Reviewer #3

This technical note explores the interest of using machine learning metamodels (surrogates) to evaluate global sensitivity indices. It compares several methods of ML (RF, NN) with linear regression and Sobol SHAPi indices. The paper is clear, includes a lot of work, and is worth published as a technical note, although too long see in HESS Manuscript types description: Manuscripts of this type should be short (a few pages only).

My major comments are thus on the length of the paper, on the hypothesis of this work, and on the chosen methods (evaluation only on the learning sampling, independance of parameters, choice of SHAPi to get Sobol indices).

We thank Reviewer #3 for the positive assessment and the constructive feedback, which we address in detail in the following responses.

Regarding the length of the manuscript, we acknowledge that it exceeds the typical length suggested for technical notes in the HESS guidelines. During submission, we considered two possibilities: submitting as a research article or as a technical note. We decided to submit as a technical note for two reasons. First, because of the technical nature of the work. Second, although the guidelines suggest that technical notes should be short, the journal has shown flexibility with this category in practice. There are recent examples of technical notes accepted by HESS that are comparable in length to or even longer than our manuscript (e.g., Lehr & Hohenbrink, 2025, 32-page preprint; Marin-Ramirez et al., 2026, 28-page preprint; and Ruzzante et al., 2026, 27-page preprint. In none of these cases did the reviewers raise any concern about the length of the manuscript during the Open Discussion). We therefore believe the current length is appropriate for the category. While we are open to a change of manuscript typology to research article if the editorial service considers this appropriate, we note that the individual sections suggested for trimming elsewhere in this review report are informative and serve distinct purposes in the manuscript rather than being redundant or superfluous.

Lehr, C., & Hohenbrink, T. L. (2025). Technical note: An illustrative introduction to the domain dependence of spatial principal component patterns. Hydrology and Earth System Sciences, 29(22), 6735–6760. https://doi.org/10.5194/hess-29-6735-2025

Marin-Ramirez, A., Mahoney, D. T., & McDaniel, G. (2026). Technical note: Analysis of concentration-discharge hysteresis loops using Self-Organizing Maps. Hydrology and Earth System Sciences, 30(5), 1359–1379. https://doi.org/10.5194/hess-30-1359-2026

Ruzzante, S. W., Knoben, W. J. M., Wagener, T., Gleeson, T., & Schnorbus, M. (2026). Technical note: High Nash–Sutcliffe Efficiencies conceal poor simulations of interannual variance in seasonal regimes. Hydrology and Earth System Sciences, 30(8), 2337–2355. https://doi.org/10.5194/hess-30-2337-2026

In particular, two things are said many times that are debatable and are linked together:

- This is said most often that the aim is not to extrapolate the surrogate outside of the learning sample, and thus not to validate it on a test (independent) sampling. This choice is much debatable, because this is sure that you will get very high R^2 on the learning sample, at least with many points, and especially if you evaluate on the same points that were used to learn (whatever the number of points in the sample). We lose the interest of a surrogate that could be used on another sampling (even in same conditions). This is said and honestly assumed by the authors, but I disagree with this point.

We appreciate the reviewer's perspective and acknowledge that evaluating surrogates on the same data used for training will naturally yield high R^2. However, we would like to clarify that the purpose of the metamodels in this study is fundamentally different from a predictive setting where generalizability to unseen data is the objective.

The metamodels serve as an intermediate tool to facilitate the estimation of PVI_i. PVI_i could be computed directly on the hydrologic model, but this would require evaluating the model on perturbed input matrices (the cross-sample matrices B and AB_i in the Sobol framework could serve for this purpose as they mimic the independent replication process followed to estimate PVI_i, see Section 2.3), which demands $n(p+1)$ additional model evaluations. The metamodel replaces these additional hydrologic model evaluations: once trained on the n samples of matrix A, the surrogate is used to predict outputs (i.e., KGE values) for the perturbed inputs needed for PVI_i computation, without running the hydrologic model again. Importantly, while these perturbed inputs do not constitute a formal independent test set, they are combinations the metamodel has not been trained on. The fact that the resulting PVI_i estimates successfully approximate T_i by virtue of Eq. 8 (Fig. 5) thus provides evidence that the surrogate is performing adequately outside the training set, effectively serving as a de facto validation of the metamodel's ability to generalize to the perturbed inputs. This is where the computational advantage arises (see also our response to the next comment), and it is the reason a train/test split is not relevant to our use case.

