Surrogate modelling and uncertainty quantification for multiscale simulation

Ye, D.

Publication date
2022

Chapter 1

Introduction

Computational modelling and simulation of real-world phenomena contributes to a better understanding of these phenomena and facilitates predicting the dynamics of the underlying systems to support decision making in sciences and engineering [1–3]. These models emulate the nature of the system from different spatial and temporal scales, and combine mechanisms from multidisciplinary perspectives. However, no matter how sophisticated the models are constructed, they are still simplifications and involve different sources of uncertainties and errors owing to the limit of knowledge and computational resources.

Uncertainty quantification (UQ) applies statistical analysis to the computational models to quantify the effect of uncertainties in initial or boundary conditions, and of other parameters of computational models on their simulated quantities of interest (QoI) [4]. Common UQ methods, such as those based on the Monte Carlo method, require a large number of simulations to provide enough data for the numerical integration of the statistical estimator. However, this can be prohibitive for computationally expensive models to achieve. One remedy to the problem is to adopt surrogate modelling. A surrogate approximates the model response at a relatively low computational cost, hence can replace the original model in the UQ experiment. To what extent the use of the surrogate model influences the estimation of the model uncertainties is an open question.

In this chapter, UQ and surrogate modelling are introduced. The research in this thesis is inspired by the need for UQ in a biomedical application, in-stent restenosis model. This chapter also shortly describes this application.

1.1 Uncertainty quantification

Uncertainty quantification studies all sources of errors and uncertainties involved in numerical simulations [3]. Generally, the uncertainties can be categorised into aleatoric uncertainties and epistemic uncertainties. The former describes the stochasticity of the model, which cannot be reduced by additional physical or experimental knowledge. Epistemic uncertainty refers to the systemic deficiency of modelling due to the lack of knowledge, such as the assumptions of the model or inadequacies of parameters [4].

There are two types of UQ problems, forward uncertainty propagation and inverse UQ [4]. Forward uncertainty propagation is to investigate the variation in the QoI of the model subject to the input uncertainties [5–7]. The inverse UQ calibrates inputs based on the measured or known data with Bayesian techniques [8, 9].

The most common UQ method for the forward propagation problem is a sampling-based algorithm, for example, the black-box Monte-Carlo method [10–13]. Sampling-based methods compute the statistics of QoI based on a collection of the responses of a model with input samples generated from uncertain input distributions. Such black-box methods can be applied to any model regardless of the nonlinearity and complexity of the model owing to the non-intrusive properties of the methods. However, the sampling-based methods are computationally expensive due to the slow convergence rates. A large number of evaluations of the model responses are required for decent estimates of the uncertainties. Various sampling techniques have been developed to reduce the number of samples needed for the evaluation, such as stratified sampling [14–16] and quasi-random sampling [17, 18]. However, for models with high computational complexity, such as large-scale applications [19] or multiscale simulations [20], it is still impractical to implement even with advanced sampling-based methods. As a result, surrogate modelling is introduced as a representation for

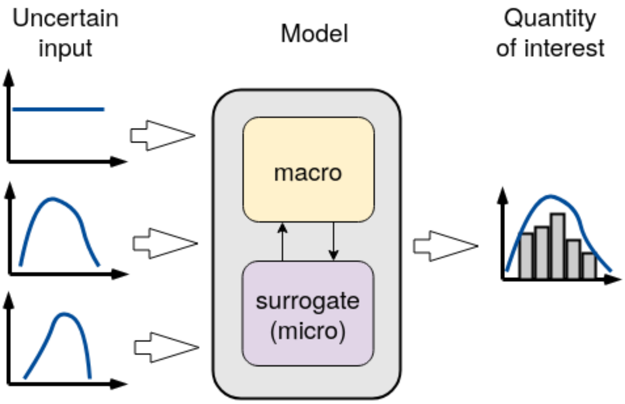

Figure 1.1: A schematic diagram of semi-intrusive UQ based on Monte Carlo method with a surrogate model for micro submodel.

efficient evaluations in the UQ.

As opposed to the non-intrusive UQ methods, intrusive UQ methods require reformulating the governing equation of the system based on polynomial chaos (PC) and Galerkin projection [21–23]. However the properties of intrusive methods typically limit their application to linear problems. The basis and the order of expansion need to be chosen such that they can present not only the output QoI but also all the intermediate variables [21]. This means that it will be non-trivial and computationally expensive to compute the solution if highly nonlinear terms are involved in the governing equation [24, 25].

