($k=6, N=10000$)					
T_{air}	B	Normal	[-0.30, +0.30]	°C	Vaisala HMP45 specified accuracy
P	B	Lognormal	[-3.0, +3.0]^E	%	RM Young 52202 specified accuracy
U	B	Normal	[-0.30, +0.30]	m s ⁻¹	RM Young 05103 specified accuracy
RH	B	Normal	[-3.0, +3.0]	%	Vaisala HMP45 specified accuracy
Q_{si}	B	Normal	[-25, +25]	W m ⁻²	Li-Cor 200X specified accuracy of ~5%
Q_{li}	B	Normal	[-15, +15]	W m ⁻²	Assumed ~5% of mean intersite values

1394 ^A B=bias, RE=random errors. Biases are additive ($b_i=0$, Eq. 5) for all forcings except P , which has multiplicative
 1395 bias ($b_i=1$).

1396 ^B Probability distributions were truncated in instances when introduction of errors caused non-physical forcing
 1397 values (see Sec. 3.3.5).

1398 ^C [The high upper \$P\$ bias \(300%\) mimics cases where snowfall data collected in an area of drift deposition are
 1399 assumed \(incorrectly\) to represent other basin locations.](#)

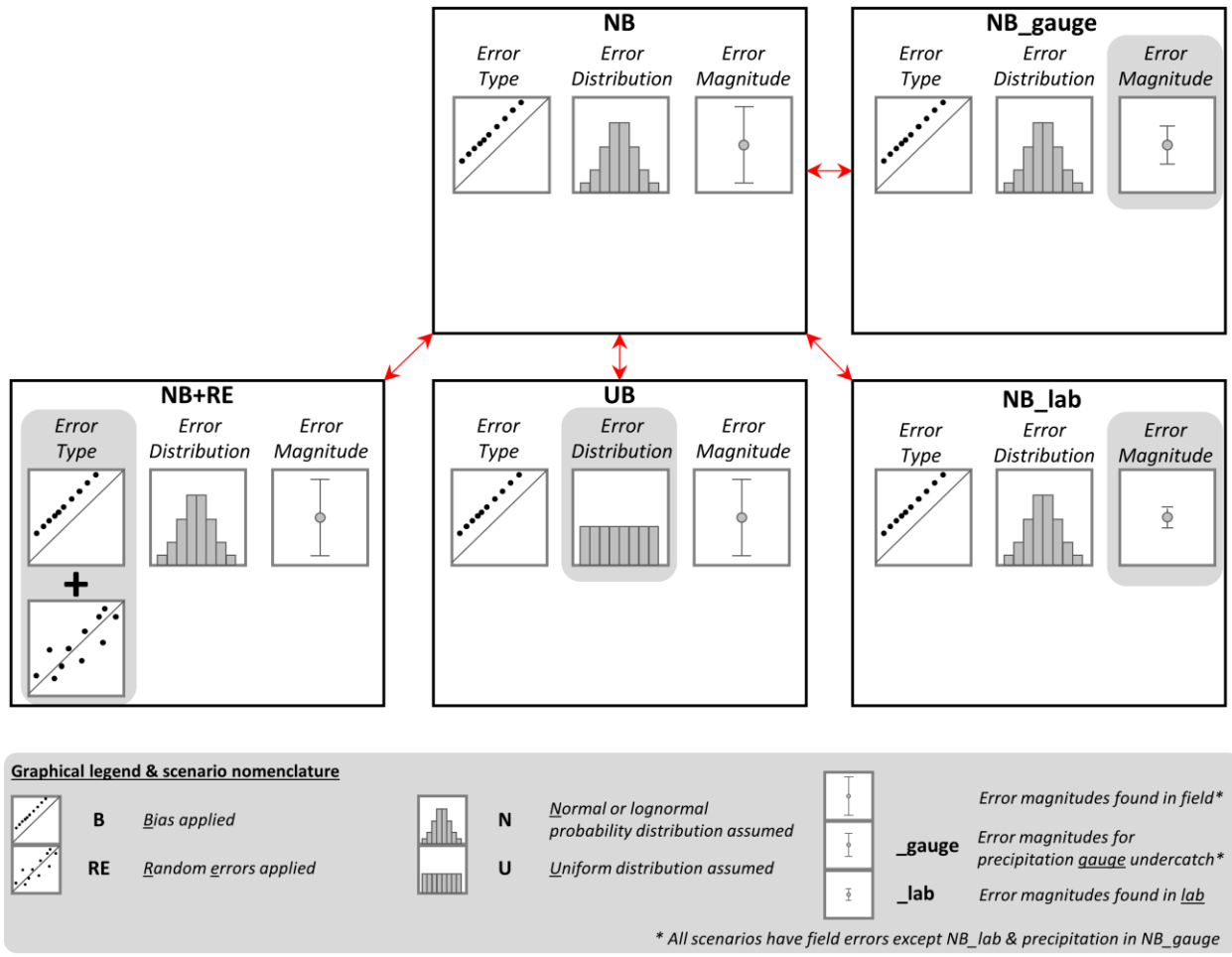
1400 ^D Uncertainty ranges in this scenario are based primarily on manufacturer's specified accuracy for typical sensors
 1401 deployed at SNOTEL sites (*NRCS Staff, personal communication, 2013*). We assume the P storage gauge has the
 1402 same accuracy as a typical tipping bucket gauge.