This approach follows established practice in the GSA–ML literature. Antoniadis et al. (2021), in their comprehensive review of random forests for GSA, do not require out-of-sample validation because the surrogate is used for sensitivity estimation within the training space, not for prediction. Similarly, Wei et al. (2015), who established the theoretical relationship between T_i and PVI_i that underpins this work, compute PVI_i from the learned model without a separate test set. In both cases, the validity of the approach is assessed by comparing the resulting sensitivity estimates against reference values, which is precisely the approach followed in our work (see Figs. 4 and 5).

We agree that if the surrogate were intended for use beyond PVI_i estimation (e.g., to explore uncertainties, support calibration, or transfer to other catchments) then independent validation would be essential. This is an interesting direction for future work, but it is beyond the scope of the present study, where the surrogate's sole purpose is to enable efficient PVI_i computation via Eq. 8 and $SHAP_i$ estimation. We will add a sentence in the conclusions acknowledging this direction for future work.

Antoniadis, A., Lambert-Lacroix, S., & Poggi, J.-M. (2021). Random forests for global sensitivity analysis: A selective review. Reliability Engineering & System Safety, 206, 107312. https://doi.org/10.1016/j.ress.2020.107312

Wei, P., Lu, Z., & Song, J. (2015). A comprehensive comparison of two variable importance analysis techniques in high dimensions: Application to an environmental multi-indicators system. Environmental Modelling & Software, 70, 178–190. https://doi.org/10.1016/j.envsoft.2015.04.015

- The interest of surrogates runs being cheaper is highlighted but, to build the surrogate, the authors use a sampling (matrix A) that was used to get the Sobol indices as well, and run it on it only (for the reason given above). So, the interest of reducing the cost is not exploited and not convincing here, since you don't use the surrogate much more to get any other indices or any other simulations, explore uncertainties, calibration or else. One convincing interest could have been to learn on one catchment, validate on another one (or an independent sampling) and/or use this surrogate on the other catchments. But if I understood well, you learnt the ML surrogates separately on each catchment and you run them on these same catchments, just to evaluate PVI.

The computational advantage of our approach, as explained in our previous response, arises from the fact that Eq. 8 allows PVI_i to serve as an estimator of T_i using the ML metamodel instead of the hydrologic model. The key comparison is as follows:

- Classical Sobol estimation of T_i requires $n(p+2)$ model evaluations: n for matrix A, n for matrix B, and np for the hybrid matrices AB_i (one per parameter).
- Our ML-based approach requires only n model evaluations (matrix A) to train the surrogate. PVI_i is then computed on the surrogate for all p parameters, replacing the $n(p+1)$

evaluations associated with matrices B and AB_i. The surrogate is used to predict the outputs of these perturbed inputs and not the hydrologic model.

Thus, the cost reduction is $n(p+2) \rightarrow n$, which is substantial and becomes as more substantial as the number of parameters p grows. The reviewer is correct that the same sampling matrix A is used for both Sobol analysis and surrogate training, but this is precisely the point: matrix A serves double duty. It provides the hydrologic model evaluations needed to train the surrogate, and it also serves as the base sample for the Sobol analysis. The savings come from eliminating the need for the hydrologic model on matrices B and AB_i.

Regarding the suggestion to learn on one catchment and validate on another: we trained metamodels separately for each catchment because the hydrologic models and their parameter spaces are catchment-specific. The surrogate for a given catchment captures the input–output relationship of the hydrologic model as applied to that catchment's forcing data. Transferring a surrogate across catchments would require either a shared parameterization or a different experimental design, which is an interesting possibility for future work but outside the scope of this study. As noted in our response to the previous comment, we will add a sentence in the conclusions acknowledging this direction for future work.

