Addressing Uncertainty in MultiSector Dynamics Research

Release v0.1.0

Patrick M. Reed, ...

Jun 02, 2021

CONTENTS

CHAPTER ONE

INTRODUCTION

This guidance text has been developed in support of the Integrated Multisector Multiscale Modeling (IM3) Science Focus Area's objective to formally integrate uncertainty into its research tasks. IM3 is focused on innovative modeling to explore how human and natural system landscapes in the United States co-evolve in response to short-term shocks and long-term influences. The project's challenging scope is to advance our ability to study the interactions between energy, water, land, and urban systems, at scales ranging from local (~1km) to the contiguous United States, while consistently addressing influences such as population change, technology change, heat waves, and drought. Uncertainty and careful model-driven scientific insights are central to IM3's key MultiSector Dynamics (MSD) science objectives shown below.

IM3 key MSD science objectives include:

Develop flexible, open-source, and integrated modeling capabilities that capture the structure, dynamic behavior, and emergent properties of the multiscale interactions within and between human and natural systems.

Use these capabilities to study the evolution, vulnerability, and resilience of interacting human and natural systems and landscapes from local to continental scales, including their responses to the compounding effects of long-term influences and short-term shocks.

Understand the implications of uncertainty in data, observations, models, and model coupling approaches for projections of human-natural system dynamics.

Addressing the objectives above poses a strong transdisciplinary challenge that heavily depends on a diversity of models and, more specifically, a consistent framing for making model-based science inferences. The term transdisciplinary science as used here formally implies a deep integration of disciplines to aid our hypothesis driven understanding of coupled human-natural systems–bridging differences in theory, hypothesis generation, modeling, and modes of inference [2]. The IM3 MSD research focus and questions require a deep integration across disciplines, where new modes of analysis can emerge that rapidly synthesize and exploit advances for making decision relevant insights that at minimum acknowledge uncertainty and more ideally promote a rigorous quantitative mapping of its effects on the generality of claimed scientific insights. More broadly, diverse scientific disciplines engaged in the science of coupled human-natural systems, ranging from natural sciences to engineering and economics, employ a diversity of numerical computer models to study and understand their underlying systems of focus. The utility of these computer models hinges on their ability to represent the underlying real systems with sufficient fidelity and enable the inference of novel insights. This is particularly challenging in the case of coupled human-natural systems where there exists a multitude of interdependent human and natural processes taking place that could potentially be represented. These processes usually translate into modeled representations that are highly complex, non-linear, and exhibit strong interactions and threshold behaviors [3, 5, 6]. Model complexity and detail have also been increasing as a result of our improving understanding of these processes, the availability of data, and the rapid growth in computing power [9]. As model complexity grows, modelers need to specify a lot more information than before: additional model inputs and relationships as more processes are represented, higher resolution data as more observations are collected, new coupling relationships and interactions as models are put together to answer multisector questions (e.g., the land-water-energy nexus). Typically, not all of this information is well known, nor is the impact of these many uncertainties on model outputs well understood. It is further especially difficult to distinguish the effects of individual as well as interacting sources of uncertainty when modeling coupled systems with multisector and multiscale dynamics [12].

Given the challenge and opportunity posed by the disciplinary diversity of IM3, we utilized a team-wide survey to allow the project's membership to provide their views on how their areas typically address uncertainty, emphasizing key literature examples and domain-specific reviews. Our synthesis of this survey information in Figure 1 summaries the team's perspectives, enabling a summary of the commonalities and differences for how different disciplinary areas are typically addressing uncertainty. Figure 1 highlights the non-trivial challenge posed by seeking to carefully consider uncertainty across an MSD focused transdisciplinary team. There are significant differences across the team's contributing disciplines in terms of the methodological approaches and tools used in the treatment of uncertainty. The horizontal axis of the figure represents a conceptual continuum of methodological approaches, ranging from deterministic (no uncertainty) modeling to the theoretical case of fully engaging in modeling all sources of uncertainty. The vertical axis of plot maps the analysis tools that are used in the disciplines' literature, spanning error-driven historical analyses to full uncertainty quantification. Given that Figure 1 is a conceptual illustration, the mapping of each discipline's boundaries is not meant to imply exactness. They encompass the scope of feedback attained in the team-wide survey responses. The color circles designate specific sources of uncertainty that could be considered. Within the mapped disciplinary approaches, the color circles distinguish those sources of uncertainty that are addressed in the bodies of literature reported by respondents. Note the complete absence of grey circles designating that at present few if any studies report results for understanding how model coupling relationships shape uncertainty. We can briefly distinguish the key terms of uncertainty quantification (UQ) and uncertainty characterization (UC). UQ refers to the formal focus on the full specification of likelihoods as well as distributional forms necessary to infer the joint probabilistic response across all modeled factors of interest [1]. Alternatively, uncertainty characterization as defined here, refers to exploratory modeling of alternative hypotheses for the co-evolutionary dynamics of influences, stressors, as well as path dependent changes in the form and function of modelled systems [8, 11]. Uncertain factors are any model component which is affected by uncertainty: inputs, resolution levels, coupling relationships, model relationships and parameters. When a model has been established as a sufficiently accurate representation of the system some of these factors may reflect elements of the real-world system that the model represents (for example, a population level parameter would reflect a sufficiently accurate representation of the population level in the system under study). As discussed in later sections, the choice of UQ or UC depends on the specific goals of studies, the availability of data, the types of uncertainties (e.g., well-characterized or deep), the complexity of underlying models as well as the computational limits. Deep uncertainty (as opposed to well-characterized) refers to situations where expert opinions consulted on a decision do not know or cannot agree on system boundaries, or the outcomes of interest and their relative importance, or the prior probability distribution for the various uncertain factors present [4, 7].

At present, there is no singular guide for confronting the computational and conceptual challenges of the multi-model, transdisciplinary workflows that characterize ambitious projects such as IM3 [10]. The primary aim of this text is to begin to address this gap and provide guidance for facing these challenges. Chapter 2 provides an overview of diagnostic modeling and the different perspectives for how we should evaluate our models. Chapter 3 the basic methods and concepts for sensitivity analysis. Chapter 4 delves into more technical applications of sensitivity analysis to support diagnostic model evaluation and exploratory modeling. Chapter 5 transitions to an overview of the key concepts and tools for UQ. Chapter 6 transitions to the use of UQ to capture risks and extremes in MSD systems. Chapter 7 provides concluding remarks across the UC and UQ topics covered in this text. The appendices of this text include a glossary of the key concepts as well as example test cases and scripts to showcase various UC and UQ related tools.

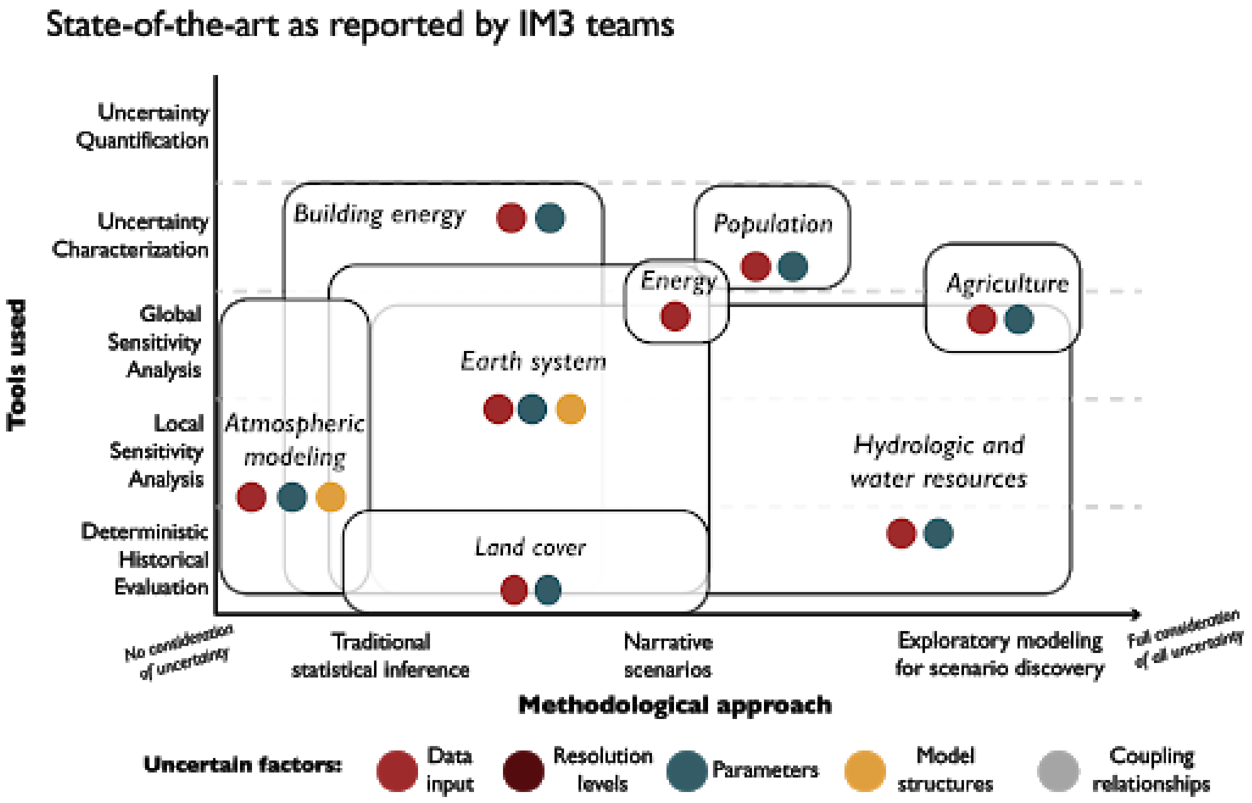
State-of-the-art as reported by IM3 teams
Tools used
Uncertainty Quantification
Uncertainty Characterization
Global Sensitivity Analysis
Local Sensitivity Analysis
Deterministic Historical Evaluation
Building energy
Population
Energy
Agriculture
Earth system
Atmospheric modeling
Hydrologic and water resources
Land cover
No consideration of uncertainty
Traditional statistical inference
Narrative scenarios
Exploratory modeling for scenario discovery
Full consideration of all uncertainty
Methodological approach
Uncertain factors:
Data input
Resolution levels
Parameters
Model structures
Coupling relationships

