

UNCERTAINTY QUANTIFICATION IN A TWO-DIMENSIONAL RIVER HYDRAULIC MODEL

**Siham EL GARROUSSI¹, Matthias DE LOZZO², Sophie RICCI¹, Didier LUCOR³,
Nicole GOUTAL^{4,5}, Cédric GOEURY⁴, Sébastien BOYAVAL⁵**

¹CECI, CERFACS/CNRS
42 Av Gaspard Coriolis 31057, Toulouse, Cedex 1
e-mail: {garroussi, ricci}@cerfacs.fr

²IRT Saint Exupéry
CS34436, 3 Rue Tarfaya, 31400 Toulouse
e-mail: matthias.delozzo@irt-saintexupery.com

³ LIMSIS
Campus Universitaire bâtiment 507, Rue John Von Neumann, 91400 Orsay
e-mail: didier.lucor@limsi.fr

⁴ EDF, LNHE
6 quai Watier, 78400 Chatou
e-mail: nicole.goutal, cedric.goeury@edf.fr

⁵ LHSV
6 quai Watier, 78400 Chatou
e-mail: sebastien.boyaval@enpc.fr

Abstract. *River hydraulic models are used to assess the environmental risk associated to flooding and consequently inform decision support systems for civil security needs. These numerical models are generally based on a deterministic approach based on resolving the partial differential equations. However, these models are subject to various types of uncertainties in their input. Knowledge of the type and magnitude of these uncertainties is crucial for a meaningful interpretation of the model results. Uncertainty quantification (UQ) framework aims to probabilize the uncertainties in the input, propagate them through the numerical model and quantify their impact on the simulated quantity of interest, here, water level field discretized over an unstructured finite element mesh over the Garonne River (South-West France) between Tonneins and La Réole simulated with a numerical solver, TELEMAC-2D. The computational cost of the sensitivity analysis with the classical Monte Carlo approach is reduced using a surrogate model instead of the numerical solver. The present study investigates one of the machine learning algorithms: A surrogate model based on Gaussian process. This latter was used to represent*

Keywords: Open-channel flow, Sensitivity analysis, Surrogate model, Gaussian process, Monte Carlo method.

the spatially distributed water level with respect to uncertain stationary flow to the model and friction coefficients. The quality of the surrogate was assessed on a validation set, with small root mean square error and a predictive coefficient equal to 1. Sobol' sensitivity indices are computed and enhance the high impact of the input discharge on the water level variation.

1 INTRODUCTION

Flood inundation models are central components in any flood risk analysis system as they transform the bulk discharge outputs from flood-frequency analyses or rainfall-runoff models into distributed predictions of flood hazard in terms of water depth, inundation extent and flow velocity. Predictions may be dynamic in time and can be derived from a range of codes which vary in complexity from non-model approaches, such as fitting a planar surface to digital elevation data, through to numerical solutions of fluid dynamics equations derived from considerations of mass and momentum conservation.

Whilst such models are parsimonious in terms of their data requirements and number of unconstrained parameters relative to other environmental physics, their underlying equations may be non-linear. Moreover, the data sets that they do require may be subject to complex, but poorly known errors that may vary markedly in time and space. As a consequence, considerable research has, in recent years, sought to understand and better estimate these uncertainties in order to improve flood risk analysis.

Typically, uncertainties in hydrodynamics models stand are classified as: parametric (or epistemic) uncertainty, arising from incomplete knowledge of the correct settings of the models parameters; input data uncertainty, arising from incomplete knowledge of the true value of the initial state and forcing, usually linked to the aleatory nature of the physics; and structural uncertainty, which is the failure of the model to represent the system, even if the correct parameters and inputs are known. Together, these three components represent a complete probabilistic description of the informativeness of the model for the underlying system. But in practice, all are extremely challenging to specify.

In this study, we consider both epistemic and aleatory uncertainties by investigating the effect of two uncertainty sources on water level calculation for extreme flood event, respectively the roughness coefficient and the upstream discharge.

On the one hand, the estimation of the roughness is difficult because it is a lumped parameter that mostly reflects the flow resistance of the river. Since the roughness coefficient has an extensive effect on flow analysis of a river, including computation of the water level and velocity, its accurate estimation is important for prediction of the water level during flooding. Because of its importance, various efforts have been made to quantify the roughness coefficients of rivers in an objective manner. Among them, an element-based method [9] and empirical equations that relate the roughness coefficient either to bed material [32] or to relative depth [5] are representative. However, owing to the diversity and irregularity of natural rivers, prediction of the roughness coefficient for a specific river reach using these methods is not simple. Thus, until now, field measurements have been made either to directly estimate the roughness coefficient [7] or to provide references [2, 18]. However, there remain uncertainties whether using the methods referred to above or using field measurements. From a practical viewpoint, water level and discharge as variables computed by numerical modeling are influenced by uncertainty in estimating the roughness coefficient. Conducting simulation of dam breakage flow for the Teton Dam, [13] showed that variation in calculated flood flow water depth was less than 5% with a 20% change in the roughness coefficient. He therefore argued that even if uncertainty in Stricklers roughness coefficient is large, its effect on the water depth might be reduced considerably in the process of computation. These conclusions should be deeply investigated in the context of flood simulation, on our own river test case, characterized by long homogeneous friction zones calibrated in high flow.

On the other hand, inundation models require the specification of boundary conditions, which

are typically the greatest source of aleatory and epistemic uncertainty when simulating the annual exceeding probability of inundation. Flow at the upstream boundary of the river is often the most important boundary condition, although most applications will require (unless using a kinematic solution) or benefit from downstream-level boundaries (e.g. tidal reaches). In locations where they are available, gauging stations are typically the most accurate source of river flow and level data. However, the ratings at these stations, that convert observed levels to flows, are usually based on low to medium flow observations, necessitating an uncertain extrapolation of the rating for high flows. Rating errors may be especially large when flow is out of bank. Where gauging stations are not available or spatially sparse, hydrological models can be used to simulate upstream discharges. However, despite much effort, rainfall-runoff models are still very uncertain, especially where calibration/validation data are lacking.

Subsequently, once the sources of uncertainties have been identified, they must be propagated in the model. The Monte Carlo (MC) methods are the most common techniques used for uncertainty propagation (UQ) [15]. This framework allows to estimate standard statistics on the model output, e.g. expectation, standard deviation, quantiles or probabilities of exceeding a given threshold. It also makes it possible to estimate sensitivity indices representing the shares of output uncertainty attributable to the different uncertain input parameters, e.g. Sobol' indices where output uncertainty is measured in terms of variance [30]. MC is simple and highly adapted to massively parallel computational resources. Yet, its convergence is slow as it scales inversely to the square root of the sample size and its cost gets prohibitive for expensive models. In this respect, surrogate models such as Gaussian process model (GP) have received tremendous attention in the last few years, as it allows one to replace the original expensive model by a surrogate which is built from an experimental design of limited size [25]. Then the surrogate can be used to compute the UQ study in negligible time. In particular, [27] have shown that, for a 1D hydraulic model, on the Garonne river section that we consider and stationary flow, it features similar performance to estimate statistics by Monte-Carlo random sampling when friction and input forcing uncertainties are taken into account. The accuracy of the water level correlation matrix and sensitivity Sobol' indices estimated with the GP surrogate was assessed with respect to a classical MC estimate based on a large data set. This article is a reference for us because it involves the same section of Garonne river, the same types of uncertain variables (friction and upstream flow) and the same family of surrogate model as those considered in our work. Our work can be seen as an extension to two-dimensional flow modelling and floodplain characterization.

The present study extends the surrogate model approach in hydraulics to 2D modeling taking into account the dynamics of the flood plain. Section 2 presents the hydrodynamics solver TELEMAC-2D, the Garonne test case used in this article and the associated uncertainties. Section 3 presents the GP surrogate strategy based on the reduction of the dimension of the output space with a Proper Orthogonal Decomposition (POD). This section also presents the metrics used to assess the quality of the surrogate along with the sensitivity indices based on output variance decomposition. Results are presented in Section 4. Conclusions and perspectives are finally given in Section 5.

