

Dear editor,

We thank you for the constructive assessment and for the additional comments, which have helped us improve the clarity and consistency of the manuscript. We have addressed all nine comments as detailed below.

In addition, we have updated the affiliations and acknowledgements of the manuscript. A third affiliation has been added for the first author (Department of Applied Physics, University of Granada, Granada, Spain), reflecting his current funding institution. The acknowledgements have been revised accordingly to include the UGR–Banco Santander postdoctoral fellowship.

On behalf of all the authors,

Patricio Yeste

- Thank you for justifying the length of the technical note. I however think that Tables 1,2,3 are not really needed to understand the analyses presented in the manuscript but they are needed for reproducibility and transparency. They could therefore be moved into the supplementary material.

We thank the editor for this suggestion and agree with it. The definitions of the model parameters and their feasible ranges are not required to follow our analyses, and we agree that the Supplementary Material is the appropriate place for this information. We have moved Tables 1–3 there, where they appear as Tables S1–S3.

In addition, we have decided to move the text in Appendix A (*Treatment of parameter dependence in the VIC model*) to the Supplementary Material. This appendix documents how the structural dependency between two VIC soil parameters was handled. As these details are required for reproducibility rather than to follow the analyses, they naturally belong in the Supplementary Material together with the parameter tables.

- paragraph starting L29: it would be worth mentioning that one of the limit of the variance-based approach is that it relies on the assumption that variance is an adequate proxy for uncertainty (as discussed e.g. in Borgonovo et al., 2011).

We thank the editor for this suggestion. We have added a sentence in the Introduction acknowledging that a limitation of variance-based approaches is their reliance on the assumption that the variance of the model output is an adequate proxy for its uncertainty, citing Borgonovo et al. (2011) in this context. The added sentence reads:

"A limitation of variance-based approaches is that they rely on the assumption that the variance of the model output is an adequate proxy for its uncertainty, which has motivated the development of moment-independent importance measures (Borgonovo et al., 2011)."

The corresponding reference has been added to the reference list.

- L74: it would be clearer to define the number of parameter "p" here (first occurrence).

We agree with the editor. We have defined p as the number of uncertain input factors at its first occurrence in Section 2.1, where the input vector $X = (X_1, \dots, X_p)$ is now described as "a vector of p uncertain input factors". The term "input factors" is widely used in the GSA literature (e.g., Saltelli et al., 2010). Note that, in the applications of this technical note, these input factors correspond to the model parameters of the hydrologic models (Section 3).

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., & Tarantola, S. (2010). Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Computer Physics Communications*, 181(2), 259–270. <https://doi.org/10.1016/j.cpc.2009.09.018>

- L195: it is probably debatable to consider VIC model as a conceptual model, it is rather physically-based.

We thank the editor for this comment. We have revised the classification of VIC throughout the manuscript, so that it is now consistently described as a physically based land-surface model, while the term "conceptual" is reserved for the lumped rainfall–runoff models HBV and HyMod.

- L256: Does this mean that the SHAP values account for the main effect? In this case, how can we interpret the results of the comparison between the SHAP value and the other total sensitivity indices in Section 4?

We thank the editor for this comment. As discussed in our responses to the reviewers, SHAP_{*i*} operates as a local prediction-explanation method that attributes the contribution of each input to individual predictions. It does not quantify main effects in the ANOVA sense. When the absolute SHAP_{*i*} values are aggregated over a representative sample, as done in our work through the mean |SHAP_{*i*}|, they provide a global measure of variable influence that reflects the importance of each parameter to the model output of interest. As noted in our response to Reviewer #1, no formal theoretical connection exists between SHAP_{*i*} and T_i , unlike the case of PVI_{*i*} (Equation 8). Therefore, mean |SHAP_{*i*}| should be interpreted as a complementary global sensitivity measure and not as a variance decomposition index. Figures 4 and 5 show that mean |SHAP_{*i*}| preserves the ranking of influential parameters across models, consistent with this complementary interpretation. A mathematical definition of SHAP_{*i*} and further details have been provided in Appendix A of the revised manuscript as per the suggestion of Reviewer #2.

- L296: It would be worth mentioning that the sub-samples are dependent, which could result in smoother variations in the sensitivity indices with the sample size compared to using independent samples of different sizes.

We thank the editor for this observation. The 10,000 Monte Carlo samples used in this study are drawn from independent uniform distributions over the parameter ranges, and therefore any subsample drawn from this set retains the independence of the individual samples. The overlap between consecutive subsamples of increasing size is an inherent property of subsampling from a finite pool rather than a statistical dependence between the draws. As the subsample size approaches the full set, this overlap naturally increases, which contributes to the smoother variations observed at larger sample sizes. The purpose of the convergence test is to verify that the full set of 10,000 samples is sufficient for the sensitivity estimates to stabilize, rather than to identify an exact sample size at which convergence is achieved. The fact that the sensitivity estimates stabilize well below 10,000, where the overlap between consecutive draws is still moderate, indicates that the convergence behaviour reflects genuine stabilization rather than an artefact of sample overlap.