Fig. 1: State-of-the-art in different modeling communities, as reported in the survey distributed to IM3 teams. Deterministic Historical Evaluation: model evaluation under fully determined conditions defined using historical observations; Local Sensitivity Analysis: model evaluation performed by varying uncertain factors around specific reference values; Global Sensitivity Analysis: model evaluation performed by varying uncertain factors throughout their entire feasible value space; Uncertainty Characterization: model evaluation under alternative factor hypotheses to explore their implications for model output uncertainty; Uncertainty Quantification: representation of model output uncertainty using probability distributions; Traditional statistical inference: use of analysis results to describe deterministic or probabilistic outcomes resulting from the presence of uncertainty; Narrative scenarios: use of a limited decision-relevant number of scenarios to describe (sets of) changing system outcomes; Exploratory modeling for scenario discovery: use of large ensembles of uncertain conditions to discover decision-relevant combinations of uncertain factors

CHAPTER

TWO

DIAGNOSTIC MODELING OVERVIEW AND PERSPECTIVES

This text prescribes a formal model diagnostic approach to IM3 computational experimentation that is a deliberative and iterative combination of state-of-the-art UC and global sensitivity analysis techniques that progresses from observed history-based fidelity evaluations to forward looking resilience and vulnerability inferences (Gupta et al., 2008; Hadjimichael et al., 2020).

2.1 Overview of model diagnostics

Model diagnostics provide a rich basis for hypothesis testing, model innovation, and improved inferences when classifying what is controlling highly consequential results (e.g., vulnerability or resilience in coupled human-natural systems). Figure 2, adapted from (Saltelli et al., 2019), presents idealized illustrations of the relationship between UC and global sensitivity analysis (GSA) for two coupled simulation models. The figure illustrates how UC can be used to address how uncertainties in various modeling decisions (data inputs, parameters, model structures, coupling relationships, and elsewhere) can be sampled and simulated to yield the empirical model output distribution(s) of interest. Monte Carlo frameworks allow us to sample and propagate (or integrate) the ensemble response of the model(s) of focus. The first step of any UC analysis is the specification of the initial input distributions as illustrated in Figure 2. The second step is to perform the Monte Carlo simulations. The question can then be raised, which of the modeling assumptions in our Monte Carlo experiment are the most responsible for the resulting output uncertainty. We can answer this question using “global sensitivity analysis” (GSA) as illustrated in Figure 2. GSA can be defined as a formal Monte Carlo sampling and analysis of modeling choices (structures, parameters, inputs) to quantify their influence on direct model outputs (or output-informed metrics). UC experiments by themselves do not explain why you get a particular uncertain outcome. The pie chart shown in Figure 2 is a conceptual representation of the results of GSA to identify those factors that are most dominantly influencing results, either individually or interactively (Saltelli et al., 2008).

UC and GSA are not independent modeling analyses. As illustrated here, any GSA requires an initial UC hypothesis in the form of statistical assumptions and representations for the modeling choices of focus (structural, parametric, and data inputs). Information from these two model diagnostic tools can then be used to inform data needs for future model runs, experiments to reduce the uncertainty present, or the simplification or enhancement of the model where necessary. Together UC and GSA provide a foundation for diagnostic exploratory modeling that has a consistent focus on the assumptions, structural model forms, alternative parameterizations, and input data sets that are used to characterize the behavioral space of one or more models.

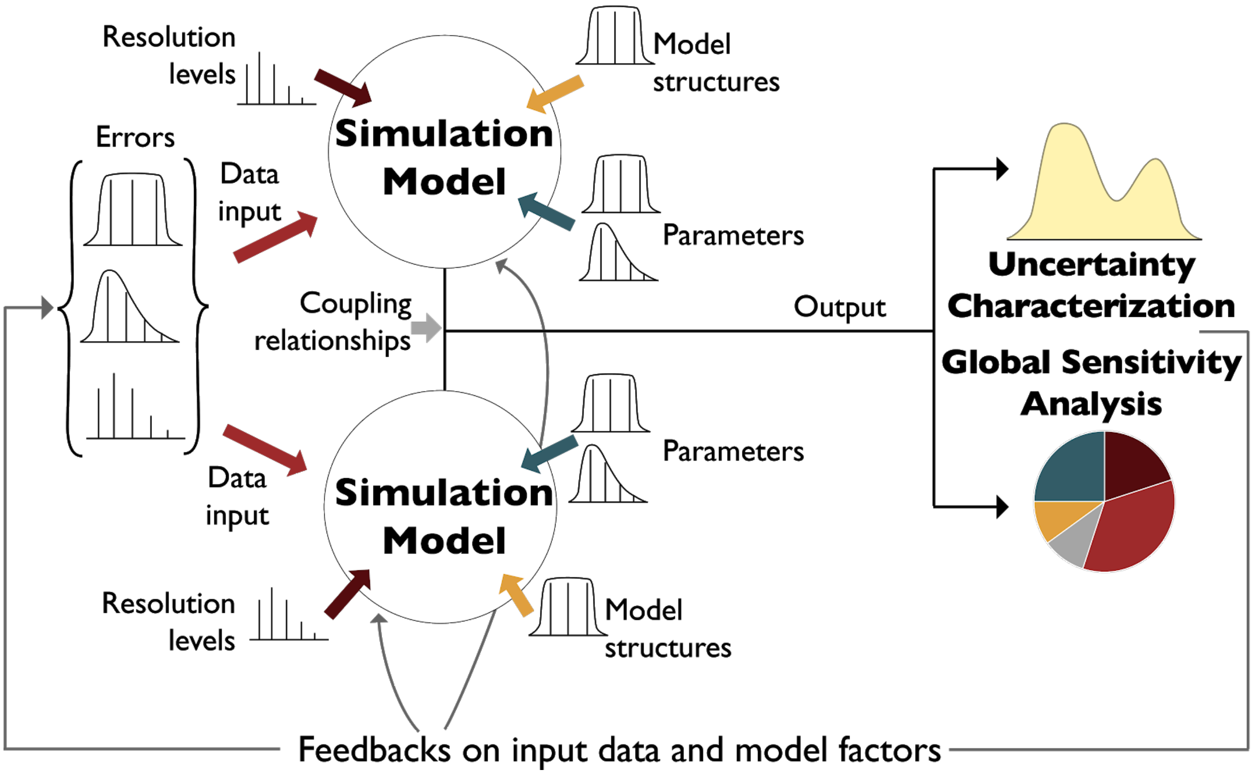

Fig. 1: Idealized uncertainty characterization and global sensitivity analysis for two coupled simulation models. Uncertainty coming from various sources (inputs, model structures, coupling relationships, and elsewhere) is propagated through the coupled model(s) to generate empirical distributions of outputs of interest (uncertainty characterization). This model output uncertainty can be decomposed to its origins, by means of sensitivity analysis. This figure has been adapted from Saltelli et al. (2019).