2 MODEL: TWO-DIMENSIONAL FLOW OF THE GARONNE RIVER UNDER UNCERTAINTY

2.1 Physical model

The Shallow Water Equations (SWE), also called depth-averaged free surface flow equations, are commonly used in environmental hydrodynamics modelling [12]. They are derived from the Navier-Stokes equations [31] and express mass and momentum conservation averaged in the vertical dimension. The non-conservative form of the equations are written in terms of the water depth (h) and the horizontal components of velocity (u and v):

$$\frac{\partial h}{\partial t} + \text{div}(hu) = 0 \quad (1)$$

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} = -g \frac{\partial H}{\partial x} + F_x + \frac{1}{h} \text{div} \left(h \nu_e \overrightarrow{\text{grad}}(u) \right) \quad (2)$$

$$\frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} = -g \frac{\partial H}{\partial y} + F_y + \frac{1}{h} \text{div} \left(h \nu_e \overrightarrow{\text{grad}}(v) \right) \quad (3)$$

where:

$$\begin{cases} F_x = -\frac{g}{K_s^2} \frac{u\sqrt{u^2+v^2}}{h^{\frac{4}{3}}} - \frac{1}{\rho_w} \frac{\partial P_{atm}}{\partial x} + \frac{1}{h} \frac{\rho_{air}}{\rho_w} C_{DZ} U_{w,x} \sqrt{U_{w,x}^2 + U_{w,y}^2} \\ F_y = -\frac{g}{K_s^2} \frac{v\sqrt{u^2+v^2}}{h^{\frac{4}{3}}} - \frac{1}{\rho_w} \frac{\partial P_{atm}}{\partial y} + \frac{1}{h} \frac{\rho_{air}}{\rho_w} C_{DZ} U_{w,y} \sqrt{U_{w,x}^2 + U_{w,y}^2} \end{cases}$$

and: ρ_w/ρ_{air} [kg.m⁻³] is the water/air density, P_{atm} [Pa] is the atmospheric pressure, $U_{w,x}$ and $U_{w,y}$ [m.s⁻¹] are the horizontal wind velocity components, C_{DZ} [-] is the wind influence coefficient, K_s [m ^{$\frac{1}{3}$} .s⁻¹] is the river bed and floodplain friction coefficient, using the Strickler formulation [?]. F_x and F_y [m.s⁻²] are the horizontal components of external forces (friction, wind and atmospheric forces), h [m] is the water depth, H [m] is the water level ($h = H - z_f$ if z_f [m] is the bottom level), u and v [m.s⁻¹] are the horizontal components of velocity and ν_e [m².s⁻¹] is the water diffusion coefficient. div and $\overrightarrow{\text{grad}}$ are respectively the divergence and gradient operators.

To solve the system of equations (Eq. (1) to Eq. (3)), initial conditions $h(x, y, t = 0) = h_0(x, y)$; $u(x, y, t = 0) = u_0(x, y)$; $v(x, y, t = 0) = v_0(x, y)$ are provided along with boundary conditions (BC) at surface, at bottom and at upstream and downstream frontiers $h(x_{BC}, y_{BC}, t) = h_{BC}(t)$.

2.2 Study area

The study area extends over a 50 km reach of the Garonne river (France) between Tonneins (upstream), downstream of the confluence with the river Lot, and La Réole (downstream) (see Figure 1). This part of the valley was equipped in the 19th century with infrastructure to protect the Garonne flood plain from flooding such as that occurred in 1875. A system of longitudinal dykes and weirs was progressively constructed after that flood event to protect the floodplains, organize submersion and flood retention areas. Protections on the Garonne river form a system of successive storage areas for the flood plain beyond the dikes.

2.3 Uncertainty characterization

The hydraulic variables are discretized on an unstructured triangular mesh over the two-dimensional study area. We note \mathbf{h} the vector of the water level over the $p = 41416$ nodes of

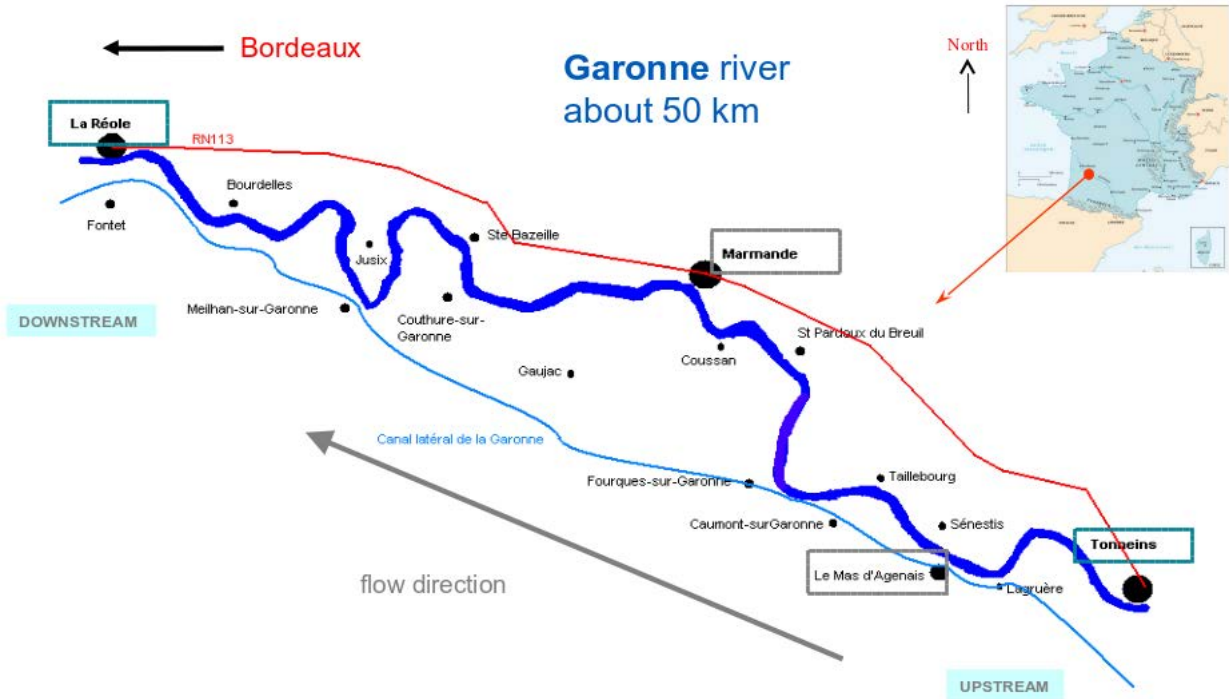


Figure 1: Study area of the Garonne river

the mesh. It represents our quantity of interest (QoI). In this study, the impact of roughness and upstream flow on the discretized water level h is quantified in the context of extreme flood event:

- The roughness coefficient defined according to 4 areas. Indeed, T2D was calibrated for high flow in [4] using steady-state water surface profiles at high discharge, from bank-full discharge in the main channel ($2\,500\text{ m}^3\cdot\text{s}^{-1}$) to bank-full discharge in the overbank flow channel between dykes. In the floodplains, the roughness coefficient $K_{s,1}$ is selected as an area with cultivated fields all around the river with a Strickler coefficient of $17\text{ m}^{1/3}\cdot\text{s}^{-1}$. Classically, according to the available expert knowledge, the friction coefficient is contained in an interval bounded by physical values depending on the roughness of soil material.

For the main channel, the Strickler roughness coefficient was split into three different areas:

- from Tonneins to upstream of Mas d'Agenais, $K_{s,2}$: $45\text{ m}^{1/3}\cdot\text{s}^{-1}$,
- from upstream of Mas d'Agenais to upstream of Marmande, $K_{s,3}$: $38\text{ m}^{1/3}\cdot\text{s}^{-1}$,
- from upstream of Marmande to La Réole, $K_{s,4}$: $40\text{ m}^{1/3}\cdot\text{s}^{-1}$.

