AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2015WR017559

Companion to *Razavi and Gupta* [2016], doi:10.1002/2015WR017558.

Key Points:
- Star-sampling (STAR) enables VARS to fully characterize global sensitivity
- Case studies show VARS is highly efficient, even for high-dimensional problems
- STAR-VARS is more robust, stable, and efficient than either Sobol or Morris

Supporting Information:
- Supporting Information S1

Correspondence to:
S. Razavi,
saman.razavi@usask.ca

Citation:
Razavi, S., and H. V. Gupta (2016), A new framework for comprehensive, robust, and efficient global sensitivity analysis: 2. Application, *Water Resour. Res.*, *52*, 440–455, doi:10.1002/2015WR017559.

Received 17 MAY 2015
Accepted 10 DEC 2015
Accepted article online 1 DEC 2015
Published online 28 JAN 2016

A new framework for comprehensive, robust, and efficient global sensitivity analysis: 2. Application

Saman Razavi[1,2] and Hoshin V. Gupta[3]

[1]Global Institute for Water Security & School of Environment and Sustainability, University of Saskatchewan, Saskatoon, Saskatchewan, Canada, [2]Department of Civil and Geological Engineering, University of Saskatchewan, Saskatoon, Saskatchewan, Canada, [3]Department of Hydrology and Water Resources, University of Arizona, Tucson, Arizona, USA

Abstract Based on the theoretical framework for sensitivity analysis called "Variogram Analysis of Response Surfaces" (VARS), developed in the companion paper, we develop and implement a practical "star-based" sampling strategy (called STAR-VARS), for the application of VARS to real-world problems. We also develop a bootstrap approach to provide confidence level estimates for the VARS sensitivity metrics and to evaluate the *reliability* of inferred factor rankings. The effectiveness, efficiency, and robustness of STAR-VARS are demonstrated via two real-data hydrological case studies (a 5-parameter conceptual rainfall-runoff model and a 45-parameter land surface scheme hydrology model), and a comparison with the "derivative-based" Morris and "variance-based" Sobol approaches are provided. Our results show that STAR-VARS provides reliable and stable assessments of "global" sensitivity across the full range of scales in the factor space, while being 1–2 orders of magnitude more efficient than the Morris or Sobol approaches.

1. Background and Objective

Sensitivity analysis (SA) plays a significant role in the development and understanding of computer simulation models by providing information regarding the "sensitivity" of model responses or state variables to factors such as parameters, forcings, boundary conditions, etc. Such knowledge can be extremely helpful in the development and application of Earth and Environmental Systems Models (EESMs); see *Razavi and Gupta* [2015] and references therein. However, as discussed in *Razavi and Gupta* [2015], major challenges to SA include: (1) there are different (even conflicting) philosophies and theoretical definitions of sensitivity; and (2) the practical implementation of SA can be highly computationally demanding, particularly for high-dimensional problems and/or computationally intensive models.

This set of companion papers presents a theoretical and practical framework for addressing the above challenges. In the companion paper, *Razavi and Gupta* [2016], we develop the theoretical basis for a comprehensive global sensitivity analysis that generates a full "spectrum" of information on sensitivity, which includes (as limiting/special cases) the commonly used sensitivity analysis metrics proposed by Sobol [*Sobol'*, 2001] and Morris [*Morris*, 1991]. Our framework is called "Variogram Analysis of Response surfaces" (VARS).

Here, we discuss the practical implementation and application of VARS to real-world problems via *sampling* of the response surface (necessary with any SA approach). We propose a carefully designed sampling strategy that enables estimates of the full suite of VARS products (including Morris and Sobol metrics) to be computed. In addition, since reliability and robustness of the inferred factor rankings are critical to model development, we implement a bootstrapping procedure to estimate confidence intervals on the results. The new framework is applied to two real-data hydrological modeling studies (a 5-parameter conceptual rainfall-runoff model and a 45-parameter land surface scheme hydrology model), thereby providing a clear demonstration of the relative effectiveness, efficiency, and robustness of VARS in comparison with the "derivative-based" Morris and "variance-based" Sobol approaches that can be considered to be the most rigorous "global" SA approaches available to-date.

The paper is organized as follows. Following a brief overview of VARS (section 2), we explain our response surface sampling strategy and bootstrapping procedure for estimating uncertainty and reliability (section 3). In section 4, we present the two real-data hydrological modeling cases studies. Finally, a summary and discussion of the contributions, methods, and results of this paper, and directions for future work, appear in section 5.

2. Overview of VARS

The VARS framework is based in an analogy to variogram analysis (see *Razavi and Gupta* [2016]). Given model response y represented as a function of a set of n factors, i.e., $y = f(\boldsymbol{x})$ where $\boldsymbol{x}=\{x_1, \ldots x_i, \ldots, x_n\}$, the "global sensitivity" of y with respect to x_i is a "scale-dependent" property that can be characterized by the use of the variogram and covariogram functions, $\gamma(\boldsymbol{h})=\frac{1}{2}\cdot V(y(\boldsymbol{x}+\boldsymbol{h})-y(\boldsymbol{x}))$ and $C(\boldsymbol{h})=\frac{1}{2}\cdot COV(y(\boldsymbol{x}+\boldsymbol{h}),y(\boldsymbol{x}))$ where $\boldsymbol{h}=\boldsymbol{x}^A-\boldsymbol{x}^B$ is the distance vector, $\boldsymbol{h}=\{h_1, \ldots, \ldots h_i, \ldots, h_n\}$, between any two points A and B in the factor space. Directional variograms $\gamma(h_i)$ and covariograms $C(h_i)$ provide a wealth of sensitivity information across a full range of scales. This information can be characterized by a set of metrics called IVARS (for Integrated Variogram Across a Range of Scales), obtained by integrating the variogram to a particular scale (H_i) of interest, $\Gamma(H_i)=\int_0^{H_i}\gamma(h_i)\,dh_i$. As discussed in *Razavi and Gupta* [2016], we recommend use of H_i values corresponding to 10%, 30%, and 50% of the factor range, thereby obtaining values for $IVARS_{10}$, $IVARS_{30}$, and $IVARS_{50}$, respectively.

Note also that there is a clear theoretical relationship between VARS and the Morris and Sobol approaches. On the one hand, as $h_i \to 0$, the variogram $\gamma(h_i) \propto E\left(\left(\frac{dy}{dx_i}\right)^2\right)$, indicating the equivalence of VARS- and Morris-based assessments at small scales. On the other, the Sobol variance-based "total-order effect" sensitivity index S_{Ti} of the ith factor is related to the variogram and covariogram functions by equation:

$$S_{Ti}=\frac{\gamma(h_i)+E[C_{\boldsymbol{x}_{\sim i}}(h_i)]}{V(y)} \tag{1}$$

where $\boldsymbol{x}_{\sim i}$ is the vector of all of n factors except x_i, and $V(y)$ is the total variance of the response surface. For details, refer to the companion paper, *Razavi and Gupta* [2016].

3. Practical (Numerical) Implementation of VARS

As with the Morris and Sobol methods, practical implementation (for all but the most trivial problems) requires that the sensitivity metrics be estimated via *sampling* of the response surface. The sampling strategy presented below is designed to provide efficient estimates of the full set of VARS-based sensitivity products listed in Table 1. In addition, we implement a bootstrapping strategy to estimates confidence intervals on the metric values, thereby enabling assessment of the reliabilities of factor rankings.

3.1. Star-Based Sampling Strategy

The "star-based" sampling strategy presented here (hereafter called STAR) is designed to facilitate computation of the full range of sensitivity-related information provided by the VARS framework; other strategies are of course possible. Without loss of generality, all factors are scaled to vary on range zero to one. The steps are as follows:

1. **Select Resolution**: Select Δh, which represents a "smallest" value for h. This enables numerical computation of the variogram at the following h values: 0, Δh, $2\Delta h$, $3\Delta h$, etc.
2. **Generate Star Centers**: Randomly choose m points, located across the factor space via, e.g., Latin hypercube sampling, and evaluate the model response at each of these "star centers." Denote their locations using $\boldsymbol{x}_j^+=\left\{x_{j,1}^+,\ldots x_{j,i}^+,\ldots,x_{j,n}^+\right\}$ where $j=1,\ldots,m$.
3. **Generate Cross Sections**: For each star center, generate a cross section of equally spaced points, Δh apart, along *each* of the n dimensions in the factor space, including and passing through that star center and evaluate the model response at each new point. The ith cross section of the jth star center, where $i=1,\ldots,n$, is obtained by fixing $\boldsymbol{x}_{\sim i}=\boldsymbol{x}_{j,\sim i}^+$ and varying x_i (see Figure 1). This results in $(1/\Delta h)-1$ new points for each cross section, and a total of $n((1/\Delta h)-1)$ new points for each star center. The set of points generated around a star center (including the star center) is called a "star."
4. **Extract Pairs**: For each dimension, extract all the pairs of points with h values of Δh, $2\Delta h$, $3\Delta h$, and so on. For each dimension, this results in $m((1/\Delta h)-1)$ pairs for $h=\Delta h$, $m((1/\Delta h)-2)$ pairs for $h=2\Delta h$, and $m((1/\Delta h)-3)$ for $h=3\Delta h$, and so on (using all of the stars).

