Ocean Modelling 95 (2015) 25–36

Contents lists available at ScienceDirect

Ocean Modelling

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

Pragmatic aspects of uncertainty propagation: A conceptual review

W. Carlisle Thacker[a,*], Mohamed Iskandarani[b], Rafael C. Gonçalves[b], Ashwanth Srinivasan[c], Omar M. Knio[d,e]

[a] Independent Scholar
[b] University of Miami, Miami, FL, United States
[c] Tendral Corporation, Miami, FL, United States
[d] Duke University, Durham, NC, United States
[e] King Abdullah University of Science and Technology, Saudi Arabia

ARTICLE INFO

Article history:
Received 7 October 2014
Revised 28 July 2015
Accepted 3 September 2015
Available online 11 September 2015

Keywords:
Polynomial chaos
Gaussian process
Optimal interpolation
Ocean modeling
Error propagation
Uncertainty quantification

ABSTRACT

When quantifying the uncertainty of the response of a computationally costly oceanographic or meteorological model stemming from the uncertainty of its inputs, practicality demands getting the most information using the fewest simulations. It is widely recognized that, by interpolating the results of a small number of simulations, results of additional simulations can be inexpensively approximated to provide a useful estimate of the variability of the response. Even so, as computing the simulations to be interpolated remains the biggest expense, the choice of these simulations deserves attention. When making this choice, two requirement should be considered: (i) the nature of the interpolation and (ii) the available information about input uncertainty. Examples comparing polynomial interpolation and Gaussian process interpolation are presented for three different views of input uncertainty.

1. Introduction

Numerical models are important tools for understanding the behavior of the ocean and atmosphere. In particular, they can be used to make quantitative predictions of future conditions. For the value of such predictions to be assessed, they should be accompanied by quantitative information about their reliability.

While the model's formulation – the physical processes and how they are treated – has a major impact on its prediction, the impact of alternative formulations is not discussed here.[1] Instead, the model is regarded as a reliable black box and the focus is on assessing the uncertainties of the numbers it produces (its *outputs* or *responses*), which result from the uncertainties of the numbers it is given (its *inputs*). When making planning decisions, only a few of a model's many outputs are generally of interest, so estimates of accuracy can be restricted to those few outputs. Computational resources limit the number of uncertain inputs that can be treated simultaneously to a manageable few, so most of the inputs must be treated as known even though their values are by no means certain. What we discuss here should be regarded as *conditional uncertainties* – uncertainties conditioned on the values assumed to be known and on the model's formulation. We are interested in how the model transforms information about the uncertainty of selected inputs into information about key outputs.

Suppose uncertainties of the inputs are quantified as a probability density centered on their most likely values, and suppose that a large ensemble of possible inputs are sampled according to this density. Then, assuming sufficient computational resources are available, for each set of inputs from this ensemble the model can be run to provide corresponding responses, and a histogram of these responses can approximate the probability density quantifying the uncertainty of the model's response stemming from the uncertainty of its inputs. In order to carry out this Monte Carlo agenda, two practical issues must be addressed: (1) how to proceed when the model's computational requirements limit the number of runs that can be made and (2) how to proceed when there is limited information about the uncertainties of the models inputs.[2]

Before computers were available propagating uncertainty had already been recognized as being important. Wiener (1938) addressed the problem within the context of a single uncertain input and a

* Corresponding author. Tel.: +1 3056651801.
E-mail addresses: carlisle.thacker@gmail.com (W.C. Thacker), miskandarani@rsmas.miami.edu (M. Iskandarani), rgoncalves@rsmas.miami.edu (R.C. Gonçalves), a.srinivasan@tendral.com (A. Srinivasan), omar.knio@duke.edu (O.M. Knio).

[1] Webster and Sokolov (2000) discuss the role of model formulation when quantifying uncertainty of climate projections.

[2] Webster and Sokolov (2000) discuss the issue of uncertainty about input uncertainty.

http://dx.doi.org/10.1016/j.ocemod.2015.09.001

single response. Rather than addressing issue (2), he assumed the input density to be a known Gaussian. His approach was to approximate the response as a polynomial function of the input expressed as a truncated expansion in Hermite polynomials,[3] and solving for the expansion coefficients became the major computational task.[4] Because of the orthogonality of Hermite polynomials when weighted by the Gaussian density,[5] statistics such as the mean and variance of the response could be expressed as simple functions of the expansion coefficients, so there was no need to sample from the input density. Wiener's approach has spawned a great deal of recent activity and is referred to as *polynomial chaos* in the engineering literature.[6]

Although Wiener (1938) did not appear to recognize it, the most important aspect of his approach was approximating a nonlinear response using an inexpensively evaluated polynomial function, which could serve as an *emulator* or *surrogate* for the original model. Because running the emulator is much less expensive than running the original model, Monte Carlo sampling becomes affordable. This reduces issue (1) to the more manageable concern of how using emulated responses impacts our view of response uncertainty. Moreover, it also provides an answer to issue (2) – how to deal with *uncertainty about input uncertainty*: simply consider several alternative descriptions of the input uncertainty and construct histograms of the response for each. Comparing the histograms to assess the impacts of the the alternative views of input uncertainty clearly must reflect how they alter practical decisions.