2.2 Perspectives on diagnostic model evaluation

When we judge or diagnose models the terms "verification and validation" are commonly used. However, their appropriateness in the context of numerical models representing complex coupled human-natural systems is questionable (Beven, 2002; Oreskes et al., 1994). The core issue relates to the fact that these systems are often not fully known or perfectly implemented when modeled. Rather, they are defined within specific system framings and boundary conditions in an evolving learning process with the goal of making continual progress towards attaining higher levels of fidelity. For example, observations used to evaluate the fidelity of parameterized processes are often measured at a finer resolution than is represented in the model and then must be scaled up for the evaluation. In other cases, numerical models may neglect or simplify system processes because the data is not available or the physical mechanisms are not fully known. If sufficient agreement between prediction and observation is not achieved, it is challenging to know whether these types of modeling choices are the cause, or if other issues, such as deficiencies in the input parameters and/or other modeling assumptions are the true cause of errors. Even if there is high agreement between prediction and observation, the model cannot necessarily be considered validated, as it is always possible that the right values were produced for the wrong reasons. For example, low error can stem from a situation where different errors in underlying assumptions or parameters cancel each other out ("compensatory errors"). Furthermore, coupled human-natural system models are often subject to "equifinality", a situation where multiple parameterized formulations can produce similar outputs or equally acceptable representations of the observed data. There is therefore no uniquely "true" or validated model, and the common practice of selecting "the best" deterministic calibration set is more of an assumption than a finding (Beven, 1993; Beven and Binley, 1992). The situation becomes even more tenuous when observational data is limited in its scope and/or quality to be insufficient to distinguish model representations or their performance differences.

These limitations on model verification undermine any purely positivist treatment of model validity: that a model should correctly and precisely represent reality to be valid. Under this perspective, closely related to empiricism, statistical tests should be used to compare the model's output with observations and only through empirical verification can a model or theory be deemed credible. A criticism to this viewpoint (besides the aforementioned challenges for model verification) is that it reduces the justification of a model to the single criterion of predictive ability and accuracy (Barlas and Carpenter, 1990). Authors have argued that this ignores the explanatory power held in models and other procedures, which can also advance scientific knowledge (Toulmin, 1977). These views gave rise to relativist perspectives of science, which instead place more value on model utility in terms of fitness for a specific purpose or inquiry, rather than representational accuracy and predictive ability (Kleindorfer et al., 1998). This viewpoint appears to be most prevalent among practitioners seeking decision relevant insights (i.e., inspire new views vs. predict future conditions). The relativist perspective argues for the use of models as heuristics that can enhance our understanding and conceptions of system behaviors or possibilities (Eker et al., 2018). In contrast, natural sciences favor a positivist perspective, emphasizing similarity between simulation and observation even in application contexts where it is clear that projections are being made for conditions that have never been observed and the system of focus will have evolved structurally beyond the model representation being employed (e.g., decadal to centennial evolution of human-natural systems).

These differences in prevalent perspectives are mirrored in how model validation is defined by the two camps: From the relativist perspective, validation is seen as a process of incremental "confidence building" in a model as a mechanism for insight (Barlas, 1996), whereas in natural sciences validation is framed as a way to classify a model as having an acceptable representation of physical reality (Oreskes et al., 1994). Even though the relativist viewpoint does not dismiss the importance of representational accuracy, it does place it within a larger process of establishing confidence through a variety of tools. These tools, not necessarily quantitative, include communicating information between practitioners and modelers, interpreting a multitude of model outputs, and contrasting preferences and viewpoints.

On the technical side of the argument, differing views on the methodology of model validation appear as early as in the 1960's. (Naylor and Finger, 1967) argue that model validation should not be limited to a single metric or test of performance (e.g., a single error metric), but should rather be extended to multiple tests that reflect different aspects of a model's structure and behavior. This and similar arguments are made in literature to this day (Beven, 2018; Gupta et al., 2012, 2008; Kumar, 2011; Nearing et al., 2020) and are primarily founded on two premises. First, that even though modelers widely recognize that their models are abstractions of the truth, they still make truth claims based on traditional performance metrics that measure the divergence of their model from observation (Nearing et al., 2020). Second, that the natural systems mimicked by the models contain many processes that exhibit significant heterogeneity at various temporal and spatial scales. This heterogeneity is lost when a single performance measure is used, as a result of the inherent loss of process information occurring when transitioning from a highly dimensional and interactive system to the dimension of a single metric (Beven, 2002). These arguments are further elaborated in section 4.1 Understanding Errors.

Multiple authors have proposed that, instead, the evaluation of several model performance signatures (characteristics) should be considered to identify model structural errors and achieve a sufficient assessment of model performance (Gupta et al., 1998). There is however a point of departure here, especially when models are used to produce inferences that can inform decisions. When agencies and practitioners use models of their systems for public decisions, those models have already met sufficient conditions for credibility (e.g., acceptable representational fidelity), but may face broader tests on their salience and legitimacy in informing negotiated decisions (Cash et al., 2003; Eker et al., 2018; White et al., 2010). This presents a new challenge to model validation, that of selecting decision-relevant performance metrics, reflective of the system's stakeholders' viewpoints, so that the most consequential uncertainties are identified and addressed (Saltelli and Funtowicz, 2014). For complex multisector models at the intersection of climatic, hydrologic, agricultural, energy, or other processes, the output space is made up of a multitude of states and variables, with very different levels of salience to the system's stakeholders and to their goals being achieved. This is further complicated when such systems are also institutionally and dynamically complex. As a result, a broader set of qualitative and quantitative performance metrics is necessary to evaluate models of such complex systems, one that embraces the plurality of value systems, agencies and perspectives present. For IM3, even though the goal is to develop better projections of future vulnerability and resilience in co-evolving human-natural systems and not to provide decision support per se, it is critical for our multisector, multiscale model evaluation processes to represent stakeholders' adaptive decision processes credibly.

As a final point, when a model is used in a projection mode, its results are also subject to additional uncertainty, as there is no guarantee that the model's functionality and predictive ability will stay the same as the baseline, where the verification and validation tests were conducted. This challenge requires an additional expansion of the scope of model evaluation: a broader set of uncertain conditions needs to be explored, spanning beyond historical observation and exploring a wide range of unprecedented conditions. This perspective on modeling, termed exploratory (Bankes, 1993), views models as computational experiments that can be used to explore vast ensembles of potential scenarios so as to identify those with consequential effects. Exploratory modeling literature explicitly orients experiments toward stakeholder consequences and decision-relevant inferences and shifts the focus from predicting future conditions to discovering which conditions lead to undesirable or desirable consequences.

This evolution in modeling perspectives can be mirrored by the IM3 family of models in a progression from evaluating models relative to observed history to advanced formalized analyses to make inferences on multisector, multiscale vulnerabilities and resilience. Exploratory modeling approaches can help fashion experiments with large numbers of alternative hypotheses on the co-evolutionary dynamics of influences, stressors, as well as path-dependent changes in the form and function of human-natural systems (Weaver et al., 2013). The aim of this text is to therefore guide the reader through the use of sensitivity analysis and uncertainty methods across these perspectives on diagnostic and exploratory modeling.