Using the sample pairs so obtained, numerical estimates of following can be computed:

1. Directional variograms, $\gamma(h_i)$, integrated variograms, $\Gamma(H_i)$, and covariogram, $C(h_i)$, where $i=1,\ldots,n$.
2. Mean ACtual Elementary effects across scales (hereafter called VARS-ACE), mean ABsolute Elementary effects across scales (hereafter called VARS-ABE), and mean SQuare Elementary effects across scales

Table 1. A Summary of the VARS, Sobol and Morris Products for Global Sensitivity Analysis Implemented in This Study

No.	Approach	SA Product	Description
1	STAR-VARS	$\gamma(h)$	Directional variogram
2		$IVARS_{10}$	Integrated Variogram Across a Range of Scales: scale range = 0-10%
3		$IVARS_{30}$	Integrated Variogram Across a Range of Scales: scale range = 0-30%
4		$IVARS_{50}$	Integrated Variogram Across a Range of Scales: scale range = 0-50%
5		VARS-TO	Variance-based Total-Order effect
6		VARS-ACE	Mean ACtual Elementary effect across scales
7		VARS-ABE	Mean ABsolute Elementary effect across scales
8		VARS-SQE	Mean SQuare Elementary effect across scales
9	Sobol	Sobol-FO	Variance-based First-Order effect
10		Sobol-TO	Variance-based Total-Order effect
11	Morris	Morris-$ACE_{\Delta x}$	Mean ACtual Elementary effect
12		Morris-$ABE_{\Delta x}$	Mean ABsolute Elementary effect
13		Morris-$SQE_{\Delta x}$	Mean SQuare Elementary effect

VARS-TO and Sobol-TO are effectively the same quantity, but calculated by different methods. For the Sobol approach, the sampling strategy (Sobol sequence) and numerical implementation described in *Saltelli et al.* [2008] were used. For the Morris approach, the conventional sampling strategy and numerical implementation described in *Campolongo et al.* [2007] were used. Morris-$ACE_{\Delta x}$, Morris-$ABE_{\Delta x}$, and Morris-$SQE_{\Delta x}$ are special cases of VARS-ACE, VARS-ABE, and VARS-SQE where the scale is specified by the step size Δx (In this study, we set $\Delta x = 5\%$ of the factor range).

(hereafter called VARS-SQE). These are equivalent to Morris-based sensitivity measures across a range of scales ($\Delta x = \Delta h,\ 2\Delta h,\ 3\Delta h,$ etc. wherein Δx is the step size for the numerical calculation of the elementary effects/partial derivatives).

3. Variograms and covariograms for any cross section, $\gamma_{\mathbf{x}_{\sim i}}(h_i)$ and $C_{\mathbf{x}_{\sim i}}(h_i)$.
4. The variance-based total-order effect (equation (1) hereafter called VARS-TO (TO for Total Order).

The computational cost of this star-based sampling strategy is $m[n((1/\ \Delta h)-1)+1]$ model runs; the approach can be easily parallelized to take advantage of multiprocessor machines. We refer to this star-based sampling implementation of VARS as "STAR-VARS."

3.2. Star-Based Bootstrapping

Bootstrapping [*Efron and Tibshirani*, 1994] can be performed in conjunction with star-based sampling to generate estimates of (a) confidence intervals on the resulting sensitivity metrics, and (b) reliability of the inferred factor rankings, as follows:

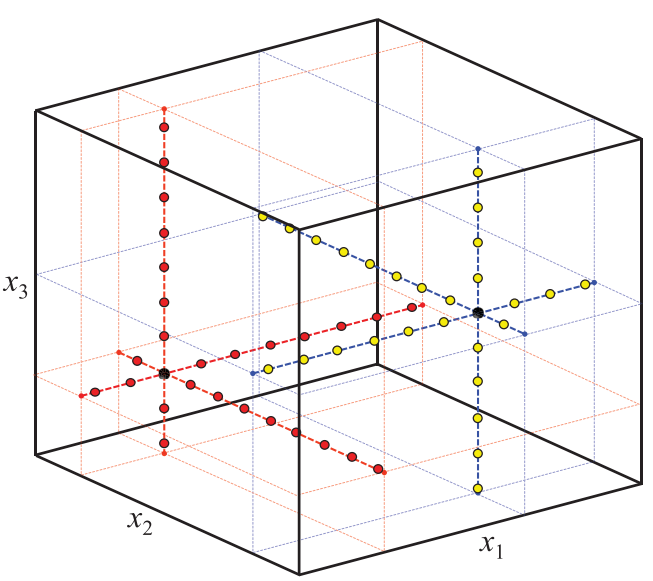

Figure 1. Three-dimensional illustration of the proposed star-based sampling strategy, with resolution $\Delta h = 0.1$ and number of stars $m = 2$—the two black markers indicate the star centers and the other markers indicate the points on the cross sections.

1. Randomly sample with replacement m of m star centers.
2. Calculate the sensitivity metrics and factor rankings using the cross sections associated with the m bootstrap-sampled star centers.
3. Repeat Steps 1 and 2 above k times (where k is some large number, e.g., hundreds or thousands), and each time, store the resulting sensitivity metrics and factor rankings. These are deemed samples of the multivariate distributions of the sensitivity metrics and factor rankings.

Then, for each factor and each sensitivity metric:

4. Calculate confidence intervals (e.g., 90% intervals) from the associated marginal distribution.
5. Count the number of times (out of the total k times sampling with

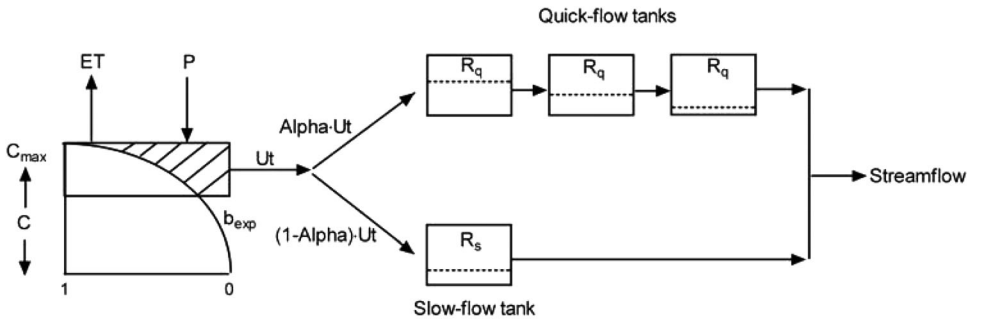

Figure 2. Conceptual structure of the 5-parameter conceptual hydrologic model (HYMOD) used in Case Study 1 (adapted from *Vrugt et al*. [2003]). For this study, the sensitivity of Nash-Sutcliffe criterion to the five model parameters is assessed via the VARS, Sobol, and Morris approaches.

replacement) that the rank of that factor is the same as the original rank obtained (based on all of the *m* star centers)—this ratio provides an estimate of the reliability of a factor ranking.

4. Real-Data Case Studies

We demonstrate the STAR-VARS approach using two real-data hydrological case studies. For benchmark comparison, we also compute the commonly used Morris and Sobol SA metrics. For Morris, we use the conventional sampling/implementation strategy as described in *Campolongo et al*. [2007] with a sampling step size of 5% of the factor range; these metrics are labeled as Morris-SQE_5, Morris-ABE_5 and Morris-ACE_5 (although Morris can also be used as a preliminary "screening" strategy, we focus here on its use as a tool for comprehensive sensitivity assessment). To be clear, these metrics are *not* computed using star-based sampling, and should not be confused with the STAR-VARS-based estimates of VARS-SQE, VARS-ABE, and VARS-ACE that are computed for various scales (sampling step sizes across the factor range). For Sobol, we use the sampling/implementation strategy proposed in *Saltelli et al*. [2008]; these metrics are labeled as Sobol-FO and Sobol-TO. Again, these are *not* computed using star-based sampling, and Sobol-TO is not to be confused with the STAR-VARS-based estimate of VARS-TO. For more details, see Table 1.

4.1. Case Study 1: 5-Parameter Conceptual Rainfall-Runoff Model

In the first case study, we apply the HYMOD model (Figure 2) to simulate the rainfall-runoff response of the 1944 km^2 Leaf River watershed, located north of Collins, Mississippi, as described by *Vrugt et al*. [2003]. Specifically, we evaluate the sensitivity of the Nash-Sutcliffe criterion (which measures goodness-of-fit between the model-simulated and observed streamflows) to variations in the five model parameters across their feasible range. The five parameters are the maximum storage capacity in the catchment, C_{max} (unit L), the degree of spatial variability of the soil moisture capacity within the catchment, b_{exp} (unitless), the factor distributing the flow between the two series of reservoirs, Alpha (unitless), and the residence times of the linear quick and slow reservoirs, R_q (unit T) and R_s (unit T). Full details regarding the model, data used, and the parameter ranges can be found in *Vrugt et al*. [2003]. All factors were scaled so that their feasible ranges correspond to [0-1]. To reiterate, the assessment presented below is specific to the use of the Nash-Sutcliffe criterion as the model performance metric.

4.1.1. Performance of Morris and Sobol: The Conflicting Benchmarks

Figure 3 shows that the Morris (Morris-SQE_5) and Sobol (Sobol-TO) methods provide conflicting assessments regarding parameter sensitivities. The Morris approach determines parameter C_{max} to be the most sensitive, whereas the Sobol approach assigns this position to parameter R_q. Further, Morris assesses parameters Alpha and b_{exp} as being almost equally sensitive, while Sobol rates parameter Alpha as being 6 times more sensitive than b_{exp}. Both approaches agree that R_s is the least sensitive. This simple example illustrates the differences in sensitivity assessment that can arise by use of these philosophically different approaches to SA [see also *Razavi and Gupta*, 2015]. The next section shows how VARS can address such issues.

4.1.2. VARS Assessment of Sensitivity Across Scales

Figure 4 shows the VARS products. Figure 4a shows that the directional variogram for R_q remains lower than that for C_{max} over the scale (i.e., *h*) range from 0 to about 0.1, and then (beyond this range) crosses over to larger values. Meanwhile, the directional variograms for b_{exp} and Alpha remain quite similar while

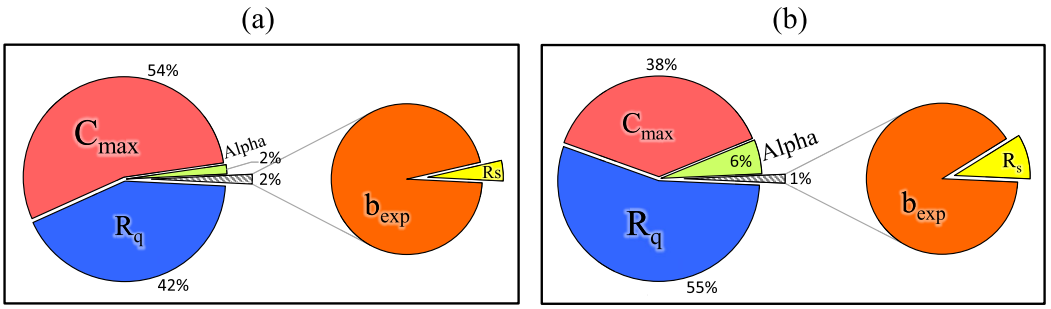

Figure 3. Results provided by the (a) Morris (Morris-SQE$_5$) and (b) Sobol (Sobol-TO) metrics for Case Study 1—the size of each slice represents the proportion of the sensitivity metric value of the respective parameter to the sum of all sensitivity metric values over all parameters.

