Discussing Issues in Simulation-Based Uncertainty Quantification. The Case of Geohazard Assessments

Ibsen Chivata Cardenas

July 25, 2022

Discussing issues in simulation-based uncertainty quantification. The case of geohazard assessments

Ibsen Chivata Cardenas

Department of Safety, Economics and Planning, University of Stavanger, Norway.
E-mail: ibsen.chivatacardenas@uis.no

By making explicit the modeling choices and assumptions made, we analyze some issues in quantifying uncertainty using geohazard models. Under the often condition of very limited data, a major problem is constraining the many parameters involved. We conclude that, despite the availability of recently developed sophistications, the quantification based on these ideal parameterized models can hardly be justified since, e.g., they will only reflect some aspects of the uncertainty involved. This calls for more insightful approaches which are yet to be developed.

Keywords: Uncertainty quantification, geohazard, parametrization.

1. Introduction

Uncertainty Quantification, UQ, helps determine how likely the responses of a system are when some quantities in the system are not known. Using models, system's responses can be calculated analytically, numerically, or by random sampling. Given the high-dimensional nature of geohazards quantities, sampling methods are frequently used because they result in a less expensive and more tractable UQ in comparison with analytical and numerical methods. When uncertainty reflects that analysts' knowledge about quantities is incomplete and focused on probabilities measuring that quantities uncertainty, we analyze some issues in UQ. We illustrate the points raised by describing critical steps in UQ, which include choices and assumptions made by analysts.

2. Quantifying uncertainty using geohazard models

Typically, a geohazard model can be described as follows. We consider a system (e.g. debris flow) with a set of input quantities $\boldsymbol{X}$ (e.g. sediment concentration, entrainment rate), whose relations to the output quantities $\boldsymbol{Y}$ (e.g. runout volume, velocity, or height of flow), can be expressed by a set of models $\boldsymbol{\text{Ɱ}}$. All $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{\text{Ɱ}}$ are *specified by analysts*. A vector $\boldsymbol{\Theta}_{\text{ɱ}}$ (including, e.g. friction, viscosity, turbulence coefficients) parameterize a model ɱ in $\boldsymbol{\text{Ɱ}}$. Parameters $\boldsymbol{\Theta}_{\text{ɱ}}$ determine specific functions among a family of potential functions modeling the system which constrain the model's responses.

Accordingly, a model ɱ can be described as a multi-output function with e.g. $\boldsymbol{Y}$ = *{runout volume, velocity, height of flow}* and we can write Eq. (1) (Based on Lu and Lermusiaux, 2021):

$$\text{ɱ}: \boldsymbol{X}_{s,t} \times \boldsymbol{\Theta}_{\text{ɱ}} \rightarrow \boldsymbol{Y}_{s,t} \tag{1}$$

$$\text{ɱ} \equiv (\boldsymbol{E}, \boldsymbol{SG}_{\text{ɱ}}, \boldsymbol{BC}_{\text{ɱ}}, \boldsymbol{IC}_{\text{ɱ}}) \tag{2}$$

where y as realizations of $\boldsymbol{Y}$ are the model responses when $\boldsymbol{X}$ take the values x at a spatial location $s \in \boldsymbol{S}$ and a specific time $t \in \boldsymbol{T}$, and parameters $\theta_{\text{ɱ}}$ in $\boldsymbol{\Theta}_{\text{ɱ}}$ are used. In Eq. (1), $\boldsymbol{X} \subset \mathbb{R}^{d_{\boldsymbol{X}}}$ is the set of input quantities, $\boldsymbol{T} \subset \mathbb{R}^{d_{\boldsymbol{T}}}$ is the time domain, $\boldsymbol{S} \subset \mathbb{R}^{d_{\boldsymbol{S}}}$ is the spatial domain, $\boldsymbol{\Theta}_{\text{ɱ}} \subset \mathbb{R}^{d_{\boldsymbol{\Theta}_{\text{ɱ}}}}$ corresponds to a parameter vector, and $\boldsymbol{Y} \subset \mathbb{R}^{d_{\boldsymbol{Y}}}$ is the set of output quantities. d = 0, 1, 2, or 3. The system is fully described if ɱ is *specified* in terms of a set of equations $\boldsymbol{E}$ (e.g. conservation equations), the spatial domain geometry $\boldsymbol{SG}_{\text{ɱ}}$ (e.g. extension), the boundary conditions $\boldsymbol{BC}_{\text{ɱ}}$ (e.g. downstream flow), and the initial conditions $\boldsymbol{IC}_{\text{ɱ}}$ (e.g. flow at $t = t_0$), see Eq. (2).

In the sampling approach to UQ, *specified* distributions of input quantities are sampled many times and the distribution of the produced outputs can be calculated. When multiple models are considered, the output probability distribution for a model ɱ can be denoted as $f(y|x,\theta_{\text{ɱ}},\text{ɱ})$ for realizations y, x, $\theta_{\text{ɱ}}$, ɱ of $\boldsymbol{Y}$, $\boldsymbol{X}$, $\boldsymbol{\Theta}_{\text{ɱ}}$, and $\boldsymbol{\text{Ɱ}}$ respectively. When observations about $\boldsymbol{Y}$, $\boldsymbol{X}$ exist i.e. $\mathcal{d} = \{\hat{\boldsymbol{Y}} = \hat{y}, \hat{\boldsymbol{X}} = \hat{x}\}$ as part of data available $\boldsymbol{\mathcal{D}}$, we can revise/update $f(y|x,\theta_{\text{ɱ}},\text{ɱ})$ to obtain $f(y|x,\theta_{\text{ɱ}},\mathcal{d},\text{ɱ})$.