The distribution of Strickler roughness coefficient is chosen uniform and the interval is set to cover the range of calibration values.

- The upstream discharge is assumed to follow a Gaussian distribution centered around the thousand return period $8\,490\text{ m}^3\cdot\text{s}^{-1}$ with a standard deviation of $700\text{ m}^3\cdot\text{s}^{-1}$. The study is thus focused on extreme flood events that activate the flood plains.

Tab. 1 summarizes the considered input uncertainties.

Variable	Distribution	Units
Q	$\mathcal{U}[2\,500, 10\,000]$	$\text{m}^3.\text{s}^{-1}$
$K_{s,1}$	$\mathcal{U}[5, 20]$	$\text{m}^{1/3}.\text{s}^{-1}$
$K_{s,2}$	$\mathcal{U}[40, 50]$	$\text{m}^{1/3}.\text{s}^{-1}$
$K_{s,3}$	$\mathcal{U}[33, 43]$	$\text{m}^{1/3}.\text{s}^{-1}$
$K_{s,4}$	$\mathcal{U}[35, 45]$	$\text{m}^{1/3}.\text{s}^{-1}$

Table 1: Distributions of the input variable uncertainties.

2.4 Computing environment

In this work, hydrodynamic is provided using TELEMAC-2D (T2D) depth-averaged hydrodynamic model¹. It solves the SWE in two dimensions with an explicit first-order time integration scheme, a finite element scheme and an iterative conjugate gradient method. In each point of the mesh, T2D gives the water depth and the vertically average horizontal velocity field [17]. T2D was developed initially by the National Hydraulics and Environment Laboratory (LNHE) of the Research and Development Directorate of EDF, and is now managed by a consortium. The software comes with an API to modify the values of the uncertain parameters in a non-intrusive way.

The surrogate model construction as well as the sensitivity analysis was carried out using the BATMAN-Open-TURNS (BATMAN-OT) library². This library (developed at CERFACS and CECILL-B licensed) provides a convenient, modular and efficient framework for design of experiments, surrogate model and uncertainty quantification [26]. It relies on open source python packages dedicated to statistics (openTURNS³ [3] and scikit-learn⁴ [23]). It also implements advanced methods for resampling, robust optimization and uncertainty visualization.

In terms of infrastructures, CERFACS’s cluster, Nemo, has been used to run T2D simulations. The Nemo cluster includes 6,912 cores distributed in 288 compute nodes. The ECU power peak is 277 Tflop/s. On this architecture, simulating the river and flood plain dynamics for the test case presented in Sect. 2.2 over 3 days, takes about 6 minutes on 15 cores.

3 AN EFFICIENT UQ FRAMEWORK FOR COSTLY TWO-DIMENSIONAL SIMULATOR

The Monte Carlo (MC) framework is the most common framework used for uncertainty quantification, due to its simplicity and good statistical results. It is theoretically applicable whatever the complexity of the deterministic model or the desired statistical estimator. However, the required sample size increases squarely with the estimator accuracy and makes this approach rather impracticable when the computational cost of each run of the model, like T2D, is non negligible. One way to lower the computationally demanding is to replace, on one side, the T2D model by a surrogate model [11], on the other side, the pure random sampling by alternative sampling methods such as the Latin Hypercube sampling approach [16].

¹More information can be found on the website www.opentelemac.org.

²BATMAN-OT can be downloaded from <https://gitlab.com/cerfacs/batman>.

³More information on: <http://www.openturns.org>.

⁴More information on: <https://scikit-learn.org>.

3.1 Build a surrogate model with spatial output

The surrogate model based on Gaussian process regression (GP) [25] has been adopted in the following. We have chosen this metamodel to the detriment of others for two reasons. The first one is its small number of hyperparameters: about one per input parameter. The second one is that it provides a measurement of its model error which would be of interest in our future work, for optimization and data assimilation problems, based on methods as expected improvement [19]. While surrogate models offer a low cost alternative to costly models, their formulation is challenging in high dimension for inputs and outputs. In the present case, the size of the uncertain input space is small and resumes to 5 scalars. Yet, the quantity of interest is 2D and discretized over more than 41 000 points. The output space is reduced in order to limit the cost of the surrogate formulation and the spatial coherence of the later, using a Proper Orthogonal Decomposition (POD) strategy [24]. POD is a post-processing technique that takes a given set of data and extracts basis functions, that contain as much "energy" as possible. The meaning of "energy" depends on which kind of POD is used [8]. Here, only POD based on snapshot method [29] is considered.

We propose to build a surrogate model combining POD and GP surrogate model. So we call it "POD+GP surrogate model". The corresponding algorithm is presented as follows:

1. build a learning dataset $\mathcal{D}_l = (\mathbf{x}^{(i)}, \mathbf{h}^{(i)})_{1 \leq i \leq N_l}$ of size N_l where the design of experiments $(\mathbf{x}^{(i)})_{1 \leq i \leq N_l}$ is a Latin hypercube sample (LHS) [21] with $\mathbf{x} = (Q, K_{s,1}, K_{s,2}, K_{s,3}, K_{s,4})$ and $\mathbf{h}^{(i)}$ is the water level computed by T2D over the mesh at $\mathbf{x}^{(i)}$;
2. decompose the sampled output vector \mathbf{h} by achieving a POD on the centered output learning matrix $\mathbf{H} = \left(h_j^{(i)} - N_l^{-1} \sum_{k=1}^{N_l} h_j^{(k)} \right)_{\substack{1 \leq i \leq N_l \\ 1 \leq j \leq p}}$ and derive the most significant components; then, any sampled local water level $\mathbf{h}^{(i)}$ can be approximated by a weighted sum of these components where weights depend on the input vector value $\mathbf{x}^{(i)}$;
3. for each component, approximate the relation between its sampled coefficient and the corresponding sampled model inputs by means of a GP surrogate model;
4. formulate the POD+GP surrogate model $\hat{\mathbf{h}}(\mathbf{x})$ as the weighted sum of the more significant POD components where weights are the GP surrogate models depending on \mathbf{x} .

3.1.1 Reduction of the output dimension by proper orthogonal decomposition (POD)

The key idea of the snapshot method [29] is to achieve a POD of the centred snapshot matrix $\mathbf{H} = \left(h_j^{(i)} - N_l^{-1} \sum_{k=1}^{N_l} h_j^{(k)} \right)_{\substack{1 \leq i \leq N_l \\ 1 \leq j \leq p}} \in \mathbb{M}_{N_l, p}(\mathbb{R})$, which gathers the water level computed at each mesh point for the N_l snapshots, from which the sample mean is substracted.

Based on many observations of a random vector, the POD gives the orthogonal directions of largest variances (or modes) in the probabilistic vector space in order to reduce the vector space dimension [6]. Note that for simplicity purpose, the adjective *centred* is dropped in the following when referring to the centred snapshot matrix \mathbf{H} .

The POD of the snapshot covariance matrix $\mathbf{C} = N_l^{-1} \mathbf{H}^T \mathbf{H} \in \mathbb{M}_p(\mathbb{R})$ is equivalent to the Singular Value Decomposition (SVD) of the snapshot matrix \mathbf{H} :

$$\mathbf{H} = \mathbf{U} \mathbf{\Lambda} \mathbf{V}^T = \sum_{k=1}^{r_p} \lambda_k \mathbf{u}_k \mathbf{v}_k^T, \quad (4)$$

where $\mathbf{U} \in \mathbb{M}_{N_l}(\mathbb{R})$ is an orthogonal matrix diagonalizing $\mathbf{H}\mathbf{H}^T$ (\mathbf{u}_k , the k^{th} column of \mathbf{U} , is a left singular vector of \mathbf{H}), where $\mathbf{V} \in \mathbb{M}_p(\mathbb{R})$ is an orthogonal matrix diagonalizing $\mathbf{H}^T\mathbf{H}$ (\mathbf{v}_k , the k^{th} column of \mathbf{V} , is a right singular vector of \mathbf{H}), and where $\mathbf{\Lambda} \in \mathbb{M}_{N_l,p}(\mathbb{R})$ is a rectangular diagonal matrix including $r_p = \min(N_l, p)$ singular values on its diagonal. The singular values $\{\lambda_k\}_{1 \leq k \leq r_p}$ are the square roots of the eigenvalues of \mathbf{C} . Note that in this study, since the size of the training set N_l is lower than the number of mesh points $p = 41\,416$, the rank of \mathbf{H} is here $r_p = N_l$.