$h \leq 0.05$, but above this value of h, the variogram of Alpha becomes significantly larger. From Figures 4c and 4d, we see that these small-scale behaviors are not captured by the VARS-based scale-dependent analogs (VARS-ACE and VARS-ABE) of the corresponding Morris metrics.

Figure 5 compares the factor sensitivity ratios provided by the different metrics. IVARS$_{30}$, IVARS$_{50}$, Sobol-FO, and Sobol-TO (corresponding to VARS-TO), all of which provide larger scale assessments of parameter sensitivity, consistently rank the parameters as follows: $R_q > C_{max} >$ Alpha $> b_{exp} > R_s$. However, IVARS$_{10}$, which corresponds to a small scale, inverts the sensitivity ranking for R_q and C_{max} and is therefore consistent with the Morris-based assessment. Overall, the assessment indicates that (a) parameters R_q and C_{max} are significantly more sensitive than the other parameters, (b) parameter R_s is almost insensitive, and (c) at smaller scales, parameters Alpha and b_{exp} demonstrate almost equal sensitivity.

4.1.3. Numerical Implementation and Robustness

In sections 4.1.1 and 4.1.2, we used large numbers of model runs (~70,000) to ensure stable results for all three approaches tested here—Morris, Sobol, and STAR-VARS; for the latter, we used $\Delta h = 0.005$ and 70 star centers. Accordingly, we consider the corresponding assessments to be accurate and, for the rest of this section, treat the resulting sensitivity metrics as "true" values. In this section, we evaluate the relative robustness of the different approaches by conducting 100 independent trials (each using a different initial random seed); for STAR-VARS we use $\Delta h = 0.1$.

Figure 6 shows the probability of failure (PF) estimated for each approach/metric, where PF represents the fractional number of trials (out of 100) in which the generated factor ranking is not consistent with the "true" factor ranking computed above. The results show STAR-VARS to be highly efficient at generating robust estimates of IVARS$_{30}$, IVARS$_{50}$, and VARS-TO. Only 322 function evaluations (seven star centers) are required for the associated PF values (0.23, 0.14, and 0.17, respectively) to all be less than 25%, and when 1610 function evaluations are used (35 star centers), the success rate becomes essentially 100% (PF = 0.03, 0, and 0).

It is, however, interesting to note that the PF for IVARS$_{10}$ remains at ~0.5 even as the number of function evaluations is increased. The reason for this is that at $h = 0.1$ the variogram values $\gamma(h)$ for parameters R_q and C_{max} are almost identical (see Figure 4a), based on which the corresponding approximated values of $\Gamma(H)$ at $H = 0.1$ are indistinguishable. If we wish to more accurately characterize the difference in the IVARS$_{10}$ values for these two parameters (i.e., if we are interested in a small scale assessment), we will need to use a smaller Δh value (and accordingly reallocate the locations of our function evaluations differently). Given the specific choice of sampling design we have made, the IVARS$_{10}$ metric has correctly ranked all of the other parameters (results not shown).

In contrast, the approaches used to calculate Sobol-TO and Morris-SQE$_5$ are significantly less efficient and robust. At 315 function evaluations, Sobol-TO has a PF as high as 0.85, and 2450 function evaluations are required to reduce PF to less than 0.2. In fact (not shown in the Figure 6) as many as 20,000 function evaluations were required to reduce PF below 0.05 (corresponding to a 95% success rate). Overall, our results indicate STAR-VARS to be some 20 times more efficient than the Sobol-TO and Morris-SQE$_5$ approaches for this case study (~10,000/500 or 20,000/1,000 where the nominator is the number of function evaluations required by Sobol or Morris to achieve a performance obtained by STAR-VARS with the number of function evaluations specified in the denominator).

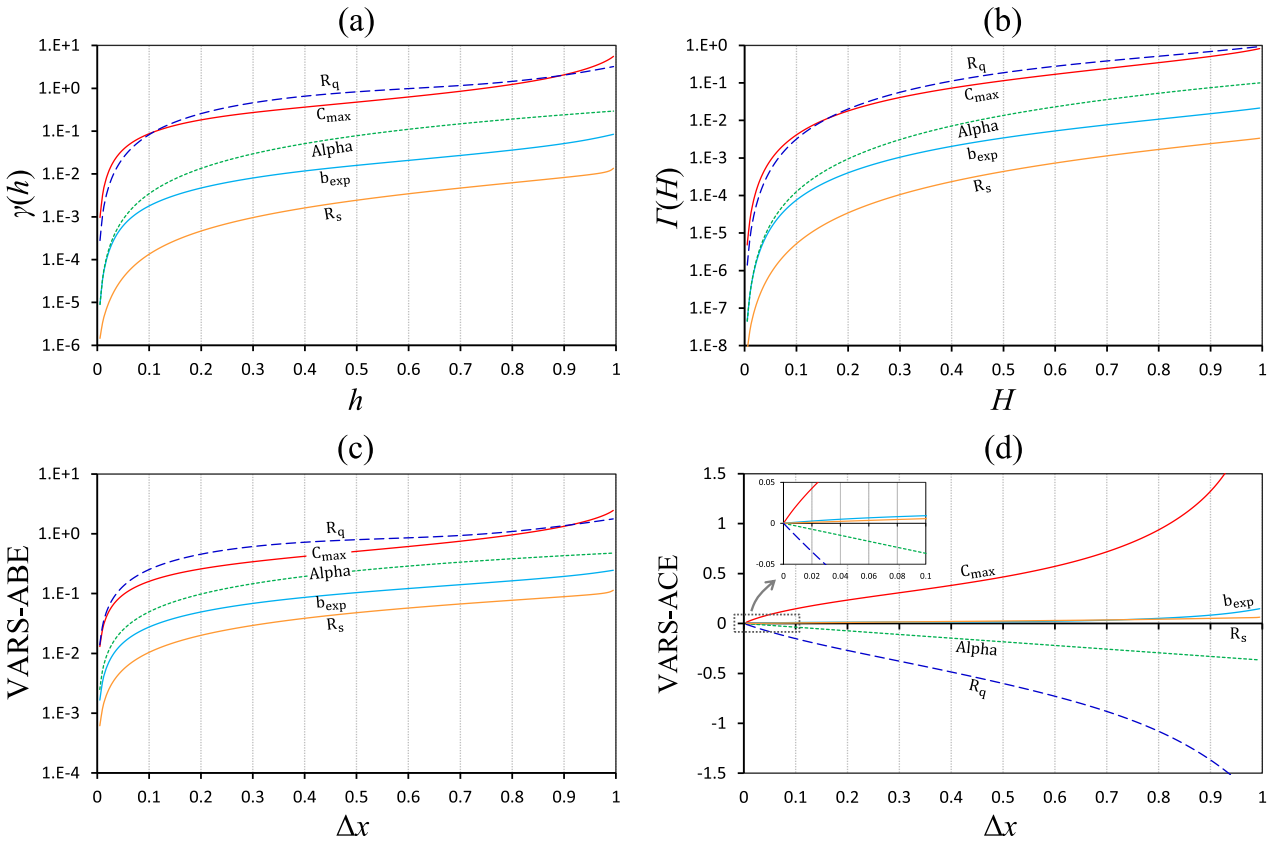

Figure 4. Demonstration of the performance of VARS on Case Study 1: (a) directional variograms, (b) integrated variograms, (c) mean absolute elementary effects across scales, and (d) mean actual elementary effects across scales.

To better understand the reasons for these results, Figure 7 shows cumulative distribution functions (CDFs) for the 100 independent estimates of $IVARS_{10}$, $IVARS_{30}$, $IVARS_{50}$, and VARS-TO obtained using 322 function evaluations, and the corresponding CDFs for the 100 independent estimates of Sobol-TO at both a comparable number (315) and a much larger number (2450) of function evaluations. From the results, we see that the CDFs for $IVARS_{10}$, $IVARS_{30}$, $IVARS_{50}$, and VARS-TO are statistically distinct (largely non-overlapping) for all of the parameters except R_q and C_{max} (the two most sensitive ones) in some cases. For these two parameters, the CDFs overlap significantly for the $IVARS_{10}$ metric (Figure 7a)—see the sampling discussion in paragraph three of this subsection—but are more distinct for the other VARS products. In contrast, the Sobol-TO estimates are significantly less robust; at 315 function evaluations the Sobol-TO CDFs for R_s and b_{exp} cannot be statistically differentiated (Figure 7e), and at

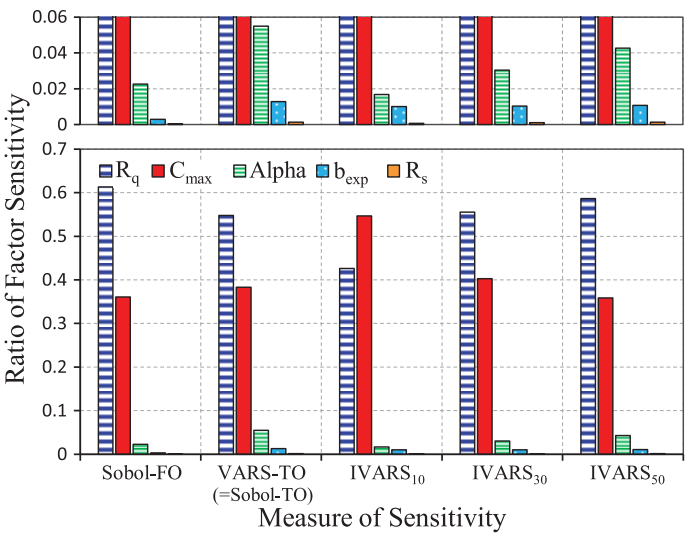

Figure 5. A comparison of the IVARS and Sobol-based metrics for Case Study 1. The vertical axis shows the value of each metric divided by the summed values of that metric over all of the factors. The top plot shows a zoom-in of the bottom plot for very small values on the vertical axis.