Apart from intrusive and non-intrusive UQ methods, a set of semi-intrusive UQ methods (SIUQ) has been proposed for multiscale simulations [26] and demonstrated their effectiveness for several multiscale UQ scenarios [27, 28]. Multiscale simulations couple mathematical models of relevant processes on different spatial or temporal scales together using suitable scale bridging methods. The term "semi-intrusive" refers to additional interventions into the code of the model compared to non-intrusive approaches: one "opens up" the black box and considers the coupling structure of the multiscale model while the embedded single scale models are still viewed as black boxes. Usually the output of a multiscale model is derived from a macroscale submodel, which in turn is implicitly determined by microscale dynamics. One of the approaches from [26] relies on performing a Monte Carlo UQ on the macroscale submodel, while replacing the most costly microscale submodel by a surrogate model. Replacing the expensive part of a model with a relatively cheap surrogate can often significantly improve computational efficiency, while reserve part of the physics of the model for analysis. Figure 1.1 shows a schematic diagram of a semi-intrusive UQ based on Monte Carlo method with a surrogate model. The effectiveness of the semi-intrusive methods has been demonstrated through the applications shown in [27, 28].

Sensitivity analysis is another important part of UQ. It measures and quantifies the relative contribution of each uncertain parameter or input to the variations of QoIs. The result of sensitivity analysis can be applied to simplify the model by fixing the insensitive inputs, or specify important regime where uncertain parameters have the most significant impact [4]. The methods of sensitivity analysis can be categorised into two types, local and global. Local sensitivity analysis is typically based on the local derivative of QoIs with respect to uncertain parameters, while the global methods measure how uncertainty in the output can be apportioned to uncertain inputs [4]. The variance-based method, also known as the Sobol method [29] is one of the most popular global sensitivity methods. It assumes that the latent function between uncertain inputs and QoIs can be decomposed into a combination of functions of individual uncertain inputs and their higher-order interactions. The variance of each corresponding part is computed and compared to the total variance, which leads to the Sobol indices. The computational efficiency of the method was later further improved by Saltelli [30, 31], yielding Sobol method computationally tractable.

1.2 Surrogate modelling

Surrogate modelling, also known as metamodelling or emulator, aims to approximate the behaviour of the system and predict high-fidelity model response at a relatively low computational cost. The construction of a surrogate model can be categorised into three types: data-fit methods, projection-based methods and simplified models [32].

The data-fit methods learn latent functions between inputs and outputs of a limited amount of available data, which is also termed regression or interpolation-based methods [4]. Various methods can be applied to perform the data-fitting under different circumstances. For example, linear regression provides a simple but interpretable prediction of linear behaviour of data [33, 34], while polynomial regression [35] and radial basis functions interpolation [36] are capable of performing nonlinear regression by presuming smoothness of the latent function with the order of the polynomial and chosen radial basis function, respectively. Apart from these point estimation methods, Bayes' theorem provides a probabilistic framework to integrate the information of a prior, and to infer the posterior distribution of the output instead of single point estimation. Bayesian inference methods, such as Bayesian linear regression [37], and Bayesian neural network [38], allow one to build confidence and reliance for the prediction [39]. Stochastic methods, such as polynomial chaos expansion [40] and stochastic collocation [41], construct the spectral expansion of random variable and directly derive the moments for QoI. Owing to the finite-dimensional representation (bases) used in the methods, they are sometimes also viewed as projection-based methods.

Gaussian process (GP) regression is one of the most commonly used data-fit methods owing to its non-parametric and Bayesian inference nature [42–44]. It was first proposed by Krige for geostatistics [45], and later extensively studied and extended to solve the regression problem under different scenarios, such as multi-task/multi-output Gaussian process for vector-valued functions [46], heteroscedastic Gaussian process for input dependent noise scenarios [47–49], sparse Gaussian process with inducing inputs for efficient training of large datasets [50, 51] or deep Gaussian process with a hierarchical structure to capture more complex processes [52]. In addition, the predictive distribution of GP provides active criteria for adaptive sampling which allow one to iteratively choose more informative training samples rather than using a one-shot sampling strategy [53]. The adaptive sampling techniques can further reduce the computational cost needed for sample generation during surrogate model construction, especially relevant for high-dimensional problems.

Another state-of-art data-fit methods are neural networks [54] which is also known as deep learning, and has gained significant attention in computational science and engineering [55–58]. A typical feed-forward neural network consists of multiple layers and the output of each layer is a function of the linear combinations of its inputs. The neural network model learns the data by minimising the error (cost function) between its prediction and desired output. It has demonstrated strong ability to learn and predict complex nonlinear dynamics in the data [59–61]. Other forms of neural networks have been proposed to solve different problems. For example, convolutional neural networks (CNNs) have shown their advantages in handling field prediction [61, 62] owing to its efficient convolution operation, and recurrent neural networks can be applied to learn time-series data by taking the information of the previous time step into account [63]. In addition, neural network models can be applied as an alternative to solve differential equations [64, 65], or even solve ill-posed problems by assimilating additional data [66].