However , I think that the aim to evaluate the validity of equation 8 in this conceptual hydrological context is much interesting and has to be done as you did. This may be a more important point of the paper than claiming cheaper cost thanks to ML wich is not convincing the way you use them. Plus, to discuss about the cost, that would be very interesting to give the numerical cost of all methods, runs, for the 3 models and 3 catchments (at least to justify the use of 3 catchments that we don't see in the results which are given only on one).

We thank the reviewer for acknowledging the interest of evaluating the validity of Eq. 8 in a hydrologic context. We agree that this is a central contribution of the paper, and we note that the computational advantage described in our previous responses is a direct consequence of this relationship. To reinforce this point, we will add clarifying text to the last paragraph of Section 2.4, making the cost reduction ($n(p+2) \rightarrow n$) explicit there rather than only in the conclusions.

Regarding the numerical cost of the methods, the description pdf of the repository accompanying our work already includes detailed information on the computational setup and execution times. We will additionally include the run times of the three hydrologic models (HBV, HyMod, and VIC). We note that, since the models are run in a lumped configuration, the computational costs are nearly identical across catchments for any given model.

As for the use of three catchments, the purpose is not related to computational cost but to testing the generalizability of the methods across different physiographic and hydroclimatic conditions. Only the results for one of the catchments is shown in the main text for conciseness, while results for the other two catchments are presented in the Supplementary Material.

I have some more detailed comments and questions below.

Section 1 Introduction

L 35-39 : true, but then why choosing only one criteria (KGE). Plus, this is the case for any model that is a bit complex, not only in hydrology.

We thank the reviewer for this observation. Regarding the choice of a single performance metric, we note that the specific choice of model output of interest is not central to the objectives of this study (see Section 3.1). The focus is on validating the relationship between PVI_i and T_i (Eq. 8) as well as the use of other feature importance approaches through ML metamodelling, not on conducting a comprehensive sensitivity analysis of a particular hydrologic model. KGE was selected as a widely used metric that integrates correlation, bias, and variability components, making it a reasonable choice for this methodological demonstration.

Regarding the second point, we agree that the challenges described in L35–39 are not exclusive to hydrologic modelling and apply to complex models in general. The text is intended to contextualize these challenges within hydrologic modelling, which is the application domain of this study, rather than to claim that they are unique to it. We will add a sentence to the introduction noting that our findings can be generalized to other non-hydrological models.

L42 : are these 3 models really expensive and high-dimensional ? We would need some values of numerical cost. Conceptual hydrological models are usually much less expensive than physically-based hydrological models.

We agree that conceptual hydrologic models are generally less computationally expensive than physically-based models. Among the three models used in this study, HBV and HyMod are indeed inexpensive conceptual models. VIC, however, is a land-surface model substantially more demanding than the other two. We will include the run times of all three models in the repository description to provide concrete numerical values.

L45 – building a high-quality surrogate is also expensive, we need a large sampling. If the model runs are not too expensive, this must be taken into consideration in the decision-making process.

We agree that building a high-quality surrogate can be expensive in general. However, in our setting the surrogate does not predict hydrologic variables (e.g., streamflow time series) but rather maps the p-dimensional parameter space to a single scalar output (KGE). Training a metamodel on this scalar input–output relationship is computationally negligible, even for large sample sizes.

Section 2

The paper could be reduced on sections describing the methodology 2,1, 2,2, 2,3, keeping a way to understand the link between the Sobol indices and the PVI from ML. I appreciate the way indices are described in section 2, to be intuitive along with the mathematical equations.

