Implementation of supervised principal component analysis for global sensitivity analysis of models with correlated inputs

Mohammad Ali Mohammad Jafar Sharbaf¹ · Mohammad Javad Abedini¹

Accepted: 12 December 2021 / Published online: 25 January 2022

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract

Global Sensitivity Analysis (GSA) plays a significant role in quantifying the tangible impact of model inputs on the uncertainty of response variable. As GSA results are strongly affected by correlated inputs, several studies have considered this issue, but most of them are computationally expensive, labor-intensive, and difficult to implement. Accordingly, this paper puts forward a novel regression-based strategy based on the Supervised Principal Component Analysis (SPCA), benefiting from the Reproducing Kernel Hilbert Space. Indeed, by conducting one kind of variance-based sensitivity analysis, a renowned method exclusively customized for models with orthogonal inputs, on SPCA regression, the impact of the correlation structure of input variables is considered. The ability of the suggested technique is evaluated with five test cases as well as three hydrologic and hydraulic models, and the results are compared with those obtained from the correlation ratio method; Taken as a benchmark solution, which is a robust but quite complicated approach in terms of programming. It is found that the proposed method satisfactorily identifies the sensitivity ordering of model inputs. Furthermore, it is proved in this study that the performance of the proposed approach is also supported by the total contribution index in the derived covariance decomposition equation. Moreover, the proposed method compared with the correlation ratio method, is found to be computationally efficient and easy to implement. Overall, the proposed scheme is appropriate for high dimensional, quite strong nonlinear or expensive models with correlated inputs, whose coefficient of determination between the original model and regression-based SPCA model is larger than 0.33.

Keywords Correlated parameters · Global sensitivity analysis · RKHS · Regression-based model · Variance-based SA · Uncorrelated parameters

1 Introduction

The identification, quantification, and propagation of uncertainty are considered to be three necessary tasks in model building and evaluation. As the process of modeling and simulation of water resources and environmental projects become more advanced and computationally expensive, the process of quantification and propagation of

uncertainty in parameter estimation and model structure identification becomes quite time-consuming and labor-intensive for high dimensional models. In a sense, in a typical modeling and simulation with a highly nonlinear model structure, a small error in input variables might result in a large uncertainty (or error) in model output (Helton et al. 2005; Iman et al. 2002). Thus, it is advantageous to investigate how the uncertainty in the output of a model can be made proportional to uncertainty in model inputs. Global Sensitivity Analysis (GSA) is devised to address these issues in model building, execution, and evaluation.

GSA is utilized in many applications in water resources engineering and environmental studies (Ciriello et al. 2013; Zheng and Wang 2015; Razavi and Gupta 2016b) such as model calibration, uncertainty reduction of model output, model validation, identification of redundant parameters,

✉ Mohammad Javad Abedini
abedini@shirazu.ac.ir

Mohammad Ali Mohammad Jafar Sharbaf
ma.mj.sharbaf@shirazu.ac.ir;
mohammad.sharbaf@gmail.com

¹ Department of Civil and Environmental Engineering, School of Engineering, Shiraz University, Engineering Building #1, Zand Street, 7134851156 Shiraz, Fars, Iran

and decision-making processes that try to explore which model inputs might have the most influential impact on output uncertainty (Crosetto and Tarantola 2001). Due to the numerous applications including the ones cited above, GSA can be effectively utilized to conduct preliminary exploratory data analysis to build the corresponding model. As an example, in a typical rainfall-runoff process, being able to implement GSA in order to build a parsimonious model structure would not only help eliminate the redundant parameters, but also help reduce the uncertainty in model output which somehow lead to saving time and capital. This task is considered as an inevitable part of any effort conducted to convert rainfall into runoff in an efficient way.

There is a wide range of GSA methods in the market based on different philosophies and theoretical definitions of sensitivity measures, such as regression technique, Morris, variance-based, Variogram Analysis of Response Surface (VARS) method, regional SA, and density-based methods. The details of these approaches can be found in relevant literature (Saltelli et al. 2004; Borgonovo 2007; Razavi and Gupta 2016a). However, the cited approaches suffer from a fundamental problem regarding the correlation structure of input variables. Nowadays, non-orthogonal model input variables can be considered a rule rather than an exception in GSA because it is apparent in the shadow of belittling and ignorance of model factor dependence; GSA outcome will be fallacious. Therefore, results of GSA should be considered with caution when there is a correlation between certain parameters in GSA, that is not considered during GSA implementation. Accordingly, several scientists have attempted to present or develop GSA methods in light of dependent factors. As an example, a few studies applied the parametric and/or non-parametric methods to decompose the variance of the response variable with regard to model correlated inputs (Saltelli et al. 2001 and 2004; Xu and Gertner 2008; Da Veiga et al. 2009; Chastaing et al. 2012; Li et al. 2010; Xu 2013; Zhang et al. 2015; Mara et al. 2015), whereas some other investigators extended the analytical formulations for non-orthogonal input variables and the associated numerical approaches to calculate the variance-based sensitivity measure and/or more advanced Sobol's sensitivity indices (Mara and Tarantola 2012; Kucherenko et al. 2012, 2017; Wang et al. 2018). As the computational budgets associated with most of the above approaches are quite expensive due to high dimensional problems, and/or more complex model structure, development of more innovative and efficient approaches with regard to non-orthogonal inputs is strongly recommended for practical applications (Ge and Menendez 2017; Zhou et al. 2019; Do and Razavi 2020; Lamboni and Kucherenko 2021). In the meanwhile, one should also acknowledge how problematic it is for

engineers and other investigators with rudimentary information on GSA to use the above methods because of their complicated procedures and concepts for practical applications on a routine basis.

Taking the above elaboration into account, this paper intended to couple the classical variance-based SA, a class of methods customized for independent inputs, with Supervised Principal Component Analysis (SPCA), based on Hilbert–Schmidt Independence Criterion (HSIC), to address issues concerning model's computational effort and complexity. To the best of the authors' knowledge, this is the first time that SPCA is implemented in the GSA arena to find the degree of importance of various correlated input variables as they affect the output response. Furthermore, in conventional literature, a number of classical variance-based approaches (e.g., classical Sobol and/or FAST as numerical approaches) had been proposed to delineate the degree of importance of models with independent input variables. Here, the authors devised an algorithm whereby SPCA is coupled with classical variance-based SA to obtain the degree of importance of models with correlated inputs. In other words, the classical variance-based approach, which had been extensively used in the case of orthogonal model inputs, can be used for models with correlated inputs if it is coupled with SPCA. This coupling exercise reduces the sample size dramatically imposed by conventional approaches such as the correlation ratio method. As a result, simplicity and efficiency are considered as two major advantages of the proposed method.

Indeed, implementation of this kind of variance-based SA on a regression model derived from the SPCA would lead to a semi-analytical method in terms of components of a dominant eigenvector to estimate the first-order sensitivity measure with no additional computational budget for a system which contains both correlated and orthogonal inputs.

The rest of the paper is organized as follows: first, an overview of the GSA based on the variance-based approach is presented in the next section. Then, the paper gives a detailed account of SPCA, followed by the proposed method along with the proof of its validity in the next section. The subsequent section is devoted to assessing the performance of the proposed method via five test functions as well as three more realistic models in hydrologic and hydraulic domains, for which the corresponding results of all test cases are compared with the correlation ratio method based on McKay's scheme. Finally, the summary and the conclusions are summarized in the last section.

2 Theoretical background

The methodology presented in this work is intended to couple SPCA with the classical variance-based SA which is limited to independent inputs. It is important to make this coupling process quite clear. In essence, the classical variance-based (e.g., Sobol and/or FAST) methodology is limited to transfer functions for which the input variables are independent. On the other hand, SPCA considers the impact of correlation among input variables and a dependent variable using regression-based analysis. Would it be possible to effectively benefit from the advantage of SPCA and combine it with the particular variance-based SA to contrive a better tool for prioritizing the input variables when the input variables are correlated? With this question in mind, first, a concise account of the basic concept of variance-based SA is provided, followed by a discussion detailing the theoretical background of SPCA to determine how this coupling exercise can be implemented in practice.

2.1 Variance-based sensitivity analysis

For a computer model with p number of input variables given by $Y = f(X)$, $X = (X_1, X_2, \dots, X_p)$, if the input variables are random and they are free to vary over their whole bound of uncertainty,¹ then the output becomes a random variable and the uncertainty of the model output can be determined by $V(Y)$. For such a model, decomposition of the total unconditional variance of output variable can be written as (Saltelli et al. 2004):

$$V(Y) = V[E(Y|X_i)] + E[V(Y|X_i)] \quad (1)$$

V conveys the notion of variance, and E represents the expectation operator. Also, $V[E(Y|X_i)]$ and $E[V(Y|X_i)]$ are called the main and residual effect, respectively (Saltelli et al. 2004). Thus, to achieve a reduction in the variance of model output, the first-order sensitivity index can be computed by defining the ratio of the main variance (i.e., V_i) as a variance of conditional expectation to variance of model output as (Saltelli et al. 2004):

$$S_i = V_i/V(Y) = V[E(Y|X_i)]/V(Y) \quad (2)$$

It is worth noting that S_i varies between zero and one, i.e., $[0,1]$. In addition, the sensitivity analysis (SA) of higher-order terms taking into the account the interaction of model factors such as X_i and X_j on model output uncertainty can be written as (Saltelli et al. 2004):

¹ The knowledge about the range of uncertainty of model inputs can be obtained from various sources such as the one elaborated by Saltelli et al. (2004), which reads as: “measurements, expert opinion, physical bounds or an analogy with factors for similar species/compounds. We may additionally have information (e.g., via observation) on the joint probability distribution of the factors”.

$$S_{ij} = V_{ij}/V(Y) = (V[E(Y|X_i, X_j)] - V_i - V_j)/V(Y) \quad (3)$$

Like S_i , S_{ij} varies in the range of zero and one. Saltelli et al. (2004, p. 112) can be consulted for further details regarding the computation of higher-order interactions. Under the independent assumption, the variance-based sensitivity indices can be estimated numerically via the two most popular techniques, e.g., Sobol (1993) and FAST (Cukier et al. 1973).

If a mathematical model under consideration is assumed to be approximately linear due to the relatively linear impact of model factors on output, the given model, i.e., $Y = f(X_1, X_2, \dots, X_p)$ can be simplified by eliminating the nonlinear or interaction terms as follows (Xu and Gertner 2008):

$$Y = a_0 + \sum_{i=1}^p a_i X_i \quad (4)$$

Later on, we will elucidate how the original nonlinear function can be converted to Eq. (4). Hence, only the main effect of each model input in the corresponding regression model is considered. When the components of this linear model's inputs are mutually independent, it can be proved that the main effect in reference to variance decomposition can be stated as follows (Saltelli et al. 2008):

$$S_i = \frac{V[E(Y|X_i)]}{V(Y)} = \theta_i^2 \quad (5)$$

In which $\theta_i = a_i \sigma_{X_i} / \sigma_Y$. The θ_i 's are standardized regression coefficients (SRCs). Actually, in the case of orthogonal model factors, the SRC is considered as a method of sensitivity analysis for the model inputs, and in accordance with the identity of Eq. (5), the squared SRCs describe the contribution of each factor to the total variance (Saltelli et al. 2008).

To evaluate the performance of the SRCs, it is necessary to appreciate the connection between SRCs and the measure of goodness, the so-called model coefficient of determination, R^2 as (Saltelli et al. 2004):

$$R^2 = \sum_{i=1}^p (\hat{Y}_i - \bar{Y}) / \sum_{i=1}^p (Y_i - \bar{Y}) \quad (6)$$

where Y_i is the output of the original model, i.e., $Y = f(X_1, X_2, \dots, X_p)$, and $\hat{Y}_i$ stands for an estimate of Y_i using the regression model, i.e., Eq. (4). This coefficient is an attempt to recognize how well the model under study can be approximated by regression in the form of Eq. (4) (Weisberg 2005). When the model is linear, using SRC as a GSA method can exactly quantify the amount of response surface variation explained by each model input.

2.2 The supervised PCA

The SPCA, initially pioneered by Barshan et al. (2011), laid the foundations for this approach in the field of supervised methods. For the cited contribution, a considerable number of dimensionality reduction techniques based on supervised methods could only consider similarities and differences for classification purposes. This property of the cited supervised methods is in contrast with Barshan et al.'s approach, which also examines the quantitative value of the target variables. As a result, it is applicable to both classification and regression problems. Their ideas can be considered as a paradigm shift in future activities of researchers carrying out prediction based on regression approaches. Their studies refer to the work of Gretton et al. (2005), who proposed an independent criterion in Reproducing Kernel Hilbert Spaces (RKHS). This criterion measures the dependency between two random variables according to the Hilbert–Schmidt Independence Criterion (HSIC). In reference to HSIC, the necessary and sufficient conditions for the two random variables to be independent, can be achieved if the value of this statistic (i.e., HSIC) is zero.

There is a theoretical relationship for HSIC, but it is impractical for actual settings. Consequently, the empirical estimate of HSIC proposed by Gretton et al. (2005) as a practical criterion, can be implemented to check the dependence or independence of two random variables with a finite number of observations. Thus, this criterion, i.e., the empirical estimate of HSIC for a series of n observations such as $\mathcal{Z} := \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, is as follows (Gretton et al. 2005):

$$\text{HSIC}(\mathcal{Z}, \mathcal{F}, \mathcal{G}) := (n-1)^{-2} \text{tr}(KHLH) \quad (7)$$

where $\mathcal{F}$ is a RKHS² in which for each point $x \in X$, there exists an element $\varnothing(x) \in \mathcal{F}$, such that $\langle \varnothing(x_i), \varnothing(x_j) \rangle_{\mathcal{F}} = k(x_i, x_j) = K_{ij}$ and K is considered to be a positive definite kernel for $n \geq 1$ and $b_1, b_2, \dots, b_n \in \mathbb{R}$, (Hein and Bousquet 2004), i.e.,

$$\sum_{i=1}^n \sum_{j=1}^n b_i b_j k(x_i, x_j) > 0 \quad (8)$$

Likewise, $\mathcal{G}$ is a RKHS in which for each point $y \in Y$, there is an element $\omega(y) \in \mathcal{G}$, such that $\langle \omega(y_i), \omega(y_j) \rangle_{\mathcal{G}} = l(y_i, y_j)$ and L is considered as a positive definite kernel. It is necessary to recall that $\mathcal{F}$ and $\mathcal{G}$ have to be separable. Indeed, they must have complete orthonormal systems.

² The necessary background of this concept is provided by Gretton et al. (2005)

Briefly, $K, L, H \in \mathbb{R}^{n \times n}$ and $K_{ij} := k(x_i, x_j)$, $L_{ij} := l(y_i, y_j)$. In addition, $H_{ij} := I - \frac{1}{n} ee^T$ (centering matrix and e is a unit vector). Finally, tr stands for the trace of a matrix.