1403 ^E We neglect P undercatch errors in the lab uncertainty scenario.

1404 **Table 4** Number of samples (model simulations) meeting the requirements for minimum peak
 1405 SWE and snow duration and valid snow disappearance dates at each site in each scenario.

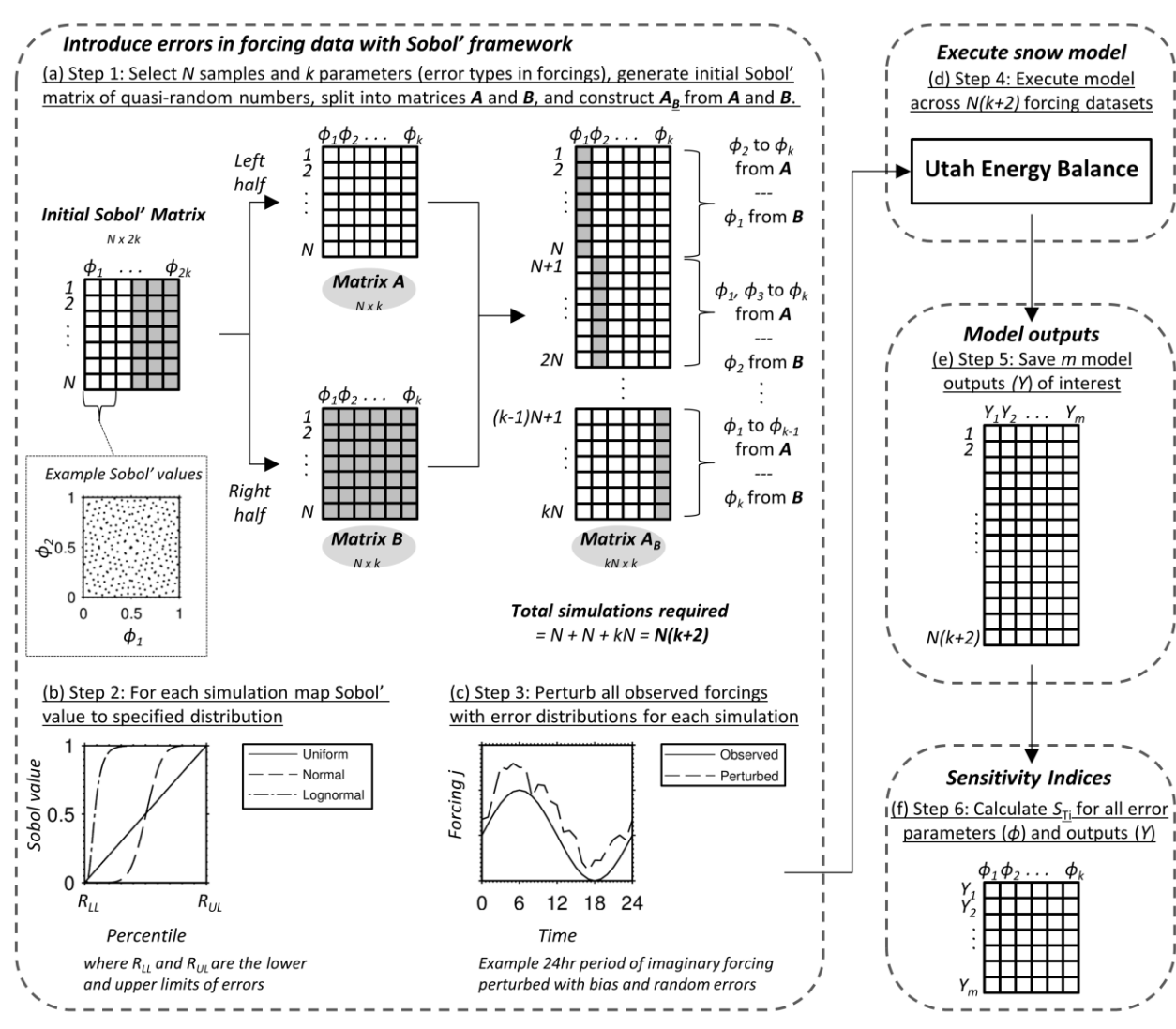
	Scenario NB	Scenario NB+RE	Scenario UB	Scenario NB_gauge	Scenario NB_lab
IC	9898 (79 184)	10 000 (140 000)	8608 (68 864)	10 000 (80 000)	10 000 (80 000)
CDP	9792 (78 336)	9869 (138 166)	8925 (71 400)	9999 (79 992)	10 000 (80 000)
RME	8799 (70 392)	9233 (129 262)	9102 (72 816)	10 000 (80 000)	10 000 (80 000)
SASP	9984 (79 872)	9984 (139 776)	3399 (27 192)	10 000 (80 000)	10 000 (80 000)