CHAPTER

THREE

SENSITIVITY ANALYSIS: THE BASICS

3.1 Global Versus Local Sensitivity

Out of the several definitions for sensitivity analysis presented in literature, the most widely used has been proposed by Saltelli et al. (2004) as “the study of how uncertainty in the output of a model (numerical or otherwise) can be apportioned to different sources of uncertainty in the model input”. In other words, sensitivity analysis explores the relationship between the model’s N input variables, x=[x1,x2,…,xN], and M output variables, y=[y1,y2,…,yM] with y=g(x), where g is the model that maps the model inputs to the outputs (Borgonovo and Plischke, 2016). Therefore, sensitivity analysis provides us with a set of alternatives to conducting large empirical experiments which are costly and often, in practice, next to impossible. Historically, there have been two broad categories of sensitivity analysis techniques: local and global. Local sensitivity analysis is performed by varying model parameters around specific reference values, with the goal of exploring how small input perturbations influence model performance. Due to its convenience, this approach has been widely used in literature, but has important limitations (Rakovec et al., 2014; Saltelli and Annoni, 2010). If the model is not linear, the results of local sensitivity analysis can be heavily biased, as they will vary depending on the range of the chosen input (e.g., (Tang et al., 2007)). If the model’s factors interact, local sensitivity analysis will underestimate their importance, as it does not account for those effects (e.g., (Hamm et al., 2006)). In general, as local sensitivity analysis only partially and locally explores the parametric space, it is not considered a valid approach for nonlinear models (Saltelli et al., 2019). This is illustrated in Fig. 3 (a-b), presenting contour plots of a model response (y1) with an additive linear model (in a) and with a nonlinear model (in b). In a linear model without interactions between the terms x1 and x2, local sensitivity analysis can produce appropriate sensitivity indices (Fig. 3 a). If however, factors x1 and x2 interact, the local and partial consideration of the space can not properly account for each factor’s effects on the model response (Fig. 3 b), as it is only informative at the base point where it is applied. In contrast, a global sensitivity analysis varies uncertain factors within the entire feasible space of variability (Fig. 3 c). This approach reveals the global effects of each parameter on the model output, including any interactive effects. For models that cannot be proven linear, global sensitivity analysis is preferred and this text is primarily discussing global sensitivity analysis methods. In general, whenever we use the term sensitivity analysis we are referring to its global application.

3.2 Why Perform Sensitivity Analysis

It is important to understand the many ways in which a SA might be of use to your modeling effort which can help shape the framing of model-informed study. Most commonly, one might be motivated to perform sensitivity analysis for the following reasons, grouped into two major categories, Model Evaluation and Fidelity Testing, and Exploratory Modeling and Scenario Discovery.

Model evaluation: Sensitivity analysis can be used to gauge model inferences when assumptions about the structure of the model or its parameterization are dubious or have changed. For instance, consider a numerical model that uses a set of calibrated parameter values to produce outputs, which we then use to inform decisions about the real-world system represented. One might like to know if small changes in these parameter values significantly change this model’s output and the decisions it informs or if, instead, our parameter inferences yield stable model behavior regardless of

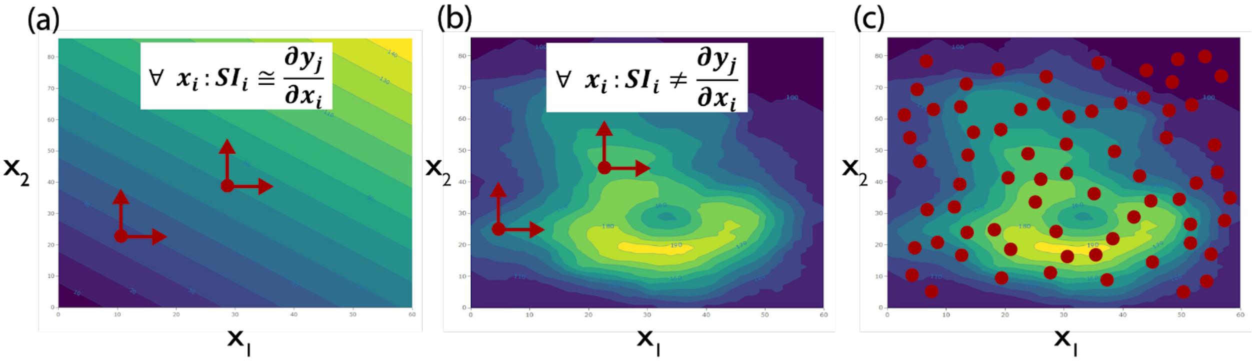

Fig. 1: Treatment of a two-dimensional space of variability by local (panels a-b) and global (panel c) sensitivity analyses. Local sensitivity analysis is only an appropriate approach to sensitivity in the case of linear models without interactions between terms (a). In the case of more complex models, local sensitivity will miscalculate sensitivity indices (b), and global sensitivity methods should be used instead (c).

the uncertainty present in the specific parameterized processes or properties. This can either discredit or lend credence to the model at hand, as well as any inferences drawn that are founded on its accurate representation of the system. Sensitivity analysis can identify which uncertain model factors cause this undesirable model behavior.

Model simplification: Sensitivity analysis can also be used to identify factors or components of the model that appear to have limited effects on direct outputs or metrics of interest. Consider a model that has been developed in an organization for the purposes of a specific research question and is later used in the context of a different application. Some processes represented in significant detail might no longer be of the same importance while consuming significant data or computational resources, as different outputs might be pertinent to the new application. Sensitivity analysis can be used to identify unimportant model components and simplify them to nominal values and reduced model forms. Model complexity and computational costs can therefore be reduced, and, by extension, monetary investments.

Model refinement: Alternatively, sensitivity analysis can reveal the factors or processes that are highly influential to the outputs or metrics of interest, by assessing their relative importance. In the context of model evaluation, this can inform which model components warrant additional investigation or measurement so the uncertainty surrounding them and the resulting model outputs or metrics of interest can be reduced.

Exploratory modeling: When sufficient credence has been established in the model, sensitivity analysis can be applied to a host of other inquiries. Inferences about the factors and processes that most (or least) control a model's outputs of interest can be extrapolated to the real system they represent and be used in a heuristic manner to inform model-based inferences. On this foundation, a model paired with the advanced techniques presented in this text can be used to "discover" decision relevant and highly consequential outcomes (i.e., scenario discovery, (Bankes, 1993; Bryant and Lempert, 2010)).

3.3 Sensitivity Analysis Applications for Model Evaluation and Fidelity Testing

Elucidation of the specific type of problem faced shapes the specific objectives of applying a sensitivity analysis, as well as methods and tools most appropriate and defensible for each application setting (Saltelli et al., 2004; Saltelli and Tarantola, 2002). The three most common sensitivity analysis applications (Factor Prioritization, Factor Fixing, and Factor Mapping) are presented below, but the reader should be aware that other uses have been proposed in the literature (e.g., (Anderson et al., 2014; Borgonovo, 2010)).

Factor prioritization: This sensitivity analysis application type (also referred to as factor ranking) refers to when one would like to identify the uncertain factors which, when fixed to their true value, would lead to the greatest reduction in output variability (Saltelli et al., 2008). Information from this type of analysis can be crucial to model improvement as these factors can become the focus of future measurement campaigns or numerical experiments so that uncertainty in the model output can be reduced. Fig. 4 (a) shows the effects of three uncertain variables (X1, X2, and X3) on the variance of output Y.V(E(Y|Xi)) indicates the variance in Y if factor Xi is left to vary freely while all other factors remain fixed to nominal values. In this case, factor X2 makes the largest contribution to the variability of output Y and it should therefore be prioritized. In the context of risk analysis, factor prioritization can be used to reduce output variance to below a given tolerable threshold (also known as variance cutting). As the number of factors of focus and their degree of interactions increases, the computational experiments and analysis techniques increase in their demands as well as sophistication (this is further elaborated in the Global versus Local Sensitivity and the Design of Experiments sections).

Factor fixing: Conversely, sensitivity analysis used in this mode (also referred to as factor screening) aims to identify the model components that have a negligible effect or make no significant contributions to the variability of the outputs or metrics of interest (usually referred to as non-influential; Saltelli et al., 2008). In the stylized example of Fig. 4 (a), X1 makes the smallest contribution to the variability of output Y suggesting that the uncertainty in its value could be negligible and the factor itself fixed in subsequent model executions. Eliminating these factors or processes in the model or fixing them to a nominal value can help reduce model complexity as well as the unnecessary computational burden of subsequent model runs, results processing, or other sensitivity analyses (the fewer uncertain factors considered, the fewer runs are necessary to illuminate their effects on the output). Significance of the outcome can be gauged in a variety of manners, depending on the application. For instance, if applying a variance-based method, a minimum threshold value of contribution to the variance could be considered as a significance 'cutoff', and factors with indices below that value can be considered non-influential. Nb: conclusions about factor fixing should be made based on total-order effects, i.e., considering all the effects a factor has, individually and in interaction with other factors (explained in more detail in the Variance-based methods section).

Factor mapping: Finally, factor mapping can be used to pinpoint which values of uncertain factors lead to model outputs within a given range of the output space (Saltelli et al., 2008). In the context of model diagnostics, it is possible that the model's output changes in ways considered impossible based on the represented processes, or other observed evidence. In this situation, factor mapping can be used to identify which uncertain model factors cause this undesirable model behavior by 'filtering' model runs that are considered 'non-behavioral' (Edwards et al., 2011; Pianosi et al., 2016; Spear and Hornberger, 1980). In Fig. 4 (b), region B of the output space denotes the set of behavioral model outcomes, which can be traced back to input space X (e.g., using Monte Carlo Filtering or pre-calibration).