In SPCA, the subspace spanned by new features is examined such that the principal components of input variables with maximum dependency on response surface are reserved through the empirical criterion, i.e., HSIC. Actually, for a model with p standardized inputs, $\tilde{X} = (\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_p)$ and l standardized outputs, $\tilde{Y} = (\tilde{Y}_1, \tilde{Y}_2, \dots, \tilde{Y}_l)$, to maximize the dependency among the projected data (i.e., $\mathcal{Z} = \tilde{X}U$)³ and the response surface (i.e., $\tilde{Y}$), maximizing the $\text{tr}(KHLH)$ is necessary. This could be justified by referring to empirical HSIC cited above (Barshan et al. 2011). To maximize $\text{tr}(KHLH)$, $\mathbf{Z}\mathbf{Z}^T$ is replaced with the kernel matrix K to get (Barshan et al. 2011):

$$\text{tr}(KHLH) = \text{tr}(\mathbf{Z}\mathbf{Z}^T HLH) = \text{tr}(\tilde{X}U U^T \tilde{X}^T HLH) \quad (9)$$

After replacing the dimension of each matrix in Eq. (9), this equation can be rewritten as follows:

$$\begin{aligned} \text{tr}(KHLH) &= \text{tr}(\tilde{X}_{n \times p} U_{p \times m} U_{m \times p}^T \tilde{X}_{p \times n}^T H_{n \times n} L_{n \times n} H_{n \times n}) \\ &= \text{tr}([\tilde{X}U]_{n \times m} [U^T \tilde{X}^T HLH]_{m \times n}) \end{aligned} \quad (10)$$

Based on the unique property of trace in matrix algebra (Strang 2016)⁴:

$$\text{tr}(KHLH) = \text{tr}([U^T \tilde{X}^T HLH]_{m \times n} [\tilde{X}U]_{n \times m}) \quad (11)$$

Consequently, we have (Barshan et al. 2011):

$$\text{tr}(KHLH) = \text{tr}(U^T \tilde{X}^T HLH \tilde{X}U) \quad (12)$$

L is a kernel matrix of $\tilde{Y}$ (e.g., $\tilde{Y}\tilde{Y}^T$), and $H_{ij} = I - \frac{1}{n} ee^T$.

Ultimately, the optimization for HSIC objective function is accompanied by the following constraint (Barshan et al. 2011):

$$\begin{aligned} \arg \max_U \text{tr}(U^T \tilde{X}^T HLH \tilde{X}U) \\ \text{Subject to } U^T U = I \end{aligned} \quad (13)$$

The optimal solution for this optimization problem is considered to be the unit eigenvectors of the real and symmetric matrix $\mathbb{Q} = \tilde{X}^T HLH \tilde{X}$. Taking the components of the unit eigenvectors as the decision variables, the

³ Matrix U is a modal matrix consisting of m orthogonal unit eigenvectors that map the data sets to a new space in which features are uncorrelated.

⁴

$$\text{tr}(\tilde{A}_{n \times m} \tilde{B}_{m \times n}) = \text{tr}(\tilde{B}_{m \times n} \tilde{A}_{n \times m})$$

optimal solution will be $U = [U_1, U_2, \dots, U_m]$ corresponding to $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m$, which are selected among p eigenvalues. Here, m is the dimension of eigenspace (Barshan et al. 2011). The beauty of the above mathematical argument can be justified as follows: If kernel L is equal to the identity matrix I , the matrix $\tilde{X}^T H L H \tilde{X}$ is equivalent to the covariance of matrix $\tilde{X}$, i.e., the PCA method (Barshan et al. 2011). Since $H^T = H$:

$$\begin{aligned} \mathbb{Q} &= \tilde{X}^T H I H \tilde{X} = (H \tilde{X})^T (H \tilde{X}) \\ &= \left(\left(I - \frac{1}{n} e e^T \right) \tilde{X} \right)^T \left(I - \frac{1}{n} e e^T \right) \tilde{X} \\ &= (\tilde{X} - \mu)^T (\tilde{X} - \mu) = \text{COV}(\tilde{X}) \end{aligned} \quad (14)$$

It is worth noting that the principal component analysis can be used for different purposes. For example, Martin-Barreiro et al. (2021) applied the modern technique in PCA, i.e., disjoint and functional principal component analysis, to classify the South American countries which, due to COVID-19, caused several infections and deaths. In addition, Ramirez-Figueroa et al. (2021) suggest a novel approach for disjoint principal component analysis based on particle swarm optimization. In reference to their proposed method, the predictive variables which have the most contributions on each principal component are selected. Based on their approach, the main problem in the multivariate statistic is solved.

3 The proposed methodology

As mentioned in the preceding section, the novelty in this study is based on implementing the classical variance-based SA (applicable under the assumption of independent inputs) on the regression model extracted from SPCA for models with correlated inputs. In subsequent subsections, derivation of regression-based SPCA and details of how to couple it with a typical variance-based SA will be provided.

3.1 Derivation of regression-based SPCA

Considering the road map offered above, consider the regression model cited in Eq. (4), as mentioned, this linear regression model is obtained in reference to the approximate linear impact of inputs on the response variable, i.e., $Y = f(X_1, X_2, \dots, X_p)$. To use this regression model, first, the samples of model inputs are required. For this purpose, Latin Hypercube Sampling (LHS) can be used to generate various realizations of the input variables based on their probability density functions⁵ for both correlated and/or uncorrelated model inputs (Iman and Conover 1982). This

method is also recognized as an inverse Nataf transformation (Nataf 1962). Then, the original model is run using these samples, and a one-dimensional response surface (Y) will be generated. As soon as the multi-inputs-single output realizations are generated, the regression model stipulated in Eq. (4) can be produced. When the independence assumption is violated, i.e., the model inputs are correlated, using ordinary least square gives a poor estimate of model parameters of this equation, i.e., a_i . Thus, performing SA on this linear model will produce an unjustified solution. Although the Principal Component Regression (PCR) method can overcome this problem, coupling it with a classical variance-based approach (e.g., Sobol and/or FAST) has a drawback that will be highlighted in the following study. As a result, the SPCA using linear kernel L on standardized response variable $\tilde{Y}$ (i.e., $\tilde{Y} \tilde{Y}^T$) was chosen to estimate the simple regression model coefficients. For this purpose, $\tilde{Y}$ is expressed as a linear combination of new features called Z_j as stipulated in Eq. (15).

$$\tilde{Y} = \beta_1 Z_1 + \beta_2 Z_2 + \dots + \beta_m Z_m + \beta_0 \quad (15)$$

In which Z_j ($j = 1, 2, \dots, m, m < p$) are a set of new features taking into account the impact of the dependent variable, $\tilde{Y}$. In addition, m is the number of supervised principal components. The procedure of deriving Eq. (15) is explained in appendix A. To express Y as a linear combination of X_i , a linear transformation was utilized to convert Z_j to X_i as follows (Barshan et al. 2011):

$$[Z_1, Z_2, \dots, Z_m] = [\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_p] \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1m} \\ u_{21} & u_{22} & \dots & u_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ u_{p1} & u_{p2} & \dots & u_{pm} \end{bmatrix} \quad (16)$$

Via eigenvalue analysis, it is proved in appendix B that after replacing matrix L with $\tilde{Y} \tilde{Y}^T$ in matrix $\mathbb{Q}$, the resulting matrix gives only one nonzero eigenvalue (λ_1). In appendix B, a unique eigenvalue and the corresponding eigenvector for matrix $\mathbb{Q}$ are rationalized in a meaningful manner. Thus, its associated unit eigenvector, i.e., $U_1 = [u_{11}, u_{21}, \dots, u_{p1}]^T$ is chosen for projecting explanatory variables, i.e., X . As the corresponding new feature (Z_1) has a maximum linear dependency on $\tilde{Y}$, this means other new features can be eliminated from the model due to zero correlation with output which is proved in appendix B. In the meanwhile, in linear regression, when model variables are uncorrelated with zero means, the regression

⁵ The interested reader can refer to Saltelli et al. (2004, p 45–46) to choose the PDF of model inputs.

coefficients of Eq. (15) can be estimated as follows (Bedford 1998; Xu and Gertner 2008):

$$\hat{\beta}_j = \frac{\text{COV}(\tilde{Y}, Z_j)}{V(Z_j)}, \quad j = 1, 2, \dots, m \quad (17)$$

As was proved in appendix B, $\text{COV}(\tilde{Y}, Z_j) = 0$, for $j \neq 1$. Thus, $\hat{\beta}_j = 0$ for $j \neq 1$. Consequently, the regression-based SPCA model will be simplified as follows:

$$\tilde{Y} = \beta_1 Z_1 + \beta_0 \quad (18)$$

After replacing Z_1 and β_1 with their equivalences [i.e., Eq. (16) and Eq. (17)] into Eq. (18), the regression-based SPCA model can be expressed in terms of standardized variables as:

$$\tilde{Y} = \frac{\text{COV}(\tilde{Y}, Z_1)}{V(Z_1)} [u_{11}\tilde{X}_1 + u_{21}\tilde{X}_2 + \dots + u_{p1}\tilde{X}_p] + \beta_0 \quad (19)$$

where u_{i1} are the entries of the U_1 eigenvector corresponding to the maximum eigenvalue, i.e., λ_1 . As a result, Eq. (19) can be expressed in terms of the original variables Y and X_i as:

$$\begin{aligned} \frac{Y - E(Y)}{\sigma_Y} = & \frac{\text{COV}(\tilde{Y}, Z_1)}{V(Z_1)} \left[u_{11} \frac{X_1 - E(X_1)}{\sigma_{X_1}} \right. \\ & \left. + u_{21} \frac{X_2 - E(X_2)}{\sigma_{X_2}} + \dots + u_{p1} \frac{X_p - E(X_p)}{\sigma_{X_p}} \right] + \beta_0 \end{aligned} \quad (20)$$

The above equation can be written as:

$$Y = \frac{\text{COV}(\tilde{Y}, Z_1)}{V(Z_1)} \sigma_Y \sum_{i=1}^p u_{i1} \frac{X_i}{\sigma_{X_i}} + a_0 \quad (21)$$

$$\begin{aligned} a_0 = & \beta_0 \sigma_Y + E(Y) - \frac{\text{COV}(\tilde{Y}, Z_1)}{V(Z_1)} \sigma_Y \sum_{i=1}^p u_{i1} \frac{E(X_i)}{\sigma_{X_i}}, i \\ = & 1, 2, \dots, p \end{aligned}$$

Hence, Eq. (21) is considered as a regression-based SPCA model for $Y = f(X)$.

3.2 Details of coupling the classical variance-based SA with SPCA

At this stage, the classical variance-based SA coupled with SPCA can be effectively utilized to investigate the impact of X_i on Y . Before mentioning this task, it might help to summarize the main point of the above mathematical manipulations for practical purposes. If we are given a functional, thereby the predictor variables are correlated. We have three choices to investigate the impact of various predictor variables on output. According to the first option, the correlation amongst inputs might be ignored and the

classical variance-based SA (applicable under the assumption of the independent inputs, e.g., Sobol and/or FAST) can be implemented directly on the original model with correlated inputs. The most important variables can then be reported; however, the results are not justified and cannot be defended. In addition, based on the second option, the methods in the literature such as the developed variance-based SA (under the assumption of correlation of model inputs, e.g., correlation ratio method) can be applied directly on the original model, which leads to having acceptable results for recognition of the importance of correlated inputs, but these methods may have complex procedures because re-coding their algorithms may not be straightforward (Saltelli et al. 2004) or may be rather time-consuming in the case of high dimensional and/or computationally expensive models⁶ (Ge and Menendez 2017). However, the regression-based SPCA can also be implemented to present Eq. (21) and then impose the classical variance-based SA on this regression⁷ model to obtain meaningful and justifiable results to prioritize the correlated inputs with respect to the uncertainty of model output. In fact, in the case of having a model with correlated inputs, we reused the classical variance-based SA method, which is proposed for models with independent inputs, but we conduct it on the regression-based SPCA model instead of the original model. We argue and prove that prioritization in reference to the third option would give rise to acceptable results for models which include correlated inputs. In the next section, we show this coupling to be more straightforward and computationally efficient as compared to the developed variance-based.

Following the third option, estimation of the sensitivity measures based on Eq. (5) using regression-based SPCA model can be written as:

⁶ Generally, the researcher is advised to prioritize the correlated variables only when necessary, as convergence of sensitivity estimates is much slower for the correlated inputs (Saltelli et al. 2004).

⁷ “The regression algorithm returns a regression meta-model, whereby the output Y is described in terms of a linear combination of the input factors” (Saltelli et al. 2004). Indeed, in a meta-modeling approach, instead of an original system, the effective model maps output and inputs. Then, the importance of model inputs is examined using the constructed surrogate model. Conversely, in a classical approach, the original system and the corresponding model output values are directly considered to estimate the sensitivity indices (Li et al. 2010).

$$S_i = \frac{V[E(Y|X_i)]}{V(Y)_{\text{unc}}} = \frac{\left[\frac{\text{COV}(\tilde{Y}, Z_1)}{V(Z_1)} \sigma_Y \right]^2 \left(\frac{u_{i1}}{\sigma_{X_i}} \right)^2 V(X_i)}{\left[\frac{\text{COV}(\tilde{Y}, Z_1)}{V(Z_1)} \sigma_Y \right]^2 \left[\left(\frac{u_{11}}{\sigma_{X_1}} \right)^2 V(X_1) + \dots + \left(\frac{u_{p1}}{\sigma_{X_p}} \right)^2 V(X_p) \right]} \quad (22)$$

where $V(Y)_{\text{unc}}$ stands for estimation of $V(Y)$ without considering the correlation among inputs. Since $U^T U = I$, the constraint of $U_1^T U_1 = 1$ is satisfied due to the orthonormal nature of eigenvectors, thus:

$$[u_{11}^2 + u_{21}^2 + u_{31}^2 + \dots + u_{p1}^2] = U_1 \cdot U_1 = U^T U = 1 \quad (23)$$

After proper simplification, we have:

$$S_i = \frac{V[E(Y|X_i)]}{V(Y)_{\text{unc}}} = u_{i1}^2, \quad i = 1, 2, \dots, p \quad (24)$$

This equation demonstrates the simplicity of the proposed method⁸ conferring a linear or nonlinear function with correlated input variables. The simplicity refers to the fact that sensitivity measures, i.e., S_i are equal to the square of i th component of the unit eigenvector corresponding to the dominant eigenvalue (i.e., first new feature).

It is worth noting that the classical Sobol or FAST as numerical methods of classical variance-based SA (applicable under the assumption of independent model inputs) can also be conducted on the Eq. (21) and the results of this coupling can be examined numerically. It is speculated that there is a small difference between the numerical results with that of the analytical solution emerging from Eq. (24).