1406 **7.8.Figures**



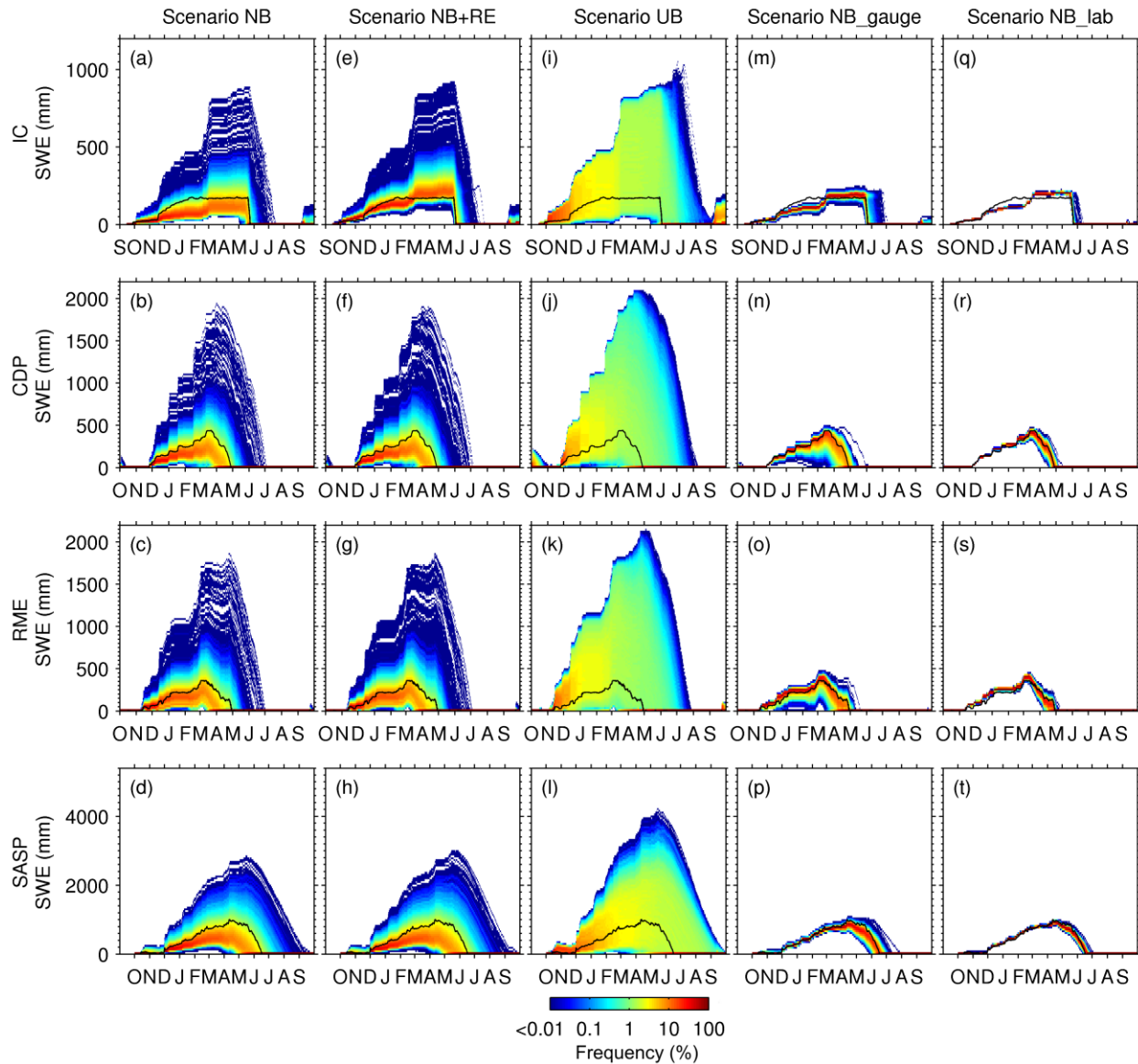
1407

1408 **Figure 1** Scenarios of interest and the type, distribution, and magnitude of errors considered in
 1409 each. NB considers normally (or lognormally) distributed biases with error magnitudes found in
 1410 the field. NB+RE is the same as NB but also considers random errors. UB is the same as NB
 1411 but considers uniformly distributed errors instead. NB_gauge is the same as NB but with
 1412 reduced precipitation uncertainty (typical difference between precipitation gauge and snow
 1413 pillow). NB_lab is the same as NB but considers laboratory error magnitudes.