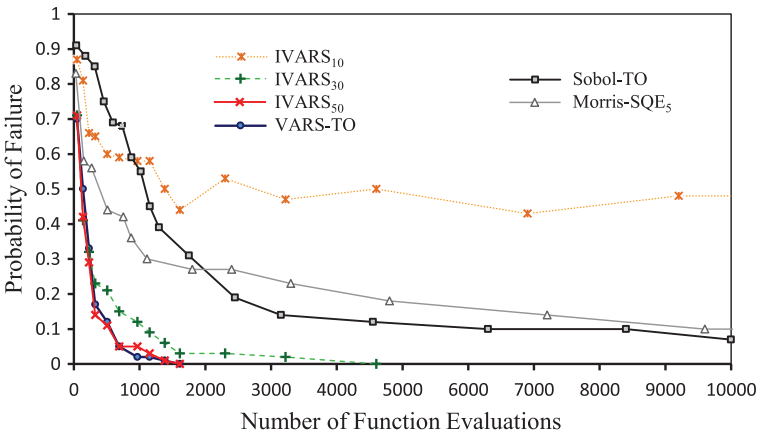

Figure 6. A comparison of the efficiency and reliability of STAR-VARS versus Sobol and Morris for Case Study 1. The figure shows probability of failure for $IVARS_{10}$, $IVARS_{30}$, $IVARS_{50}$, and VARS-TO (total-order effect based on the VARS implementation) along with Sobol-TO (total-order effect based on the Sobol implementation of *Saltelli et al*. [2008]) and Morris-SQE_5 (mean squared elementary effects with $\Delta x = 0.05$). Probability of failure is assessed using 100 independent trials of each approach with different initial random seeds and represents the portion of trials that each metric generated factor rankings inconsistent with the "true" factor ranking based on that metric.

2450 function evaluations the Sobol-TO CDFs for R_s, b_{exp}, and Alpha have considerably larger spreads than the corresponding VARS-TO estimates obtained using only 322 function evaluations (compare Figures 7f and 7d).

Equally, if not more, important, in a significant number of the trials (as high as 50%) the Sobol-TO estimates for various parameters take on negative values; to show the full scope of this problem, supporting information Figure S1 replots Figures 7e and 7f using a linear scale for horizontal axis. Since a negative value for variance is fundamentally impossible, these results are an unfortunate artifact of numerical implementation. This is in spite of the fact that the numerical implementation of Sobol used in this study is the state-of-the-art implementation presented by *Saltelli et al*. [2008].

4.1.4. Bootstrapping, Reliability, and Confidence Intervals

Since the SA results are necessarily dependent on computational effort expended, Figure 8 shows bootstrap-based confidence intervals for the directional variograms corresponding to parameters R_q and Alpha, for the cases of 10 star centers (Figure 8a) and 50 star centers (Figure 8b). The plots indicate that the confidence intervals become progressively narrower with increasing numbers of star centers (cross sections). A further assessment of the reliability that can be associated with parameter sensitivity rankings revealed that only 10 star centers are required for $IVARS_{30}$, $IVARS_{50}$, and VARS-TO to provide success rates better than or equal to 85% (PF $<$ 25%) for all of the factors. And when 50 star centers are used, the corresponding reliability (except for R_q and C_{max} using $IVARS_{10}$ as per discussion above) increases to over 95%.

4.1.5. Demonstration of the Link Between VARS and Sobol

Finally, Figure 9 illustrates how the VARS and Sobol analyses are linked through equation (17) presented in the companion paper, *Razavi and Gupta* [2016]. The variograms $\gamma(h_i)$, $i = 1,\ldots, 5$ shown in Figures 9a–9e for the five parameters are reproduced from Figure 4a. These figures further show $E[C_{\boldsymbol{x}_{\sim i}}(h_i)]$, the covariograms of the response surface in directions $i = 1,\ldots, 5$ averaged across the factor space, computed using equation (4) in the companion paper; consistent with the constant mean assumption, the cross section average is used for the covariance calculation.

As discussed in the companion paper, the sum of $\gamma(h_i)$ and $E[C_{\boldsymbol{x}_{\sim i}}(h_i)]$ should be independent of h_i and is equivalent to the variance-based (Sobol) total-order effect for factor *i*. However, as we use the constant mean assumption to ease the calculation of variograms and covariograms and due to the fact that validity of this assumption may degrade for larger h_i, the sum $\gamma(h_i)+E[C_{\boldsymbol{x}_{\sim i}}(h_i)]$ for factor *i* may not remain constant with h_i as shown in Figure 9. Accordingly, during numerical implementation of VARS, the most accurate estimate of the variance-based total effect (VARS-TO) is generally obtained from $\gamma(\Delta h_i)+E[C_{\boldsymbol{x}_{\sim i}}(\Delta h_i)]$ computed at the smallest available h_i.

Of course, the variance, $V(y)$, of the response surface can either be estimated directly using all of the points sampled or, alternatively, using equation (13) from the companion paper (which requires computing $C(\boldsymbol{h})$

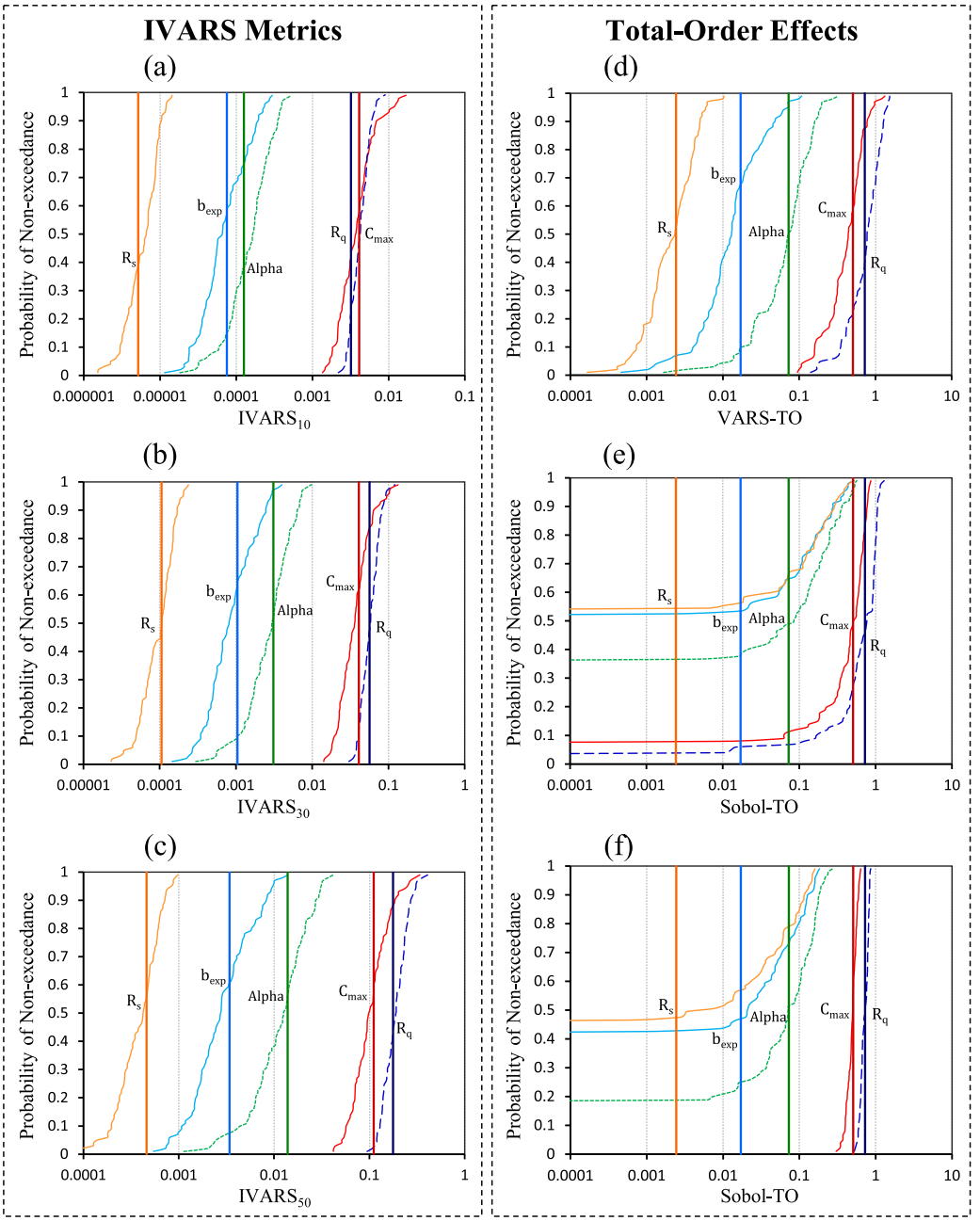

Figure 7. An assessment of the robustness of STAR-VARS against sampling variability, compared with the Sobol method. Each subplot shows the cumulative distribution functions over 100 independent trials (with different initial random seeds) of a sensitivity metric for different factors. The left column (subplots a, b, and c) shows $IVARS_{10}$, $IVARS_{30}$, and $IVARS_{50}$ obtained using a lower computational budget of 322 model runs (seven star centers). In the right column, subplot (d) shows the STAR-VARS-based estimates of Total-Order effects (TO) obtained with 322 model runs, and subplots (e) and (f) show the Saltelli-based TO obtained with 315 and 2450 model runs, respectively. The vertical lines on each subplot represent the corresponding "true" sensitivity metrics obtained using much larger numbers of model runs (100,000+). STAR-VARS is seen to be significantly more robust than the Saltelli-based Sobol approach. Further, subplots (e) and (f) show that in a large proportion of trials, the Sobol-TO estimate for different factors is (unfeasibly) negative, and therefore cause for concern.

for the entire factor space). The latter can be done using variograms and covariograms at any direction (including the directional variograms). Results of such calculations are not shown here.