At the k^{th} mesh point, the snapshot $h_k(\mathbf{x}^{(i)})$ can then be retrieved as a linear combination of r_p modes $\{\Psi_i\}_{1 \leq i \leq r_p}$:

$$h_k(\mathbf{x}^{(i)}) = (\mathbf{U} \mathbf{\Lambda} \mathbf{V}^T)_{ki} = U_{k,:} (\mathbf{\Lambda} \mathbf{V}^T)_{:,i} = \sum_{j=1}^{r_p} \gamma_{k,j} \Psi_j(\mathbf{x}^{(i)}), \quad (5)$$

where for any $j \in \{1, \dots, N_l\}$, $\gamma_{p,j} := U_{p,j}$ and $\Psi_j(\mathbf{x}^{(i)}) := (\mathbf{\Lambda} \mathbf{V}^T)_{j,i}$.

From that, we want to approximate each relation $\mathbf{x} \rightarrow \Psi_j(\mathbf{x})$ by a GP surrogate model $\Psi_{gp,j}$ from the dataset $(\mathbf{x}^{(i)}, \Psi_j(\mathbf{x}^{(i)}))_{1 \leq i \leq N_l}$ in order to propose the following POD+GP surrogate model:

$$\hat{h}_k(\mathbf{x}) = \sum_{i=1}^{r_p} \gamma_{k,i} \Psi_{gp,i}(\mathbf{x}), \quad (6)$$

This POD+GP surrogate model requires the construction of r_p GP surrogate models.

3.1.2 Learning of the significant POD modes by Gaussian process (GP) modelling

As stated by [25], a GP is a random process (here the mode Ψ_i) indexed over a domain (here \mathbb{R}^d), for which any finite collection of process values (here $\{\Psi_i(\mathbf{x}^{(j)})\}_{1 \leq j \leq N_l}$) has a joint Gaussian distribution. Concretely, let $\tilde{\Psi}_i$ be a Gaussian random process fully described by its zero mean and its correlation π_i :

$$\tilde{\Psi}_i(\mathbf{x}) \sim \text{GP}(0, \sigma_i^2 \pi_i(\mathbf{x}, \mathbf{x}')), \quad (7)$$

with $\pi_i(\mathbf{x}, \mathbf{x}') = \mathbb{E}[\tilde{\Psi}_i(\mathbf{x})\tilde{\Psi}_i(\mathbf{x}')]]$. In our case, the correlation function π (or kernel) is chosen as a squared exponential:

$$\pi_i(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2 \ell_i^2}\right), \quad (8)$$

where ℓ_i is a length scale describing dependencies of model output between two input vectors \mathbf{x} and \mathbf{x}' , and where σ_i^2 is the variance of the output signal. Squared exponential kernel leads to satisfying results but other kernel functions could have been considered, such as a decreasing exponential one or a Matérn one – with their associated hyper-parameters. The choice of

the kernel is still an open problem and can be mitigated using the available information on the problem. The square exponential kernel leads to very smooth, thus stable results. Furthermore, it implies that the model is exact at sample points; it does not introduce any other strong assumptions, hence its wide usage among practitioners.

Then the surrogate model of interest is the mean of the GP resulting of conditioning $\tilde{\Psi}_i$ by the training set $\{\Psi_i(\mathbf{x}^{(k)})\}_{1 \leq k \leq N_i}$. For any $\mathbf{x}^* \in \mathbb{R}^d$,

$$\Psi_{\text{gp},i}(\mathbf{x}) = \sum_{k=1}^N \beta_{k,i} \pi_i(\mathbf{x}, \mathbf{x}^{(k)}), \quad (9)$$

where $\beta_{k,i} = (\mathbf{\Pi}_i + \tau^2 \mathbf{I}_{N_i})^{-1} (\Psi_i(\mathbf{x}^{(1)}) \dots \Psi_i(\mathbf{x}^{(N_i)}))^T$ with $\mathbf{\Pi}_i = (\pi_i(\mathbf{x}^{(j)}, \mathbf{x}^{(k)}))_{1 \leq j, k \leq N_i}$, and where τ (referred to as the nugget effect) avoids ill-conditioning issues for the matrix $\mathbf{\Pi}$. The hyperparameters $\{\ell_i, \sigma_i, \tau\}$ are optimized by maximum likelihood applied to the data set \mathcal{D}_N using the L-BFGS-B algorithm [33].

3.1.3 Quality measures for the POD+GP surrogate model

In the present study, two common error metrics are used to assess the quality of the surrogate water level both on the entire mesh (global approach) and at each point of the mesh (local approach): the root mean square error (RMSE) and the predictive coefficient (Q_2). This validation is carried out over an input-output validation dataset \mathcal{D}_v of size N_v .

Root mean square error (RMSE)

The RMSE is used to measure the accuracy of the model, it should be 0 when the model is perfect. At the k^{th} given mesh node, it is defined as the square root of the mean square error (MSE) measuring the square distance between the surrogate model and the reference model:

$$\text{MSE}_k(\mathcal{D}_v) = N_v^{-1} \sum_{i=1}^{N_v} \left(h_k^{(i)} - \hat{h}_k^{(i)} \right)^2 \quad \text{and} \quad \text{RMSE}_k(\mathcal{D}_v) = \sqrt{\text{MSE}_k(\mathcal{D}_v)} \quad (10)$$

Their global counterpart are: $\text{MSE}(\mathcal{D}_v) = p^{-1} \sum_{k=1}^p \text{MSE}_k(\mathcal{D}_v)$ and $\text{RMSE}(\mathcal{D}_v) = \sqrt{\text{MSE}(\mathcal{D}_v)}$.

Predictive coefficient (Q_2)

At the k^{th} mesh node, the Q_2 predictive coefficient is defined as:

$$Q_{2,k} = 1 - \frac{\text{MSE}_k(\mathcal{D}_v)}{\text{MSE}_k(\mathcal{D}_v; \text{mean})} \quad (11)$$

where $\text{MSE}_k(\mathcal{D}_v; \text{mean}) = N_v^{-1} \sum_{i=1}^{N_v} \left(h_k^{(i)} - \bar{h}^{(i)} \right)^2$ is the MSE of the ‘‘averaging model’’ returning the mean of the learning outputs whatever the input parameter value.

The global counterpart of $\text{MSE}_k(\mathcal{D}_v; \text{mean})$ is $\text{MSE}(\mathcal{D}_v; \text{mean}) = p^{-1} \sum_{k=1}^p \text{MSE}_k(\mathcal{D}_v; \text{mean})$. Thus, the global counterpart of $Q_{2,k}$ is:

$$Q_2 = 1 - \frac{\text{MSE}(\mathcal{D}_v)}{\text{MSE}(\mathcal{D}_v; \text{mean})}. \quad (12)$$

The predictive coefficient measures the performance of the surrogate model with respect to the simplest one which consists in averaging the learning output values. When Q_2 is lower than (resp. equal to) zero, the surrogate is worse than (resp. equal to) the learning output values average. When Q_2 is equal to one, the surrogate interpolates the validation dataset. In practice, the surrogate is deemed appropriate when Q_2 is greater than 0.8. The predictive coefficient is also found under the name of Nash-Sutcliffe model efficiency coefficient in the hydrological literature where it assesses the predictive capacity of the simulated discharge over a time window with respect to observed discharges [22].