3.4 Sensitivity Analysis Applications for Exploratory Modeling and Scenario Discovery

The language used above reflects a use of sensitivity analysis for model fidelity evaluation and improvement (top right panel in Fig. 3). However, as previously mentioned, when a model has been established as a sufficiently accurate representation of the system, sensitivity analysis can produce additional inferences (bottom right panel in Fig. 3). For instance, under the factor mapping use, the analyst can now focus on undesirable system states and discover which factors are most responsible for them: for instance, "population growth of above 25% would be responsible for unacceptably high energy demands". Factor prioritization and factor fixing can be used to make equivalent inferences, such as "growing populations and increasing temperatures are the leading factors for changing energy demands" (prioritizing of factors) or "changing dietary needs are inconsequential to increasing energy demands for this region" (a factor that can be fixed in subsequent model runs). All these inferences hinge on the assumption that the real system's stakeholders consider the model states faithful enough representations of system states. As elaborated in the Perspectives on model evaluation section, this view on sensitivity analysis is founded on a relativist perspective on modeling, which tends to place more value on model usefulness rather than strict accuracy of representation in terms of error. As such, sensitivity analysis performed with decision-making relevance in mind will focus on model outputs or metrics that are consequential and decision relevant (e.g., energy demand in the examples above).

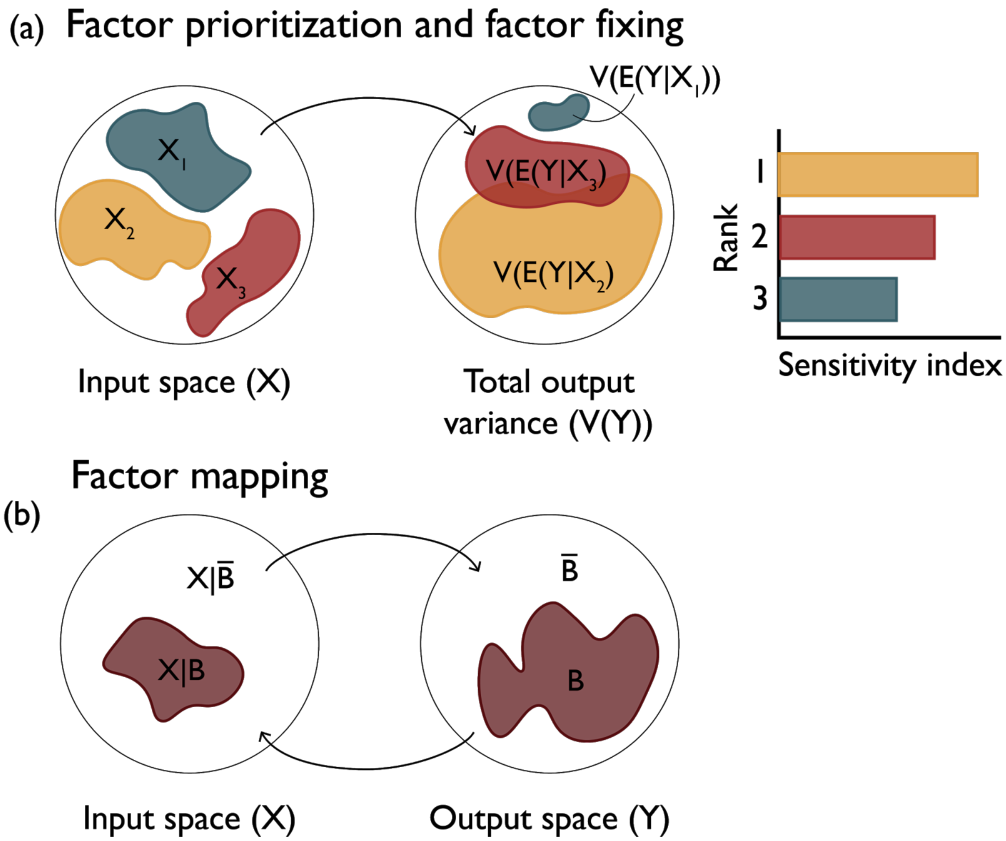

Fig. 2: Factor prioritization, factor fixing and factor mapping settings of sensitivity analysis.

3.5 Design of Experiments

Before embarking on any sensitivity analysis, the first element that needs to be clarified is the uncertainty space of the model (Helton et al., 2006; Pianosi et al., 2016). In other words, how many and which factors making up the mathematical model are considered uncertain and can potentially affect the model output and the inferences drawn from it. Uncertain factors can be model parameters, model structures, inputs, or alternative model resolution levels (scales), all of which can be assessed through the tools presented in this text. Depending on the kind of factor, its variability can be elicited through various means: expert opinion, values reported in the literature, its physical meaning (e.g., population values in a city can never be negative), or through the use of more formal UQ methods, elaborated in later sections of this text. The model uncertainty space represents the entire space of variability present in each of the uncertain factors of a model. The complexity of most real-world models means that the response function, y=g(x), mapping inputs to outputs, is hardly ever available in an analytical form and therefore analytically computing the sensitivity of the output to each uncertain factor becomes impossible. In these cases, sensitivity analysis is only feasible through numerical procedures that employ different strategies to sample the uncertainty space and calculate sensitivity indices. A sampling strategy is often referred to as a design of experiments and represents a methodological choice made before conducting any sensitivity analysis. Experimental design was first introduced by (Fisher, 1960) in the context of laboratory or field-based experiments. Its application in sensitivity analysis is similar to setting up a physical experiment in that it is used to discover the behavior of a system under specific conditions. An ideal design of experiments should provide a framework for the extraction of all plausible information about the impact of each factor on the output of the numerical model. The design of experiments is used to set up a simulation platform with the minimum computational cost to answer specific questions that cannot be readily drawn from the data through analytical or common data mining techniques. Models representing coupled human-natural systems usually have a large number of inputs, state variables and parameters, but not all of them exert fundamental control over the numerical process, despite their uncertainty, nor have substantial impacts on the model output, either independently or through their interactions. Each factor influences the model output in different ways that need to be discovered. For example, the influence of a parameter on model output can be linear or non-linear and can be continuous or only be active during specific times or at particular states of the system (Herman et al., 2013; Massmann et al., 2014). An effective and efficient design of experiments allows the analyst to explore these complex relationships and evaluate different behaviors of the model for various scientific questions (Van Schepdael et al., 2016). There are a few different approaches to the design of experiments. The selection of design is closely related to the chosen sensitivity analysis approach, which is in turn shaped by the research motivations, scientific questions, and computational constraints at hand (additional discussion of this can be found at the end of the Sensitivity Analysis Methods section). For example, in a sensitivity analysis using perturbation and derivatives methods, the model input parameters vary from their nominal values one at a time, something that the design of experiments needs to reflect. If, instead, one were to perform sensitivity analysis using a multiple-starts perturbation method, the design of experiments needs to consider that multiple points across the factor space are used. The design of experiments specifically defines two key characteristics of samples that are fed to the numerical model: the number of samples and the range of each factor. Generally, sampling can be performed randomly or by applying a stratifying approach. In random sampling, such as Monte Carlo (Metropolis and Ulam, 1949), samples are randomly generated by a pseudo-random number generator with an a-priori assumption about the distribution of parameters and their possible ranges. Random seeds can also be used to ensure consistency and higher control over the random process. However, this method could leave some holes in the parameter space and cause clustering in some spaces, especially for a large number of parameters (Norton, 2015). Most sampling strategies use stratified sampling to mitigate these disadvantages. Stratified sampling techniques divide the domain of each factor into subintervals, often of equal lengths. From each subinterval, an equal number of samples is drawn randomly, or based on the specific locations within the subintervals (Saltelli et al., 2008). The rest of this section overviews some of the most commonly used designs of experiments. Table 1 summarizes the designs discussed.

Table 1: Summary of designs of experiments overviewed in this section.
* Depends on the sample size.