Running the emulator is simple. The bulk of the computation effort is in building it, as the model must be run repeatedly to generate enough responses to interpolate. A major point of this paper is that it is best to organize computational effort in such a way that allows alternative input densities to be explored in a flexible and cost-effective manner. For each input density, sampling should reflect its more likely values, so that the response is most accurate where it is most important. When using quadrature to compute the expansion coefficients, the polynomial chaos formulation that links the input density to the orthogonality of the polynomials guarantees this sort of sampling as long as the input density is correct, but changing the input density requires redoing the sampling.[7] Our point is that it is important to be able to interpolate more flexibly so that sampling and interpolating can be incremental, so that as the view of input uncertainty changes existing samples need only be supplemented by a few more to improve accuracy for input values previously considered unlikely. Solving an algebraic system for the expansion coefficients, rather than using quadrature, provides this needed flexibility.[8]

An emulator need not be restricted to polynomial interpolation. Gaussian process interpolation has also been used for this purpose.[9] Rather than treating the response as a sum of specified functions of the random inputs, the response function itself is regarded as a *random function*. More specifically, for each input, the output has a specified mean and variance, and for each pair of inputs there is a specified covariance. It is easy to recognize this as a novel application of optimal interpolation, as it corrects a prior description of the response surface using an assumed covariance function to interpolate the information provided by model simulations. Using Gaussian processes shifts the focus from determining coefficients of a polynomial approximation to choosing appropriate mean and covariance functions. Once these choices have been made and the responses of the exploratory ensemble have been assimilated, the updated mean function serves as the approximate response function and the updated variance function provides a measure of its accuracy.

We use a single input and a single output for illustrating these ideas, as that allows the underlying issues to be discussed more clearly. To illustrate our points, we use a model that simulates the fate of oil droplets emerging from a deep underwater source and rising due to their buoyancy as they are advected by a prescribed velocity field. This model was chosen because a database of its simulations was available and new simulations could be avoided. Uncertainty in droplet size is the single uncertain input of interest and the single response is the surface concentration of oil in a region some distance from the source. The details of this model are unimportant here. What matters is that the response is a highly nonlinear function of the input. In the examples discussed below, polynomials and Gaussian processes provide alternative interpolations for the same sets of simulated responses and thus a framework in which the separate roles of the choice of input density and the choice of the interpolation method can be addressed.

These simulations had previously been used at an early stage of model development to check how the uncertainty of surface oil concentration within a restricted region depended on the uncertainty of droplet size at the spill site. These simulations are from two different quadrature ensembles, one providing a 6th-degree polynomial approximation to the response function and the other providing a 20th-degree polynomial approximation, which might be regarded as being exact over the range of inputs considered. When combined and then sub-sampled they provide the possibility of exploring different approaches to interpolation and different views of input uncertainty.

This paper takes a step-by-step approach to illustrate the above ideas incrementally. Section 2 describes the data from the database of quadrature simulations and discusses the differences in the resulting 6th- and 20th-degree polynomial approximations to the response curve, both of which indicate that the response is a highly nonlinear function of droplet size. Section 3 discusses the histograms characterizing the uncertainty in surface-concentration response, which reflect the differences between these two views of the response curve, under the assumption that the uniform probability density for droplet size on which the quadrature ensembles were based is the correct density. Section 4 illustrates how the existing 20th-degree polynomial approximation to the response can be used to assess the uncertainty of the surface oil concentration when the initial assumption of uniform probability of droplet size is replaced with alternate assumptions – *without the need of any new simulations*. Section 5 illustrates how decoupling the polynomial approximation from the input probability density can provide a more flexible choice of simulations. And Section 6 explores Gaussian process interpolation as an alternative to polynomial interpolation and illustrates its impact on the propagation of uncertainty from droplet size to surface concentration. As the focus is on practicality, methodological details are confined to appendices. The paper ends with a few concluding remarks.

2. Simulated data

The context of this study is provided by a model being constructed for simulating the effects of the Deepwater Horizon oil spill in the Gulf of Mexico. While that model has many uncertain inputs and many responses worthy of study, here we focus on one particular uncertain input and one particular response, as they had been examined in the early stages of model construction and those quadrature

[3] His use of Hermite polynomials was tied to his use of a Gaussian probability density, since Hermite polynomials are orthogonal when weighted by a Gaussian density. If he had expanded in Legendre polynomials, the method requires a uniform input density, and Laguerre polynomials require an exponential density. See, for example, Eldred et al. (2008) or Xiu (2009).

[4] See Appendix A for a brief discussion of how the coefficients might be computed.

[5] Recall its use in the context of the meridional structure of equatorial Rossby waves.

[6] The reviews of Najm (2009) and Xiu (2009) provide brief introductions to polynomial chaos and references to much of its literature.

[7] See Appendix B.

[8] See Appendix C.

[9] Rougier et al. (2009) have used Gaussian processes for characterizing responses of climate models. See also the articles by Sacks et al. (1989) and by Kennedy and O'Hagan (2000) and the excellent book by Rasmussen and Williams (2006). Appendix D provides a brief discussion of Gaussian process interpolation. A reviewer recommended the folowing articles as potentially inspiring: French (2003), Ratto et al. (2009), Yang (2011), Borgonovo et al. (2012), Castaings et al. (2012).