Environmental Modelling & Software 72 (2015) 77–91

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Employing statistical model emulation as a surrogate for CFD

P. Moonen [a, *], J. Allegrini [b, c]

[a] UPPA, Université de Pau et Des Pays de l'Adour, Pau, France
[b] Empa, Swiss Federal Laboratories for Materials Science and Technology, Dübendorf, Switzerland
[c] ETH, Swiss Federal Institute of Technology Zürich, Zürich, Switzerland

ARTICLE INFO

Article history:
Received 7 January 2015
Received in revised form
24 June 2015
Accepted 24 June 2015
Available online xxx

Keywords:
Bayesian statistics
Gaussian process
Uncertainty quantification
Emulation

ABSTRACT

This work focuses on substituting a computationally expensive simulator by a cheap emulator to enable studying applications where running the simulator is prohibitively expensive. The procedure consists of two steps. In a first step, the emulator is calibrated to closely mimic the simulator response for a number of pre-defined cases. In a second step the calibrated emulator is used as surrogate for the simulator in the otherwise prohibitively expensive application. An appealing feature of the proposed framework contrary to other approaches is that the uncertainty on the emulator prediction can be determined. While the proposed framework is applicable in virtually all areas of natural sciences, we discuss the approach and evaluate its performance based on a typical example in the realm of computational wind engineering, namely the determination of the wind field in an urban area.

1. Introduction

In virtually all areas of natural sciences, we encounter a rapid increase in model sophistication, directed at predicting the behavior of progressively larger and more complex systems. Prime examples related to fluid dynamics are weather forecasting, climate change assessment and reservoir management. State-of-the-art models or simulators thus have a tremendous potential, yet this potential cannot always be fully exploited due to the associated computational cost. Attempts to reduce the computational cost of the simulator by reducing the spatial or temporal resolution, the spatial or temporal extent of the study, or the involved physics lead inevitably to concessions in terms of accuracy or representativeness. A statistical emulator can provide results of similar accuracy as the original simulator, yet at a much lower computational cost, and provides uncertainty estimates on these predictions.

Essentially, an emulator is a cheap model, which mimics the output of a more demanding model. Statistical emulators, also known as surrogate models or meta-models, date back to the work on Design and Analysis of Computer Experiments in the 1980s (O'Hagan, 2006). The central idea is to build a simple statistical approximation of the simulator, which maps the simulator inputs on values which are sufficiently close to the simulator outputs. Provided that the emulator is simpler than the original simulator – and thus faster – the emulator can be used in cases where running the simulator would be unfeasible. Moreover, by rooting the emulator on a statistical basis, its output is a probability distribution. This distribution approximates the distribution obtained by a Monte-Carlo analysis with the original simulator. Obtaining the latter is nevertheless often prohibitively expensive.

There are several ways to construct an emulator. The principal tool in the statistical community is based on Gaussian Processes, while the applied mathematics community relies on Polynomial Chaos expansions. As indicated by O'Hagan (2013), Gaussian Processes offer advantages over Polynomial Chaos expansions in terms of efficiency and flexibility, which is the main reason why we will employ the former approach in this paper. The main limitation of Gaussian Processes in their standard form is related to their (in) ability to handle large datasets. Among the large number of workarounds which have been proposed (e.g. Csató and Opper, 2002; Rougier et al., 2009a; Snelson, 2007 and the references cited herein), we opt for an approach based on dimensional reduction (Higdon et al., 2008) which allows for highly multivariate output while maintaining computational tractability. We remark that Gaussian Processes are closely related to the limit case of large neural networks, and that the method is equivalent to kriging, a technique popular in geostatics, and Kalman filters, popular in speech analysis (MacKay, 1998).

Statistical emulators are well known and extensively used in many fields, including porous media (e.g. Lee et al., 2002; Li and

* Corresponding author.
E-mail address: peter.moonen@univ-pau.fr (P. Moonen).

http://dx.doi.org/10.1016/j.envsoft.2015.06.007
1364-8152/

Zhang, 2007; Rajabi et al., 2015), water resources (e.g. Galelli and Soncini-Sessa, 2010; Razavi et al., 2012; Laloy et al., 2013; Wang et al., 2014; Tsoukalas and Makropoulos, 2015), environmental engineering (Ordóñez et al., 2012; Parry et al., 2013; Margvelashvili et al., 2013; Petropoulos et al., 2013), petroleum engineering (e.g. Hegstad and Omre, 2001; Craig et al., 2001), atmospheric dispersion (e.g. Politis and Robertson, 2004; Konda et al., 2010; Reggente et al., 2014) and climatology (Rougier et al., 2009b; Qin et al., 2013; Castruccio et al., 2014; Plouffe et al., 2015). In contrast, statistical emulation is relatively unknown to the computational wind engineering (CWE) community. The few published studies do however illustrate the wide range of potential applications. Bayesian calibration was for instance used to assess the impact of the model constants of the k-ε turbulence model on the flow in a street canyon (Guillas et al., 2014) and on flat-plate boundary-layers (Edeling et al., 2014). Polynomial Chaos Expansions were used to model uncertainties in CFD calculations (Knio and Le Maître, 2006; Xiu and Karniadakis, 2003). Neural networks were successfully employed to predict speed-up ratios for wind speeds over different topographies including hills, valleys and escarpments (Bitsuamlak et al., 2007). On a larger scale, Tsegas et al. (2011) demonstrated that meta-modeling enables a two-way coupling between meso-scale and microscale Computational Fluid Dynamics (CFD) simulations. Also coupling between different types of models can be established. Yi and Malkawi (2011) for instance used a neural network, trained based on CFD simulations, to enhance building energy simulations by incorporating site-specific wind conditions. Goethals et al. (2012) investigated based on surrogate modeling the sensitivity of night cooling performance to room and system design. Finally, Tagade et al. (2013) conducted multizone-CFD simulations to train a Gaussian process emulator, and used the emulator to rapidly localize and characterize multiple sources after contaminant detection by sensors.