The projection-based methods of surrogate modelling identify a low dimensional subspace that is constructed to retain the essential patterns of the system input-output mapping [32]. The numerical solutions, such as velocity and pressure fields of fluid dynamics, are generally correlated over time and space. Computational cost can be significantly reduced if the hidden low-dimensional pattern or basis could be found. To construct the reduced-order basis, an ensemble of numerical solutions (snapshots) based on independent variable values is required. Such an ensemble of data is also known as snapshot set, through which a compressed description of the data can be extracted. One of the widely applied reduced-order methods is Proper Orthogonal Decomposition (POD) [67]. The POD can be realized using principal component analysis (PCA), or the singular value decomposition (SVD). The projection coefficients can be computed either in an intrusive manner by manipulating the governing equations with Galerkin methods [68, 69] or in a non-intrusive manner by formulating it as an interpolation problem [60, 70]. The intrusive methods preserve the physics behind the model and require to solve reduced order system for prediction. However it may still suffer from the high computational cost when computing the projection of the operators of governing equations from a high dimensional full order model [60]. The non-intrusive methods, in contrast, are entirely data-driven and assume that the projection coefficients change continuously over space, time or parameters. The interpolator is trained based on the existing data and subsequently applied for the prediction.

Dynamic mode decomposition is another projection-based method for the temporal decomposition of computational fluid dynamics [71, 72]. Empirical interpolation method [73] and discrete empirical interpolation method [74] were proposed to recover an affine expansion for the nonlinear problem. Operator inference methods can directly approximate the reduced-order differential operator from data without knowing full-order operators [75, 76].

Surrogate modelling based on simplified models can be an approximation on a coarse grid [77], or a simplification by neglecting non-linear terms [78], which could e.g. be the case for low Reynolds number flow, simplifying the Navier-Stokes equations to a linear Stokes equation. However, such simplification usually requires in-depth knowledge of the original model to ensure the validity of the model.

1.3 In-stent restenosis model

The research in this thesis is inspired by the need for UQ in a biomedical application, in-stent restenosis model. The in-stent retenosis process involves multiple uncertain physiological and mechanical parameters, such as re-endothelialization rate, blood flow velocity and stent deployment depth. UQ and sensitivity analysis therefore are applied to quantify and study the contributions of these uncertain parameters to the QoI, neointima growth or cross-sectional area of the lumen. Surrogate modelling is required owing to the high computational cost of the model evaluation, especially the three-dimensional model. Here we briefly introduce in-stent restenosis process and the in-stent restenosis model.

Coronary heart disease is mainly caused by the accumulation and development of atherosclerotic plaques, which narrow the vessel lumen and reduce the flow of blood. It can cause ischaemia or further evolve into a myocardial

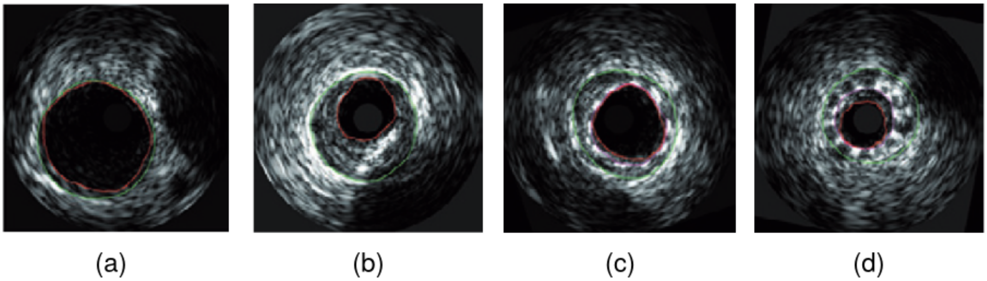

Figure 1.2: Intravascular ultrasound images of the stenosis and restenosis from [81]. a) normal coronary artery, b) coronary stenosis, c) after percutaneous coronary intervention (stenting), d) restenosis.

infarction. The most common treatment is percutaneous coronary intervention with stent deployment [79, 80]. However, in addition to displacing the plaque from the lumen and restoring the blood flow, the balloon dilation and stent placement also denude the endothelium layer and damage the vessel wall. The damage then triggers smooth muscle cell (SMC) activation, proliferation and migration and extracellular matrix formation, as well as other processes, e.g. inflammation and platelet aggregation [81, 82]. This leads to the growth of neointima, which is newly formed tissue composed mainly of smooth muscle cells and extracellular matrix, in the vessel lumen. The excessive growth of neointima can result in a renarrowing of the vessel, a condition known as in-stent restenosis (ISR). Figure 1.2 shows a series of intravascular ultrasound images of normal coronary artery, stenosis, and restenosis.