We appreciate the reviewer's positive assessment of Section 2. As noted in our first response, we deliberately chose to submit as a technical note given the technical nature of the work, and we believe the current length is appropriate for this category according to recent technical notes published in HESS. We consider that each section in Section 2 serves a distinct purpose: Sections 2.1 and 2.3 provide the theoretical foundations and Sections 2.2 and 2.4 the practical estimation methods. Therefore, reducing them would compromise the clarity and self-contained nature of the presentation.

L91 : Why calling total Sobol index Ti and not Sti ?

The notation T_i was introduced by Homma & Saltelli (1996) for total sensitivity indices, initially as ST_i. In later work, Saltelli adopted the more compact notation T_i, which has since become standard (e.g., Lo Piano et al., 2020; Puy et al., 2021, 2022). We follow this convention.

Homma, T., & Saltelli, A. (1996). Importance measures in global sensitivity analysis of nonlinear models. Reliability Engineering & System Safety, 52(1), 1–17. https://doi.org/10.1016/0951-8320(96)00002-6

Lo Piano, S., Ferretti, F., Puy, A., Albrecht, D., & Saltelli, A. (2021). Variance-based sensitivity analysis: The quest for better estimators and designs between explorativity and economy. Reliability Engineering & System Safety, 206, 107300. https://doi.org/10.1016/j.ress.2020.107300

Puy, A., Lo Piano, S., & Saltelli, A. (2021). Is VARS more intuitive and efficient than Sobol' indices? Environmental Modelling & Software, 137, 104960. https://doi.org/10.1016/j.envsoft.2021.104960

Puy, A., Becker, W., Piano, S. L., & Saltelli, A. (2022). A COMPREHENSIVE COMPARISON OF TOTAL-ORDER ESTIMATORS FOR GLOBAL SENSITIVITY ANALYSIS. International Journal for Uncertainty Quantification, 12(2), 1–18. https://doi.org/10.1615/Int.J.UncertaintyQuantification.2021038133

L95-100 : Basing the GSA on independent parameters is a huge assumption but often made. However, for a more hydrological complex model with, for ex. hydraulic characteristics curves, it is

not correct to assume independence. GSA with dependent parameters is explored in papers that may be cited, e.g. Kucherenko et al., 2012, Mara et al., 2015, Il Idrissi et al. 2021, or during the sampling process like in Lambert et al., 2025, e.g.).

We thank the reviewer for these additional references. We note that the manuscript already acknowledges the independence assumption in L95–100 and cites several works on dependent-input GSA, including Chastaing et al. (2015), Jacques et al. (2006), Kucherenko et al. (2017), Li et al. (2010), and Mara et al. (2015). We will additionally cite Il Idrissi et al. (2021) in this section. The case of dependent inputs, while important, is beyond the scope of this work, as the validity of Eq. 8 is established for independent inputs.

Chastaing, G., Gamboa, F., & Prieur, C. (2015). Generalized Sobol sensitivity indices for dependent variables: numerical methods. Journal of Statistical Computation and Simulation, 85(7), 1306–1333. https://doi.org/10.1080/00949655.2014.960415

Jacques, J., Lavergne, C., & Devictor, N. (2006). Sensitivity analysis in presence of model uncertainty and correlated inputs. Reliability Engineering & System Safety, 91(10), 1126–1134. https://doi.org/10.1016/j.ress.2005.11.047

Kucherenko, S., Klymenko, O. V., & Shah, N. (2017). Sobol' indices for problems defined in non-rectangular domains. Reliability Engineering & System Safety, 167, 218–231. https://doi.org/10.1016/j.ress.2017.06.001

Li, G., Rabitz, H., Yelvington, P. E., Oluwole, O. O., Bacon, F., Kolb, C. E., & Schoendorf, J. (2010). Global Sensitivity Analysis for Systems with Independent and/or Correlated Inputs. The Journal of Physical Chemistry A, 114(19), 6022–6032. https://doi.org/10.1021/jp9096919

Mara, T. A., Tarantola, S., & Annoni, P. (2015). Non-parametric methods for global sensitivity analysis of model output with dependent inputs. Environmental Modelling & Software, 72, 173–183. https://doi.org/10.1016/j.envsoft.2015.07.010

L105 This is a bit surprising to refer to Puy at al. (2022) which is a R package paper (that you don't even use in this study) more than a seminal article in GSA methods (such as Saltelli et al., Da Veiga et al, 2021, etc.).