At this stage, a serious question may arise. While the classical variance-based SA (e.g., Sobol and/or FAST) cannot be implemented on correlated data emerging from a linear or nonlinear model, how and why can such a tool be safely implemented on the regression-based SPCA model for evaluating the importance of correlated predictor variables? The following mathematical elaboration is intended to address this question. To this end, first, it was shown in appendix B that there is a relationship between the first eigenvalue of matrix $\mathbb{Q}$ assuming $L = \tilde{Y}\tilde{Y}^T$, the unit eigenvector components associated with the first eigenvalue, u_{i1} , and $\text{COV}(Y, X_i)$ as follows:

$$u_{i1} = \frac{\text{COV}(\tilde{Y}, \tilde{X}_i)}{(\lambda_1)^{.5}} = \frac{1}{(\lambda_1)^{.5} \times \sigma_Y \times \sigma_{X_i}} \text{COV}(Y, X_i), \quad i = 1, 2, \dots, p \quad (25)$$

Equation (25) is substituted into Eq. (24). As a result, the first-order sensitivity measure can also be computed as follows:

$$S_i = \frac{V[E(Y|X_i)]}{V(Y)_{\text{unc}}} = \frac{1}{\lambda_1 \times V(Y) \times V(X_i)} \text{COV}(Y, X_i)^2, \quad i = 1, 2, \dots, p \quad (26)$$

Equation (26) can be restated as:

$$S_i = \frac{V[E(Y|X_i)]}{V(Y)_{\text{unc}}} = \frac{\text{COV}(Y, X_i)}{V(Y)} \left[\frac{\text{COV}(Y, X_i)}{\lambda_1 \times V(X_i)} \right], \quad i = 1, 2, \dots, p \quad (27)$$

$$b_i = \left[\frac{\text{COV}(Y, X_i)}{\lambda_1 \times V(X_i)} \right]$$

In what follows, we managed to prove that the coefficient b_i is approximately equivalent to the corresponding coefficient obtained from the regression-based SPCA model, i.e., a_i . Therefore, it may be more practical to start with the coefficients of the regression-based SPCA model in reference to Eq. (21):

$$a_i = \frac{\text{COV}(\tilde{Y}, Z_1)}{V(Z_1)} \sigma_Y \left[\frac{u_{i1}}{\sigma_{X_i}} \right], \quad i = 1, 2, \dots, p \quad (28)$$

The coefficient a_i , can be simplified via the following mathematical manipulations. After substituting the standardized variables into Z_1 , $\text{COV}(\tilde{Y}, Z_1)$ can be written as:

$$\begin{aligned} \text{COV}(\tilde{Y}, Z_1) &= \text{COV}(\tilde{Y}, u_{11}\tilde{X}_1 + u_{21}\tilde{X}_2 + \dots + u_{p1}\tilde{X}_p) \\ &= \text{COV}(\tilde{Y}, u_{11}\tilde{X}_1) + \text{COV}(\tilde{Y}, u_{21}\tilde{X}_2) \\ &\quad + \dots + \text{COV}(\tilde{Y}, u_{p1}\tilde{X}_p) \\ &= u_{11}\text{COV}(\tilde{Y}, \tilde{X}_1) + u_{21}\text{COV}(\tilde{Y}, \tilde{X}_2) \\ &\quad + \dots + u_{p1}\text{COV}(\tilde{Y}, \tilde{X}_p) \end{aligned} \quad (29)$$

Following Eq. (25), the $\text{COV}(\tilde{Y}, Z_1)$ becomes:

$$\begin{aligned} \text{COV}(\tilde{Y}, Z_1) &= \frac{[\text{COV}(\tilde{Y}, \tilde{X}_1)]^2}{(\lambda_1)^{.5}} \\ &\quad + \frac{[\text{COV}(\tilde{Y}, \tilde{X}_2)]^2}{(\lambda_1)^{.5}} + \dots + \frac{[\text{COV}(\tilde{Y}, \tilde{X}_p)]^2}{(\lambda_1)^{.5}} \end{aligned} \quad (30)$$

⁸ It should be remembered that the proposed method is based on conducting the classical variance-based on the regression-based SPCA, and the SPCA approach is in reference to optimization of the HSIC method, i.e., Eq. (13). In addition, the HSIC approach proposed by Gretton et al. (2005) is based on using covariance operators in reproducing kernel Hilbert spaces (RKHSs). The details of HSIC and its philosophy can be found in Gretton et al. (2005).

As proved in appendix B, the dominant eigenvalue, λ_1 , is:

$$\lambda_1 = \sum_{i=1}^p [\text{COV}(\tilde{Y}, \tilde{X}_i)]^2 \quad (31)$$

After combining Eq. (30) and (31), $\text{COV}(\tilde{Y}, Z_1)$ can be written as:

$$\text{COV}(\tilde{Y}, Z_1) = (\lambda_1)^{.5} \quad (32)$$

Now, to further simplify the coefficients of the regression-based SPCA model, the computation of $V(Z_1)$ can be accomplished. For this purpose, considering Eq. (18) once again and taking variance of both sides results in:

$$\begin{aligned} \tilde{Y} &= \beta_1 Z_1 + \beta_0 \\ V(\tilde{Y}) &= V(\beta_1 Z_1 + \beta_0) \\ V(\tilde{Y}) &= \beta_1^2 V(Z_1) \\ V(\tilde{Y}) &= \left(\frac{\text{COV}(\tilde{Y}, Z_1)}{V(Z_1)} \right)^2 V(Z_1) \end{aligned} \quad (33)$$

Since $\tilde{Y}$ is a standardized variable, $V(\tilde{Y}) = 1$. Therefore, by substituting Eq. (32) into Eq. (33) we have:

$$\begin{aligned} 1 &= \frac{\lambda_1}{V(Z_1)} \\ V(Z_1) &= \lambda_1 \end{aligned} \quad (34)$$

Equation (34) is very applicable to linear models. However, as one departs from linearity, then $V(Z_1)$ departs from the dominant eigenvalue, λ_1 . As a result, $V(Z_1) \approx \lambda_1$.

After replacing u_{i1} and $\text{COV}(\tilde{Y}, Z_1)$ with their equivalences, i.e., Eq. (25) and Eq. (32), respectively into Eq. (28), the coefficients a_i are simplified to the following relation:

$$\begin{aligned} a_i &= \frac{\text{COV}(\tilde{Y}, Z_1)}{V(Z_1)} \sigma_Y \left[\frac{u_{i1}}{\sigma_{X_i}} \right] \\ &= \frac{(\lambda_1)^{.5}}{V(Z_1)} \sigma_Y \left[\frac{\text{COV}(Y, X_i)}{(\lambda_1)^{.5} \times \sigma_Y \times \sigma_{X_i} \times \sigma_{X_i}} \right] \\ a_i &= \frac{1}{V(Z_1)} \left[\frac{\text{COV}(Y, X_i)}{V(X_i)} \right], i = 1, 2, \dots, p \end{aligned} \quad (35)$$

Based on Eq. (34) and subsequent comment, the coefficients a_i can be further simplified to:

$$a_i \approx \frac{\text{COV}(Y, X_i)}{\lambda_1 \times V(X_i)}, i = 1, 2, \dots, p \quad (36)$$

Therefore, in reference to the above equation, after implementing the classical variance-based SA on the regression-based SPCA model, the first-order sensitivity measure can be written as:

$$\begin{aligned} S_i &= \frac{V[E(Y|X_i)]}{V(Y)_{\text{unc}}} = u_{i1}^2 \\ &= \frac{\text{COV}(Y, X_i)}{V(Y)} \left[\frac{\text{COV}(Y, X_i)}{\lambda_1 \times V(X_i)} \right] \approx \frac{\text{COV}(Y, X_i)}{V(Y)} a_i \end{aligned} \quad (37)$$

Finally:

$$\begin{aligned} S_i &= \frac{V[E(Y|X_i)]}{V(Y)_{\text{unc}}} = u_{i1}^2 \\ &\approx \frac{\text{COV}(Y, a_i X_i)}{V(Y)}, i = 1, 2, \dots, p \end{aligned} \quad (38)$$

Moreover, in appendix C, first-order sensitivity measure, [i.e., Eq. (38)] was related to two sources of variability, i.e., the variability due to each variable and the variability due to co-variation among the variable under consideration and other predictor variables as stipulated below:

$$\begin{aligned} S_i^{tc} &= \frac{\text{COV}(Y, a_i X_i)}{V(Y)} \\ &= \frac{V(a_i X_i)}{V(Y)} + \frac{\text{COV}\left(a_i X_i, \sum_{j=1, j \neq i}^p a_j X_j\right)}{V(Y)}, i = 1, 2, \dots, p \end{aligned} \quad (39)$$

It is worth noting that the derived covariance decomposition is in line with the covariance decomposition, first proposed by Li et al. (2010). In short, referring to the question raised in the theoretical background section, the above mathematical manipulations imply that while the classical variance-based SA (under the assumption of the independent factors) cannot be used directly on the original model with correlated input, it is possible to effectively benefit from the advantage of SPCA and combine it with this type of variance-based SA to devise a better tool for prioritizing the input variables when the input variables are correlated.

To dig further into the proposed method, the following algorithm intends to document the coupling of the classical variance-based SA with regression-based SPCA in more detail.

Algorithm 1: the proposed method, i.e., performing variance-based SA on regression-based SPCA

1. Read the observation of model inputs
 2. Estimate the correlation matrix of model inputs
 3. Determine the probability density function of model inputs
 4. Generate the correlated samples for correlated inputs based on the correlation matrix and their probability density function using Iman and Conover (1982) approach
 5. Execute the model using the correlated samples of inputs to generate the output
 6. Scale the generated inputs and output to get $\tilde{X}$ and $\tilde{Y}$
-

7. Compute the centering matrix: $H_{ij} = I - \frac{1}{n}ee^T$, and linear kernel imposed on output: $L = \tilde{Y}\tilde{Y}^T$
8. Construct the matrix $\mathbb{Q} = \tilde{X}^T H L H \tilde{X}$
9. Estimate the eigenvalues and the corresponding eigenvectors of matrix $\mathbb{Q}$
10. Derive the first new feature ($Z_1 = \tilde{X}U_1$) associated with dominant eigenvalue as other eigenvalues are zero
11. Construct the regression-based SPCA model based on Eq. (21)
12. Estimate the coefficient of determination (R^2) based on Eq. (6)
13. Compute the first-order sensitivity measure based on the square of the unit eigenvector components corresponding to the dominant eigenvalue [Eq. (24)] as a semi-analytical method or perform classical FAST or Sobol on the regression-based SPCA-model as a numerical method
14. Report the importance of correlated input based on step 13
15. Derive the approximated equivalence of the proposed method (S^c) [Eq. (39)]
16. Compare the proposed method with the approximated equivalence of the proposed method and the developed variance-based SA approach, i.e., correlation ratio method as a benchmark solution

It is worth noting that, if we have a system for which the distributions of model input and their correlation structure are given (i.e., a system with simulated data), then omit the first three steps of the algorithm and start from the next step.

After proposing the above approach, an important question can be raised by the reader: is it possible to conduct the classical variance-based SA (applicable under the assumption of independent inputs, e.g., Sobol and/or FAST) on PCR (a special form of the SPCA) to differentiate between the important and irrelevant variables in models with correlated inputs? It depends on the nature of the problem at hand. The correlation of new predictor variables with the output and/or lack of it could be the cause of success or failure. The process to address this issue is very similar to what was done while coupling the classical variance-based SA with the regression-based SPCA model. In order to keep the integrity of the material in place, interested readers are referred to appendix D for further details regarding the coefficients in Eq. (40). In summary, the PCR model can be expressed as a linear combination of original input variables as follows:

$$\begin{aligned}
 Y &= \frac{c_1\sigma_Y}{\sigma_{X_1}}X_1 + \frac{c_2\sigma_Y}{\sigma_{X_2}}X_2 + \cdots + \frac{c_p\sigma_Y}{\sigma_{X_p}}X_p \\
 &+ E(Y) + \beta_0\sigma_Y - \sum_{i=1}^p c_i \frac{E(X_i)}{\sigma_{X_i}} \\
 c_i &= \frac{\text{COV}(\tilde{Y}, Z'_1)}{\lambda_1} u'_{i1} + \frac{\text{COV}(\tilde{Y}, Z'_2)}{\lambda_2} u'_{i2} + \cdots \\
 &+ \frac{\text{COV}(\tilde{Y}, Z'_m)}{\lambda_m} u'_{im}, i = 1, 2, \dots, p
 \end{aligned} \quad (40)$$

Thus, after coupling the variance-based SA (under the assumption of the independent inputs) with PCR, the first-order sensitivity measure can be written as:

$$\begin{aligned}
 S_i &= \frac{V[E(Y|X_i)]}{V(Y)_{\text{unc}}} \\
 &= \frac{\left(\frac{c_i\sigma_Y}{\sigma_{X_i}}\right)^2 V(X_i)}{\left(\frac{c_1\sigma_Y}{\sigma_{X_1}}\right)^2 V(X_1) + \left(\frac{c_2\sigma_Y}{\sigma_{X_2}}\right)^2 V(X_2) + \cdots + \left(\frac{c_p\sigma_Y}{\sigma_{X_p}}\right)^2 V(X_p)}
 \end{aligned} \quad (41)$$

After some manipulations, the above equation is simplified as:

$$S_i = \frac{V[E(Y|X_i)]}{V(Y)_{\text{unc}}} = \frac{c_i^2}{[c_1^2 + c_2^2 + \cdots + c_p^2]}, i = 1, 2, \dots, p \quad (42)$$

In subsequent sections, a few numerical test cases are devised to either confirm or refute the validity of the above coupling exercise.

The procedure of coupling the variance-based SA with PCR is shown in the following algorithm:

Algorithm 2: The details of coupling classical variance-based SA with PCR

1. Read the observation of model inputs
2. Estimate the correlation matrix
3. Determine the distribution of model inputs
4. Generate the correlated samples for correlated inputs based on the correlation matrix and their probability density function
5. Run the model using the correlated samples of inputs to generate the output
6. Scale the generated inputs and output
7. Estimate the eigenvalues and corresponding eigenvectors of the correlation matrix $\tilde{X}^T \tilde{X}$, i.e., U and λ
8. Compute the random vector of new features i.e., $Z' = \tilde{X}U$
9. Construct the principal component regression like Eq. (40)
10. Compute the first-order sensitivity measure by conducting the classical variance based on PCR. It can be done numerically by performing classical Sobol or Fast on PCR or analytically using Eq. (42)
11. Report the importance of correlated input based on step 10

Similar to algorithm 1, ignore the first three steps of this algorithm for a system with simulated data.

To better appreciate the proposed method, i.e., coupling the classical variance-based SA with regression-based SPCA model, and comparing it with the variance-based PCR method, the following flowchart can be used (See Fig. 1):

4 Applications and results

In this section, the proposed method based on coupling the classical variance-based SA (applicable under the assumption of independent inputs) with either regression-based SPCA or PCR is applied to five test cases as well as three simple hydrologic and hydraulic models to evaluate the effectiveness of the corresponding modified model. From the theoretical background section, it is quite clear that implementation of either PCR and/or regression-based SPCA calls for the generation of realizations for various input variables. This task, i.e., generation of correlated realizations, is performed via SIMLAB2.2 (The SAMO team, 2003). As the governing regression model (e.g., regression-based SPCA) is linear, the number of realizations to build this regression can be kept quite small (Song et al. 2015), e.g., 500 samples for models with low number of dimensions and 2000 samples in the case of high dimensionality. In the meanwhile, a few test cases were run with different sample sizes to examine the impact of various number of realizations on the results of the proposed scheme.