Design of experiments	*Factor interactions considered*	*Treatment of factor domains*
One-At-a-Time (OAT)	No - main effects only	Continuous (distributions)
Full Factorial Sampling	Yes - including total effects	Discrete (levels)
Fractional Factorial Sampling	Yes - only lower-order effects*	Discrete (levels)
Latin Hypercube (LH) Sampling	Yes - including total effects*	Continuous (distributions)
Quasi-Random Sampling with Low-Discrepancy Sequences	Yes - including total effects*	Continuous (distributions)

3.5.1 One-At-a-Time (OAT)

In this approach, only one model factor is changed at a time while all others are kept fixed across each iteration in a sampling sequence. The OAT method assumes that model factors of focus are linearly independent (i.e., there are no interactions) and can analyze how factors individually influence model outputs or metrics of interest. While highly popular given its ease of implementation, OAT is ultimately highly limited in its exploration of a model's sensitivities (Saltelli and Annoni, 2010). It is primarily used with local sensitivity techniques with similar criticisms: applying this sampling scheme on a system with nonlinear and interactive processes will miss important information on the effect uncertain factors have on the model. OAT samplings can be repeated multiple times in a more sophisticated manner and across different locations of the parameter space to overcome some of these challenges, which would increase computational costs and negate the main reasons for its selection.

3.5.2 Full and Fractional Factorial Sampling

In full factorial sampling, each factor is treated as being discrete, by considering two or more levels (or intervals). The sampling process then generates samples within each possible combination of levels, corresponding to each parameter. This scheme produces a more comprehensive sampling of the factors' variability space, as it accounts for all candidate combinations of factor levels (Fig. 6 (a)). If the number of levels is the same across all factors, the number of generated samples is estimated using nk, where n is the number of levels and k is the number of factors. For example, Fig. 6 (a) presents a full factorial sampling of three uncertain factors (x1, x2, and x3), each considered as having four discrete levels. The total number of samples necessary for such an experiment is 43=64. As the number of factors increases, the number of simulations necessary can also grow exponentially, making full factorial sampling computationally burdensome (Fig. 6 (b)). As a result, literature has commonly applied full factorial sampling at only two levels per factor, typically the two extremes (Montgomery, 2017). This significantly reduces computational burden but is only considered appropriate in cases where factors can indeed only assume two discrete values (e.g., when testing the effects of epistemic uncertainty and comparing between model structure A and model structure B). In the case of physical parameters on continuous distributions (e.g., when considering the effects of measurement uncertainty in a temperature sensor), discretizing the range of a factor to only extreme levels can bias its estimated importance. Fractional factorial design is a widely used alternative to full factorial that allows the analyst to significantly reduce the number of simulations by confounding the main effects of a factor with its interactive effects (Saltelli et al., 2008). In other words, if one can reasonably assume that higher-order interactions are negligible, information about the main effects and lower-order interactions can be obtained using a fraction of the full factorial design. Traditionally, fractional factorial design has also been limited to two levels (Montgomery, 2017), referred to as Fractional Factorial designs 2k-p (Box and Hunter, 1961). Recently, Generalized Fractional Factorial designs have also been proposed that allow for the structured generation of samples at more than two levels per factor (Surowiec et al., 2017). Consider a case where the modeling team dealing with the problem in Fig. 6 (a) cannot afford to perform 64 simulations of their model. They can afford 32 runs for their experiment and instead decide to fractionally sample the variability space of their factors. A potential design of such a sampling strategy is presented in Fig. 6 (c).

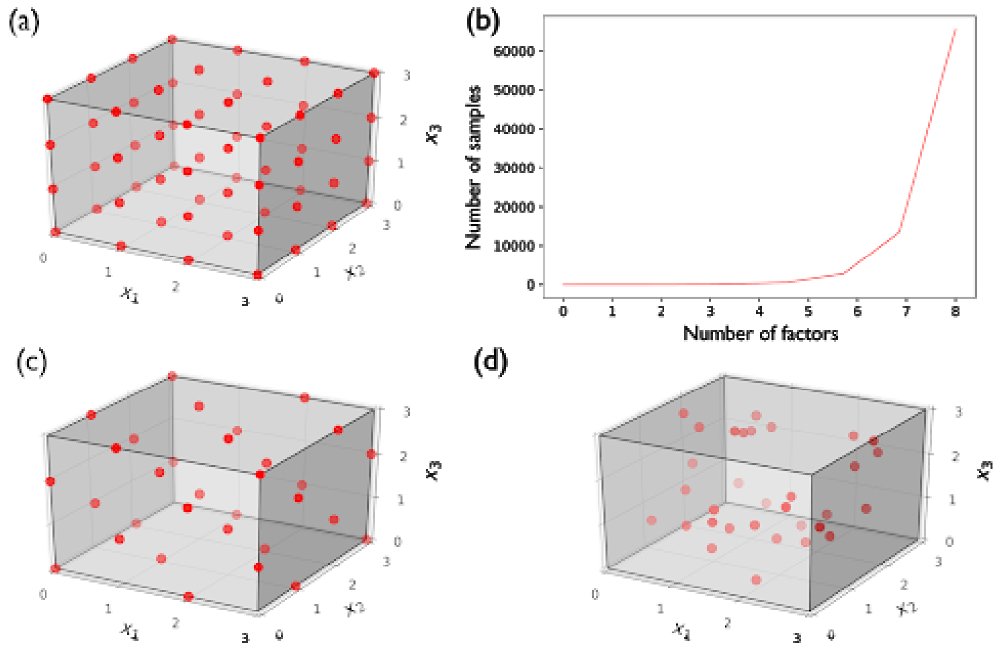

Fig. 3: Alternative designs of experiments and their computational costs for three uncertain factors (x1, x2, and x3). (a) Full factorial design sampling of three factors at four levels, at a total of 64 samples; (b) exponential growth of necessary number of samples when applying full factorial design at four levels; (c) fractional factorial design of three factors at four levels, at a total of 32 samples; and (d) Latin Hypercube sample of three factors with uniform distributions, at a total of 32 samples.

3.5.3 Latin Hypercube (LH) Sampling

Latin hypercube sampling (McKay et al., 1979) is one of the most common methods in space-filling experimental designs. With this sampling technique, for N uncertain factors, an N-dimensional hypercube is generated, with each factor divided into an equal number of levels depending on the sample to be generated. Equal numbers of samples are then randomly generated at each level, across all factors. In this manner, LH design guarantees sampling from every level of the variability space and without any overlaps. When the number of samples generated is much larger than the number of uncertain factors, LH sampling can be very effective in examining the effects of each factor (Saltelli et al., 2008). LH sampling is an attractive technique, because it guarantees a diverse coverage of the space, through the use of subintervals, without being constrained to discrete levels for each factor - compare Fig. 6 (c) with Fig. 6 (d) for the same number of samples.

LH sampling is less effective when the number of samples is not much larger than the number of uncertain factors, and the effects of each factor cannot be appropriately distinguished. The samples between factors can also be highly correlated, biasing any subsequent sensitivity analysis results. To address this, the sampling scheme can be modified to control for the correlation in parameters while maximizing the information derived. An example of such modification is through the use of orthogonal arrays (Tang, 1993).

3.5.4 Low-Discrepancy Sequences

Low-discrepancy sequences is another sampling technique that employs a pseudo-random generator for Monte Carlo sampling (Dalal et al., 2008; Zaremba, 1968). These quasi-Monte Carlo methods eliminate potential gaps and clusters between samples by minimizing discrepancy when generating uniformly distributed random samples within the hypercube. The discrepancy property is mathematically measured by characterizing the lumpiness of a sequence of samples in a multidimensional space, which results in evenly distributed samples (Dalal et al., 2008). Discrepancy can be quantitatively measured using the deviations of sampled points from the uniform distribution (Kucherenko et al., 2015). Low-discrepancy sequences ensure that the number of samples in any subspace of the variability hypercube is approximately the same. This is not something guaranteed by LH sampling, and even though the design can be improved through optimization with various criteria, it is limited to small sample sizes and low dimensions (Iooss et al., 2010; Jin et al., 2008; Kucherenko et al., 2015; Morris and Mitchell, 1995; Park, 1994). The Sobol sequence (Sobol, 1976; Sobol', 1967), one of the most widely used sampling techniques, utilizes the low-discrepancy approach to uniformly fill the sampled factor space. A core advantage of this style of sampling is that it takes far fewer samples (i.e., simulations) to attain a much lower level of error in estimating model output statistics (e.g., the mean and variance of outputs).