To study the mechanism of restenosis, a multiscale model for ISR was proposed [83] and a first two-dimensional version of that model (named ISR2D) was developed and studied in detail [84–86]. The model consists of three submodels: an initial condition (IC) model, an agent-based SMC model and a blood flow (BF) model. The IC model simulates balloon expansion and stent deployment and provides the input configuration for the other two models. The agent-based SMC model simulates the biological and mechanical states of each cell of the vessel, while the BF model provides the haemodynamics information as a function of the current vessel lumen shape. For every time step of the SMC model, the BF simulation is run to convergence for the current geometry, and the resulting wall shear stress values are sent back to the SMC model, which uses those in the model of nitric oxide production by endothelial cells that in turn regulates the SMCs.

Sufficiently high wall shear stress (WSS) at the arterial wall triggers endothelium to produce nitric oxide, which in turn inhibits the growth of the SMCs if the WSS exceeds a threshold value. Blood flow thus affects SMC growth, but in turn is also affected by it, as the proliferating SMCs change the geometry of the artery. The main output of the model is the cross-sectional area of the lumen as a function of time after stenting. A clinically recognised ISR occurs if more than 50% of the original cross-sectional area of the artery is covered by the neointima [87].

The ISR2D model has been applied to investigate the effect of functional endothelium regeneration and the impact of stent deployment and design on restenosis [6, 28, 84, 85]. Most recently, the effects of local blood flow dynamics with scenarios of adaptive and non-adaptive coronary vasculature on restenosis were studied based on the ISR2D model [88]. The two-dimensional model is, however, a simplification of the actual pathology. Therefore, a more comprehensive three-dimensional model (named ISR3D) was developed and compared with in vivo experimental scenarios [89, 90]. More detailed descriptions of ISR2D and ISR3D can be found in Chapter 2 and Chapter 3 respectively.

1.4 Thesis outline

This PhD thesis focuses on the non-intrusive and semi-intrusive UQ analysis of the biomedical ISR model, with surrogate modelling techniques. We mainly focus on the forward propagation problems and apply various surrogate modelling techniques, such as GP regression and CNN to reduce the computational cost of the expensive multiscale simulation and the large number of evaluations required in the UQ.

Chapter two presents uncertainty estimations of the ISR2D model with both non-intrusive and semi-intrusive methods with surrogate modelling. A surrogate model based on GP regression for non-intrusive UQ takes the whole model as a black-box and directly maps the three uncertain inputs to the quantity of interest, the neointimal area. In the semi-intrusive UQ, the most expensive submodel is replaced with a surrogate model. We developed a surrogate model for the blood flow simulation using CNN. The results on uncertainty propagation with non-intrusive and semi-intrusive metamodelling methods allow us to draw conclusions on the advantages and limitations of these methods.

A UQ of ISR3D with four uncertain parameters, including endothelium regeneration time, the threshold strain for smooth muscle cells bond breaking, blood flow velocity and the percentage of fenestration in the internal elastic lamina, is presented in Chapter three. Two quantities of interest were studied, namely the average cross-sectional area and the maximum relative area loss in a vessel. Owing to the high computational cost required for UQ, a surrogate model, based on Gaussian process regression with proper orthogonal decomposition, was developed and subsequently used for model response evaluation in the UQ. A detailed analysis of the uncertainty propagation and sensitivity analysis is presented.

In chapter four, a data-driven surrogate model for blood flow simulations in unparameterised vessels is presented. The surrogate model is based on a non-intrusive reduced-order method and surface registration. The surface registration is applied to parameterise the shapes and offer a mapping between the reference domain and target domain. With the coordinate mapping, all the evaluations of FOM are performed on a reference domain which ensures the spatial compatibility of snapshots. The non-intrusive reduced order model is subsequently constructed using POD and the RBF interpolator is trained for predicting the reduced coefficients of ROM based on reduced coefficients of geometric parameters of the shape. Two examples of blood flowing through a stenosis and a bifurcation are presented and analysed.

Chapter five introduces a series of UQ patterns (UQPs) for efficient UQ of multiscale models, and categorises them by the level of intrusiveness and optimization method. These UQPs provide the basic building blocks to create tailored UQ for multiscale models. We show how these patterns can be implemented in multiscale models using the formalism of the multiscale modelling and simulation framework (MMSF) and corresponding coupling toolkit, MUSCLE3.

The results of the work presented in the thesis are summarised and discussed in Chapter six.