It should be remembered that after conducting eigenvalue analysis on $\mathbb{Q}$ extracted from SPCA method, the aforementioned matrix has only one dominant eigenvalue and associated unit eigenvector. However, when it comes to PCR, it has more than one dominant eigenvalue. For this

reason, the impact of a different number of components on model performance was investigated. It is worth noting that if all components are included in the regression model, the PCR coefficients are equivalent to ordinary least squares (Jolliffe 1986).

Although we analytically touched on the validity of the proposed scheme, the correlation ratio method was also used as a benchmark solution implemented on the original model to further evaluate the hypothesis incorporated into various test cases. This benchmark, which is based on McKay's approach (McKay 1997; Saltelli et al. 2001), is always recognized as a valid approach due to its non-parametric nature (Xu and Gertner 2008). For this reason, this benchmark solution, which is extracted from SIMLAB 2.2, is suitable for nonlinear models even in the presence of strong nonlinear effects. However, as the method uses replicated LHS, a large sample size is required to acquire suitable accuracy. In this work, like Xu and Gertner (2008), we suggest 100 replications with each replication having a sample size of 500 (a total of 50,000 model runs) for most of the test functions and 150,000 samples (150 replication with each replication having a sample size of 1000) for the Sobol G function with 35 number of correlated input variables. Finally, the negative impact of ignoring the correlation between inputs is examined by performing Sobol numerical variance-based SA (which assumed the model inputs are independent) on all test functions.

Fig. 1 Flowchart of the proposed method and the variance-based PCR method

It is quite helpful to note that the proposed method was written in R programming environment.⁹

4.1 Test case 1: A linear model with constant coefficients

This test case, which had also been used by Li et al. (2010), is a simple and additive model with five inputs:

$$Y = X_1 + X_2 + X_3 + X_4 + X_5 \quad (43)$$

In this case, the marginal distribution of each input variable is normal with a mean of 0.5 and a unit standard deviation, i.e. $X_i \sim N(0.5, 1)$. The Spearman Rank Correlation Coefficient (SRCC), or it is better to say the Spearman Rank Correlation matrix, of input variables is defined as follows:

$$\begin{bmatrix} 1 & 0.6 & 0.2 & 0 & 0 \\ 0.6 & 1 & 0.2 & 0 & 0 \\ 0.2 & 0.2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0.2 \\ 0 & 0 & 0 & 0.2 & 1 \end{bmatrix}$$

Clearly, in the case of independent inputs, it is easy to deduce each variable has the same impact on model output uncertainty due to their equal SRCs. Taking into account the correlation structure of input variables, each variable has a different impact. Table 1 summarizes the result of sensitivity analysis using four different approaches including the Sobol method implemented on the original model, the proposed approach, three flavors of variance-based PCR, and the benchmark solution implemented on the original model as well as the approximated equivalence of the proposed method (S^{tc}) obtained from covariance decomposition implemented on the regression-based SPCA model.

In reference to the numerical values of S_i for various approaches and different input variables in Table 1, after clustering the input variables into three clusters, the solution associated with the benchmark solution has X_1, X_2 in cluster (1), X_3 in cluster (2), and X_4, X_5 in cluster three. It is evident that the sorting in the proposed approach and its equivalence (S^{tc}) is consistent with the benchmark solution. However, neither the Sobol implemented on the original model, nor three flavors of the variance-based PCR managed to reproduce the benchmark solution.

The change in numerical values of S_i from one approach to another is highly coined with the approach itself and one should not expect similar values emerging from each approach. Indeed, for sensitivity analysis purposes,

⁹ Implementation of both the proposed approach and the benchmark software was carried out with the following HW and SW facilities: Processor: Intel(R) Core (TM) i5-2450 M, CPU@ 2.50 GHz, Cache: 3 MB, Installed memory (RAM): 4.00 GB DDR 3, System type: 64 bit operating system, Windows: 7.

detecting the importance of input variables and their rankings in each approach will become quite important.

Figure 2 investigates the impact of the number of realizations on sensitivity measures based on the proposed method. As the figure shows, the number of realizations has minimal effect on S_i (i.e., the aforementioned clustering) and one can safely choose a minimum number of realizations (e.g., 500) for sensitivity analysis.

4.2 Test case 2: A linear model with variable coefficients

Like the preceding test case, the second test case is also adopted from Li et al. (2010) as follows:

$$Y = 5X_1 + 4X_2 + 3X_3 + 2X_4 + X_5 \quad (44)$$

Except the equation, all other conditions of this case are very similar to the first test case. The results of the estimated sensitivity measures based on the aforementioned methods cited in Test case 1 are summarized in Table 2.

According to the content of this table, the rankings of model variables based on the proposed method are very similar to the correlation ratio scheme. Furthermore, the importance of inputs in reference to the proposed method are the same as those obtained from S^{tc} . Likewise, the performance of coupling PCR and classical variance-based in recognizing the sensitivity ordering of input variables agrees with the benchmark. In this test case, it seems the correlation structure has a minimal impact on prioritizing the input variables. However, the model structure seems to change the destiny of variables with regard to their importance.

4.3 Test case 3: A simple nonlinear model

To evaluate the performance of the proposed method when the transfer function is nonlinear with correlated inputs, the simplest nonlinear model with three-variable inputs, first proposed by Xu and Gertner (2008), is considered in this study:

$$Y = \frac{X_1 X_2}{X_3} \quad (45)$$

The probability density function of each input is uniform, i.e., $X_i \sim U(1, 10)$, and the SRCC of the model inputs are:

$$\begin{bmatrix} 1 & 0.4 & 0.2 \\ 0.4 & 1 & 0.4 \\ 0.2 & 0.4 & 1 \end{bmatrix}$$

Table 3 shows the results of SA regarding the methods cited earlier.

As the table shows, the most important input variable according to the Sobol scheme is X_3 . However, both the

Table 1 Summary of SA–Test case 1

	Variance-based (Sobol)	Proposed method $R^2 = 0.9836$	Coupling variance-based SA with PCR			Benchmark	S^{tc}
			(2-comp) $R^2 = 0.9977$ $\sum_5 \lambda_i = 0.614$	(3-comp) $R^2 = 0.9995$ $\sum_5 \lambda_i = .776$	(4-comp) $R^2 = 1$ $\sum_5 \lambda_i = .924$		
X_1	0.2035	0.2663	0.2200	0.2009	0.1999	0.4641	0.2889
X_2	0.2011	0.2705	0.2250	0.2057	0.2047	0.4820	0.2933
X_3	0.2021	0.1938	0.1507	0.1905	0.2021	0.3128	0.1940
X_4	0.2041	0.1323	0.2061	0.1826	0.2041	0.2381	0.1093
X_5	0.2039	0.1371	0.2130	0.2202	0.2038	0.2406	0.1146

Fig. 2 Sensitivity measure (S_i) versus input variables for various number of realizations–Test case 1

proposed scheme as well as the benchmark solution found the most important variable to be X_1 . Moreover, the result is quite consistent with S^{tc} . By contrast, conducting the classical variance-based approach on PCR using either one and/or two components, cannot capture the degree of importance of input variables.

4.4 Test case 4: A nonlinear model—a typical version of the portfolio model

This nonlinear test function, which is a typical version of the portfolio model, has four variables with the following equation:

$$Y = X_1 X_3 + X_2 X_4 \quad (46)$$

This relatively strong test case with respect to the degree of nonlinearity, as relates to interactions among the input variables, can be examined in the following three scenarios.

4.4.1 The first scenario

This scenario was utilized by Kucherenko et al. (2012). The marginal distributions of variables are normal, i.e., $X_i \sim N(\mu, \sigma)$ with $\mu \equiv (0, 0, 250, 400)$, $\sigma \equiv (4, 2, 200, 300)$ and Spearman rank correlation among factors are: $\rho_{12}^s = 0.8$, $\rho_{34}^s = 0.8$ and the SRCC for the remaining variables is assumed to be zero. The importance of each input based on the mentioned approaches is quantified and documented in Table 4.

As Table 4 clearly demonstrates, the suggested implemented technique along with the results summarized in the last column (S^{tc}) and yardstick solution are in good

Table 2 Summary of SA of–Test case 2

	Variance-based (Sobol)	Proposed method $R^2 = 0.997$	Coupling variance-based SA with PCR			Benchmark	S^{tc}
			(2-comp) $R^2 = 0.9934$ $\sum_5 \lambda_i = 0.614$	(3-comp) $R^2 = 0.9953$ $\sum_5 \lambda_i = .776$	(4-comp) $R^2 = 0.9979$ $\sum_5 \lambda_i = .9243$		
X_1	0.4607	0.3897	0.3673	0.3837	0.3680	0.7128	0.3831
X_2	0.2927	0.3588	0.3683	0.3871	0.3763	0.6629	0.3662
X_3	0.1653	0.1792	0.1929	0.1443	0.1646	0.3018	0.1827
X_4	0.0741	0.0457	0.0397	0.0480	0.0722	0.0894	0.0399
X_5	0.0188	0.0268	0.0435	0.0368	0.0188	0.0509	0.0280

Table 3 Summary of SA–Test case 3

	Variance-based (Sobol)	Proposed method $R^2 = 0.6096$	Coupling variance-based SA with PCR		Benchmark	S^{tc}
			(1-comp) $R^2 = 0.1131$, $\sum_3 \lambda_i = .579$	(2-comp) $R^2 = 0.5932$, $\sum_3 \lambda_i = .837$		
X_1	0.1706	0.5369	0.2611	0.7541	0.3183	0.6228
X_2	0.1789	0.3309	0.3032	0.0265	0.2057	0.3975
X_3	0.4176	0.1322	0.4351	0.2186	0.1236	0.0203

agreement. This could be attributed to the fact that all schemes take care of the impact of correlation structure among input variables effectively. Furthermore, this ranking can be achieved using PCR based on two components. In contrast, performing the classical variance-based SA on PCR regression using three components only ranks the inputs correctly and cannot differentiate between the importance of X_1 and X_2 . Similar to the test case 2, adding the correlation structure cannot change the rank of the inputs, although with respect to the correlation among inputs, X_2 becomes slightly less important than X_1 .

4.4.2 The second scenario

Except for the correlation structure, all conditions of this test case are the same as the preceding scenario. The correlation structure is: $\rho_{14}^s = 0.8$, $\rho_{34}^s = 0.3$ and the remaining pairs are assumed to be independent. Figure 2 demonstrates the scatter plot matrix for this test case. The results of SA are summarized in Table 5.

In reference to the content of Table 5, as far as the coupling of PCR with the classical variance-based SA is concerned, this coupling cannot successfully delineate the most important variables and their proper rankings. However, the proposed method managed to capture the benchmark solution and the results are quite compatible with S^{tc} summarized in the last column of the table, albeit with the rather small Multiple R-squared: 0.52. The low value of the coefficient of determination stems from the

nonlinear relationship between inputs and output as well as the interaction among variables. The scatterplot depicted in Fig. 3 clearly demonstrates this moderate trend in data which somehow leads to the low value of R^2 .

4.4.3 The third scenario

This scenario is proposed by Ge and Menendez (2017), they assumed the PDF associated with each random variable has a different form. In other words, in this case, the distributions of X_1, X_2, X_3 and X_4 are respectively, normal $N(0, 1)$, gamma $\Gamma(2, 1)$, uniform $\mathcal{U}(0, 1)$, and lognormal $\log N(0, 1)$. The rank correlation coefficients are the same as the preceding scenario, i.e., $\rho_{14}^s = 0.8$, $\rho_{34}^s = 0.3$ and $\rho_{ij}^s = 0$ for $i \neq 1, 3$ and $j \neq 4$. The first-order sensitivity measures based on the adopted approaches can be seen in Table 6.

Concerning the different governing PDF for each predictor variable, as Table 6 demonstrates, the proposed method managed to mimic the variation of sensitivity measures documented by the benchmark solution as well as S^{tc} , while conducting the classical variance-based on PCR has failed to reproduce the proper ranking of model inputs.

4.5 Test case 5: the high-dimensional model similar to Sobol G function

As an additional more comprehensive example, this test case is designed to evaluate the ability of the proposed

Table 4 Summary of SA of the first scenario–Test case 4

	Variance-based (Sobol)	Proposed method $R^2 = .591$	Coupling variance-based SA with PCR		Benchmark	S^{tc}
			(2-comp) $R^2 = .593$, $\sum_4 \lambda_i = .905$	(3-comp) $R^2 = .594$, $\sum_4 \lambda_i = .955$		
X_1	0.3802	0.5140	0.5024	0.5564	0.5791	0.5045
X_2	0.2399	0.4822	0.4975	0.4412	0.5550	0.4880
X_3	0.0074	0.0012	0.0001	0.0009	0.0105	0.0029
X_4	0.0049	0.0026	2.6871e-07	0.0015	0.0181	0.0047

Table 5 Summary of SA of the second scenario–Test case 4

	Variance-based (Sobol)	Proposed method $R^2 = 0.512$	Coupling variance-based SA with PCR		Benchmark	S^{tc}
			(2- comp) $R^2 = 0.371, \sum_4 \lambda_i = .730$	(3- comp) $R^2 = 0.554, \sum_4 \lambda_i = .965$		
X_1	0.3802	0.4296	0.2539	0.2832	0.3755	0.4375
X_2	0.2399	0.2622	0.1298	0.4857	0.2548	0.1813
X_3	0.0074	0.0014	0.2113	0.0343	0.0104	0.0064
X_4	0.0050	0.3068	0.4051	0.1967	0.3200	0.3748

Fig. 3 Scatterplot matrix for the second scenario of test case 4

scheme in the presence of a high dimensional problem. This model has the following equation:

$$F = \prod_{i=1}^p f_i(X_i), f_i(X_i) = \frac{2X_i + d_i}{1 + d_i} \quad (47)$$

The variables incorporated into this test case have uniform standard marginal distribution and the parameter d_i is always nonnegative. Equation (47) is able to accommodate the interaction. The Sobol G function, a strongly nonlinear and non-monotonic model, can be generated by replacing X_i with $|2T_i - 1|$. The test case considered in this study has a lower degree of nonlinearity and is not as complex as the Sobol G function. In the Sobol G function, by increasing d_i , the importance of the corresponding variable will be reduced (Saltelli et al. 2008). This mathematical feature is very applicable to the proposed test function as well. In this test case, we assume the F -function is comprised of 35

variables and their associated coefficients are listed in Table 7.

As the content of this table shows, the importance of the orthogonal inputs based on their associated coefficients (d_i) are divided into four groups. Indeed, here we want to examine the ability of the proposed method in the classification of inputs in which most of them are correlated. For this reason, the SRCCs between strongly influential variables and non-influential inputs, i.e., (X_1, X_{35}), (X_2, X_{34}), (X_3, X_{33}), (X_4, X_{32}) and (X_5, X_{31}) are 0.8. In addition, it is assumed that the SRCCs between moderately influential variables and non-influential inputs, i.e., (X_6, X_{30}), (X_7, X_{29}) and (X_8, X_{28}) are 0.7. Furthermore, this kind of correlation between (X_{11}, X_{12}), (X_{12}, X_{13}), (X_{13}, X_{14}),...and (X_{24}, X_{25}) is 0.2. By considering this correlation structure and other pertinent information, the estimated first-order sensitivity measures can be seen in Figs. 4–8 for all schemes considered in this study.