4.2. Case Study 2: 45 Parameter Land Surface Scheme-Hydrology Model

Next we evaluate parameter sensitivities for the MESH modelling system [*Pietroniro et al.*, 2007], which couples the Canadian Land Surface Scheme (CLASS, see Figure 10) [*Verseghy*, 2000] with land-surface

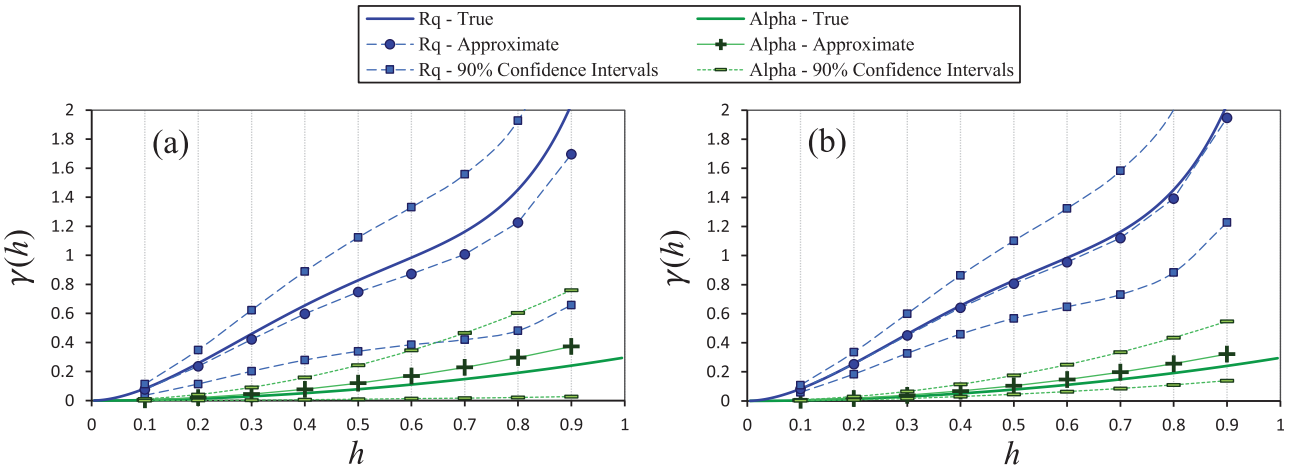

Figure 8. Illustration of how the bootstrap-based uncertainties associated with STAR-VARS for Case Study 1 change with computational budget. The subplots show directional variograms along with their 90% confidence intervals when using (a) 10 star centers (460 function evaluations) and (b) 50 star centers (2300 function evaluations)—True variograms are reproduced from Figure 4a - here in linear scale on vertical axis.

parameterization and hydrological routing schemes used by WATFLOOD [*Kouwen et al.*, 1993]. The study area is the White Gull basin with a drainage area of 603km^2, a research site of Boreal Ecosystems Atmosphere Study (BOREAS) located in Saskatchewan, Canada. The 45 (surface and subsurface) parameters of the model were calibrated by maximizing the Nash-Sutcliffe criterion with regards to streamflow [see *Mamo*, 2015, for details]; definitions of the parameters and their feasible ranges are reported in supporting information Table S1.

This case study provides a rigorous test of the *efficiency* of our approach under computationally intensive conditions, with each model run requiring approximately 30 s of computer time. To provide a fair assessment, we implement the VARS, Sobol, and Morris approaches using comparable computational budgets. For VARS, we use the star-based sampling strategy with 20,300 model runs (50 star centers and $\Delta h = 0.1$) and 101,500 model runs (250 star centers and $\Delta h = 0.1$). For Sobol, we correspondingly use 20,304 and 100,016 model runs, and for Morris we use 20,010 and 100,004 model runs and a 5% step size.

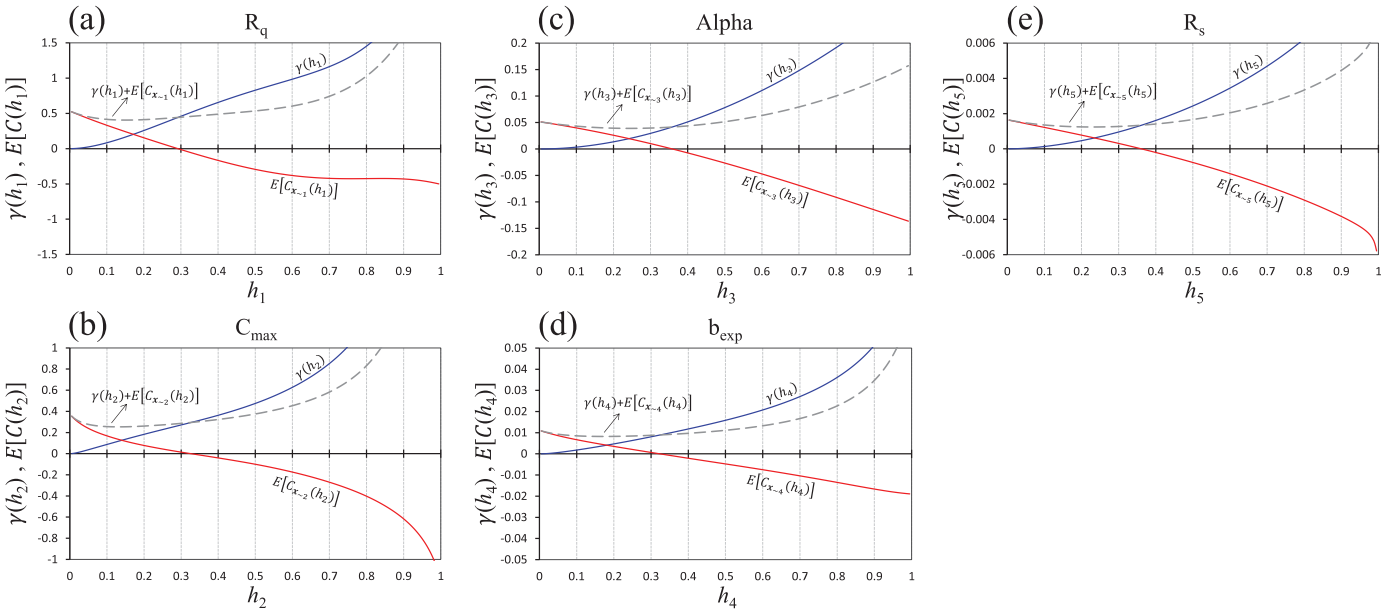

Figure 9. Demonstration of the link between VARS and Sobol for Case Study 1. Subplots are for factors (a) R_q, (b) C_{max}, (c) Alpha, (d) b_{exp}, and (e) R_s. The blue (dark) lines represent variograms, the red (light) lines represent average cross sectional covariograms, and the dashed lines represent their sum of their individual variance contributions to the total variance of response surface. Under the "constant mean" assumption (equation (2) of the companion paper), the dashed lines should be perfectly horizontal; in practical implementation this assumption may degrade at larger scales and so the variance contributions should be estimated using smaller values of h (in the limit as $h \to 0$ the assumption perfectly holds).

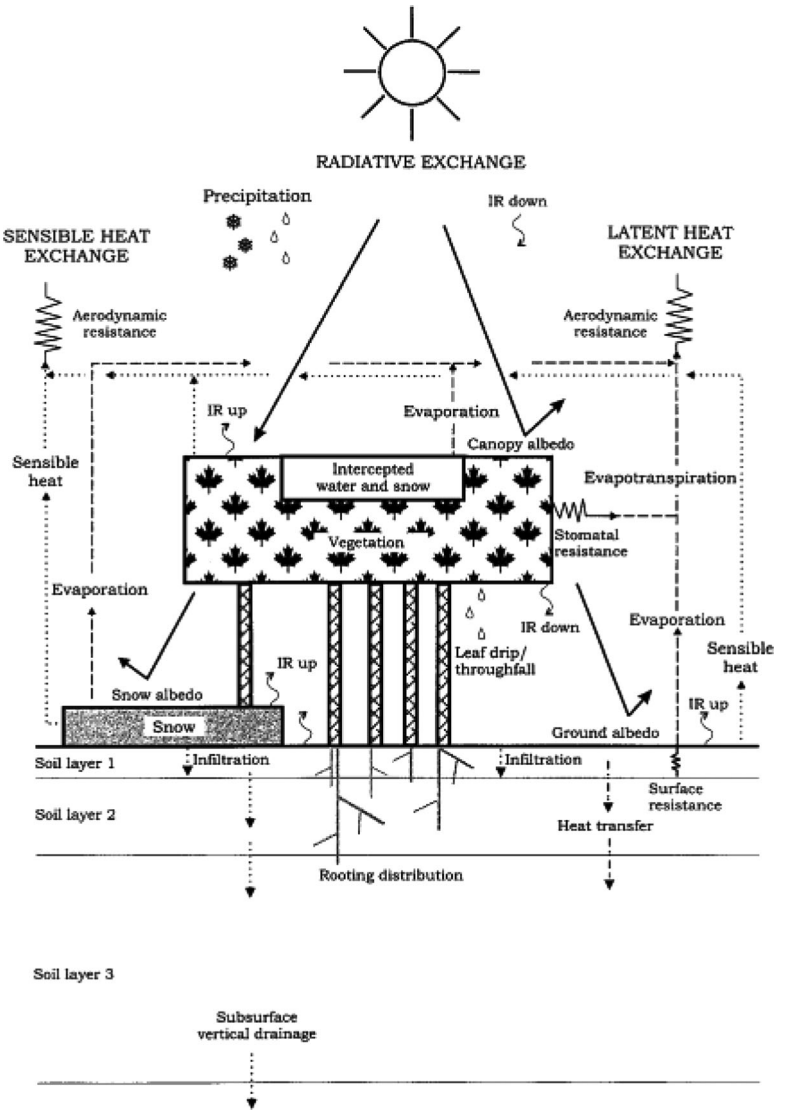

Figure 10. Schematic diagram of the Canadian Land Surface Scheme (CLASS) adapted from *Verseghy* [2000].