From the preceding literature overview, it is clear that the concept "model emulation" has enabled a wide scope of otherwise unfeasible studies, ranging from sensitivity analysis and uncertainty assessment, over multi-scale modeling and model coupling, up to system design and inverse identification. Applications in other fields suggest that experimental model validation (e.g. Bayarri et al., 2007), the identification of systematic model bias (e.g. Williams et al., 2006; Dietzel and Reichert, 2012; Del Giudice et al., 2015) and objective selection of the "best" model (e.g. Dettmer et al., 2010; Del Giudice et al., 2015) can be added to this list. The large potential of Bayesian methods for the CWE community is however contrasted by their limited application. In part, this can be attributed to the fact that the sizeable literature on the topic is published in specialist statistical journals. This paper aims to introduce model emulation to the CWE community at a general level, and to provide the entry points to the more specialized literature for the interested reader. To avoid abstraction, the paper is structured around a typical CFD example, namely the determination of the wind field in an urban area. Section 2 describes the classical CFD approach, starting from deciding on the size of the computational domain and assigning the boundary conditions, over finding a suitable discretization, up to the selection of the model and the algorithms to be used. In Section 3, the same case is treated from the point of view model emulation. The different steps in the process are discussed, choices are proposed and motivated, and alternatives are suggested. The performance of the approach is assessed in several steps, ranging from a qualitative comparison of simulated and emulated wind fields, over statistical data analysis by means of scatter plots and validation metrics (Section 4). In Section 5, we analyze how the performance of the emulator approach depends on the choices made when building the emulator (i.e. the choices outlined in Section 3) and we demonstrate that these choices were not ad hoc. Section 6 gives a glance at the type of applications which can be tackled using emulators. A discussion and conclusions round up the paper.

2. Simulator approach

This paper aims at demonstrating the large potential of model emulation for the fluid dynamics community. Rather than a theoretical treatment, we opt to introduce the technique based on a typical application in the field of wind engineering. More precisely, we focus on the determination of the local wind conditions in an urban area as function of the wind conditions at a nearby meteorological station. Possible applications are e.g. assessment of the local wind comfort, determination of the wind loads on a structure, or estimation of the natural ventilation potential of a building. In the current section, we discuss how such questions can be adequately answered using a simulator approach, i.e. by means of CFD. The subsequent section then discusses the same problem statement using an emulator approach and it will be shown that this approach can significantly reduce the analysis time in case many simulator runs are required.

2.1. Urban configuration

As an illustrative example we focus on the idealized urban area depicted in Fig. 1. The urban area consists of a number of stand-alone cubical buildings with height $H = 10$ m, regularly distributed around a short street canyon. In view of wind comfort assessment, we could be interested in the wind conditions at 2 m height as function of the wind speed and direction at a nearby meteorological station. Although the wind direction can vary continuously, it is common to divide the full range of wind directions into 12 bins of 30° and to solely perform a simulation for the center value of each of these bins. Making use of the double symmetry of the urban layout, the number of directions-to-be-considered can be further reduced to 4 for the case at hand (0, 30, 60 and 90°). The wind speed at the location of interest (U) is strongly related to the wind speed at a nearby meteorological station (U_{pot}). The ratio of both is termed the total wind amplification factor or normalized wind speed γ:

$$\gamma = \frac{U}{U_{pot}} \tag{1}$$

The amplification factor can be considered independent from the actual wind speed, as long as Reynolds–independence holds, i.e. when the wind speed is sufficiently high and other effects, such as thermal convection, are comparatively small.

2.2. Computational domain and boundary conditions

We opt for a hexahedral computational domain. Governing best practice guidelines for the use of CFD in wind engineering specify that the distance between the building group ($7H \times 7H$) and the inlet surface should be at least $5H$, with H the characteristic height of the buildings (COST 732, 2007; Tominaga et al., 2008; Franke et al., 2011). The outlet surface should be positioned at least at $10H$ downstream of the building group. As the wind will enter the domain in the quadrant 0°–90°, these guidelines have to be applied in both the 0° and the 90° direction, resulting in a square footprint of the domain of with an edge length of $5H + 7H + 10H = 22H$. In vertical direction, a dimension of $10H$ is chosen to minimize the blockage ratio and to avoid artificial speed-up.

At the both inflow surfaces of the computational domain vertical profiles of the mean horizontal wind speed (U in m/s), the turbulent