It is worth mentioning the variance and coefficient of determination for the number of components in PCR method, i.e., 17 to 28 are documented in Tables 8 and 9.

Figure 4 shows that both the proposed scheme as well as the S^{tc} approach can reproduce the grouped ranking stipulated by the benchmark solution. In other words, the proposed method can successfully delineate the most important variables ($[X_1, X_2, \dots, X_5]$ and $[X_{31}, X_{32}, \dots, X_{35}]$), moderately important variables ($[X_6, X_7, \dots, X_{11}]$ and $[X_{28}, X_{29}, X_{30}]$), weakly important ($[X_{12}, X_{13}, \dots, X_{25}]$) and non-influential variables ($[X_{26}, X_{27}]$). It is worth mentioning that the coefficient of determination between the original model and regression-based SPCA model is 0.33. This implies that the proposed method can be used for quite strong nonlinear models. In addition, as the number of samples of the proposed method, i.e., 2000 samples in comparison with the benchmark solution i.e., 150,000 samples is low, the suggested technique can be considered as a computationally efficient approach.

If one ignores the correlation structure inherent in input variables, the conventional scheme, Sobol, fails to highlight the importance of the last eight variables. However, the correlation structure inherent in input variables triggers

Table 6 Summary of SA of the third scenario–Test case 4

	Variance-based (Sobol)	Proposed method $R^2 = 0.685$	Coupling variance-based SA with PCR		Benchmark	S^{tc}
			(2- comp.) $R^2 = 0.613, \sum_4 \lambda_i = .691$	(3- comp.) $R^2 = 0.664, \sum_4 \lambda_i = .925$		
X_1	0.0087	0.2447	0.2148	0.2933	0.3181	0.3014
X_2	0.1695	0.1632	0.2979	0.3591	0.1890	0.1211
X_3	0.0014	0.0352	0.1666	0.0280	0.0562	0.0496
X_4	0.5057	0.5569	0.3207	0.3197	0.5835	0.5280

Table 7 Parameter allocation to various groups in high dimensional test function (Sheikholeslami et al. 2019)

Group	coefficient
Considerably important	$d_1 = d_2 = 0$ $\{d_3, d_4, d_5\} = \{0.005, 0.01, 0.02\}$
Moderately important	$\{d_6, d_7, \dots, d_{11}\} = \{2, 2.75, 3, 3.15, 3.25, 3.5\}$
Weakly important	$\{d_{12}, d_{13}, \dots, d_{25}\} = \{8, 8.5, 9, 10, 10.5, 11, 12, 12.5, 13, 13.5, 14, 14.5, 15, 16\}$
Not important	$\{d_{26}, d_{27}, \dots, d_{35}\} = \{70, 71, 72, 73, 75, 80, 85, 90, 95, 99\}$

both the proposed scheme as well as the benchmark solution to give appropriate grouped ranking to input variables. Upon coupling the classical variance-based SA with PCR (see Fig. 5, 6, 7, 8), the first-order sensitivity measures of the last eight variables become more distinct but the coupling exercise did not manage to regenerate the ranking associated with those variables compared to the benchmark solution.

4.6 Practical test cases

In this part, three practical test cases are considered. Two of them are based on two nonlinear hydrologic models which used real observed data but one of them is based on a low dimensional model and another regards a high dimensional model. In addition, the last test case considers a hydraulic model to simulate flood inundation using the simulated data. The details of these models are discussed below:

4.6.1 Test case 6: A simple hydrologic rainfall-runoff model

This test case represents a real hydrologic model that relates instantaneous peak discharge due to rainfall to watershed characteristics in Indiana by a power model as given below (McCuen and Snyder 1986):

$$Q_{10} = 5.23A^{0.6965}S^{0.6565}L_c^{0.223}P_i^{0.7009}H_r^{0.1497}d^{-0.1053}C^{1.1976} \quad (48)$$

When runoff starts, the basin characteristics control the pattern of the flood and affect the flood peaks (Benson 1964). For this reason, using this model is reasonable. This model is a real application but with low dimension in operational hydrology. In this model, real data is used to calibrate a nonlinear model with 7 decision variables. Based on the observed and measured data of watershed characteristics in 58 watersheds in Indiana and independent estimation of 10-year instantaneous peak discharge, i.e., Q_{10} , using historic data for each watershed, the parameters of the power model are estimated by McCuen and Snyder (1986). Then, the probability density function of the model inputs is chosen based on these observed data (i.e., 58 instances) via EasyFit software. Finally, the basin characteristics and their probability density functions are described in Table 10.

The correlation structure of the model inputs based on SRCC measure, using the observed data, is as follows:

	A	S	L_c	P_i	H_r	D	C
A	1	-.44	0.75	-0.01	0.09	-.01	.05
S	-.44	1	-0.22	0.61	0.63	0.57	0.52
L_c	0.75	-0.22	1	0.15	0.40	0.10	0.15
P_i	-0.01	0.61	0.15	1	0.81	0.73	0.79
H_r	0.09	0.63	0.40	0.81	1	0.76	0.62
D	-0.01	0.57	0.10	0.73	0.76	1	0.55
C	0.05	0.52	0.15	0.79	0.62	0.55	1

After implementing the adopted approaches, including Sobol on the original model, the proposed approach, five

Fig. 4 Summary of SA—Test case5

Table 8 The variance and the coefficient of determination for number of components in PCR method (17–22)

	17- comp	18-comp	19- comp	20- comp	21- comp	22- comp
$\sum_{35} \lambda_i$	0.7288	0.7559	0.7824	0.8086	0.8334	0.8561
R^2	0.3256	0.3257	0.3273	0.3298	0.3304	0.3304

Table 9 The variance and the coefficient of determination for number of components in PCR method (23– 28)

	23-comp	24-comp	25-comp	26-comp	27-comp	28-comp
$\sum_{35} \lambda_i$	0.8760	0.8950	0.9133	0.9303	0.9465	0.9553
R^2	0.3306	0.3307	0.3307	0.3307	0.3308	0.3309

flavors of variance-based PCR, and the benchmark solution implemented on the original model as well as the approximated equivalence of the proposed method (S^{tc}), on the respective governing models, the first-order sensitivity measures are summarized in Tables 11 and 12.

As can be seen in Table 11, the most influential parameters are H_r , P_i , and C based on the proposed method. Nevertheless, it is difficult to differentiate among

variables such as P_i and C when it comes to the sensitivity measures. In addition, parameters such as (D), (S , L_c), and (A) are divided into the third, fourth, and fifth groups in terms of their influences, respectively. This classification, borne out by the proposed approach is very consistent with the correlation ratio method and S^{tc} . It seems in coupling the classical variance-based SA with PCR (Table 12), the number of principal components considered is quite

Fig. 5 Summary of SA—Test case5 **Fig. 6** Summary of SA—Test case5 **Fig. 7** Summary of SA—Test case5 **Fig. 8** Summary of SA—Test case 5

Table 10 Description of output-inputs–Test case 6 (McCuen and Snyder 1986)

Variable	Meaning	Distribution
A	Drainage area (mi ²)	Beta (0.93, 1.28, 14.6, 203)
S	Channel longitudinal slope (ft/ft)	Gamma (1.66, 0.2421, 1.05)
L_c	Channel length (mi), distance along a stream from the watershed divide to the mouth of the watershed	Beta (1.84, 5.31, 5.64, 70.1)
P_i	Precipitation index (in), mean annual precipitation minus the sum of average annual evapotranspiration and mean annual snowfall (water equivalent)	Beta (2.21, 2.63, 3.69, 18.5)
H_r	Watershed relief (ft), the difference in elevation between the highest point on the watershed divide and the stream at the point of discharge	Gamma (1.6, 0.0089, 39.89)
D	Drainage density (mi/mi ²), the total stream length divided by watershed area	Triangular (1.29, 2.5, 14.35)
C	Runoff coefficient, rainfall volume divided by the total volume of runoff	Triangular (0.19, 0.7, 1.06)
Q_{10}	Instantaneous peak discharge (ft ³ /s) for a 10-year return period	–

influential in delineating the important parameters. In our case, beyond the two principal components, the PCR approach is unable to reproduce the sensitivity ordering inherent in model input variables. Interestingly, upon ignoring the correlation structure of input variables, the Sobol approach, implemented on the original model, found the drainage area to be the most influential model input. Actually, according to the Sobol approach, important variables are found to be unimportant (i.e., type II error) while less influential parameters are considered to be quite important (i.e., type I error).

Figure 9 considers the impact of the number of realizations on sensitivity measures based on the proposed scheme. As the figure clearly demonstrates, the number of realizations has minimal effect on S_i and one can safely choose a minimum number of realizations (e.g., 500) for sensitivity analysis.

4.6.2 Test case 7—Snowmelt flood model

This model is similar to the previous real test function which relates the flood peak to hydrologic factors, but in

this model as a high dimensional model, we want to examine the importance of the 14 correlated factors on the uncertainty of snowmelt-flood as a response surface of this model. As the timing of flood hydrograph will be controlled by the basin characteristics, the observed data of 35 watersheds in Northern New Mexico and Colorado are extracted from the Benson report (1964). The main runoff mechanism in those watersheds is from snowmelt. The nonlinear model can be stated as follows:

$$Q_{10} = 1.36 \times 10^{15} \frac{A^{0.4798} L_b^{0.2466} P_m^{1.5347} t_6^{0.3670} S_n^{0.2797} R_a^{0.6525}}{S^{0.208} S_t^{0.3154} E_a^{2.1466} H_r^{0.0985} I_h^{1.8067} N^{3.5268} t_1^{1.1149} W_e^{0.0480}} \quad (49)$$

The basin characteristics along with the corresponding PDF for each feature, derived from observed data, are listed in Table 13.

In addition, the correlation matrix of the model inputs based on SRCC measure, using the observed data, can be seen as follows:

	A	S	S_t	E_a	L_b	H_r	P_m	I_h	N	t_1	t_6	S_n	W_e	R_a
A	1	-0.84	-0.14	-0.74	0.91	-0.20	-0.53	-0.61	-0.22	-0.57	0.74	-0.61	-0.59	-0.48
S	-0.83	1	-0.05	0.59	-0.90	0.53	0.53	0.76	0.37	0.42	-0.63	0.48	0.35	0.32
S_t	-0.14	-0.06	1	0.07	-0.09	-0.30	0.49	0.03	0.10	0.05	-0.19	0.32	0.26	0.54
E_a	-0.74	0.59	0.07	1	-0.72	-0.01	0.44	0.39	-0.24	0.91	-0.94	0.83	0.84	0.47
L_b	0.91	-0.91	-0.09	-0.72	1	-0.20	-0.58	-0.71	-0.24	-0.56	0.76	-0.60	-0.51	-0.43
H_r	-0.20	0.53	-0.29	-0.01	-0.21	1	0.13	0.48	0.37	-0.04	-0.09	0.01	-0.09	-0.01
P_m	-0.52	0.52	0.49	0.44	-0.58	0.13	1	0.56	0.35	0.34	-0.58	0.59	0.40	0.69
I	-0.61	0.76	0.03	0.39	-0.71	0.48	0.56	1	0.63	0.22	-0.50	0.40	0.15	0.31
N	-0.22	0.37	0.10	-0.24	-0.24	0.37	0.35	0.63	1	-0.46	0.15	-0.19	-0.39	0.15
t_1	-0.57	0.42	0.05	0.91	-0.56	-0.04	0.34	0.23	-0.46	1	-0.92	0.86	0.87	0.47
t_6	0.74	-0.63	-0.19	-0.94	0.76	-0.09	-0.58	-0.50	0.15	-0.92	1	-0.91	-0.87	-0.63
S_n	-0.61	0.48	0.32	0.83	-0.60	0.01	0.59	0.40	-0.19	0.86	-0.92	1	0.92	0.77
W_e	-0.59	0.36	0.25	0.84	-0.51	-0.09	0.40	0.15	-0.39	0.87	-0.87	0.92	1	0.71
R_a	-0.48	0.32	0.54	0.46	-0.43	-0.01	0.69	0.31	0.14	0.47	-0.63	0.77	0.71	1

Table 11 Summary of SA-Test case 6

	Variance-based (Sobol)	Proposed method $R^2=0.738$	Benchmark	S^{tc}
A	0.2697	0.0484	0.1940	0.0313
S	0.3010	0.1046	0.3008	0.1023
L_c	0.0164	0.0814	0.2947	0.0707
P_i	0.0677	0.1936	0.6399	0.2054
H_r	0.0139	0.2453	0.6923	0.2360
D	0.0050	0.1432	0.4385	0.1640
C	0.1607	0.1835	0.6392	0.1902

Table 12 Summary of SA-Test case 6

Coupling variance-based SA with PCR					
	(2-comp) $R^2 = 0.7374$ $\sum_7 \lambda_i = 0.807$	(3-comp) $R^2 = 0.7376$ $\sum_7 \lambda_i = 0.88$	(4-comp) $R^2 = 0.7714$ $\sum_7 \lambda_i = 0.93$	(5-comp) $R^2 = 0.7816$ $\sum_7 \lambda_i = 0.96$	(6-comp) $R^2 = 0.7847$ $\sum_7 \lambda_i = 0.986$
A	0.0803	0.0767	0.0558	0.2032	0.2540
S	0.0435	0.0511	0.2415	0.3959	0.3696
L_c	0.1405	0.1509	0.2345	0.1056	0.0599
P_i	0.1874	0.1746	0.0881	0.0394	0.1297
H_r	0.2191	0.2384	0.1814	0.0872	0.1039
D	0.1675	0.1751	0.0040	0.0004	0.0042
C	0.1722	0.1435	0.2063	0.1790	0.0887

Fig. 9 Sensitivity measure (S_i) versus input variables for various number of realizations—Test case 6

The first-order sensitivity measures based on the adopted approaches can be seen in Figs. 10, 11.

The variance and coefficient of determination for the number of components in PCR method, i.e., 3 to 8 are documented in Table 14.

As Fig. 10 shows, the proposed method is quite consistent with the benchmark solution but is very similar to the approximated equivalence of the proposed method (S^{tc}). It is important to mention that the coefficient of determination between the original model and the regression-based SPCA model is 0.46. As can be seen, if the correlation structure of model inputs is ignored and the

Sobol approach is implemented on the original model, the results clearly depict the Type 1 error. In other words, the N and R_a as unimportant variables are considered to have a significant impact. In addition, six flavors of variance-based PCR cannot capture the importance of correlated inputs as compared to the mentioned methods (see Fig. 11). In addition, as the number of samples in the proposed method (i.e., 2000 samples) in comparison with the benchmark solution (i.e., 50,000 samples) is low, the suggested technique can be considered as a computationally efficient approach (Table 14).