We respectfully note that Puy et al. (2022) provides not only an R package implementation but also a comprehensive overview of variance-based sensitivity analysis estimators and sampling designs, which is why it is cited in our manuscript. We consider it an appropriate reference regardless of its framing as a software paper.

L 140 – 144 : This relation was given, to my opinion, by Gregorutti et al. (2017). Please check or precise the role of each of these two references.

We thank the reviewer for prompting this clarification. The proof of the relationship between T_i and PVI_i (Eq. 8) was established by Wei et al. (2015), as cited in the manuscript. The probabilistic definition of PVI_i (Eq. 6) was given by Gregorutti et al. (2013), whose arXiv preprint was later published as Gregorutti et al. (2017). We opted to cite the published version rather than the preprint. Both references are cited in the manuscript, each in its appropriate context: Gregorutti et al. (2017) for the definition of PVI_i (Eq. 6) and Wei et al. (2015) for the relationship to T_i (Eq. 8).

Gregorutti, B., Michel, B., & Saint-Pierre, P. (2013). Correlation and variable importance in random forests (Version 5). https://doi.org/10.48550/ARXIV.1310.5726

Gregorutti, B., Michel, B., & Saint-Pierre, P. (2017). Correlation and variable importance in random forests. Statistics and Computing, 27(3), 659–678. https://doi.org/10.1007/s11222-016-9646-1

Wei, P., Lu, Z., & Song, J. (2015). A comprehensive comparison of two variable importance analysis techniques in high dimensions: Application to an environmental multi-indicators system. Environmental Modelling & Software, 70, 178–190. https://doi.org/10.1016/j.envsoft.2015.04.015

Section 3. Numerical Exp.

I think objective 1 is particularly new and interesting in this study, the obj. 2 should be more convincing, by adding numerical cost comparisons and using the surrogate in other contexts : as said before, I don't feel like the authors gained any cost here, since they use the same sampling for SHAP indices and to build the metamodels.

We refer to our responses to the previous comments regarding the computational advantage of our approach ($n(p+2) \rightarrow n$ via Eq. 8) and the numerical costs to be added to the repository description. The clarifying text to be added to Section 2.4 will also help reinforce this objective. We would also like to clarify that Objective 2 is not solely about cost reduction: it also concerns evaluating the capabilities of different ML metamodels to accurately approximate T_i, which is a methodological question independent of computational efficiency.

L192 : the choice of hydrological models and of the daily time step which smooths many processes may be important if you find contradictory results and rules comparing to Rouzies 2023, where the model is more complex / semi-conceptual? The models you chose are conceptual

We note that the validity of Eq. 8 is theoretical and does not depend on the choice of model. What may differ is how well the surrogate can approximate the relationship in practice, particularly for more complex settings such as distributed models with both horizontal and vertical parameterizations. We will enrich the discussion by comparing our findings with those of Rouzies et al. (2023), as also suggested by the reviewer in a later comment, addressing how differences in model complexity may influence the practical applicability of the approach.

L200 : with p the number of model parameters

We confirm that p is defined in the manuscript as the number of free parameters of each model (13 for HBV, 5 for HyMod, and 9 for VIC) in Section 3.2.

L210+Table 4 : is it useful since in the end you only show one catchment in the rest of the paper ?

Table 4 provides the physiographic and hydroclimatic characteristics of all three catchments, which are used to test the generalizability of the methods across different conditions. Results for one catchment are shown in the main text and for the other two in the Supplementary Material. We therefore consider Table 4 useful as it contextualises the three catchments referenced throughout the study.

L224 : why looking at the sensitivity of KGE if you don't plan to do calibration after ? Why not studying a model output directly ?