If the parameters $\boldsymbol{\Theta}_{ɱ}$ are poorly known, a distribution $\pi(\theta_{ɱ}|ɱ)$ that weighs each parameter value $\theta_{ɱ}$, can be *specified* for each $ɱ$. If some data is available e.g., in the form of corresponding measurements $d=\{\hat{y},\hat{x}\}$, $d \in \boldsymbol{\mathcal{D}}$, such distribution can, therefore, be constrained and described as $\pi(\theta_{ɱ}|d,ɱ)$. However, Betz (2017) suggested that the parameter system is fully described by a parameter vector $\boldsymbol{\Theta}=\{\boldsymbol{\Theta}_{ɱ},\boldsymbol{\Theta}_{X},\boldsymbol{\Theta}_{\varepsilon},\boldsymbol{\Theta}_{o}\}$, in which, $\boldsymbol{\Theta}_{ɱ}$ relates to parameters of the model $ɱ$, $\boldsymbol{\Theta}_{X}$ are parameters linked to the input $\boldsymbol{X}$, $\boldsymbol{\Theta}_{\varepsilon}$ is the vector of the prediction-error model, $\varepsilon=y-y^*$, (y^* are future non-observed system's responses) and $\boldsymbol{\Theta}_{o}$ is the vector associated with measurement errors. More explicitly, to compute an overall output probability distribution for $ɱ$, in a rather general description, we may have the following. $f(y|x,\theta_{ɱ},ɱ)$ which expresses the probability of the output, $f(x|\theta_{X},ɱ)$ is a distribution reflecting the uncertainty of the input, $f(y^*|y,\theta_{\varepsilon},ɱ)$ which is linked to model output error, $f(x|\hat{x},\theta_{o},ɱ)$ for measurement errors, plus $\pi(\theta|ɱ)$ reflecting overall parameters plausibility. In an attempt to exhaustively quantify uncertainty, different models $ɱ$ are considered, and the overall output probability distribution, for models *assumed* mutually independent, is computed as (Betz, 2017):

$$f(y|x,\boldsymbol{\Theta},\boldsymbol{\mathcal{D}},\boldsymbol{\mathcal{M}}) = \sum f(y|x,\theta,d,ɱ)\,\omega(ɱ|\boldsymbol{\mathcal{D}},\boldsymbol{\mathcal{M}}) \quad (3)$$

$$f(y|x,\theta,d,ɱ) = \int f(y|x,\theta,ɱ)\,\pi(\theta|d,ɱ)\,d\theta \quad (4)$$

In Eq. (3), $\omega(ɱ|\boldsymbol{\mathcal{D}},\boldsymbol{\mathcal{M}})$ is another distribution weighing each model $ɱ$ in $\boldsymbol{\mathcal{M}}$.

Back-analysis helps in constraining elements θ, yet in a limited fashion given the considerable number of parameters. See in Eq. (3-4). Back-analysis is further challenged by the potential dependency among θ or $ɱ$ and between θ and $\boldsymbol{SG}_{ɱ}$, $\boldsymbol{BC}_{ɱ}$, $\boldsymbol{IC}_{ɱ}$. With considerable data and using Bayesian networks encoding *assumptions* (e.g., independence, linear relationships, normality) we may specify a joint distribution $f(x,\theta,ɱ)$ to be sampled (e.g. Albert, Callies, and von Toussaint, 2022). Yet a more conventional case is that the majority of x or θ can only be *specified* using the maximum entropy principle subject to the system's physical constraints and based on analysts' *credence*, as well as *assuming* mutual independence.

Options to address the parametrization problem are surrogate models, parameters reduction, and model learning. Remarkably, versions of the latter option do not need any prior information about $\boldsymbol{E}$ but require local verification of conservation laws in d (Lu and Lermusiaux, 2021). Yet such approaches are still data demanding. More importantly, we note that, like many satisfactorily *validated* models, the *credibility* of not observed surrogate model outputs can always be questioned, since, e.g., records may miss crucial events or the models fail to reproduce outputs caused by recorded abrupt changes (e.g., extreme velocities turbidity currents).

3. Discussion and conclusion

By making explicit the modeling choices and assumptions made in UQ, we show that in geohazard assessments, it is very difficult to meet data requirements for *ideal* parameterization of models. In the reported analysis, we also note that, if models can be fully parameterized, these could be accurate at reproducing data from past events but may turn out to be inadequate for missed input data or quantities or unobserved outputs. Under these circumstances, such models can hardly be *justified*. As presented, model outputs are also not only conditional on the *choice* of $\boldsymbol{\Theta}$ (including priors, likelihood functions and linked hyperparameters) but on $\boldsymbol{SG}_{ɱ}$, $\boldsymbol{BC}_{ɱ}$, $\boldsymbol{IC}_{ɱ}$, as well as on the complete *specification* of $\boldsymbol{X}$, and the many *assumptions made* by analysts. Unfortunately, only some elements in $\boldsymbol{\Theta}$, $\boldsymbol{SG}_{ɱ}$, $\boldsymbol{BC}_{ɱ}$, $\boldsymbol{IC}_{ɱ}$, or $\boldsymbol{X}$ can be constrained by data, thus, considering all the above, the quantification will only reflect some aspects of the uncertainty involved. This calls for more thoughtful approaches which are yet to be developed.

Acknowledgement
This research is funded by ARCEx partners and the Research Council of Norway (grant number 228107).

References

Albert, C.G., U., Callies, and U. von Toussaint (2022). A Bayesian approach to the estimation of parameters and their interdependencies in environmental modeling. *Entropy 24(2)*, 231-262.

Betz, W. (2017). *Bayesian inference of engineering models*. Technische Universität München.

Lu, P., and P.F. Lermusiaux (2021). Bayesian learning of stochastic dynamical models. *Physica D: Nonlinear Phenomena 427*, 133003.