4.6.3 Test case 8—A hydraulic model: flood inundation in a diversion channel

This practical example is a simplified version of a diversion channel subjected to uniform flow under a flood inundation scenario. In order to protect the service road from flood inundation, a dyke is built in between the diversion channel and the service road. This simple application, which is used as an instructive test in Iooss and Lemaître (2015), and Chastaing et al. (2012), tries to simulate the height of water in the diversion channel with respect to the height of dyke intended to protect the service road, agricultural and industrial sites adjacent to the diversion channel bank from inundation. This model that comprises the characteristics of a typical channel reach can be written as (see Fig. 12):

$$S = Z_v + h - H_d - C_b \quad (50)$$

Upon assuming uniform flow in a wide rectangular channel, h can be derived as (De Rocquigny, 2006):

Table 13 Description of output-inputs—Test case 7 (Benson report 1964)

Variable	Meaning	Distribution
A	Drainage area (mi^2)	Weibul (0.724, 140.5, 10)
S	Channel longitudinal slope (ft/ft)	Lognormal (5.031, 0.734)
S_r	Percentage of Area covered with lakes	Exponential (6.731, 1)
E_a	Index of elevation (ft)	Weibul (5.34, 3684.5, 5760.6)
L_b	Length of Basin	Lognormal (2.967, 0.641)
H_r	Rise of Basin (ft)	Normal (2320.3, 600.99)
I_h	10-year, 24-h rainfall intensity (in)	Weibul (1.266, 0.227, 2.186)
P_m	Mean annual long-term precipitation (in)	Gamma (56.81, 0.364)
N	Mean annual number of thunderstorm days	Gamma (2.348, 3.468, 44.941)
R_a	Mean annual runoff (in)	Lognormal (1.7229, 0.8079)
t_1	Mean January degree-days below freezing (F)	Weibul (3.986, 20.812)
t_6	Mean June temperature (F)	Triangular (40.575, 47.1, 60.72)
S_n	Mean total annual snowfall (in)	Gamma (5.852, 34.637)
W_e	Equivalent snow water content (in)	Gamma (3.2556, 3.770)
Q_{10}	Instantaneous peak discharge (ft^3/s) for a 10-year return period	—

Fig. 10 Summary of SA—Test case 7

$$h = \left(\frac{Q}{BK_s \sqrt{\left(\frac{Z_m - Z_v}{L} \right)}} \right)^6 \quad (51)$$

where S is the maximum overflow (in meter) being a function of eight inputs. Symbols used in Eq. (50) and (51) were defined in Table 15. Computation of maximum overflow calls for monitoring these eight variables. As these field variables are highly corrupted by noise and some of them also exhibit both spatial and temporal variability (e.g., Z_m and Z_v), the dependent variable S has to be considered as a random variable of its own. In Table 15, after defining each variable, the probability density function associated with each variable is also documented.

Generally, the flow rate is highly correlated with channel roughness. As the channel roughness increases, the flow discharge decreases. The Strickler coefficient is inversely proportional to channel roughness. As a result, a correlation coefficient of 0.45 is assumed for subsequent computations. Furthermore, Z_v and Z_m can be correlated with correlation coefficient of 0.3. From an educational point of view, other parameters are considered uncorrelated. After implementing the adopted approaches on the governing equations, the first-order sensitivity measures are summarized in Table 16.

As Table 16 illustrates, the results of the proposed method are consistent with the assessment reported in

benchmark and S^{tc} solutions. According to the table content, H_d has an influential impact on overflow depth followed by Q and Z_v . In fact, in practice setting, that would be the case as far as protection of adjacent agricultural as well as industrial sites is concerned. In light of this, the height of dyke along with the maximum annual flow rate should be chosen with great care in practice. If the classical variance-based SA is implemented on PCR, the methodology can trigger the height of dyke and leave the maximum annual flow rate intact. In addition, this coupling compared to the proposed method, cannot effectively demonstrate the other variables' importance. Once again, the above procedural scheme was also implemented for the various number of realizations. As Fig. 13 shows, it seems the number of realizations has minimal impact on the ranking process.

5 Summary, conclusions, limitations, and future research

Nowadays, the GSA is considered a powerful approach to organize the input variables in terms of their degree of importance, as they affect the dependent variable. Over the last two decades, one of the most important challenging tasks in GSA is to consider the impact of correlated input variables on the uncertainty evolved in model output.

Fig. 11 Summary of SA—Test case 7

Recent scientific endeavors and the associated literature have tried to develop approaches to address this issue. In light of this, when it comes to quantitative approaches (e.g., correlation ratio method), the methodology is usually accompanied by a great deal of computational cost particularly in the case of having high dimensionality and can be quite labor intensive. For this reason and due to the complexity of most of the developed approaches, the majority of research activities assume the decision variables to be independent for prioritization purposes. In this paper, an innovative methodology is developed whereby the conventional variance-based approach (under the assumption of orthogonal input variables) is coupled with a regression-based SPCA model to evaluate and assess the impact of correlated input variables.

To evaluate the effectiveness of the proposed scheme, eight test cases are considered. In all test cases, the input variables are equipped with common information content leading to a correlation among input variables. However, the model structure (i.e., linear versus nonlinear and/or low versus high dimensional) could differ from one test case to another. With respect to the governing probability density function, a variety of scenarios are assumed. After implementing the proposed scheme on various test cases, the following results are obtained:

1. The scheme managed to capture the ranking structure in input variables with even less computational cost

and being quite consistent with the results obtained from the benchmark solution.

2. The proposed scheme can mimic the variation among the first-order sensitivity measures, similar to the benchmark solution. This success of the proposed method can be attributed to the fact that the result of coupling the classical variance-based with regression-based SPCA can easily be proven approximately equivalent to the total contribution index in the covariance decomposition equation, i.e., S^{tc} which was done in this study. S^{tc} can be related to variance-covariance structure for comparison purposes. In addition, this equivalency can be justified by the fact that the regression-based SPCA considers the impact of the output variable on the new features in the eigenspace.
3. In reference to test case 5, the proposed method can be recommended when the coefficient of determination is larger than 0.33. This implies the proposed method is appropriate for models whose degree of nonlinearity is more distinct than moderately nonlinear models, i.e., $R^2 \geq 0.7$.
4. To conclude, the advantage of the proposed method is that the developed methodology for models with correlated inputs is remarkably simple to implement, computationally efficient, particularly for high dimensional problems as compared to the benchmark solution. Indeed, the case of correlated inputs presents the significant complexities for the developed variance-based SA such as the benchmark solution, including the difficulty of re-coding the benchmark solution as well as a significant increase of computational cost (number of samples). However, it seems, the efficiency of the proposed method is more tangible in the case of a model with high dimensional problems that are correlated.

Table 14 The variance and the coefficient of determination for number of components in PCR method (3–8)

	3-comp	4-comp	5-comp	6-comp	7-comp	8-comp
$\sum_{i=1}^n \lambda_i$	0.7650	0.8414	0.8809	0.9150	0.9419	0.9655
R^2	0.462	0.6139	0.6398	0.6401	0.6435	0.6705

Fig. 12 Flood inundation in a diversion channel **Table 15** Description of input variables–Test case 8 (Limbourg and De Rocquigny 2010; Chastaing et al. 2012; Iooss and Lemaître 2015)

Random variables	Meaning	Unit	Marginal distribution
h	Maximum annual water height in a diversion channel	m	–
Q	Maximum annual flow rate	m ³ /s	Gumbel G(1013,558) truncated to [500, 3000]
K_s	Strickler coefficient	m ^{1/3} /s	Normal N(30, 8) truncated to [15, ∞]
Z_v	Downstream bed channel	m	Normal N(50.19, 0.38)
Z_m	Upstream bed channel	m	Normal N(55.03, 0.45)
H_d	Dyke height	m	Uniform U(7, 9)
C_b	Pedestrian level	m	Triangular T(55, 55.5, 56)
L	Length of channel	m	Triangular T(4990, 5000, 5010)
B	Channel width	m	Triangular T(295, 300, 305)

Table 16 Summary of SA–Test case 8

	Variance-based (Sobol)	Proposed method $R^2=0.9048$	Coupling variance-based SA with PCR			Benchmark	S^{rc}
			(5-comp) $R^2=0.3482$ $\sum_8 \lambda_i = 0.7364$	(6-comp) $R^2=0.5634$ $\sum_8 \lambda_i = 0.8509$	(7-comp) $R^2=0.693$ $\sum_8 \lambda_i = 0.9339$		
Q	0.3648	0.2579	0.0673	0.0374	0.0360	0.2335	0.2314
K_s	0.1298	0.0118	0.0823	0.0450	0.0217	0.0110	0.0120
Z_v	0.1653	0.2089	0.1001	0.0693	0.3399	0.1630	0.2146
Z_m	0.0054	0.00003	0.0984	0.0610	0.0225	0.0017	0.0005
H_d	0.2903	0.4560	0.2693	0.6851	0.5175	0.3764	0.4897
C_b	0.0362	0.0596	0.0448	0.1002	0.0599	0.0485	0.0652
L	0.0000	0.0007	0.3087	0.0006	0.0002	0.0013	0.0011
B	0.0001	0.0050	0.0293	0.0011	0.0008	0.0007	0.0092

By considering the above advantages, possible implications of the current research can be stated as follows:

- The proposed method can be implemented to reduce the uncertainty of model output based on recognition of the most important correlated variables and their contributions to the uncertainty of the response surface. By eliminating or reducing the uncertainty of the most important correlated factors and/or improving factor characterizations that have significant contributions to the uncertainty of output, the uncertainty of output will be reduced.
- Considering the first item and to be more specific, nowadays, data assimilation is considered the rule rather than the exception in rainfall-runoff modeling. A variety of methods are available for data assimilation, namely input updating, parameter updating, state updating, output updating, model structure updating, etc. All these assimilation schemes are developed to

Fig. 13 Sensitivity measure (S_i) versus input variables for various number of realizations—Test case 8

achieve the best match between observations and simulations. In light of this, one is quite curious to see how and in what aspect he or she has to invest to get the best possible result. Should he or she go for say, parameter or state, or output updating to obtain the best match with minimum investment in model implementation? One basic implication of the current study is to implement the proposed methodology and prioritize the nature of data assimilation for a better result, particularly, when inputs, parameters, and states are correlated.

- Research is underway to implement the proposed methodology to design a typical rain gauge network to maximize its information content and/or reduce the variance of the estimated long-term mean over a block as large as the study area.

Overall, the proposed scheme can be considered an enlightening approach for sensitivity analysis modelers.

In contrast to the proposed method, after coupling the variance-based approach (under the assumption of orthogonal input variables) with PCR, it was found that implementation of this scheme on most of the test functions gave rise to solutions which are incompatible with the benchmark solution, the correlation ratio method. On the other hand, based on the second test function and the first scenario of the fourth test function, it seems that if results of Sobol, which are conducted directly on the original model, and benchmark solutions are approximately the same, then this coupling may appear somewhat successful. In general, contrary to the coupling of the classical variance-based SA with the regression-based SPCA, the coupling of the classical variance-based SA with PCR [Eq. (42)] cannot accommodate the impact of the

covariance structure of model inputs on the uncertainty of the response surface.

Apart from the mentioned advantages, it is necessary to touch on some disadvantages of the proposed method. As a drawback, the proposed approach cannot replicate the acceptable results in light of a more complex model. This implies that this approach cannot reliably differentiate and rank the input variables when the coefficient of determination between the original model and regression-based SPCA model is very small. In such scenarios, it is recommended to use the benchmark solution or developed approaches cited in the literature (e.g., Ge and Menendez 2017). In addition, the proposed method is not appropriate for the case when a higher-order correlation of model inputs can be traced to be among the input variables. In such instances, it is necessary to modify the proposed approach in the presence of complex dependency or higher-order correlation among model inputs.

In future, it is recommended to modify and implement the proposed methodology on more complex transfer functions and monitor the model efficiency in comparison with the developed approaches in the literature such as the correlation ratio method. In other words, as in our proposed method, we have coupled the classical variance-based SA with regression-based SPCA model and this model cannot regard the interaction terms in its equation, consequently, the higher-order interaction impacts cannot be quantified. For this reason, future research could focus on presenting a modified model based on using SPCA which considers the interaction and non-linear terms, and couple this SPCA nonlinear model with the classical variance-based SA. Furthermore, independent component analysis (ICA) is considered as an important tool for the consideration of higher-order correlation. Therefore, we recommend obtaining an algorithm for supervised independent component analysis (SICA) similar to SPCA and coupling it with classical variance-based SA to derive a new method for GSA in the case of dependent inputs for examination of the higher-order correlation in the uncertainty of model output.

Appendix A

The linear regression model, which can be used in place of the original model, i.e., $Y = f(X_1, X_2, \dots, X_p)$ can be written as follows:

$$Y = \sum_{i=1}^p a_i X_i + a_0 \quad (52)$$

In the SPCA approach, it is recommended to standardize the original variables (Y, X_i). Subsequently, the functional relationship between the standardized variables ($\tilde{Y}, \tilde{X}_i$) is:

$$\tilde{Y} = \sum_{i=1}^p \theta_i \tilde{X}_i + \theta_0 \quad (53)$$

$$\tilde{Y} = [\tilde{X}_1, \dots, \tilde{X}_p] \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix} + \theta_0$$

In brief, we consider:

$$\tilde{X} = [\tilde{X}_1, \dots, \tilde{X}_p] \quad \text{and} \quad \theta = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix} \quad (54)$$

Thus,

$$\tilde{Y} = \tilde{X}\theta + \theta_0 \quad (55)$$

$\tilde{X}$ is a standardized random vector that comprises of p standardized random variables, i.e., $\tilde{X}_i$ and θ is a vector with p scalar values, i.e., θ_i . Indeed, the dimension of $\tilde{X}$ and θ are $1 \times p$ and $p \times 1$, respectively. This means:

$$\tilde{Y} = \tilde{X}_{1 \times p} \theta_{p \times 1} + \theta_0 \quad (56)$$

Since the identity Matrix can be multiplied by a matrix without any changes in results, we have:

$$\tilde{Y} = \tilde{X}_{1 \times p} I_{p \times p} \theta_{p \times 1} + \theta_0 \quad (57)$$

Since $U^T_{p \times p} U_{p \times p} = U_{p \times p} U^T_{p \times p} = I$ (Jolliffe 1986),

$$\tilde{Y} = \tilde{X}_{1 \times p} U_{p \times p} U^T_{p \times p} \theta_{p \times 1} + \theta_0 \quad (58)$$

In the equation above, $\tilde{X}_{1 \times p} U_{p \times p}$ are the new features. In other words, $Z_{1 \times p} = \tilde{X}_{1 \times p} U_{p \times p}$ or (Barshan et al. 2011)

$$[Z_1, Z_2, \dots, Z_p] = [\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_p] \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1p} \\ u_{21} & u_{22} & \dots & u_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ u_{p1} & u_{p2} & \dots & u_{pp} \end{bmatrix} \quad (59)$$