We have clarified in Section 3.5 that each subsample was drawn as an independent random sample without replacement from the full set of Monte Carlo simulations, and we have revised the description of the convergence test to better explain its purpose, which was to verify that the full set of 10,000 samples was sufficient for the sensitivity estimates to stabilize. We have also acknowledged the overlap between consecutive subsamples as an inherent property of subsampling from a finite pool.

- Figure 3: is the convergence metric the average of the Euclidean distance across the model parameters?

We thank the editor for this question. The convergence metric is not the average of the Euclidean distance across model parameters. As described in Section 3.5, the sensitivity estimates corresponding to each subsampling matrix were normalized by dividing by the sum of all sensitivities, and the Euclidean distance between the resulting normalized sensitivity vectors at consecutive subsample sizes was used as the convergence metric. This is also stated in the caption of Figure 3: "The convergence metric is computed as the Euclidean distance between

normalized sensitivity vectors obtained from consecutive subsample sizes. Normalization was performed by dividing by the sum of all sensitivities."

- L325: What is the sample size at which convergence is reached?

We thank the editor for this comment. As discussed in our response to the comment regarding the subsampling strategy (L296), the purpose of the convergence test was to verify that the full set of 10,000 samples was sufficient for the sensitivity estimates to stabilize, rather than to identify an exact sample size at which convergence was achieved. As shown in Figure 3, the sensitivity estimates stabilized at subsample sizes well below 10,000 for all models, methods, and catchments considered, confirming that the sample size used in this study was sufficient. We have added a statement in Section 4.1 to clarify this.

- Some clarification is needed regarding the runtime. In the response to Reviewer #3 (comment on L45), it is written that "Training a metamodel on this scalar input–output relationship is computationally negligible, even for large sample sizes." However, I can see from the description document of the data and code (Table 3) that running the scripts for RF and NN takes much longer than for Sobol'.

Note: the editor referred to "RF" and "Table 3" in the description; in the revised version of the repository description, the metamodel is labelled as GBDT (Gradient Boosting Decision Tree), and the table reporting the GSA script execution times is Table 4.

We thank the editor for pointing out this apparent inconsistency. The statement that training a metamodel on the scalar input–output relationship is "computationally negligible" refers specifically to the training of the metamodel itself: fitting a GBDT, NN, or LM on 10,000 samples with p input variables to predict a single scalar output (KGE). The wall-clock execution times reported in Table 4 of the repository description correspond to the complete GSA scripts, which include not only metamodel training but also the computation of PVI_i (permutation importance across all parameters and samples), $SHAP_i$ values for all samples, and the estimation of confidence intervals through 1000 bootstrap resamples. The latter was introduced in response to Reviewer 3 and is currently the dominant component of the overall runtime, especially for the NN metamodel, where PVI_i is recomputed via scikit-learn on CPU during each bootstrap iteration. The convergence scripts are even more demanding, as they require repeated training and evaluation of the metamodels across 100 subsample sizes, but these serve as a validation exercise rather than part of the proposed workflow.

We note that the execution time reported for the Sobol' script in Table 4 corresponds solely to the application of the Sobol' estimator to the pre-computed Monte Carlo outputs, and does not include the cost of the model simulations themselves. The actual computational cost of Sobol' GSA is dominated by the Monte Carlo experiment, which requires $n(p+2)$ evaluations of the original hydrologic model and is reported in Table 3 of the repository description. For VIC, this experiment took 53.74 hours on an HPC cluster, whereas the GBDT-based GSA script, including all PVI_i and $SHAP_i$ computations and the bootstrap confidence intervals, completed in 12.69 minutes on a GPU. For HBV and HyMod, the Monte Carlo experiments were considerably cheaper (7.26 and 2.06 minutes, respectively), so the computational advantage of the metamodel is less pronounced for these models. This is consistent with the fact that the advantage of the proposed framework scales with the computational cost of the original model: running a metamodel is substantially more efficient than running a physically based model such as VIC, whereas for simpler conceptual models such as HBV and HyMod the difference is smaller.

References

Borgonovo, E., Castaings, W. and Tarantola, S. (2011), Moment Independent Importance Measures: New Results and Analytical Test Cases. Risk Analysis, 31: 404-428. <https://doi.org/10.1111/j.1539-6924.2010.01519.x>