3.2 Quantify and explain the output uncertainty due to input uncertainty propagation

Once the model is built and validated, it can be used instead of the reference model in an uncertainty quantification study. After propagating the input uncertainties through the surrogate model by means of specific Monte Carlo methods, we can conduct a statistical analysis on the output uncertainty as well as a sensitivity analysis to explain how the uncertain input parameters contribute to this output variability.

3.2.1 Statistical analysis on the output

Using a standard MC approach on the validation data set \mathcal{D}_v , the mean value and standard deviation of the water level at the k^{th} mesh point are formulated as:

$$\mu_k = \frac{1}{N_v} \sum_{i=1}^{N_v} \hat{h}_k^{(i)} \quad \text{and} \quad \sigma_k = \sqrt{\frac{1}{N_v - 1} \sum_{i=1}^{N_v} \left(\hat{h}_k^{(i)} - \mu_k \right)^2}. \quad (13)$$

3.2.2 Sensitivity analysis on the output with respect to the inputs

Sobol' indices [30] are commonly used for sensitivity analysis. They provide the shares of the QoI variance \mathbb{V} attributable to the different uncertain inputs. Under the hypotheses that random input variables are independent, here the roughness coefficients and the upstream flow, and the random QoI is square integrable, here the water level h , the decomposition of the QoI reads:

$$\mathbb{V} = \sum_{i=1}^d \mathbb{V}_{\{i\}} + \sum_{j=i+1}^d \mathbb{V}_{\{i,j\}} + \cdots + \mathbb{V}_{\{1,2,\dots,d\}} = \sum_{J \subset \{1,2,\dots,d\}} \mathbb{V}_J, \quad (14)$$

where $\mathbb{V} := \text{Var}[\text{QoI}]$, $\mathbb{V}_i := \mathbb{V}[\mathbb{E}[\text{QoI}|X_i]]$, $\mathbb{V}_{ij} := \mathbb{V}[\mathbb{E}[\text{QoI}|X_i X_j]] - \mathbb{V}_i - \mathbb{V}_j$ and more generally, for any $I \subset \{1, \dots, d\}$, $\mathbb{V}_I := \mathbb{V}[\mathbb{E}[\text{QoI}|x_I]] - \sum_{J \subset I \text{ s.t. } J \neq I} \mathbb{V}_J$. Then, we obtain:

$$1 = \sum_{i=1}^d S_{\{i\}} + \sum_{j=i+1}^d S_{\{i,j\}} + \cdots + S_{\{1,2,\dots,d\}} = \sum_{J \subset \{1,2,\dots,d\}} S_J, \quad (15)$$

where for any $J \subset \{1, 2, \dots, d\}$, $S_J = \frac{\mathbb{V}_J}{\mathbb{V}}$ is called a Sobol' index and belongs to the interval $[0, 1]$. $S_{\{i\}}$ is the first order Sobol' index corresponding to the ratio of output variance due to the i^{th} input parameter uniquely, and $S_{\{i,j\}}$ is the second-order Sobol' index describing the ratio of output variance due to the i^{th} parameter in interaction with the j^{th} parameter. Also the total Sobol' index that corresponds to the whole contribution of the i^{th} input parameter reads:

$$S_{T_i} = \sum_{\substack{I \subset \{1,\dots,d\} \\ I \ni i}} S_I. \quad (16)$$

The computation of first order Sobol' indices requires simple integration, those of the second order requires double integration, and so on. Many Monte Carlo techniques exist to estimate these integrals. In this study, the Sobol' indices are estimated using the algorithm proposed in [28].

Lastly, we note that the expression (14) is defined for a scalar QoI, such as the water level h_k at mesh node $k \in \{1, 2, \dots, p\}$ where p is the mesh size. Consequently, we can easily plot the different Sobol' indices over the mesh on which is defined the model output. Furthermore, this information can be summarized using the generalized Sobol' indices [20]:

$$\forall J \subset \{1, 2, \dots, d\}, S_J = \frac{\sum_{k=1}^p \mathbb{V}_{[k]} S_{[k],J}}{\sum_{\ell=1}^p \mathbb{V}_{[k]}^{(\ell)}}. \quad (17)$$

4 RESULTS

4.1 Learning and test samples

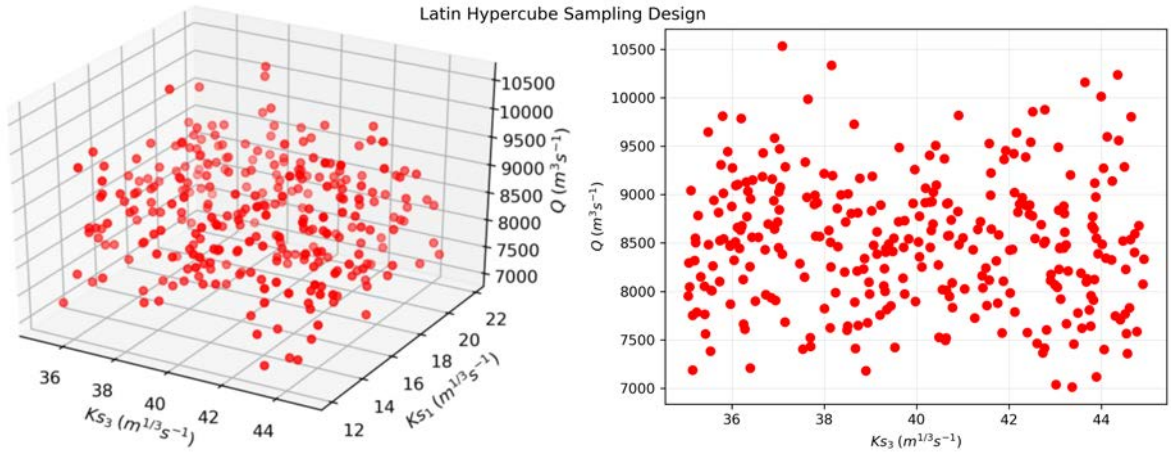


Figure 2: Latin Hypercube Sampling (LHS) DoE for a 300-sample data set, along $(K_{s,1}, K_{s,3}, Q)$ directions on the left panel and along $(K_{s,3}, Q)$ directions on the right panel.

The design of experiment (DoE) for the training and validation data set was generated using Latin Hypercube Sampling (LHS) [21] which is a statistical method for generating a near-random sample of parameter values from a multidimensional distribution. Considering d the number of input variables, LHS strategy scales as $o(d)$ while other strategies require a larger number of samples; for instance to insure the convergence of first order statistics [10]. The LHS space-filling experimental design is shown in Fig. 2, it is associated with a limited computational cost. While more advanced sampling method could be used, LHS strategy was deemed efficient for the present study.

4.2 Surrogate model

The LHS strategy has been applied twice. A first time to build a 2000-sample training set and a second one to create a 1000-sample validation set. Here, the GP kernel was prescribed to

a Matern(2.5) function. The validation set was only used to assess the quality of the surrogate model with RMSE and Q_2 error metrics.

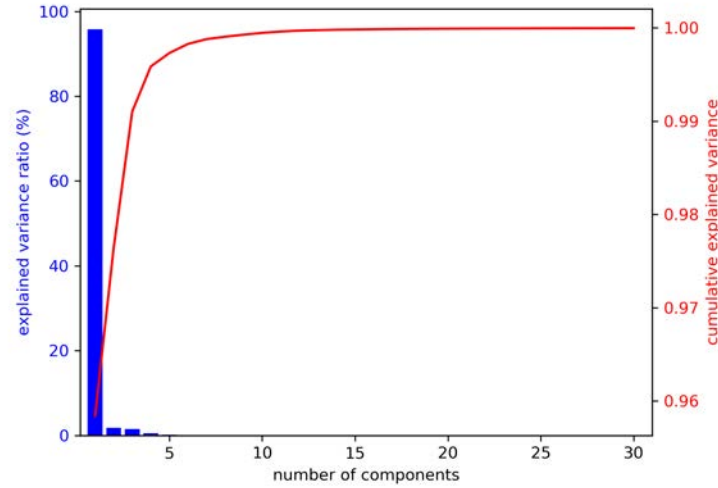


Figure 3: Principal component analysis

The dimension of the output space was reduced with a POD in order to limit the cost of the GP surrogate. The number of modes, also called principal components, to be retained is justified by two criteria taken into account:

- Elbow criterion: on the scree of the POD modes, there is a decrease (elbow) followed by a more regular decrease. In our case, as shown in figure 3, a decrease occurs at the fourth mode, then a regular decrease from the fifth mode. Thus only the first four modes are of interest.
- Kaiser's criterion: only those modes whose inertia is greater than the average inertia should be retained. This criterion leads us to select 4 modes, explaining 99.6% of the total inertia. Indeed, the first principal component explains 95.84% of the total inertia, the second 1.80%, the third 1.46% and the fourth 0.48%.