As stated in Section 3.1, the specific choice of model output of interest is not central to the objectives of this study. The methodology is applicable to any model output, whether a performance metric such as KGE or a model output such as mean streamflow. We chose KGE as a widely used and comprehensive metric. Using a performance metric allows us to illustrate how global sensitivity measures (T_i, PVI_i) can mask varying parameter importance across different performance levels, and how distributed sensitivity measures such as $SHAP_i$ can reveal this information by evaluating sensitivities across the parameter and target variable spaces.

L226 : uniform distributions is not so obvious for hydrological parameters. For example Ksat is usually a lognormal distribution. At least, it must be said that this is a high hypothesis and not the most common choice in hydrology.

We agree that uniform distributions are a simplification and that some hydrologic parameters, such as saturated hydraulic conductivity, are more naturally described by skewed distributions such as lognormal. However, as discussed by Mai (2023), the uniform distribution is the most common choice in hydrologic sensitivity analysis and model calibration, as it only requires specifying lower and upper bounds without prior knowledge of the distribution shape, and most calibration toolkits

support it by default. We also note that the validity of Eq. 8, which is the central focus of this study, is established for independent inputs regardless of their marginal distributions. Using uniform distributions thus provides a clean setting for testing this relationship without introducing additional assumptions about parameter distributions.

Mai, J. (2023). Ten strategies towards successful calibration of environmental models. Journal of Hydrology, 620(PA), 129414. https://doi.org/10.1016/j.jhydrol.2023.129414

L244 “Accordingly, model configurations that would typically be regarded as overfitting in a predictive setting are acceptable here, provided that they do not distort the resulting sensitivity measures.” => You check this because you have the SHAPi values, but the aim is to show that ML methods can replace the classical Sobol indices calculations (obj. 2 of the paper), and in the case you d’ont have the “real” Sobol indices, you are not able to assess that.

We understand the reviewer's concern, but this reasoning applies to any validation effort: the reference is needed to establish confidence in the approach, but once this confidence is established, the reference is no longer required. The objective of this study is precisely to validate the relationship between PVI_i and T_i (Eq. 8) across different models and catchments. Once this validation is performed, the approach can be applied in new settings without the need to recompute the classical Sobol indices, as the surrogate-based PVI_i estimates can be trusted based on the evidence provided here.

L251 : “cheap framework” : no since you have run the sampling from the matrix A? to me, unless I miss something, this is the same cost.

We refer to our responses to the previous comments regarding the computational advantage of our approach ($n(p+2) \rightarrow n$ model evaluations via Eq. 8).

L256-270 : The properties of SHAP that you cite are indeed interesting for the KGE criteria (so, in a context of calibration, that may be said explicitely). However, the choice of SHAP confused me a bit. To me they are specifically adapted in GSA to deal with dependent inputs (Owen and Prieur, 2019) and you choose these indices, assuming voluntarily independent inputs here. Can you add a sentence on this ?

We thank the reviewer for raising this important point, which allows us to clarify a common source of confusion. $SHAP_i$ values (Lundberg & Lee, 2017) and Shapley effects (Owen, 2014) are related but distinct concepts. Shapley effects are a sensitivity analysis measure designed to handle dependent inputs by allocating contributions to the variance of the model output among possibly correlated variables. $SHAP_i$ values, by contrast, constitute a XAI method that decomposes individual predictions into additive feature contributions. While both share the same theoretical root in the Shapley value framework from cooperative game theory, they address different attribution problems.

In our study, $SHAP_i$ values are used as a complementary feature importance measure to provide a distributed evaluation of sensitivities across the parameter and target variable spaces, not as a method to handle dependent inputs. The formal mathematical definition of $SHAP_i$ will be provided in Appendix B of the revised manuscript, following the suggestion of Reviewer #2. We refer to our response to Reviewer #1 (Q6) for a detailed discussion of the relationship between $SHAP_i$ and Shapley effects, and the connection to Sobol indices through Owen (2014).