3.5.5 Other types of sampling

The sampling techniques mentioned so far are general sampling methods useful for a variety of applications beyond sensitivity analysis. There are however techniques that have been developed for specific sensitivity analysis methods. Examples of these methods include the Morris One-At-a-Time (Morris, 1991), Fourier Amplitude Sensitivity Test (FAST; (Cukier et al., 1973)), Extended FAST (Saltelli et al., 1999), and Extended Sobol methods (Saltelli, 2002). For example, the Morris sampling strategy builds a number of trajectories (usually referred to as repetitions and denoted by r) in the input space each composed of N+1 factor points, where N is the number of uncertain factors. The first point of the trajectory is selected randomly and the subsequent N points are generated by moving one factor at a time by a fixed amount. Each factor is perturbed once along the trajectory, while the starting points of all of the trajectories are randomly and uniformly distributed. Several variations of this strategy also exist in the literature; for more details on each approach and their differences the reader is directed to (Pianosi et al., 2016).

3.5.6 Synthetic generation of input time series

Models often have input time series or processes with strong temporal and/or spatial correlations (e.g., streamflow, energy demand, price of commodities, etc.) that, while they might not immediately come to mind as factors to be examined in sensitivity analysis, can be treated as such. Synthetic input time series are used for a variety of reasons, for example, when observations are not available or are limited, or when past observations are not considered sufficiently representative to capture rare or extreme events of interest (Herman et al., 2016; Milly et al., 2008). Synthetic generation of input time series provides a valuable tool to consider non-stationarity and incorporate potential stressors, such as climate change impacts into input time series (Borgomeo et al., 2015). For example, a century of record will be insufficient to capture very high impact rare extreme events (e.g., persistent multi-year droughts). A large body of statistical literature exists focusing on the topics of synthetic weather (Herrera et al., 2017; Wilks and Wilby, 1999) and streamflow (Lamontagne and Stedinger, 2018; Medda and Bhar, 2019) generation that provides a rich suite of approaches for developing history-informed, well-characterized stochastic process models to better estimate rare individual or compound (hot, severe drought) extremes. It is beyond the scope of this text to review these methods. Readers are encouraged to explore the studies cited above in this section as well as the following publications for discussions and comparisons of these methods: (Borgomeo et al., 2015; Herman et al., 2016; Kirsch et al., 2013; Loucks and Beek, 2017; Steinschneider et al., 2015; Vogel, 2017; Vogel and Stedinger, 1988). The use of these methods for the purposes of exploratory modeling, especially in the context of well-characterized versus deep uncertainty, is further discussed in the Consequential Scenarios section.

3.6 Sensitivity Analysis Methods

In this section, we describe some of the most widely applied sensitivity analysis methods along with their mathematical definitions. We also provide a detailed discussion on applying each method, as well as a comparison of and their features and limitations.

3.6.1 Derivative-based Methods

Derivative-based methods explore how model outputs are affected by perturbations in a single model input around a particular input value. These methods are local and are performed using OAT sampling. For simplicity of mathematical notations, let us assume that the model *g(X)* only returns one output. Following Borgonovo (2008) and Pianosi et al. (2016), the sensitivity index, S_i, of the model's *i*-th input factor, x_i, can be measured using the partial derivative evaluated at a nominal value, $\underline{x}$, of the vector of inputs:

$$S_i(\underline{x}) = \frac{\partial g}{\partial x}|_{\underline{x}^c{}_i}$$

where c_i is the scaling factor.

3.6.2 Elementary Effect Methods

3.6.3 Regression-based Methods

3.6.4 Regional Sensitivity Analysis

3.6.5 Variance-based Methods

3.6.6 Analysis of Variance (ANOVA)

3.6.7 Moment-Independent (Density-Based) Methods

3.7 How To Choose A Sensitivity Analysis Method: Model Traits And Dimensionality

Figure 8 presents a graphical synthesis of the methods overviewed in this section, with regards to their appropriateness of application based on the complexity of the model at hand and the computational limits on the number of model evaluations afforded. The bars below each method also indicate the sensitivity analysis purposes they are most appropriate to address, which are in turn a reflection of the motivations and research questions the sensitivity analysis is called to address. Computational intensity is measured as a multiple of the number of model factors that are considered uncertain (d). Increasing model complexity mandates that more advanced sensitivity analysis methods are applied to address potential nonlinearities, factor interactions and discontinuities. Such methods can only be performed at increasing computational expense. For example, computationally cheap linear regression should not be used to assess factors' importance if the model cannot be proven linear and the factors independent, because important relationships will invariably be missed (recall the example in Fig. 5). When computational limits do constrain applications to make simplified assumptions and sensitivity techniques, any conclusions in such cases should be delivered with clear statements of the appropriate caveats.

The reader should also be aware that the estimates of computational intensity that are given here are indicative of magnitude and would vary depending on the sampling technique, model complexity and the level of information being asked. For example, a Sobol sensitivity analysis typically requires a sample of size n × d+2 to produce first- and total-order indices, where d is the number of uncertain factors and n is a scaling factor, selected ad hoc, depending on model complexity (Saltelli, 2002a). The scaling factor n is typically set to at least 1000, but it should most appropriately be set on the basis of index convergence. In other words, a prudent analyst would perform the analysis several times with increasing n and observe at what level the indices converge to stable values (Nossent et al., 2011). The level should be the minimum sample size used in subsequent sensitivity analyses of the same system. Furthermore, if the analyst would like to better understand the degrees of interaction between factors, requiring second-order indices, the sample size would have to increase to n × 2d+2 (Saltelli, 2002a).

3.8 Software Toolkits

CHAPTER

FOUR

SENSITIVITY ANALYSIS: DIAGNOSTIC & EXPLORATORY MODELING

4.1 Understanding Errors: What Is Controlling Model Performance?

4.2 Consequential Dynamics: What is Controlling Model Behaviors of Interest?

4.3 Consequential Scenarios: What is Controlling Consequential Outcomes?

CHAPTER

FIVE

UNCERTAINTY QUANTIFICATION: THE BASICS

As described in the previous sections, uncertainty characterization (UC) can be defined as exploratory modeling where alternative hypotheses for the co-evolutionary dynamics of influences, stressors, as well as path-dependent changes in the form and function of systems are explored (Marchau et al., 2019). UC exploratory modeling has a consistent focus on the assumptions, structural model forms, alternative parameterizations, and input data sets that are used to characterize the behavioral space of one or more models. The focus of UC is not to exactly quantify and predict probabilistic likelihoods for all possible quantities, but instead to inform which modeling choices yield the most consequential behavioral changes or outcomes, especially when considering deeply uncertain, scenario-informed projections (Moallemi et al., 2020b; Walker et al., 2013).

In comparison, uncertainty quantification (UQ) refers to the representation of uncertainties using probability distributions. The act of quantification requires specific assumptions about distributional forms and likelihoods, which may be more or less justified depending on prior information about the system or model behavior (Frankignoul and Hasselmann, 1977; Zellner and Tian, 1964). Without this justification, alternative specifications may yield substantially different inferences.

5.1 Why is Uncertainty Quantification Important for Understanding MultiSector System Dynamics?

5.2 Uncertainty Quantification for Exploratory Modeling

5.3 Bayesian Uncertainty Quantification

5.4 Uncertainty Quantification Under (Deep) Uncertainty

5.5 Integrating Model Diagnostics and Uncertainty Quantification

CHAPTER

SIX

UNCERTAINTY QUANTIFICATION: A TOOL FOR CAPTURING RISKS & EXTREMES

6.1 Understanding Risk: How Probable Are Extreme Events?

6.2 Understanding Tails: Statistical Modeling of Extreme Events

6.3 How to Choose an Appropriate Method?

6.4 How to Select a Prior Distribution?

6.5 Posterior Predictive Checking

6.6 Model Selection and Comparison

6.7 What are Common Methods?

There are many methods to quantify uncertainty. Each method has advantages and disadvantages for a particular analysis. Here we focus on parametric uncertainty quantification, as a discussion of structural uncertainty quantification is beyond the scope of this review. Moreover, we prefer to think about structural uncertainty from the perspective of exploratory modeling and deep uncertainty, rather than from the perspective of quantification and selection or averaging.

Uncertainty quantification methods can be broadly classified as Markov Chain Monte Carlo (MCMC) approaches, particle-based approaches, and emulation-based approaches, though there are some hybrid methods. Several of the most common approaches for uncertainty quantification are described below. In all cases, the computational and conceptual challenges associated with parametric uncertainty quantification grow rapidly with the number of model parameters. As noted in the prior sections, sensitivity analyses are useful for dimensionality reduction prior to conducting parametric uncertainty quantification. Both factor fixing and factor prioritization can be used to limit the number of parameters which are treated as uncertain.