The cost of the GP surrogate significantly decreases when the dimension of the output is reduced⁵ (6 times smaller) applying the POD, as presented in Tab. 2. The output dimension is indeed reduced from 41 416 elements to 4 components that explain 99.6% of the variance of the QoI. But, the physical interpretation of the different modes is not always obvious.

As displayed in Fig. 4, the first mode, which explains 95.84% of the output variability, seems to represent the effect of the upstream discharge on the average water level height. Indeed, this component is essentially negative thus its weighting will increase everywhere the average water level height if negative coefficient or decrease everywhere the average water level if positive coefficient. While the second mode, which explains 1.80% of the output variability, seems to

⁵The remaining cost can be considered significant compared to linear surrogate models such as polynomial chaos expansion. This situation is well-known and naturally explained by the learning sample size increasing the cost of inverting the covariance matrix. For prediction, this surrogate model is as fast as the others and also provides a measure of its error.

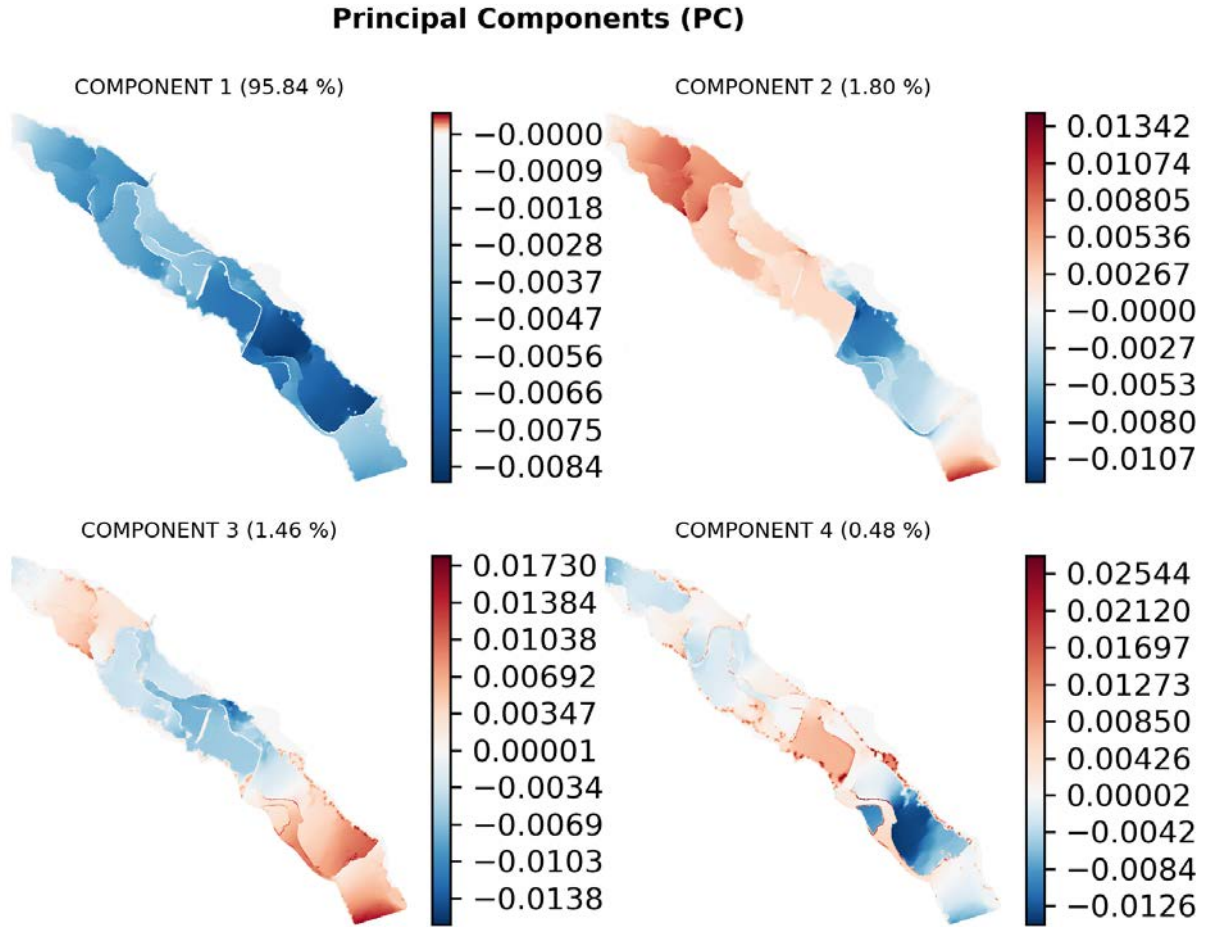


Figure 4: Principal component analysis

represent the effect of main channel flow Strickler friction coefficients on the average water level height as it allows to distinguish the three friction areas defined in section 2.3.

To give more meaning to the principal modes of the decomposition, perspective of our work could stand in the representation of the learning data set on the bi-dimensional sub-spaces spanned by couple of modes, e.g. visualizing the learning data set in a plot with the first mode on the x-axis and the second one on the y-axis.

	GP	POD+GP
CPU run time (h)	≥ 12	2.5

Table 2: CPU run time comparison between GP without and with POD.

The POD+GP surrogate quality is very good with respect to global error metrics $RMSE = 0.8 \text{ cm}$ and $Q_2 = 0.99748$. Locally, the quality deteriorates near the boundary of the catchment area as well as along the dikes as shown in Fig. 5. The heterogeneity of the mesh with small cells in the river bed ($\leq 40 \text{ m}$), near the dikes ($\leq 80 \text{ m}$) and larger cells in the flood plain ($\leq 150 \text{ m}$) should be noted and may hide some local failures of the surrogate in the global RMSE and Q_2 criteria.

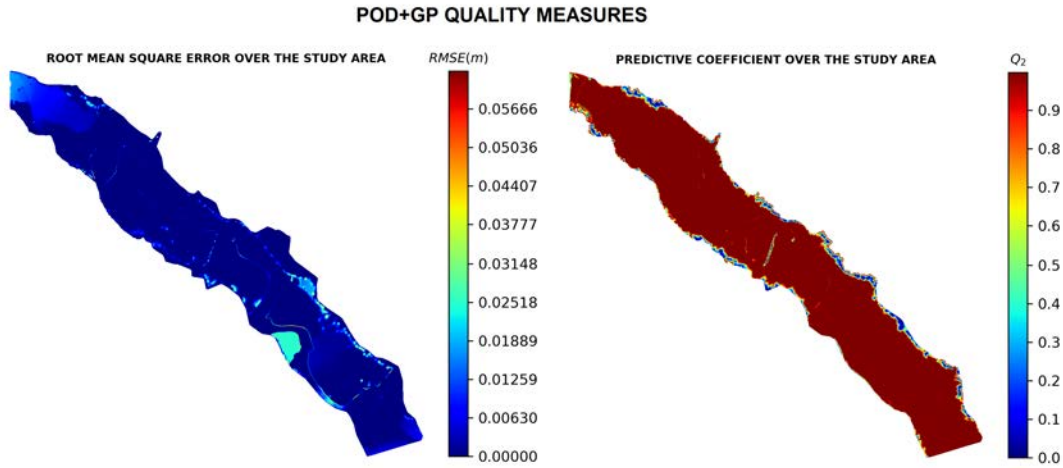


Figure 5: Root mean square error

4.3 Sensitivity analysis with the POD+GP surrogate

The POD+GP surrogate is used to carry out a variance-based sensitivity analysis (SA) over the entire simulated area, with a focus on mesh node 29 515 where Marmande, a city prone to flooding, is located. The POD+GP surrogate allows for a reliable estimation of first and second order statistical moments at Marmande as shown in Tab. 3: the water level mean and standard deviation estimated from the direct model T2D and surrogate are in good agreement with an under estimation of 1.3% for the mean computed with the surrogate.