4.2.1. VARS Metrics and Their Consistency

Figure 11 shows the VARS assessment conducted using the larger computational budget (101,500 model runs). The first point to note (Figure 11a) is that all 45 of the estimated directional parameter variograms show quite simple forms; it is also clear that LAMIN4 is the least sensitive parameter. A second important point is that most of the parameter variograms cross one or more of the other parameter variograms, indicating different sensitivity rankings at different scales.

Next, Figure 11b shows the degree of correspondence between the IVARS sensitivity metrics evaluated at different scales and the VARS-based estimate of the Total-Order effect (VARS-TO). As expected, the ranking provided by the larger-scale metric $IVARS_{50}$ corresponds very well with that provided by VARS-TO, and the correspondence diminishes at shorter ranges of scale ($IVAR_{30}$ and $IVARS_{10}$). Interestingly, however, Figure 11c shows that VARS-TO and Sobol-TO, supposedly equivalent metrics but obtained via different numerical implementations, do not correspond well; in fact this figure indicates that the *Saltelli et al.* [2008] implementation is unable to effectively differentiate between more than half of the parameters, in spite of the large computational budget used (>100,000 model runs). Further, as discussed earlier, the Sobol total-order effect estimates are (unfeasibly) negative for many of the parameters, suggesting problems of numerical stability.

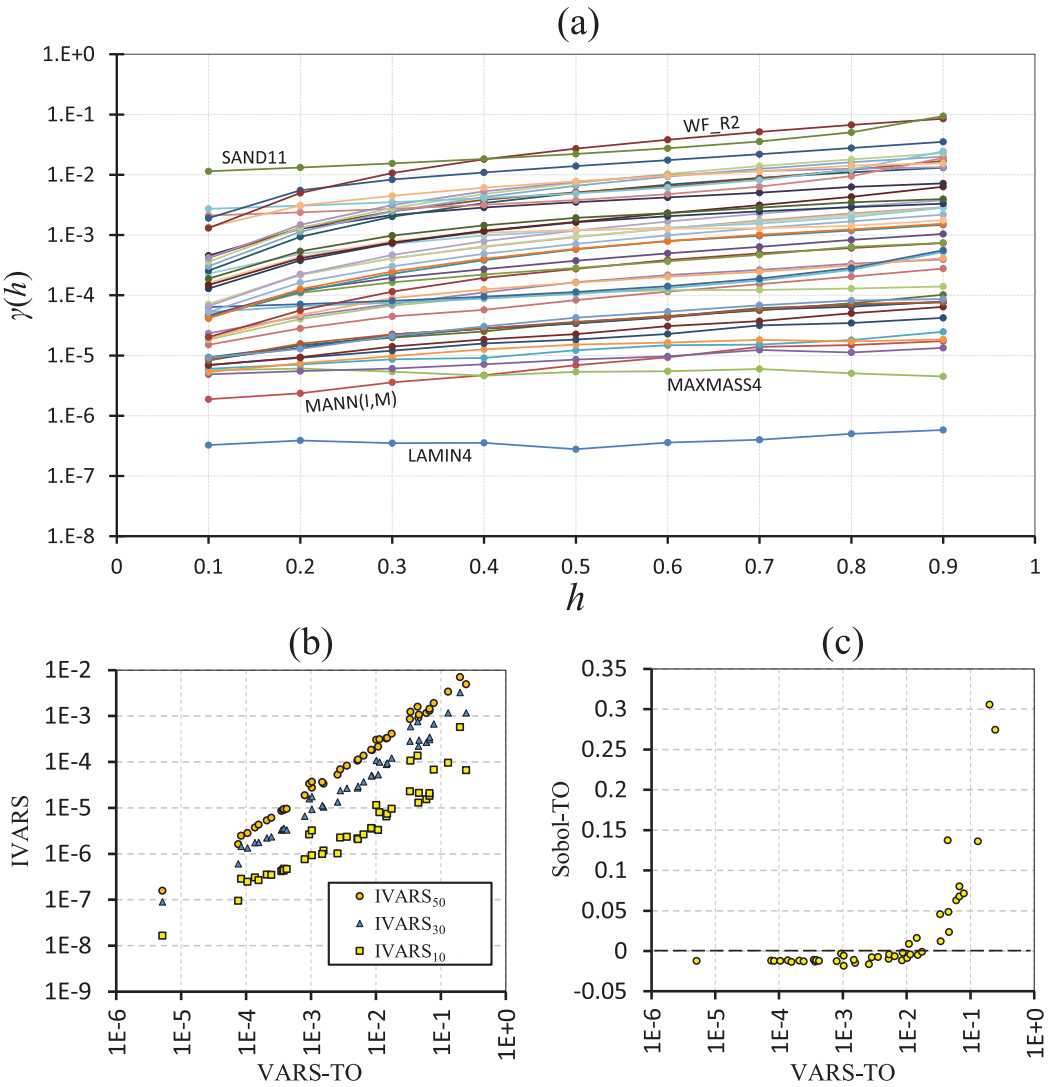

Figure 11. Illustration of the performance of STAR-VARS for the 45-factor computationally intensive physically based land surface hydrology model of Case Study 2. Subplot (a) shows the estimated directional variograms. Labels shown for some of the variograms correspond to parameter names. Subplot (b) is a log-log scatter plot of the IVARS metrics at various scales ($IVARS_{10}$, $IVARS_{30}$, and $IVARS_{50}$) versus VARS-TO. Subplot (c) is a log-linear scatter plot of Sobol-TO versus VARS-TO. The VARS and Sobol experiments were conducted with 101,500 function evaluations (250 star centers and Δh=0.1) and 100,015 function evaluations, respectively.

4.2.2. Uncertainty and Reliability of VARS Metrics

To assess reliability, Figures 12a and 12c show bootstrap-based 90% confidence intervals for the VARS-based sensitivity metrics for each parameter (using parameter ordering based on VARS rankings at the larger computational budget). We see that the confidence intervals are relatively narrow, and that their widths decrease with increased computational budget. Figures 12b and 12d show that the reliabilities of the parameter rankings improves from 4 to 100% (median = 23%) for the smaller computational budget to 10–100% (median = 33%) for the larger budget.

Together, these results indicate that while the metric estimates are fairly precise, the large number of parameters having similar sensitivities can cause the parameter rankings to vary considerably; i.e., one should not be too particular about the actual ranking obtained, and instead consider whether a parameter is high or low (in a relative sense) in the entire group. For example, it is clear that LAMIN4 (minimum leaf area index of Grass) and MANN(I,M) (Manning's constant for overland flow) are among the least sensitive parameters, while WF_R2 (river roughness factor) and SAND11 (% of sand soil layer 1) are among the most sensitive ones.

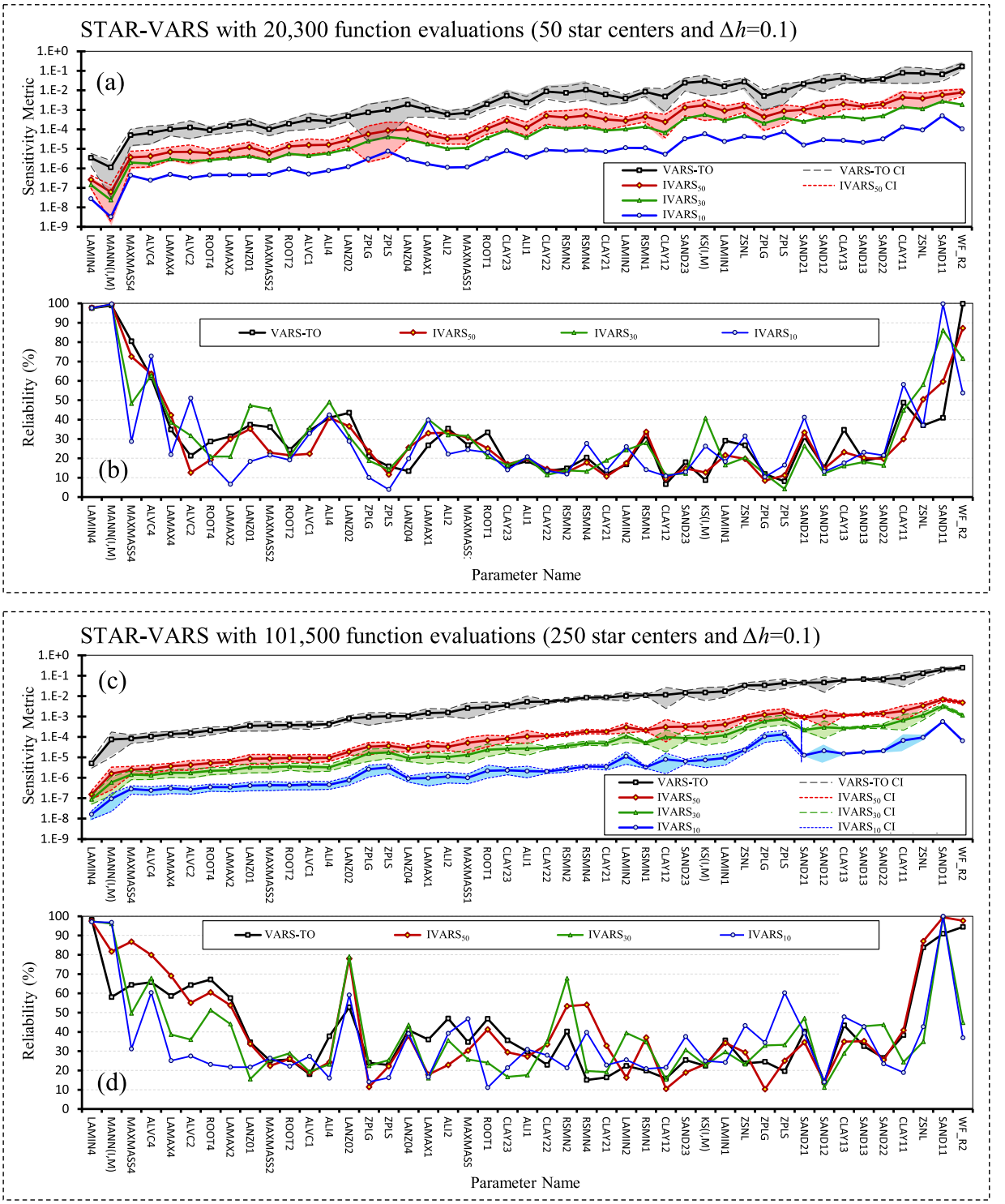

Figure 12. An assessment of uncertainty and reliability of STAR-VARS for Case Study 2. The first two rows are based on a computational budget of 20,300 function evaluations (50 star centers, $\Delta h = 0.1$) and the next two rows are based on a computational budget of 101,500 function evaluations (250 star centers, $\Delta h = 0.1$). Subplots (a) and (c) show VARS metrics for the 45 factors and their bootstrap-based 90% confidence intervals (CI). Subplots (b) and (d) show the reliability estimates of factor rankings. The factors are ordered according to their ranks based on VARS-TO obtained with the larger computational budget.