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 4768–4777). Red Hook, NY, USA: Curran Associates Inc.

Owen, A. B. (2014). Sobol’ Indices and Shapley Value. SIAM/ASA Journal on Uncertainty Quantification, 2(1), 245–251. https://doi.org/10.1137/130936233

Section 4. Results and discussion

I would suggest to put Figure 1 BEFORE figure 2 : we first validate the surrogates before using them for SA.

We agree with the reviewer that it is logical to first validate the surrogates before using them for sensitivity analysis. We will reorder Figs. 1 and 2 in the revised manuscript, placing the predictive performance comparison (currently Fig. 2) before the sanity check of feature importance estimates (currently Fig. 1), and update all cross-references accordingly.

L323 again, this is logical to have very good performance since the ML is evaluated on the points the surrogate was built. This has to be explicited when discussing Figure 2.

We agree that the high predictive performance is expected given that the surrogates are evaluated on the same points used for training. As discussed in our response to the comment on the absence of a test set, this does not diminish the validity of the approach, since the purpose of the metamodels is not predictive generalization but the estimation of PVI_i. We will add a clarifying sentence in the discussion of Fig. 2 (in the revised manuscript, this will be Fig. 1) to explicitly state that the high R^2 values are a consequence of the in-sample evaluation.

On Fig 1 I am surprised that the ratio is not larger due to equation 8 for any model ? what does it mean in terms of variance of Y?

We appreciate the reviewer's observation. Fig. 1 shows PVI_i values (panels a–f) and mean absolute $SHAP_i$ values (panels g–l). According to Eq. 8, the relationship between T_i and PVI_i is linear, with a slope equal to 1/(2Var(Y)). For the Garlitz catchment shown in the main text, Var(Y) is 0.66 (HBV), 0.54 (HyMod), and 0.07 (VIC), so the expected slopes are approximately 0.76, 0.93, and 7.28, respectively. The slopes observed in Fig. 5 (panels a–c) are consistent with these values. For HBV and HyMod, the slope is close to 1, so PVI_i and T_i are of similar magnitude, while for VIC, the slope is well above 1 and PVI_i is correspondingly smaller than T_i.

In Fig 2 caption, please put the configurations of NN and RF you plot. OR, it would be even better to put here all the configurations to see the interest of the one you choose and the sensitivity of the ML to the configuration. This can be done by boostrap easily. Again, I am surprised that the ratio is so small between them all.

The configurations of those metamodels are detailed in Table 5. Configuration 1 was selected for Fig. 2 and the subsequent analysis as it is computationally more efficient while still accurately reproducing the original model response (see Section 3.3). We will add a reference to Table 5 in the Fig. 2 caption to clarify which configuration is shown. Regarding the suggestion to show all configurations or use bootstrap, the strong overlap of importance estimates across all five configurations in Fig. 1 already demonstrates that the results are insensitive to the configuration choice, so we consider showing a single configuration in Fig. 2 sufficient. Regarding the reviewer's comment on the ratio, we refer to our response to the previous comment.

Rouzies et al (2023) compared Sobol (from Polynomial Chaos Expansion) vs RF (and HSIC) measures on a hydrological model as well. They look at the ratio from equation 8 too and find much less evidence of it. This is a more complex hydrological model and it may allow to enrich your discussion, since you compare several complexities of hydrological models.

We thank the reviewer for suggesting this comparison. Rouzies et al. (2023) applied GSA to the PESHMELBA model, a distributed, physically-based pesticide transfer model with 145 input parameters, and compared Sobol' indices (via polynomial chaos expansion), HSIC dependence measures, and RF feature importance. Notably, they found that the relationship between RF feature importance and Sobol' total-order indices (Eq. 8) was not well respected in their study, which they attributed to the poor quality of the RF surrogate for their complex model. This contrasts with our results, where Eq. 8 holds across all three conceptual models and metamodels, and

highlights that the practical applicability of the approach depends on the fidelity of the surrogate. We will enrich the discussion by comparing our findings with those of Rouzies et al. (2023), addressing how differences in model complexity, parameter dimensionality, and surrogate quality may influence the validity of Eq. 8 in practice.