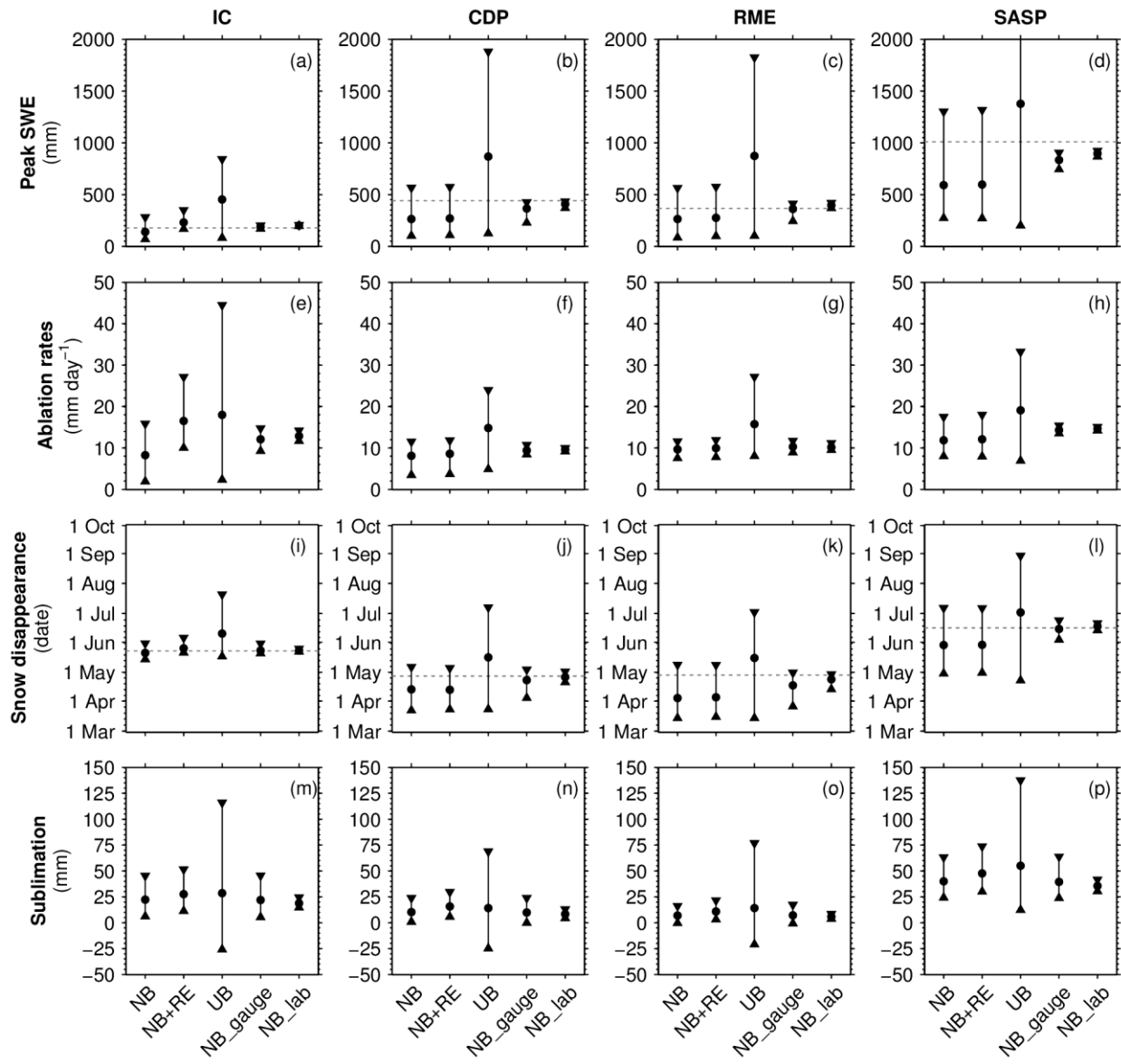
1414

1415 **Figure 2** Conceptual diagram showing methodology for imposing errors on the forcings with
 1416 error parameters (ϕ) within the Sobol' sensitivity analysis framework, and workflow for model
 1417 execution and calculation of sensitivity indices on model outputs (Y).