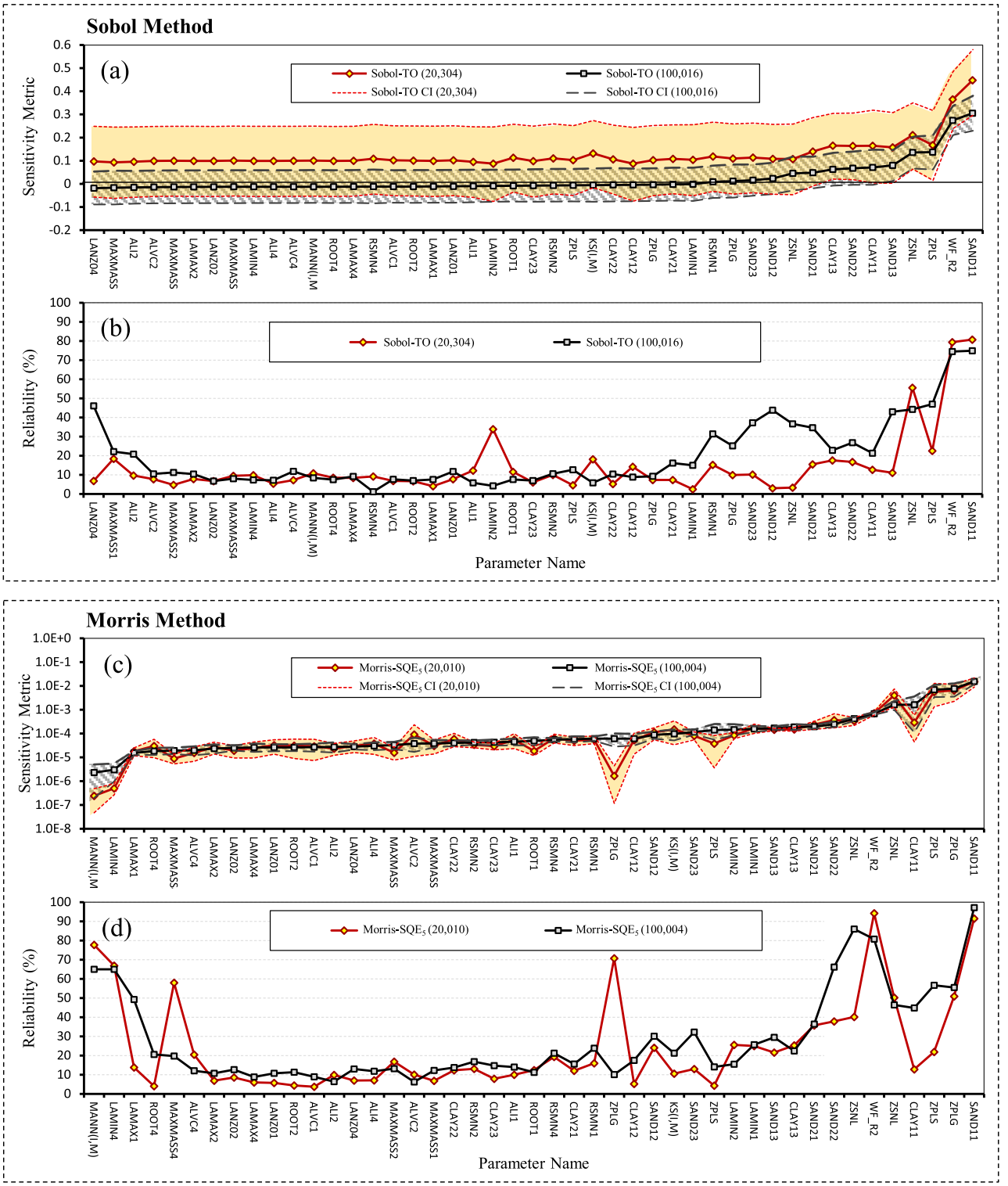

Figure 13. An assessment of uncertainty and reliability of the Sobol and Morris methods for Case Study 2. Subplot (a) shows Sobol-TO for the 45 factors and their bootstrap-based confidence intervals using the two computational budgets of 20,304 and 100,016 function evaluations. Subplot (b) shows corresponding reliability estimates for the Sobol-TO factor rankings. Subplot (c) shows Morris-SQE_5 for the 45 factors and their bootstrap-based confidence intervals using the two computational budgets of 20,010 and 100,004 function evaluations. Subplot (d) shows corresponding reliability estimates for the Morris-SQE_5 factor rankings. For each metric, the factors are ordered according to their ranks based on the larger function evaluations.

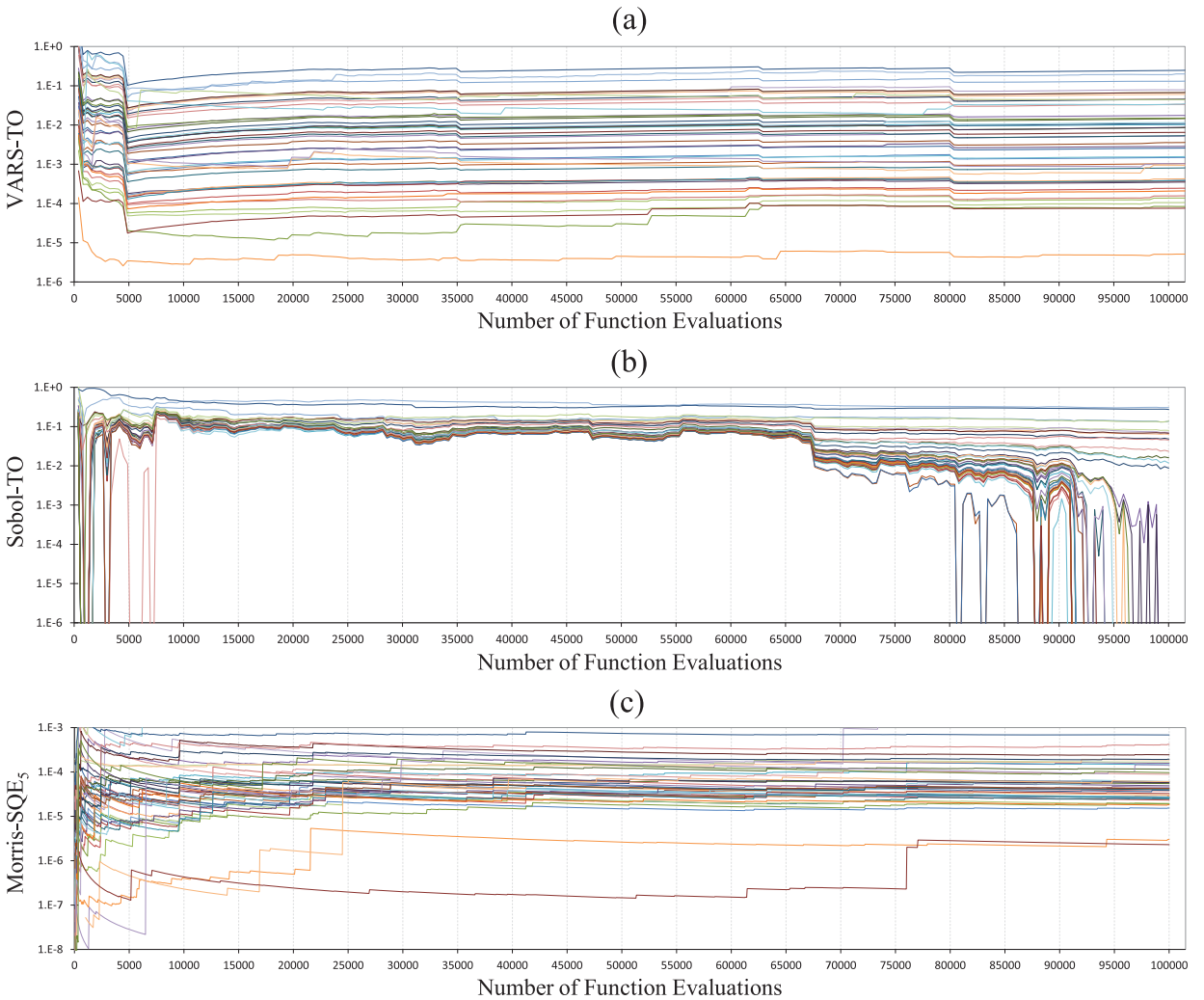

Figure 14. Illustration of the relative stability and convergence of (a) STAR-VARS, (b) Sobol, and (c) Morris methods with increasing computational budget for the 45 factors in Case Study 2. In subplot (b) the values of Sobol-TO that go below 10^-6 were computed as being negative.