Fig 4: uncertainties on the estimated indices would help to see the confidence we have by choosing one method again another. It may allow to prefer ML methods of the uncertainty of indices is much smaller, for example.

We agree that adding uncertainty estimates would strengthen the comparison. We will include 95% confidence intervals for all sensitivity indices shown in Fig. 4 in the revised manuscript. For consistency, we will also add them to Fig. 6a. For T_i, bootstrap resampling will be performed using the implementation available in SALib (1000 resamples). For PVI_i and mean $|SHAP_i|$, confidence intervals will be obtained by evaluating the trained metamodels on resampled versions of the training data (1000 resamples). For the GBDT metamodel (which was mistakenly identified as a Random Forest in our initial submission but is in fact a Gradient Boosted Decision Tree; see our response to Reviewer #1), subsampling without replacement with a subsample size of 80% will be used, following a similar approach to Rouzies et al. (2023) for random forests, to avoid interference with the internal subsampling mechanism of XGBoost. For the NN and LM metamodels, standard bootstrap with replacement will be used. The methodology will be described in Section 3.4 of the revised manuscript.

Last part on SCHAPi with Fig 6 may be removed. It is very interesting and shows the interest of using SCHAPi against other (aggregated) indices, but the paper is too long, and this is independent from the rest. This gives a zoom on SCHAPI specifities although the rest of the paper focuses on the interest of ML to evaluate sensitivity. Plus, text and caption repeat the information.

We consider the $SHAP_i$ section, including Fig. 6 and its associated discussion, an integral part of the paper, as it demonstrates the value of $SHAP_i$ for distributed sensitivity evaluation, which is one of the contributions of this study. This section will be enhanced in response to comments from Reviewer #1, and a formal mathematical definition of $SHAP_i$ will be added as Appendix B following the suggestion of Reviewer #2. We therefore consider that this section should be further developed rather than removed, in particular considering that the suggestion to trim it is motivated by the length of the manuscript, which we address in our first response.

References of this review report

M. Il Idrissi, V. Chabridon and B. Iooss, Developments and applications of Shapley effects to reliability-oriented sensitivity analysis with correlated inputs, Environmental Modelling & Software, 143, 105115, 2021

Da Veiga, S.; Gamboa, F.; Iooss, B. & Prieur, C. Basics and Trends in Sensitivity Analysis Society for Industrial and Applied Mathematics, 2021, doi:10.1137/1.9781611976694

Kucherenko, S., A. Tarantola, and P. Annoni (2012). Estimation of global sensitivity indices for models with dependent variables. Comput. Phys. Comm. 183, 937–946. doi:10.1016/j.cpc.2011.12.020

Lambert, G.; Helbert, C. & Lauvernet, C. Quantization-Based Latin Hypercube Sampling for Dependent Inputs With an Application to Sensitivity Analysis of Environmental Models Applied Stochastic Models in Business and Industry, 2025, 41, e2899 , doi:10.1002/asmb.2899

Mara, T. A. & Tarantola, S. Variance-based sensitivity indices for models with dependent inputs Reliability Engineering & System Safety, 2012, 107, 115-121, doi: 10.1016/j.ress.2011.08.008

Owen, A.B., Prieur, C., 2017. On Shapley value for measuring importance of dependent inputs. SIAM/ASA J. Uncertain. Quantification 5, 986–1002. doi:10.1137/16M1097717.

Rouzies, E., Lauvernet, C., Sudret, B., and Vidard, A.: How is a global sensitivity analysis of a catchment-scale, distributed pesticide transfer model performed? Application to the PESHMELBA model, Geosci. Model Dev., 16, 3137–3163, doi:10.5194/gmd-16-3137-2023, 2023.