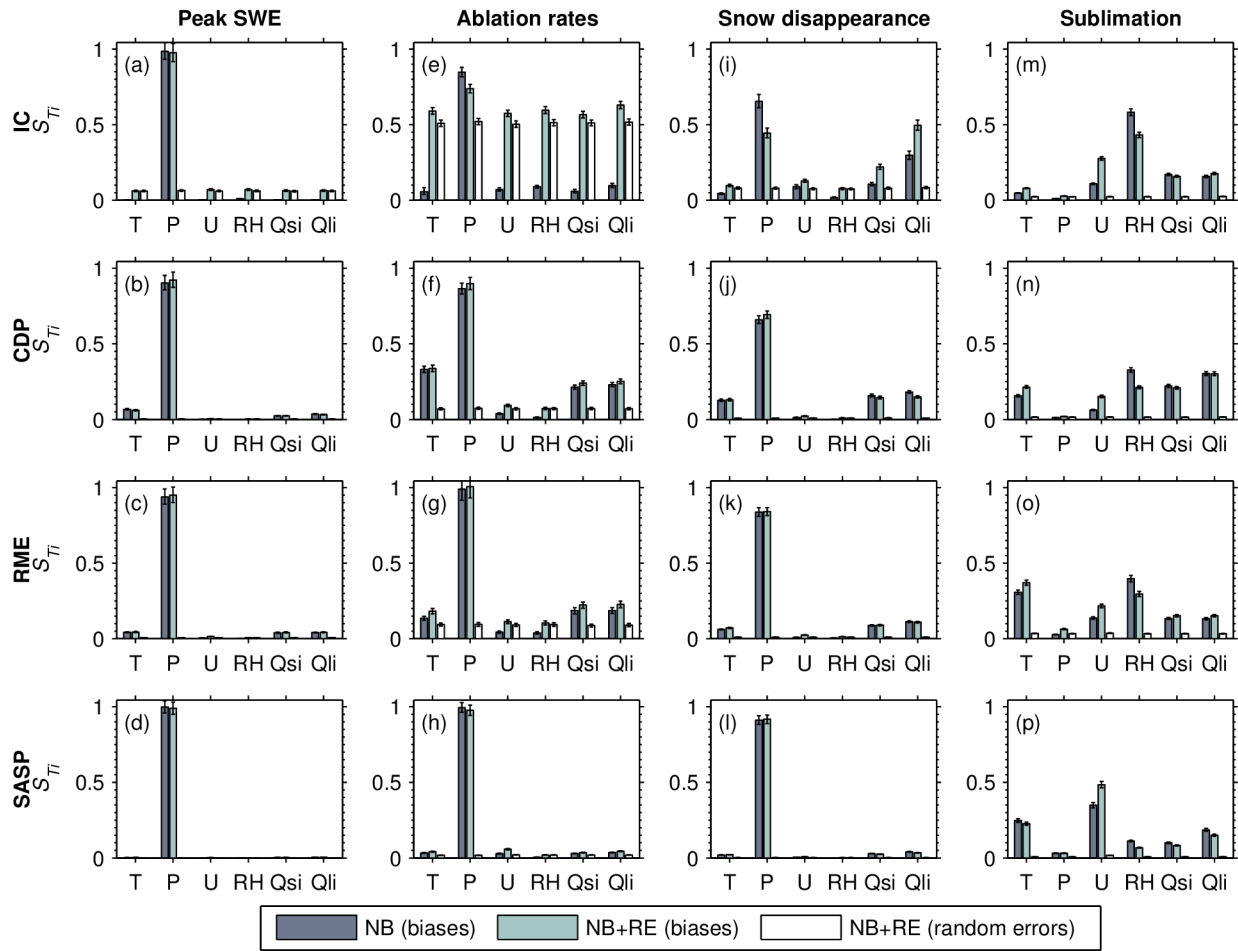
1418

1419 **Figure 3** Observed (black line) and modeled SWE (color density plot) at the four sites across the
 1420 [five](#) uncertainty scenarios (see Figure 1 and Table 3). The number of model simulations in the
 1421 density plots varies with the site and scenario (see Table 4). The density plots were constructed
 1422 using 100 bins in the SWE dimension with relative frequency tabulated in each bin each day.
 1423 Note the frequency colorbar is on a logarithmic scale. Sites are arranged from top to bottom in
 1424 order of increasing elevation and decreasing latitude. Scenarios are defined as normally
 1425 distributed bias (NB), normally distributed bias and random errors (NB+RE), uniformly
 1426 distributed bias (UB), [normally distributed bias with precipitation gauge uncertainty NB_gauge](#),
 1427 and normally distributed bias at laboratory error magnitudes (NB_lab).



1428

1429 **Figure 4** Distributions of model outputs (rows) at the four study sites (columns) arranged by
 1430 scenario. For each scenario, the circle is the mean and the whiskers show the range
 1431 encompassing 95% of the simulations (see Table 4 for number of simulations for each site and
 1432 scenario). The dashed lines in (a-d) and (i-l) are the observed values. Axes are matched between
 1433 sites for a given model output; note that the range in scenario UB in (d) is truncated by the axes
 1434 limits (upper value = 3030 mm). Scenarios are defined as normally distributed bias (NB),
 1435 normally distributed bias and random errors (NB+RE), uniformly distributed bias (UB),
 1436 [normally distributed bias with precipitation gauge uncertainty NB_gauge](#), and normally
 1437 distributed bias at laboratory error magnitudes (NB_lab).