Given the statistical distributions for the input variables, the SA at Marmande highlights that most of the water level variance is explained by the upstream discharge Q and to a lesser extend, by the Strickler friction coefficient $K_{s,4}$ prescribed between Marmande and La Role as displayed in Fig. 6. At this location, the floodplain friction coefficient $K_{s,1}$ and the friction coefficients upstream of Mas d’Agenais ($K_{s,2}$) and upstream of Marmande ($K_{s,3}$) have barely no impact on the water level. It should be noted that the bootstrap method [1] is used to estimate the variance of the Sobol’ indices, this variance is represented by the black error bars in Fig. 6. These indicate that the computation of the SA indices is converged and reliable. It should also be noted that for each input variable, the first (S) and total Sobol’ (S_T) indices at Marmande are equal, meaning that, at this location, the multivariate impact of the input on the water level is minimal.

	POD+GP	T2D
Mean (m)	21.57	21.54
Standard deviation (m)	0.24	0.24

Table 3: Statistical moments of the water level height in Marmande.

Fig. 7 displays the mean and the standard deviation of the water level over the 2D domain estimated with the POD+GP surrogate. The mean varies between 0 m near the limits of the domain and 21.57 m at Marmande, where it reaches its maximum. In the floodplain, the mean water level ranges from 3.21 m and 7.8m close to the dikes. The water level standard deviation ranges from 0 m to 0.4 m in the floodplain between Mas d’Agenais and Marmande, where the flow is highly bi-dimensional.

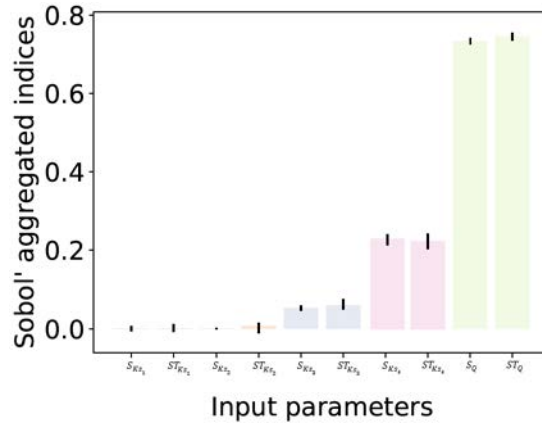


Figure 6: First (S) and total (S_T) Sobol' indices estimated with POD+GP surrogate at Marmande.

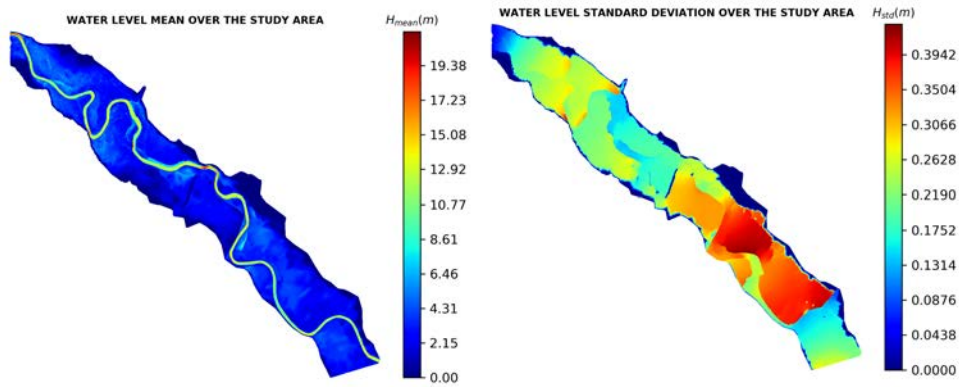


Figure 7: a- Mean water level, b- standard deviation of the water level, estimated with the POD+GP surrogate.

The 2D computation and representation of first Sobol' indices confirms that the variance of the water level is mostly explained (81 % on average) by the input discharge as shown in Fig. 8 and Fig. 9. The floodplain friction coefficient $K_{s,1}$ has no impact on the analysis, the upstream friction coefficient in the river bed $K_{s,2}$ has a small impact on the water level close to Mas d'Agenais, the friction coefficient $K_{s,3}$ between Mas d'Agenais and Marmande explains up to 10 % of the water level variance close to Marmande and the downstream friction coefficient $K_{s,4}$ has an impact over the entire domain with most significance at the upstream and downstream boundaries. As the sum of the first Sobol' indices is smaller than 1, higher order Sobol' indices are non zero, meaning that multivariate effects between Q and K_s explain the remaining part of the water level variance.

5 CONCLUSION AND PERSPECTIVES

In this paper, an uncertainty quantification study was carried out with a 2D numerical solver for the Shallow Water Equations on a section of the Garonne river. It consisted in building a Gaussian process surrogate model on a POD-reduced 2D water level output field.

The surrogate model was formulated with respect to friction coefficients and input discharge, the distribution for friction is supposed to be uniform and centered around calibration values while the discharge distribution is supposed to be Gaussian, centered around a high flood value. The construction of the surrogate was achieved over a 2000-sample training data set and it was

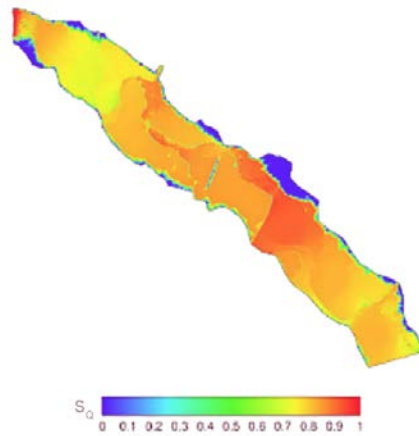


Figure 8: First order Sobol' sensitivity indices computed with the POD+GP surrogate with respect to the input discharge Q .

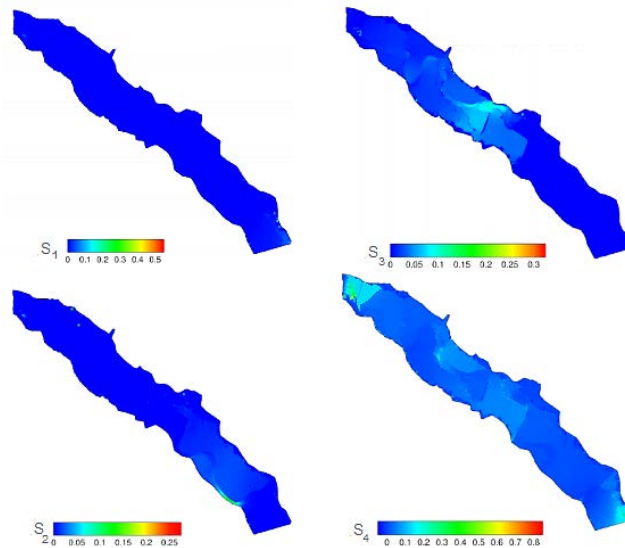


Figure 9: First order Sobol' sensitivity indices computed with the POD-GP surrogate with respect to the friction coefficients $K_{s,1}$ (flood plain), $K_{s,2}$ upstream Mas d'Agenais, $K_{s,3}$ (between Mas d'Agenais and Marmande and $K_{s,4}$ (downstream of Marmande).

validated over a 1000-sample data set. The dimension of the quantity of interest was reduced from 41416 elements to 4 principal components using the POD which has resulted in a significant reduction of the computational cost of the surrogate. The correlation kernel was here prescribed as a Matern(2.5) function. The quality of the POD+GP surrogate model was assessed, the surrogate was deemed satisfying with Q_2 metrics close to 1 for the entire domain and RMSE smaller than $0.01m$. The quality of the surrogate decreases near the dikes. The surrogate was used to perform a global sensitivity analysis based on variance decomposition. It was demonstrated that the upstream discharge is the predominant input variable and explains more than 80 % of the water level variance. The downstream friction coefficient is also a significant input with heterogeneous influence.