Z is a random vector consisting of p random variables of new feature, i.e., Z_i , and $U^T_{p \times p} \theta_{p \times 1}$ is a new vector, i.e., β that includes p values, i.e.,:

$$\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad (60)$$

As a result:

$$\tilde{Y} = Z\beta + \theta_0 \quad (61)$$

Thus:

$$\tilde{Y} = \beta_1 Z_1 + \beta_2 Z_2 + \dots + \beta_p Z_p + \beta_0 \quad (62)$$

where $\beta_0 = \theta_0$

Since the new features are uncorrelated, we can use variable reduction and define the model above based on the reduced model:

$$\tilde{Y} = \beta_1 Z_1 + \beta_2 Z_2 + \dots + \beta_m Z_m + \beta_0 \quad (63)$$

Appendix B

After replacing L with $\tilde{Y}\tilde{Y}^T$ in the matrix $\mathbb{Q} = \tilde{X}^T H L H \tilde{X}$, we have:

$$\mathbb{Q} = \tilde{X}^T H \tilde{Y} \tilde{Y}^T H \tilde{X} \quad (64)$$

Since $\tilde{X}$ and $\tilde{Y}$ are standardized, and matrix H is symmetric, i.e., $H = H^T$, we have (Barshan et al. 2011):

$$\begin{aligned} \tilde{X}^T H &= (H \tilde{X})^T = \left(\left(I - \frac{1}{n} e_{n*1} e_{1*n}^T \right) \tilde{X} \right)^T = (\tilde{X} - \mu_{\tilde{X}})^T \\ &= \tilde{X}^T \end{aligned} \quad (65)$$

And

$$\begin{aligned} \tilde{Y}^T H &= (H \tilde{Y})^T = \left(\left(I - \frac{1}{n} e_{n*1} e_{1*n}^T \right) \tilde{Y} \right)^T = (\tilde{Y} - \mu_{\tilde{Y}})^T \\ &= \tilde{Y}^T \end{aligned} \quad (66)$$

As a result, the matrix $\mathbb{Q}$ is simplified as follows:

$$\mathbb{Q} = \tilde{X}^T \tilde{Y} \tilde{Y}^T \tilde{X} \quad (67)$$

The terms $\tilde{X}^T \tilde{Y}$ and $\tilde{Y}^T \tilde{X}$ are two-column and row vectors, respectively for which their arrays are $\text{COV}(\tilde{Y}, \tilde{X}_i)$:

$$\begin{aligned} \tilde{X}^T_{p*n} \tilde{Y}_{n*1} &= [0pt0pt\text{COV}(\tilde{Y}, \tilde{X}_1) \text{COV}(\tilde{Y}, \tilde{X}_2) \dots \text{COV}(\tilde{Y}, \tilde{X}_p)], \tilde{Y}^T_{1*n} \tilde{X}_{n*p} \\ &= [\text{COV}(\tilde{Y}, \tilde{X}_1), \text{COV}(\tilde{Y}, \tilde{X}_2), \dots, \text{COV}(\tilde{Y}, \tilde{X}_p)] \end{aligned} \quad (68)$$

In reference to Eq. (68), Eq. (67) implies that the matrix $\mathbb{Q}$ can be considered as multiplication of a vector i.e., $\tilde{X}^T_{p*n} \tilde{Y}_{n*1}$ (with p entries) by its transpose.

Generally, multiplication of a vector (with dimension of $p \times 1$) by its transpose would always result in a symmetric matrix with rank¹⁰ of one. In the following, the eigenvalues and eigenvectors of this matrix are further discussed (Personal Communication, Strang 2021). For this reason, consider the vector W as follows:

¹⁰ The definition of matrix rank can be found in almost all linear algebra textbook.

$$W = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_p \end{bmatrix} \quad (69)$$

And a Matrix:

$$\tilde{A} = WW^T \quad (70)$$

For this matrix, $\|W\|^2$ can be claimed to be an eigenvalue of matrix $\tilde{A}$ corresponding to eigenvector W . This can be proved as follows:

$$\tilde{A}U = \lambda U \quad (71)$$

After replacing U with W and $\tilde{A}$ with WW^T , we have:

$$\tilde{A}W = (WW^T)W = W(W^TW) = W\|W\|^2 = \|W\|^2W \quad (72)$$

In summary, based on the above equation:

$$\tilde{A}W = \|W\|^2W \quad (73)$$

As for the remaining eigenvalues and corresponding eigenvectors, such as:

$$M = \{M_2, M_3, \dots, M_p\} \quad (74)$$

As these eigenvectors are perpendicular to W , one can show that:

$$W.M_i = 0 \rightarrow W^TM_i = 0, i = 2, 3, \dots, p \quad (75)$$

Thus,

$$\tilde{A}M_i = WW^TM_i = W(W^TM_i) = W \times 0 = 0 = 0 \times M_i \quad (76)$$

This implies that the remaining eigenvalues are zero.

Based on the above theory, since the matrix $\mathbb{Q}$ is the multiplication of a vector i.e., $\tilde{X}^T\tilde{Y}$ (with p entries) by its transpose, its dominant eigenvalue can be written as:

$$\begin{aligned} \lambda_1 &= \|\tilde{X}^T\tilde{Y}\|^2 \\ &= [COV(\tilde{Y}, \tilde{X}_1)]^2 + [COV(\tilde{Y}, \tilde{X}_2)]^2 + \dots \\ &\quad + [COV(\tilde{Y}, \tilde{X}_p)]^2 \end{aligned} \quad (77)$$

Thus:

$$\lambda_1 = \|\tilde{X}^T\tilde{Y}\|^2 = \sum_{i=1}^p [COV(\tilde{Y}, \tilde{X}_i)]^2 \quad (78)$$

And other eigenvalues are zero. In addition, the eigenvector of this matrix corresponds to the dominant eigenvalue, λ_1 is $\tilde{X}^T\tilde{Y}$ with its components as follows:

$$u'_{i1} = COV(\tilde{Y}, \tilde{X}_i) \quad (79)$$

As in subsequent applications, we will be using the standard eigenvector (unit eigenvector), i.e. $U_1^TU_1 = 1$, each component of the dominant eigenvector will be divided by the magnitude of the eigenvector to make it standard as:

$$u_{i1} = \frac{COV(\tilde{Y}, \tilde{X}_i)}{\|COV(\tilde{Y}, \tilde{X}_i)\|}, i = 1, 2, \dots, p \quad (80)$$

Thus:

$$u_{i1} = \frac{COV(\tilde{Y}, \tilde{X}_i)}{\left(\sum_{i=1}^p [COV(\tilde{Y}, \tilde{X}_i)]^2\right)^{.5}} \quad (81)$$

After using Eq.(78), the eigenvector u_{i1} can be simplified to:

$$u_{i1} = \frac{COV(\tilde{Y}, \tilde{X}_i)}{(\lambda_1)^{.5}}, i = 1, 2, \dots, p \quad (82)$$

The $\tilde{Y}$ and $\tilde{X}_i$ are standardized values of the original variables, thus the u_{i1} can be expressed in terms of the original variables (Y, X_i) as:

$$u_{i1} = \frac{1}{(\lambda_1)^{.5} \times \sigma_Y \times \sigma_{X_i}} COV(Y, X_i), i = 1, 2, \dots, p \quad (83)$$

In addition, in the following, we prove that if L is replaced by $\tilde{Y}\tilde{Y}^T$ in matrix $\mathbb{Q} = \tilde{X}^T H L H \tilde{X}$, the correlation between standardized output and all new features with the exception of the dominant new feature (i.e., $Z_{ij} \neq 1$) is zero. For this purpose, we substitute the standardized variables into Z_j , thus, $COV(\tilde{Y}, Z_j) (j = 2, 3, \dots, m)$ can be expanded as:

$$\begin{aligned} COV(\tilde{Y}, Z_j) &= COV(\tilde{Y}, u_{1j}\tilde{X}_1 + u_{2j}\tilde{X}_2 + \dots + u_{pj}\tilde{X}_p) \\ &= COV(\tilde{Y}, u_{1j}\tilde{X}_1) + COV(\tilde{Y}, u_{2j}\tilde{X}_2) + \dots + COV(\tilde{Y}, u_{pj}\tilde{X}_p) \\ &= u_{1j}COV(\tilde{Y}, \tilde{X}_1) + u_{2j}COV(\tilde{Y}, \tilde{X}_2) + \dots + u_{pj}COV(\tilde{Y}, \tilde{X}_p) \\ &= U_j.COV(\tilde{Y}, \tilde{X}_i) = \langle U_j, COV(\tilde{Y}, \tilde{X}_i) \rangle, i = 1, 2, \dots, p \\ &\text{and } j = 2, 3, \dots, m. \end{aligned} \quad (84)$$

Where $\langle \cdot, \cdot \rangle$ denotes the inner product of two vectors.

Based on Eq. (82), Eq. (84) can be written as:

$$\begin{aligned} COV(\tilde{Y}, Z_j) &= \langle U_j, COV(\tilde{Y}, \tilde{X}_i) \rangle \\ &= \langle U_j, \lambda_1^{.5} \frac{COV(\tilde{Y}, \tilde{X}_i)}{\lambda_1^{.5}} \rangle, j \neq 1 \end{aligned} \quad (85)$$

Since the eigenvector U_1 is perpendicular to other eigenvectors $(U_j, j \neq 1), \langle U_j, U_1 \rangle = 0$. Thus, $COV(\tilde{Y}, Z_j) = 0$ for $j \neq 1$. This implies only the first new feature has maximum linear dependency on output.

Appendix C

For a linear model such as:

$$Y = a_0 + \sum_{i=1}^p a_i X_i \quad (86)$$

The variance of model output with correlated inputs can be computed as:

$$\begin{aligned} V(Y) &= \sum_{i=1}^p \sum_{j=1}^p a_i a_j \text{COV}(X_i, X_j) \\ &= \sum_{i=1}^p \sum_{j=1}^p \text{COV}(a_i X_i, a_j X_j) \\ &= V(a_1 X_1) + \text{COV}(a_1 X_1, a_2 X_2) + \cdots \\ &\quad + \text{COV}(a_1 X_1, a_p X_p) \\ &\quad + \text{COV}(a_2 X_2, a_1 X_1) + V(a_2 X_2) + \cdots \\ &\quad + \text{COV}(a_2 X_2, a_p X_p) + \cdots \\ &\quad + \text{COV}(a_p X_p, a_1 X_1) + \text{COV}(a_p X_p, a_2 X_2) \\ &\quad + \cdots + V(a_p X_p) \end{aligned} \quad (87)$$

Based on the following equation:

$$\text{COV}\left(X, \sum_{i=1}^M X_i\right) = \sum_{i=1}^M \text{COV}(X, X_i) \quad (88)$$

where X and X_i are random variables. Thus, the $V(Y)$ can be rewritten as follow:

$$\begin{aligned} V(Y) &= V(a_1 X_1) + \text{COV}(a_1 X_1, a_2 X_2 + a_3 X_3 + \cdots + a_p X_p) \\ &\quad + V(a_2 X_2) + \text{COV}(a_2 X_2, a_1 X_1 + a_3 X_3 + \cdots + a_p X_p) \\ &\quad + \cdots + V(a_p X_p) + \text{COV}(a_p X_p, a_1 X_1 + a_2 X_2 + \cdots + a_{p-1} X_{p-1}) \end{aligned} \quad (89)$$

Finally, the $V(Y)$ is decomposed by the following equation:

$$V(Y) = \sum_{i=1}^p \left[V(a_i X_i) + \text{COV}\left(a_i X_i, \sum_{j=1, j \neq i}^p a_j X_j\right) \right] \quad (90)$$

One possible interpretation of the above equation is that the impact of each variable on the uncertainty of model output is related to variance of each factor and its correlation structure with other variables. This equation can be rewritten as follows:

$$\begin{aligned} V(Y) &= \sum_{i=1}^p \left[\text{COV}(a_i X_i, a_i X_i) + \text{COV}\left(a_i X_i, \sum_{j=1, j \neq i}^p a_j X_j\right) \right] \\ &= \sum_{i=1}^p \left[\text{COV}\left(a_i X_i, \sum_{j=1, j \neq i}^p a_j X_j + a_i X_i\right) \right] \\ &= \sum_{i=1}^p \left[\text{COV}\left(a_i X_i, \sum_{j=1}^p a_j X_j\right) \right] \\ &= \sum_{i=1}^p [\text{COV}(a_i X_i, Y - a_0)] V(Y) = \sum_{i=1}^p [\text{COV}(Y, a_i X_i)] \end{aligned} \quad (91)$$

This equation indicates that the variance of linear model output is partitioned by all $\text{COV}(Y, a_i X_i)$. Finally, we can express the variance of model output via the following equation in reference to (90) and (91):

$$\begin{aligned} V(Y) &= \sum_{i=1}^p [\text{COV}(Y, a_i X_i)] \\ &= \sum_{i=1}^p \left[V(a_i X_i) + \text{COV}\left(a_i X_i, \sum_{j=1, j \neq i}^p a_j X_j\right) \right] \end{aligned} \quad (92)$$

In reference to the above equation, we have:

$$\sum_{i=1}^p [\text{COV}(Y, a_i X_i)] = \sum_{i=1}^p \left[V(a_i X_i) + \text{COV}\left(a_i X_i, \sum_{j=1, j \neq i}^p a_j X_j\right) \right] \quad (93)$$

The main conclusion which can be drawn from the above equation is:

$$\begin{aligned} \frac{\text{COV}(Y, a_i X_i)}{V(Y)} &= \frac{V(a_i X_i)}{V(Y)} + \frac{\text{COV}\left(a_i X_i, \sum_{j=1, j \neq i}^p a_j X_j\right)}{V(Y)}, i \\ &= 1, 2, \dots, p \end{aligned} \quad (94)$$

It is worth mentioning, $\text{COV}(Y, a_i X_i)$ is the total contribution of the random variable X_i composed of variance structure i.e., $V(a_i X_i)$ and covariance structure i.e., $\text{COV}\left(a_i X_i, \sum_{j=1, j \neq i}^p a_j X_j\right)$.