For comparison, Figure 13 facilitates a similar assessment for the Sobol and Morris approaches (note that the parameter orders are different, each being based on their respective rankings at the larger computational budget). For Sobol, although the uncertainty decreases with increasing computational budget, it remains relatively large compared to both VARS and Morris, and the lower bounds of the confidence intervals are (undesirably) at negative values for the majority of the parameters. This translates into very low reliability estimates for the parameter rankings (median = 9% and 11% at lower and higher computational budgets, respectively). Accordingly, it seems clear that *much larger* computational budgets will be needed for the Sobol-based assessment to become sufficiently credible. For Morris, the uncertainty is substantially less than that obtained using Sobol, but remains larger than that obtained using VARS. Accordingly, the Morris-based reliability estimates for parameter rankings (median = 13% and 17% at lower and higher computational budgets, respectively) are generally lower than those obtained using VARS (medians = 23% and 33% respectively).

4.2.3. Stability and Convergence: The Evolution of STAR-VARS

Finally Figures 14a–14c compare the stability and robustness of the tested implementations of VARS, Sobol, and Morris with increasing computational budget. While the VARS-based estimate of total-order effect (VARS-TO) stabilizes after about 5000 function evaluations, the Saltelli-based estimate (Sobol-TO) has yet to stabilize even after 100,000 function evaluations. These results suggest that the Saltelli-based implementation of the Sobol approach may not provide reliable results for high-dimensional problems unless excessively large computational budgets are used. Overall, we find that the STAR-VARS approach is at least 20

times (100,000/5,000) more efficient than the Saltelli-based implementation of Sobol—and probably even much more so. Similarly, the implementation of Morris is more robust and stable than the Saltelli-based Sobol. Strangely, however, its factor rankings change significantly until around 40,000 model runs, and continue to change for the entire range of computational budgets tested (see also supporting information Figure S2). In practice of course (but depending on the application), small differences in parameter ranking may be of little concern, particularly for high-dimensional problems.

5. Concluding Remarks

In this paper, we develop and test a practical (numerical) implementation of the VARS framework for global sensitivity analysis for which the theoretical basis was presented in the companion paper, *Razavi and Gupta* [2016]. Our implementation includes (1) a star-based sampling strategy (called STAR), and (2) a bootstrap strategy to compute estimates of confidence intervals for each sensitivity measure and to provide reliability assessments for the inferred factors rankings.

Effectiveness, efficiency, and robustness of the STAR-VARS approach was demonstrated via two carefully selected case studies, the first involving a five parameter conceptual Rainfall-Runoff model, and the second involving a 45 parameter Land Surface Scheme Hydrology model. For benchmark comparison, we also provided results using the *Saltelli et al*. [2008] state-of-the-art implementation of the Sobol approach and the conventional implementation of the Morris approach described in *Campolongo et al*. [2007].

Our results show the STAR-VARS implementation provides consistent, reliable, and robust estimates of factor sensitivity across a range of scales, and at relatively low computational cost. As an added bonus, it also provides accurate estimates of the Sobol total-order effects and the Morris elementary effects. For these two cases, STAR-VARS was found to be 20+ times more efficient than the tested implementations of Sobol and Morris. Case Study 2, in particular, provided a powerful illustration of the relative strength of STAR-VARS over Sobol for high-dimensional EESMs, demonstrating the ability of the former to provide relatively reliable estimates of parameter sensitivity with relatively low computational budgets. Our results also suggest that, in practice, the Saltelli-based implementation of the Sobol method may not be able to provide reliable results for high-dimensional problems of this kind.

The following points summarize the major contributions of the VARS framework:

1. VARS provides a comprehensive spectrum of information about the underlying sensitivities of a response surface to its factors.
2. This framework is *unique* in that it characterizes a variety of sensitivity-related properties of response surfaces including local sensitivities and their global distribution, the global distribution of model responses, and the structural organization of the response surface.
3. VARS effectively tackles the issue of scale by providing sensitivity information spanning a range of scales, from small-scale features such as roughness and noise, to large-scale features such as multimodality.
4. VARS has a clear theoretical relationship with the variance-based (Sobol) and derivative-based (Morris) approaches to sensitivity analysis, both of which are shown to be special cases.
5. STAR-VARS (VARS implemented using star-based sampling) can efficiently provide reliable estimates of the IVARS sensitivity metrics as well as of the variance-based total-order effects and derivative-based elementary effects across a range of scales.
6. STAR-VARS is both computationally efficient and statistically robust, even for high-dimensional response surfaces, providing stable estimates with relatively small computational budgets. This computational efficiency is, in part, due to VARS being based on the information contained in *pairs* of points, rather than in *individual* points.

While the results presented here are promising, it is clear that much more work needs to be done to better understand the strengths and limitations of the VARS framework, and how well it functions in a variety of situations. There may be strategies that can be implemented to further enhance its relative efficiency, an issue that will be increasingly important as EESMs become progressively more complex and realistic and the number of factors (e.g., parameters) to be studied increases. We mention a few of these issues below:

1. The approach needs to be tested on a broad range of other models and problems (something we hope the community will help us to explore).

2. It will be useful to investigate other sampling strategies (experimental designs) to see if the accuracy, efficiency and robustness of the sensitivity analysis results can be improved (work in progress).
3. To assess the accuracy and reliability of the bootstrapping strategy to providing efficient assessment of confidence, the strategy should be compared to the use of actual resampling from the original problem space, which is much more computationally intensive (work in progress).
4. VARS should be compared with other sensitivity analysis approaches, including the recently proposed DELSA [*Rakovec et al*., 2014] and PAWN [*Pianosi and Wagener*, 2015].

Equally, or perhaps more important, the effectiveness of an EESM parameter sensitivity analysis approach depends critically on the careful design/choice of the metric (or metrics) used to assess model performance [see *Gupta et al*., 2008, for a critical discussion]. In this context, *Rosolem et al*. [2012] recently demonstrated the importance of implementing parameter sensitivity analysis in a multiple-criteria context, particularly when the EESM generates a variety of different fluxes as outputs. Our future work will explore the issue of metrics in more depth.

In conclusion, we hope that this two-paper contribution will help prompt a greater level of interest in the problem of sensitivity analysis, and in issues of model development and system identification. As always we invite dialog with others interested in these and related issues. A copy of the VARS computer code can be obtained from the first author upon request for use in noncommercial applications (for commercial use, please contact The University of Arizona's Tech Launch Arizona).

Acknowledgments
The first author is thankful to the University of Saskatchewan's Global Institute for Water Security and Howard Wheater, the Canada Excellence Research Chair in Water Security, for encouragement and support. The second author received partial support from the Australian Research Council through the Centre of Excellence for Climate System Science (grant CE110001028), and from the EU-funded project "Sustainable Water Action (SWAN): Building Research Links Between EU and US" (INCO-20011-7.6 grant 294947). We are thankful to Moges Mamo, Andrew Ireson, and Bruce Davison for providing us with the second case study. The data used to support this paper are available upon request from the first author.

References

Campolongo, F., J. Cariboni, and A. Saltelli (2007), An effective screening design for sensitivity analysis of large models, *Environ. modell. software*, 22(10), 1509–1518.

Efron, B., and R. J. Tibshirani (1994), *An Introduction to the Bootstrap*, Taylor and Francis, N. Y.

Gupta, H. V., T. Wagener, and Y. Q. Liu (2008), Reconciling theory with observations: Elements of a diagnostic approach to model evaluation, *Hydrol. Processes*, *22*(18), 3802–3813.

Kouwen, N., E. Soulis, A. Pietroniro, J. Donald, and R. Harrington (1993), Grouped response units for distributed hydrologic modeling, *J. Water Resour. Plann. Manage.*, *119*(3), 289–305.

Mamo, M. T. (2015), *Exploring the ability of a distributed hydrological land surface model in simulating hydrological processes in the boreal forest environment*, MSc thesis, University of Saskatchewan, Saskatoon, SK, Canada, 153 pp.

Morris, M. D. (1991), Factorial sampling plans for preliminary computational experiments, *Technometrics*, *33*(2), 161–174.

Pianosi, F., and T. Wagener (2015), A simple and efficient method for global sensitivity analysis based on cumulative distribution functions, *Environ. Modell. Software*, *67*, 1–11.

Pietroniro, A., V. Fortin, N. Kouwen, C. Neal, R. Turcotte, B. Davison, D. Verseghy, E. Soulis, R. Caldwell, and N. Evora (2007), Development of the MESH modelling system for hydrological ensemble forecasting of the Laurentian Great Lakes at the regional scale, *Hydrol. Earth Syst. Sci.*, *11*(4), 1279–1294.

Rakovec, O., M. C. Hill, M. P. Clark, A. H. Weerts, A. J. Teuling, and R. Uijlenhoet (2014), Distributed Evaluation of Local Sensitivity Analysis (DELSA), with application to hydrologic models, *Water Resour. Res.*, *50*, 409–426, doi:10.1002/2013WR014063.

Razavi, S., and H. V. Gupta (2015), What do we mean by sensitivity analysis? The need for comprehensive characterization of 'Global' sensitivity in Earth and Environmental Systems Models, *Water Resour. Res.*, *51*, 3070–3092, doi:10.1002/2014WR016527.

Razavi, S., and H. V. Gupta (2016), A new framework for comprehensive, robust, and efficient global sensitivity analysis: 1. Theory, *Water Resour. Res.*, *52*, doi:10.1002/2015WR017558.

Rosolem, R., H. V. Gupta, W. J. Shuttleworth, X. Zeng, and L. G. G. Gonçalves (2012), A fully multiple-criteria implementation of the Sobol' method for parameter sensitivity analysis, *J. Geophys. Res.*, *117*, D07103, doi:10.1029/2011JD016355.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola (2008), *Global Sensitivity Analysis: The Primer*, John Wiley, Hoboken, N. J.

Sobol', I. M. (2001), Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates, *Math. Comput. Simul.*, *55*(1–3), 271–280.

Verseghy, D. L. (2000), The Canadian land surface scheme (CLASS): Its history and future, *Atmos. Ocean*, *38*(1), 1–13.

Vrugt, J. A., H. V. Gupta, W. Bouten, and S. Sorooshian (2003), A shuffled complex evolution metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters, *Water Resour. Res.*, *39*(8), 1201, doi:10.1029/2002WR001642.