It is essential to mention that the conclusions for this study are strongly related to the hypothesis made for the statistical distribution of the inputs. For instance, further study should

investigate wider ranges for flood plain coefficients that are highly unknown and may significantly over time as flood events occur on the catchment.

The results of the sensitivity analysis allows for a better understanding of the physics as well as classification of major sources of uncertainty. The latter is of great importance in the context of data assimilation where the control vector should be properly defined to include key factor to improve the model outputs. It was here highlighted that in order to improve water level at Marmande, the control vector should include at least the upstream discharge and the downstream friction coefficient. A perspective for this study thus stands in the implementation of an ensemble-based data assimilation algorithm to improve input discharge and friction assimilation water level observations in the system. Additionally, the cost of the ensemble integration should be reduced using the surrogate model in place of the direct hydraulic solver.

6 ACKNOWLEDGEMENTS

We would like to thank Nicole Goutal from EDF R&D's LNHE department for providing us with the case study, Pamphile Roy from CERFACS for his algorithmic developments in the BATMAN-OpenTURNS library and the EDF R&D's PRISME department for their support in using OpenTURNS. We thank all these people for the enriching discussions we were able to have on uncertainty methods for two-dimensional river hydraulics issues.

REFERENCES

- [1] G. Archer, A. Saltelli, I. Sobol, Journal of Statistical Computation and Simulation. *Sensitivity measures, anova-like techniques and the use of bootstrap.*, **58**, 99–120, 1997.
- [2] H.H. Barnes, Journal of Hydrology. *Roughness characteristics of natural channels*, **7**, 354, 1969.
- [3] M. Baudin, R. Lebrun, B. Iooss, A.L. Popelin, Handbook of Uncertainty Quantification. *OpenTURNS: An Industrial Software for Uncertainty Quantification in Simulation*, 2001–2038, 2017.
- [4] A. Besnard, N. Goutal, La Houille Blanche. *Comparaison de modèles 1D à casiers et 2D pour la modélisation hydraulique d'une plaine d'inondation - Cas de la Garonne entre Tonneins et La Réole*, 42–47, 2011.
- [5] F.G. Charlton, P.M. Brown, R.W. Benson, Hydraulics Research Station. *The hydraulic geometry of some gravel rivers in Britain*, 1978.
- [6] A. Chatterjee, Current Science. *An introduction to the proper orthogonal decomposition*, 808–817, 2000.
- [7] W.F. Coon. *Estimates of roughness coefficients for selected natural stream channels with vegetated banks in New York*, U.S. Geological Survey, 1995.
- [8] L. Cordier, M. Bergmann. *Proper Orthogonal Decomposition: an overview*, Von Karman Institute for Fluid Dynamics, 2002.
- [9] W.L. Cowan, Agricultural Engineering. *Estimating hydraulic roughness coefficients*, **7**, 473–475, 1956.

- [10] G. Damblin, Journal of simulation. *Numerical studies of space filling designs: optimization of Latin Hypercube Samples and subprojection properties*, 276–289, 2013.
- [11] M. De Lozzo, Journal de la société française de Statistique. *Surrogate modeling and multifidelity approach in computer experimentation*, **156**, 21–55, 2015.
- [12] J.C. de Saint-Venant, C. R. Acad. Sc. Paris. *Théorie du mouvement non-permanent des eaux, avec application aux crues des rivières et à l'introduction des marées dans leur lit*, **73**, 147–154, 1871.
- [13] D.L. Fread. *BREACH: an Erosion Model for Earthen Dam Failures (Model description and User Manual)*, National Oceanic and Atmospheric Administration, National Weather Service, Silver Spring, MD, 1988.
- [14] P. Gauckler, Gauthier-Villars. *Etudes Théoriques et Pratiques sur l'Écoulement et le Mouvement des Eaux*, 1867.
- [15] R. Ghanem, D. Higdon, H. Owhadi, Springer. *Handbook of Uncertainty Quantification*, 2017.
- [16] C. Goeury, T. David, R. Ata, Y. Audouin, N. Goutal, A.L. Popelin, M. Couplet, M. Baudin, R. Barate. *Uncertainty quantification on a real case with Telamac-2D*, 2015.
- [17] J.M. Hervouet, Wiley. *Hydrodynamics of Free Surface Flows: Modelling with the finite element method*, 2007.
- [18] D.M. Hicks, P.D. Mason, Institute of Water and Atmospheric Research (N.Z.). *Roughness characteristics of New Zealand rivers*, 1991.
- [19] D.R. Jones, M. Schonlau, W.J. Welch, Journal of Global Optimization. *Efficient Global Optimization of Expensive Black-Box Functions*, **13**, 455–492, 1998.
- [20] M. Lamboni, H. Monod, D. Makowski, Reliability Engineering & System Safety. *Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models*, **96**, 450–459, 2011.
- [21] M. McKay, J. Beckman, W. Conover, Technometrics. *A comparison of three methods for selecting values of input variables in the analysis of output from a computer code.*, **21**, 239–245, 1979.
- [22] J.E. Nash, J.V. Sutcliffe, Journal of Hydrology. *River flow forecasting through conceptual models part I: A discussion of principles*, **10**, 282–290, 1970.
- [23] F. Pedregosa, V. Gaël, A. Gramfort, V. Michel and B. Thirion, O. Grisel, Journal of machine learning research. *Scikit-learn: Machine learning in Python*, **12**, 2825–2830, 2011.
- [24] J.O. Ramsay, B.W. Silverman, Springer-Verlag. *Functional data analysis*, New York, 1997.
- [25] C.E. Rasmussen, C.K.I Williams, MIT Press. *Gaussian Processes for Machine Learning*, 248, Cambridge, MA, USA, 2006.

- [26] P.T. Roy, S. Ricci, R. Dupuis, R. Campet, J.C. Jouhaud, C. Fournier, The Journal of Open Source Software. *BATMAN: Statistical analysis for expensive computer codes made easy*, 2018.
- [27] P.T. Roy, N. El Moçayd, S. Ricci, J.C. Jouhaud, N. Goutal, M. De Lozzo, M. Rochoux, Stochastic Environmental Research and Risk Assessment. *Comparison of polynomial chaos and Gaussian process surrogates for uncertainty quantification and correlation estimation of spatially distributed open-channel steady flows*, **32**, 1723–1741, 2018.
- [28] A. Saltelli, Computer Physics Communications. *Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index.*, **181**, 259–270, 2010.
- [29] A. Siade, M. Putti, W. Yeh, Water resources research. *Snapshot selection for groundwater model reduction using proper orthogonal decomposition*, **46**, 2010.
- [30] I.M. Sobol, Math. Modeling Comput. Experiment. *Sensitivity estimates for nonlinear mathematical models*, **4**, 407–414, 1993.
- [31] H. Sohr, Birkhuser Basel. *The Navier-Stokes Equations*, 2001.
- [32] A. Strickler, Berna. *Beitrge zur Frage der Geschwindigkeitsformel und der Rauhigkeitszahlen fur Strme, Kanle und Geschlossene Leitungen*, 1923.
- [33] D.J. Wales, J.P.K. Doye, The Journal of Physical Chemistry A. *Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms*, **101**, 5111–5116, 1997.