1438

1439

1440

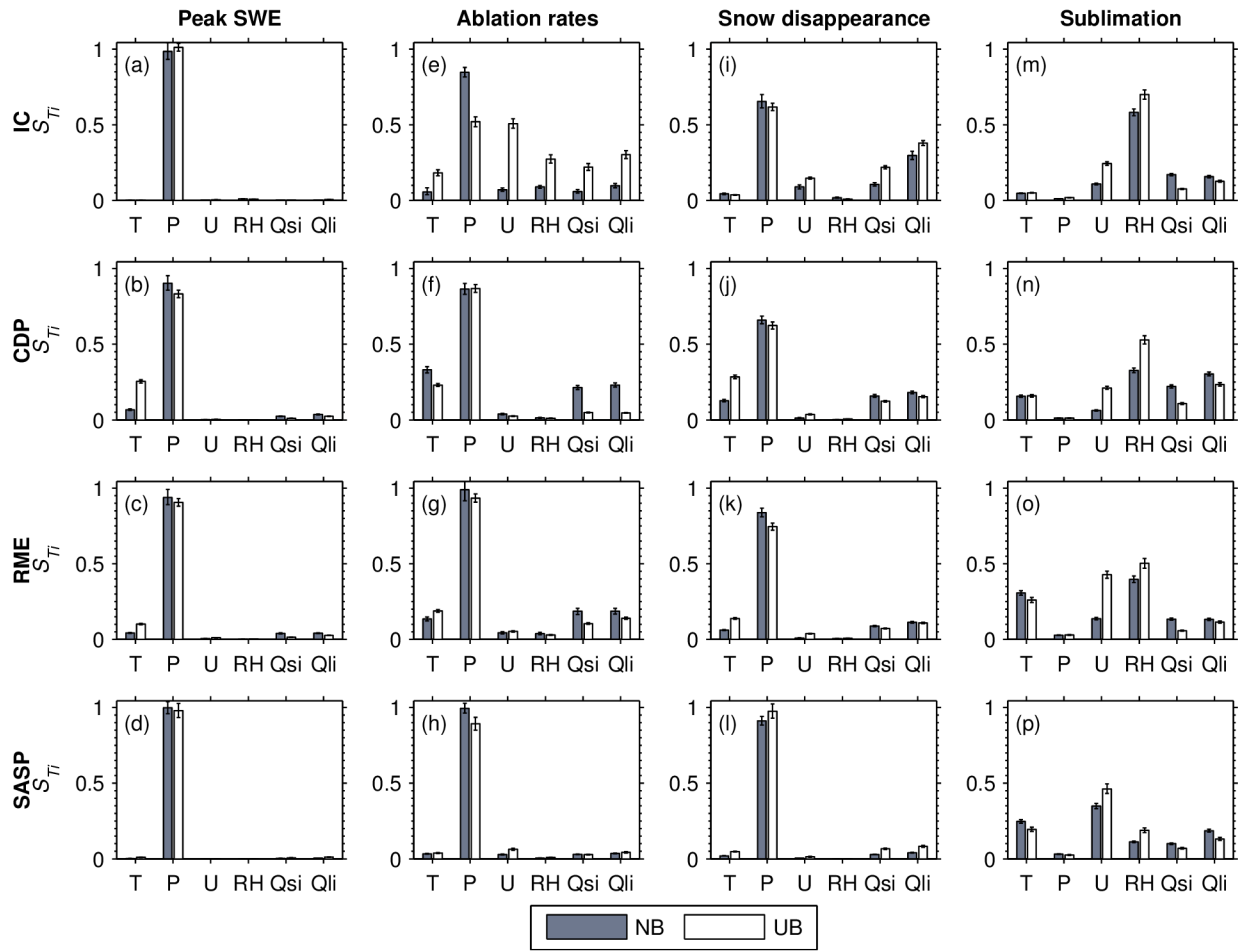
1441

1442

1443

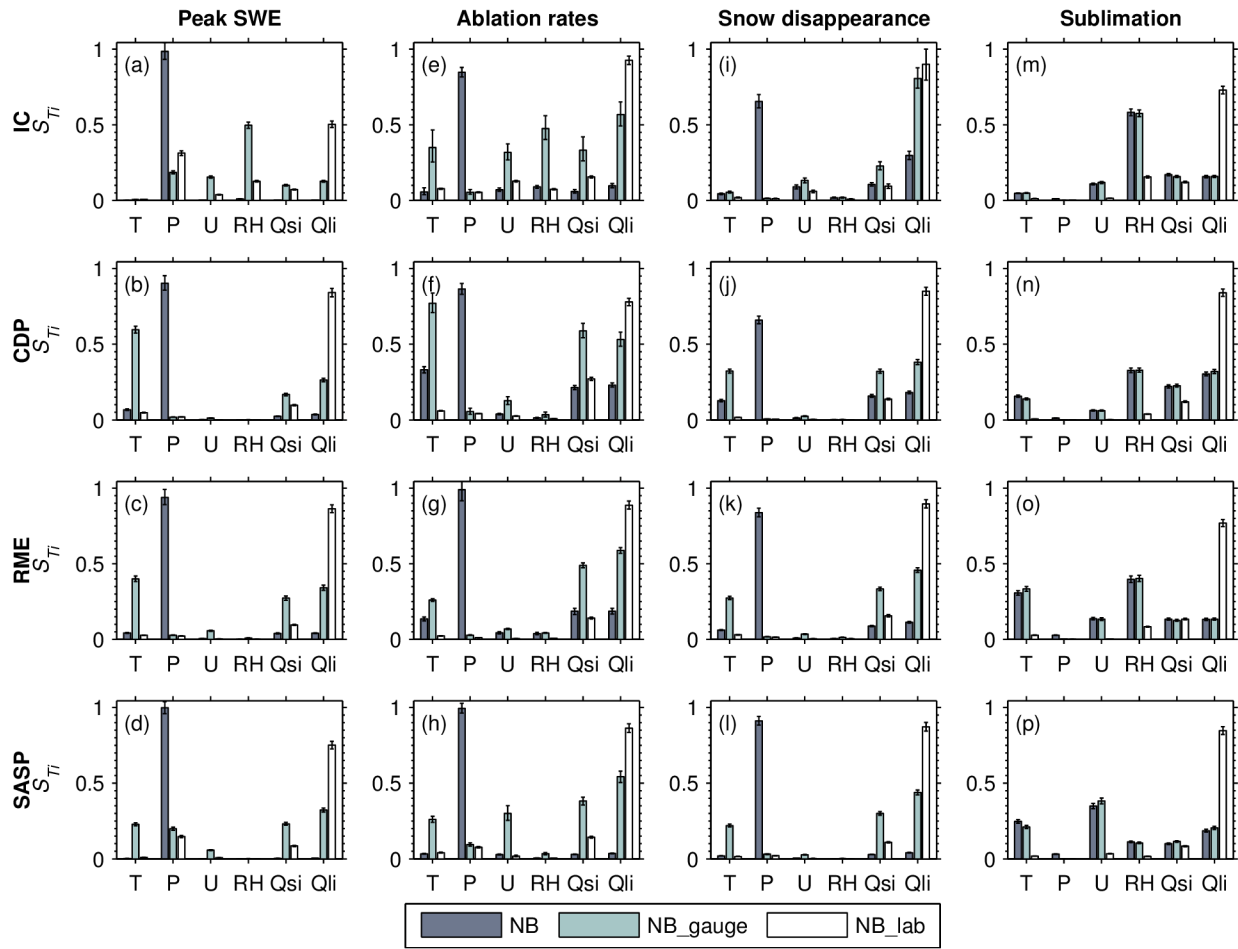
1444

Figure 5 Model sensitivity as a function of forcing error type. Shown are the total-order sensitivity indices (S_{Ti}) of four model response variables (columns) at the four sites (rows) from scenarios NB and NB+RE. In NB+RE, bias and random error parameters are shown separately. NB+RE considers normally distributed bias and random errors, while NB considers normally distributed bias only. The bar indicates the mean (bootstrapped) sensitivity indices and associated 95% confidence intervals.