Appendix D

In this appendix, a rationale is offered to prove Eq. (40) in the text. After regressing the $\tilde{Y}$ on the new uncorrelated features based on PCR, we have (Jolliffe 1986):

$$\tilde{Y} = \beta_1 Z'_1 + \beta_2 Z'_2 + \cdots + \beta_m Z'_m + \beta_0 \quad (95)$$

where Z'_j is a new feature on the basis of PCR. As mentioned in the main text, the coefficient of linear regression when the inputs are uncorrelated with zero means can be estimated by Eq. (17). Thus:

$$\hat{\beta}_i = \frac{\text{COV}(\tilde{Y}, Z'_i)}{V(Z'_i)}, i = 1, 2, \dots, m \quad (96)$$

Since in PCA, $V(Z'_j) = \lambda_j$, the PCR is:

$$\tilde{Y} = \frac{\text{COV}(\tilde{Y}, Z'_1)}{\lambda_1} Z'_1 + \frac{\text{COV}(\tilde{Y}, Z'_2)}{\lambda_2} Z'_2 + \dots + \frac{\text{COV}(\tilde{Y}, Z'_m)}{\lambda_m} Z'_m + \beta_0 \quad (97)$$

where:

$$[Z'_1, Z'_2, \dots, Z'_m] = [\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_p] \begin{bmatrix} u'_{11} & u'_{12} & \dots & u'_{1m} \\ u'_{21} & u'_{22} & \dots & u'_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ u'_{p1} & u'_{p2} & \dots & u'_{pm} \end{bmatrix} \quad (98)$$

Then, it can be easily deduced that the expression of the PCR in terms of the standardized variables i.e. $\tilde{X}_i$ is:

$$\begin{aligned} \tilde{Y} = & \frac{\text{COV}(\tilde{Y}, Z'_1)}{\lambda_1} (u'_{11}\tilde{X}_1 + u'_{21}\tilde{X}_2 + \dots + u'_{p1}\tilde{X}_p) \\ & + \frac{\text{COV}(\tilde{Y}, Z'_2)}{\lambda_2} (u'_{12}\tilde{X}_1 + u'_{22}\tilde{X}_2 + \dots + u'_{p2}\tilde{X}_p) + \dots \\ & + \frac{\text{COV}(\tilde{Y}, Z'_m)}{\lambda_m} (u'_{1m}\tilde{X}_1 + u'_{2m}\tilde{X}_2 + \dots + u'_{pm}\tilde{X}_p) + \beta_0 \end{aligned} \quad (99)$$

Therefore:

$$\begin{aligned} \tilde{Y} = & \left(\frac{\text{COV}(\tilde{Y}, Z'_1)}{\lambda_1} u'_{11} + \frac{\text{COV}(\tilde{Y}, Z'_2)}{\lambda_2} u'_{12} + \dots + \frac{\text{COV}(\tilde{Y}, Z'_m)}{\lambda_m} u'_{1m} \right) \tilde{X}_1 \\ & + \left(\frac{\text{COV}(\tilde{Y}, Z'_1)}{\lambda_1} u'_{21} + \frac{\text{COV}(\tilde{Y}, Z'_2)}{\lambda_2} u'_{22} + \dots + \frac{\text{COV}(\tilde{Y}, Z'_m)}{\lambda_m} u'_{2m} \right) \tilde{X}_2 + \dots \\ & + \left(\frac{\text{COV}(\tilde{Y}, Z'_1)}{\lambda_1} u'_{p1} + \frac{\text{COV}(\tilde{Y}, Z'_2)}{\lambda_2} u'_{p2} + \dots + \frac{\text{COV}(\tilde{Y}, Z'_m)}{\lambda_m} u'_{pm} \right) \tilde{X}_p + \beta_0 \end{aligned} \quad (100)$$

The PCR can be expressed as regards to the original variables (Y, X_i), after some manipulations:

$$\begin{aligned} Y = & \frac{c_1 \sigma_Y}{\sigma_{X_1}} X_1 + \frac{c_2 \sigma_Y}{\sigma_{X_2}} X_2 + \dots \\ & + \frac{c_p \sigma_Y}{\sigma_{X_p}} X_p + E(Y) + \beta_0 \sigma_Y - \sum_{i=1}^p c_i \frac{E(X_i)}{\sigma_{X_i}} \\ c_i = & \frac{\text{COV}(\tilde{Y}, Z'_1)}{\lambda_1} u'_{i1} + \frac{\text{COV}(\tilde{Y}, Z'_2)}{\lambda_2} u'_{i2} \\ & + \dots + \frac{\text{COV}(\tilde{Y}, Z'_m)}{\lambda_m} u'_{im}, i = 1, 2, \dots, p \end{aligned} \quad (101)$$

Acknowledgements We would like to thank Prof. G. Strang from MIT for his enlightening remark on symmetric matrix and uniqueness of dominant eigenvalue of matrix $\mathbb{Q}$ documented in appendix B. Furthermore, extensive implementation of SIMLAB 2.2 in this research is greatly acknowledged.

Funding This manuscript has been prepared without receiving support from any organization.

Availability of data and material In this manuscript, eight test functions were used, six of which are based on the simulated data, and two that are based on the real data. These data are extracted from McCuen and Snyder's book (1986) and Benson's report (1964) documented in the reference list.

Code availability Source code of the proposed method in R programming is available at the following address: <https://bit.ly/3niTUwU>

Declarations

Conflict of interest No potential conflict of interest was reported by the authors.

References

- Barshan E, Ghodsi A, Azimifar Z, Zolghadri Jahromi M (2011) Supervised principal component analysis: visualization, classification and regression on subspaces and submanifolds. *Pattern Recognit* 44(7):1357–1371. <https://doi.org/10.1016/j.patcog.2010.12.015>
- Bedford T (1998) Sensitivity indices for (tree)-dependent variables. In: Proceedings of the second international symposium on sensitivity analysis of model output, Venice (Italy), pp 17–20
- Benson MA (1964) Factor affecting the occurrences of floods in the southwest U.S.. *Geol Surv Water supply pap* 1580-D
- Borgonovo E (2007) A new uncertainty importance measure. *Reliab Eng Syst Saf* 92(6):771–784. <https://doi.org/10.1016/j.res.2006.04.015>
- Chastaing G, Gamboa F, Prieur C (2012) Generalized Hoeffding-Sobol decomposition for dependent variables-application to sensitivity analysis. *Electron J Statist* 6:2420–2448. <https://doi.org/10.1214/12-EJS749>
- Ciriello V, Di Federico V, Riva M et al (2013) Polynomial chaos expansion for global sensitivity analysis applied to a model of radionuclide migration in a randomly heterogeneous aquifer. *Stoch Environ Res Risk Assess* 27(4):945–954. <https://doi.org/10.1007/s00477-012-0616-7>
- Crosetto M, Tarantola S (2001) Uncertainty and sensitivity analysis: tools for GIS-based model implementation. *Int J Geogr Inf Sci* 15(5):415–437. <https://doi.org/10.1080/13658810110053125>
- Cukier R, Fortuin C, Shuler K, Petschek A, Schaibly J (1973) Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I Theory *J Chem Phys* 59(8):3873–3878. <https://doi.org/10.1063/1.1680571>
- Da Veiga S, Wahl F, Gamboa F (2009) Local polynomial estimation for sensitivity analysis on models with correlated inputs. *Technometrics* 51(4):452–463. <https://doi.org/10.1198/TECH.2009.08124>
- De Rocquigny E (2006) La maîtrise des incertitudes dans un contexte industriel-Iere partie: une approche méthodologique globale basée sur des exemples. *Journal de la Société Française de Statistique* 147(3):33–71
- Do NC, Razavi S (2020) Correlation Effects? A Major but Often Neglected Component in Sensitivity and Uncertainty Analysis. *Water Resour Res* 56(3): e2019WR025436. Doi: <https://doi.org/10.1029/2019WR025436>
- Ge Q, Menendez M (2017) Extending Morris method for qualitative global sensitivity analysis of models with dependent inputs.

- Reliab Eng Syst Saf 162:28–39. <https://doi.org/10.1016/j.res.2017.01.010>
- Gretton A, Bousquet O, Smola A J, Scholkopf B (2005) Measuring statistical dependence with hilbert-schmidt norms. In: *Proceedings Algorithmic Learning Theory (ALT)*, springer- Berlin, pp 63–77
- Hein M, Bousquet O (2004) *Kernels, Associated Structures and Generalizations (127)*. Tübingen, Germany: Max Planck Institute for Biological Cybernetics. <http://hdl.handle.net/11858/00-001M-0000-0013-D8B3-0>
- Helton JC, Davis FJ, Johnson JD (2005) A comparison of uncertainty and sensitivity analysis results obtained with random and Latin hypercube sampling. *Reliab Eng Syst Saf* 89(3):305–330. <https://doi.org/10.1016/j.res.2004.09.006>
- Iman RL, Conover WJ (1982) A distribution-free approach to inducing rank correlation among input variables. *Commun Stat Simul Comput* 11(3):311–334. <https://doi.org/10.1080/03610918208812265>
- Iman RL, Johnson ME, Schroeder TA (2002) Assessing hurricane effects. Part 1. Sensitivity analysis. *Reliab Eng Syst Saf* 78(2):131–145. Doi: [https://doi.org/10.1016/S0951-8320\(02\)00133-3](https://doi.org/10.1016/S0951-8320(02)00133-3)
- Iooss B, Lemaître P (2015) A review on global sensitivity analysis methods. In: Meloni C, Dellino G (Eds) *Uncertainty management in simulation optimization of complex systems: algorithms and applications*. Springer, Boston MA, pp 101–122. https://doi.org/10.1007/978-1-4899-7547-8_5
- Jolliffe IT (1986) *Principal component analysis*. Springer-Verlag, New York
- Kucherenko S, Klymenko OV, Shah N (2017) Sobol' indices for problems defined in non-rectangular domains. *Reliab Eng Syst Saf* 167:218–231. <https://doi.org/10.1016/j.res.2017.06.001>
- Kucherenko S, Tarantola S, Annoni P (2012) Estimation of global sensitivity indices for models with dependent variables. *Comput Phys Commun* 183(4):937–946. <https://doi.org/10.1016/j.cpc.2011.12.020>
- Lamboni M, Kucherenko S (2021) Multivariate sensitivity analysis and derivative-based global sensitivity measures with dependent variables. *Reliab Eng Syst Saf* 212:107519. <https://doi.org/10.1016/j.res.2021.107519>
- Li G, Rabitz H, Yelvington PE, Oluwale OO, Bacon F, Kolb CE, Schoendorf J (2010) Global sensitivity analysis for systems with independent and/or correlated inputs. *J Phys Chem A* 114(19):6022–6032. <https://doi.org/10.1021/jp9096919>
- Limbourg P, De Rocquigny E (2010) Uncertainty analysis using evidence theory—confronting level-1 and level-2 approaches with data availability and computational constraints. *Reliab Eng Syst Saf* 95(5):550–564. <https://doi.org/10.1016/j.res.2010.01.005>
- Mara TA, Tarantola S (2012) Variance-based sensitivity indices for models with dependent inputs. *Reliab Eng Syst Saf* 107:115–121. <https://doi.org/10.1016/j.res.2011.08.008>
- Mara TA, Tarantola S, Annoni P (2015) Non-parametric methods for global sensitivity analysis of model output with dependent inputs. *Environ Model Softw* 72:173–183. <https://doi.org/10.1016/j.envsoft.2015.07.010>
- Martin-Barreiro C, Ramirez-Figueroa JA, Cabezas X, Leiva V, Galindo-Villardón MP (2021) Disjoint and functional principal component analysis for infected cases and deaths due to COVID-19 in South American countries with sensor-related data. *Sensors* 21(12):4094. <https://doi.org/10.3390/s21124094>
- McCuen RH, Snyder WM (1986) *Hydrologic modeling: statistical methods and applications*. Prentice-Hall, Englewood, NJ
- McKay MD (1997) Nonparametric variance-based methods of assessing uncertainty importance. *Reliab Eng Syst Saf* 57(3):267–279. [https://doi.org/10.1016/S0951-8320\(97\)00039-2](https://doi.org/10.1016/S0951-8320(97)00039-2)
- Nataf A (1962) Détermination des distributions dont les marges sont données. *C R Acad Sci* 225:42–43
- Ramirez-Figueroa JA, Martin-Barreiro C, Nieto-Librero AB, Leiva V, Galindo-Villardón MP (2021) A new principal component analysis by particle swarm optimization with an environmental application for data science. *Stoch Environ Res Risk Assess* 35(10):1969–1984. <https://doi.org/10.1007/s00477-020-01961-3>
- Razavi S, Gupta HV (2016a) A new framework for comprehensive, robust, and efficient global sensitivity analysis: 1. Theory *Water Resour Res* 52(1):423–439. <https://doi.org/10.1002/2015WR017558>
- Razavi S, Gupta HV (2016b) A new framework for comprehensive, robust, and efficient global sensitivity analysis: 2. Application *Water Resour Res* 52(1):440–455. <https://doi.org/10.1002/2015WR017559>
- Saltelli A, Ratto M, Tarantola S (2001) Model-free importance indicators for dependent input. In: *Proceedings of SAMO 2001, third international symposium on sensitivity analysis of model output*, Madrid
- Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, Saisana M, Tarantola S (2008) *Global sensitivity analysis: the primer*. Wiley
- Saltelli A, Tarantola S, Campolongo F, Ratto M (2004) *Sensitivity analysis in practice: a guide to assessing scientific models*. Wiley
- SAMO team (2003) *Simlab 2.2-Reference Manual*. Joint Research Center of the European Commission. Ispra, Italy.
- Sheikholeslami R, Razavi S, Gupta HV, Becker W, Haghnegahdar A (2019) Global sensitivity analysis for high-dimensional problems: How to objectively group factors and measure robustness and convergence while reducing computational cost. *Environ Model Softw* 111:282–299. <https://doi.org/10.1016/j.envsoft.2018.09.002>
- Sobol' IM (1993) Sensitivity analysis for non-linear mathematical models, *Math Modeling Comput Exp* 1(4): 407–414 (Translated from Russian)
- Sobol' IM (1990) Sensitivity estimates for nonlinear mathematical models, *Matematicheskoe Modelirovanie* 2:112–118 (in Russian)
- Song X, Zhang J, Zhan C, Xuan Y, Ye M, Xu C (2015) Global sensitivity analysis in hydrological modeling: Review of concepts, methods, theoretical framework, and applications. *J Hydrol* 523:739–757. <https://doi.org/10.1016/j.jhydrol.2015.02.013>
- Strang G (2016) *Introduction to linear algebra*. Cambridge University Press, New York
- Wang P, Lu Z, Zhang K, Xiao S, Yue Z (2018) Copula-based decomposition approach for the derivative-based sensitivity of variance contributions with dependent variables. *Reliab Eng Syst Saf* 169:437–450. <https://doi.org/10.1016/j.res.2017.09.012>
- Weisberg S (2005) *Applied linear regression*. Wiley, New York
- Xu C (2013) Decoupling correlated and uncorrelated parametric uncertainty contributions for nonlinear models. *Appl Math Model* 37(24):9950–9969. <https://doi.org/10.1016/j.apm.2013.05.036>
- Xu C, Gertner GZ (2008) Uncertainty and sensitivity analysis for models with correlated parameters. *Reliab Eng Syst Saf* 93(10):1563–1573. <https://doi.org/10.1016/j.res.2007.06.003>
- Zhang K, Lu Z, Cheng L, Xu F (2015) A new framework of variance based global sensitivity analysis for models with correlated inputs. *Struct Saf* 55:1–9. <https://doi.org/10.1016/j.strusafe.2014.12.005>
- Zheng C, Wang Q (2015) Spatiotemporal pattern of the global sensitivity of the reference evapotranspiration to climatic variables in recent five decades over China. *Stoch Environ Res Risk Assess* 29(8):1937–1947. <https://doi.org/10.1007/s00477-015-1120-7>

Zhou Y, Lu Z, Xiao S, Yun W (2019) Distance correlation-based method for global sensitivity analysis of models with dependent inputs. *Struct Multidiscip Optim* 60(3):1189–1207. <https://doi.org/10.1007/s00158-019-02257-z>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.