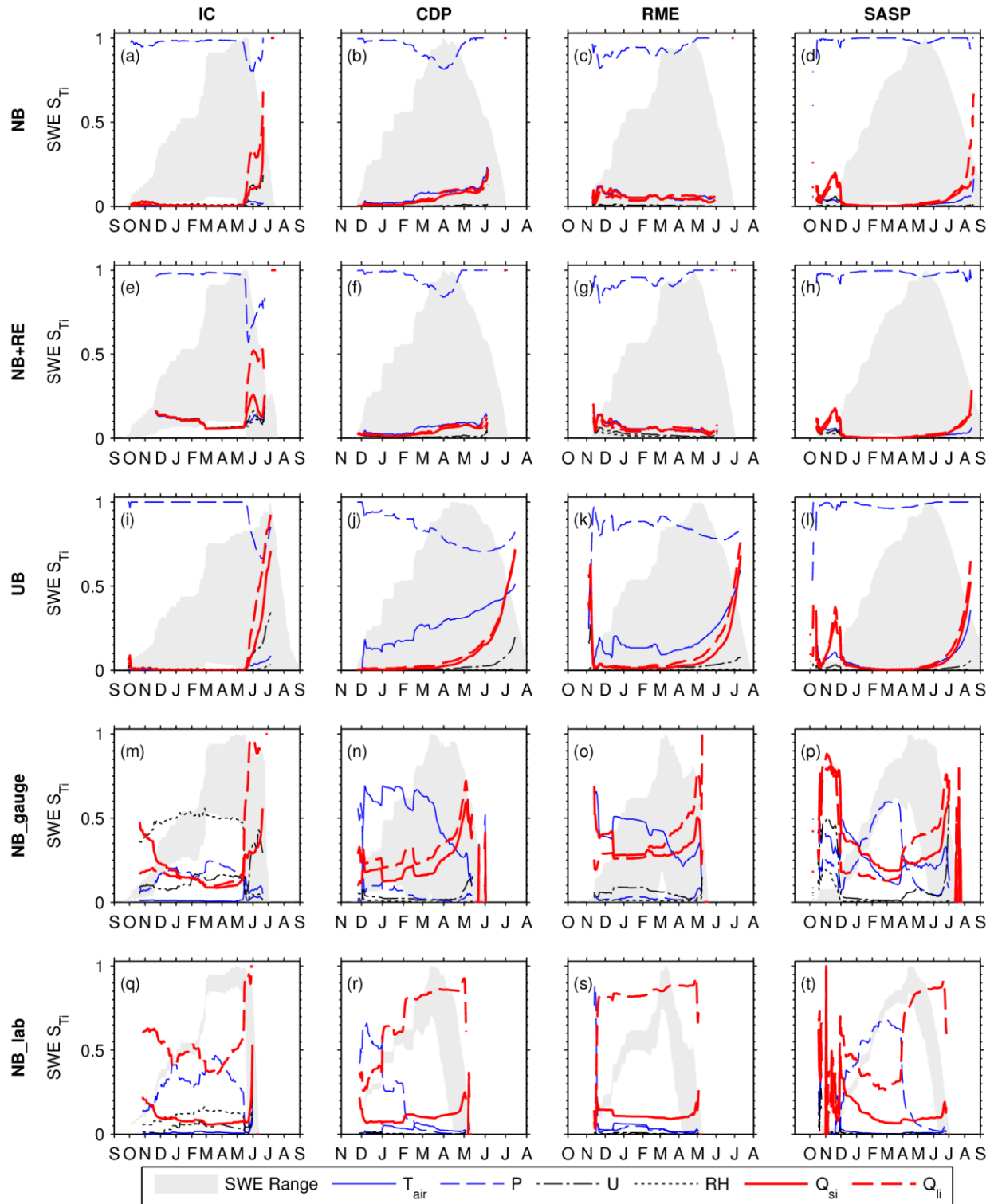
1445
1446
1447
1448

Figure 6 Same as Fig. 5, but comparing S_{Ti} values from scenarios NB and UB to test model sensitivity as a function of error distribution. UB considers uniformly distributed bias, while NB considers normally distributed bias.



1449
 1450
 1451
 1452
 1453
 1454

Figure 7 Same as Fig. 5, but comparing S_{Ti} values from scenarios NB, NB_gauge, and NB_lab to test model sensitivity as a function of error magnitudes. NB considers normally distributed bias at error magnitudes found in the field. NB_gauge has lower precipitation uncertainty (gauge undercatch) than NB but is otherwise identical. NB_lab considers normally distributed bias at error magnitudes found in the laboratory.



1455

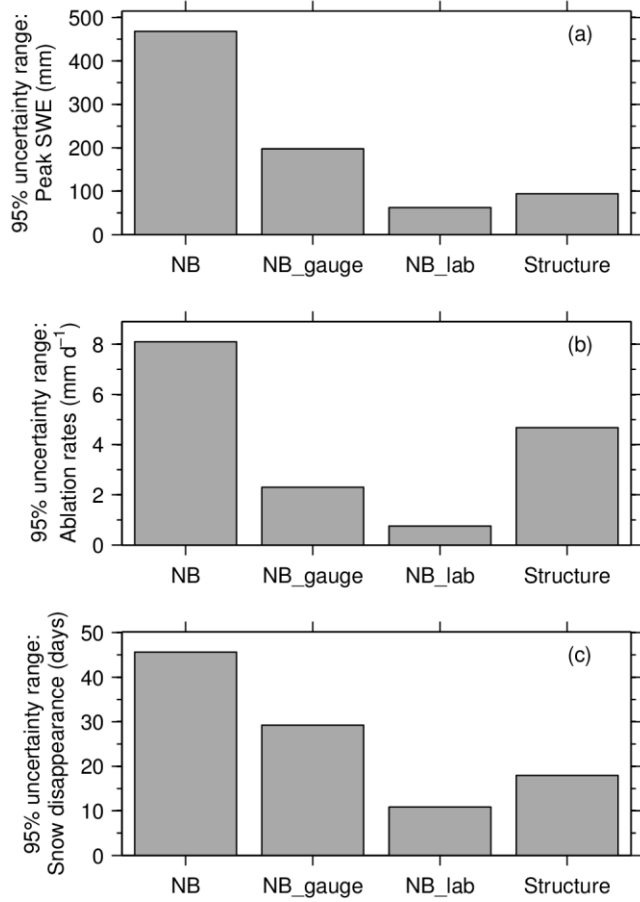
1456

1457

1458

1459

Figure 8 Variation of daily SWE sensitivity to forcing bias based on site (columns) and error scenario (rows). The normalized range (where 1 = maximum SWE) in modeled SWE is shown (gray area) for context. Sensitivity indices in the early and late part of the snow season were screened out, as a high number of simulations with SWE=0 yielded invalid sensitivity indices.



1460

1461 **Figure 9** [Uncertainty ranges \(95% intervals\) in \(a\) peak SWE, \(b\) ablation rates, and \(c\) snow](#)
 1462 [disappearances date at CDP in WY2006 for three forcing uncertainty scenarios and the](#) Essery et
 1463 [al. \(2013\) structural uncertainty.](#)