6.7.1 Scenario Discovery

6.7.2 Pre-calibration/GLUE

6.7.3 Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC) is a "gold standard" approach to full uncertainty quantification. MCMC refers to a category of algorithms which systematically sample from a target distribution (in this case, the posterior distribution) by constructing a Markov chain. MCMC algorithms rely on the mixing properties of the resulting Markov chain to guarantee asymptotic convergence to the posterior distribution, as the chain is constructed so that the posterior is its stationary distribution. It should be stressed that this guarantee exists only asymptotically. Studies use heuristics to test for signs of misconvergence and to assess the skill of the approximation (xx)

MCMC algorithms begin with the choice of some initial value for the Markov chain. This value can be randomly determined, or can be some other quantity such as a maximum likelihood or maximum a posteriori estimates. While the Markov chain will eventually converge to the posterior regardless of the choice of initial value, the amount of time required to escape the transient dynamics of the Markov chain is dependent on this value. Typically, transient samples are discarded as burn-in, as they may skew the sample distribution if the burn-in is relatively long compared to the number of iterations spent exploring the posterior, though this practice is not universal and has been questioned by some statisticians (Geyer, 2011). However, when not discarding the transient area, the chain must be run for a larger number of iterations to ensure that these samples do not bias the sample distribution.

Diagnosing the convergence of the Markov chain to the posterior is more art than science, relying on heuristics and judgement. One example heuristic is to run many Markov chains from different initial conditions, ideally well-dispersed across the parameter space; one may be able to conclude that the chains have not yet converged if the resulting marginal parameter distributions are sufficiently different when plotted. The Gelman-Rubin diagnostic formalizes this idea by comparing the within-chain and pooled variances of multiple chains (Gelman and Rubin, 1992). The ratio of these two quantities, called the potential scale reduction factor, can diagnose a lack of convergence if it is sufficiently far from 1 (typically using a threshold such as 1.1 or 1.05). Thus, it is generally good practice to use several MCMC runs to facilitate the diagnoses of non-convergence.

Another key value is the effective sample size (ESS). Due to the Markovian property, the samples obtained using MCMC are autocorrelated, and therefore not independent. As a result, the number of samples obtained using MCMC are not directly useful when interpreting the extent of exploration (or computing quantities such as the Monte Carlo standard error (Flegal et al., 2008)). For example, it may not be appropriate to draw inferences about tail properties for a small ESS.

Many MCMC algorithms exist, with varying strengths and weaknesses, discussed below. For example, some require more tuning to improve the ESS than others. All of these algorithms involve the evaluation of the model at various parameter settings. Once a Markov chain is constructed and deemed to suitably represent the posterior distribution, parameter values can be sampled from it with replacement as a proxy for directly sampling from the posterior.

6.7.3.1 Metropolis-Hastings

6.7.3.2 Gibbs Sampling

6.7.3.3 Hamiltonian Monte Carlo

6.7.4 Particle-based Methods

6.8 What are Example Software Implementations?

There exist many software platforms to implement uncertainty assessment. Each implementation is built upon a specific programming language including, but not limited to R, Python, C++, Fortran, MATLAB, and Julia. Two key considerations are the user's preferred programming language and the computer model's native code. For instance, a computer model running in C++ may be better suited for a software implementation based on the same language. For inconsistencies, please see the discussion on wrappers below.

Here, we present an overview of popular packages inherent to R, Python, and Julia. The user is free to code the UQ implementation without incorporating these existing packages; however, it may require more effort to code the pertinent subroutines (e.g., MCMC and building surrogate models). Uncertainty quantification for computer models typically operates within the Bayesian framework (see What are Common Methods?). Each implementation includes a mechanism that enables Bayesian inference using MCMC, Gaussian process emulation, or Sequential Monte Carlo. We focus on a subset of the common approaches.

6.8.1 Markov Chain Monte Carlo with the True Model

6.8.2 Markov Chain Monte Carlo with Surrogate Models

CHAPTER SEVEN

CONCLUSION

As noted in the Introduction (Section 1.0), the computational and conceptual challenges of the multi-model, transdisciplinary workflows that characterize ambitious projects such as IM3 have limited UC and UQ analyses. Moreover, the very nature and purpose of modeling and diagnostic model evaluation can have very diverse philosophical framings depending on the disciplines involved (see Figure 1 and Section 2.0). The guidance provided in this text can be used to frame consistent and rigorous experimental designs for better understanding the consequences and insights from our modeling choices when seeking to capture complex human-natural systems. The progression of sections of this text provide a thorough introduction of the concepts and definitions of diagnostic model evaluation, sensitivity analysis, UC, and UQ. In addition, we comprehensively discuss how specific modeling objectives and applications should guide the selection of appropriate techniques; broadly, these can include model diagnostics, in-depth analysis of the behavior of the abstracted system, and projections under conditions of deep uncertainty. This text also contains a detailed presentation of the main sensitivity analysis, UC, and UQ analysis methods and a discussion of their features and main limitations. Readers are also provided with an overview of computer tools and platforms that have been developed and could be considered in addressing IM3 scientific questions. The appendices of this text include a terminology glossary of the key concepts as well as example test cases and scripts to showcase various UC related capabilities.

Although we distinguish the UC and UQ model diagnostics, the reader should note that we suggest an overall consistent approach to both in this text by emphasizing "exploratory modeling" (see review add citation). Although data support, model complexity, and computational limits strongly distinguish the feasibility and appropriateness of the UC and UQ diagnostic tools (e.g., see Figure 18), we overall recommend that modelers view their work through the lens of cycles of learning. Iterative and deliberative exploration of model-based hypotheses and inferences for transdisciplinary teams is non-trivial and ultimately critical for mapping where innovations or insights are most consequential. Overall, we recommend approaching modeling with an openness to the diverse disciplinary perspectives such as those mirrored by the IM3 family of models in a progression from evaluating models relative to observed history to advanced formalized analyses to make inferences on multi-sector, multi-scale vulnerabilities and resilience. Exploratory modeling approaches can help fashion experiments with large numbers of alternative hypotheses on the co-evolutionary dynamics of influences, stressors, as well as path-dependent changes in the form and function of coupled human-natural systems (Weaver et al., 2013). This text guides the reader through the use of sensitivity analysis and uncertainty methods across the diverse perspectives that have shaped modern diagnostic and exploratory modeling.

CHAPTER

EIGHT

GLOSSARY

CHAPTER

NINE

INDICES AND TABLES

- genindex
- modindex
- search

BIBLIOGRAPHY

[1] Roger Cooke and others. *Experts in uncertainty: opinion and subjective probability in science*. Oxford University Press on Demand, 1991.

[2] National Research Council and others. *Convergence: Facilitating transdisciplinary integration of life sciences, physical sciences, engineering, and beyond*. National Academies Press, 2014.

[3] Sondoss Elsawah, Tatiana Filatova, Anthony J Jakeman, Albert J Kettner, Moira L Zellner, Ioannis N Athanasiadis, Serena H Hamilton, Robert L Axtell, Daniel G Brown, Jonathan M Gilligan, and others. Eight grand challenges in socio-environmental systems modeling. *Socio-Environmental Systems Modelling*, 2:16226–16226, 2020.

[4] Saul I Gass and Carl M Harris. Encyclopedia of operations research and management science. *Journal of the Operational Research Society*, 48(7):759–760, 1997.

[5] Yacov Y Haimes. Risk modeling of interdependent complex systems of systems: theory and practice. *Risk analysis*, 38(1):84–98, 2018.

[6] Dirk Helbing. Globally networked risks and how to respond. *Nature*, 497(7447):51–59, 2013.

[7] Jan H Kwakkel, Warren E Walker, and Marjolijn Haasnoot. Coping with the wickedness of public policy problems: approaches for decision making under deep uncertainty. 2016.

[8] Enayat A Moallemi, Jan Kwakkel, Fjalar J de Haan, and Brett A Bryan. Exploratory modeling for analyzing coupled human-natural systems under uncertainty. *Global Environmental Change*, 65:102186, 2020.

[9] Andrea Saltelli, Ksenia Aleksankina, William Becker, Pamela Fennell, Federico Ferretti, Niels Holst, Sushan Li, and Qiongli Wu. Why so many published sensitivity analyses are false: a systematic review of sensitivity analysis practices. *Environmental modelling & software*, 114:29–39, 2019.

[10] Andrea Saltelli, Philip B Stark, William Becker, and Pawel Stano. Climate models as economic guides scientific challenge or quixotic quest? *Issues in Science and Technology*, 31(3):79–84, 2015.

[11] Warren E Walker, Poul Harremoës, Jan Rotmans, Jeroen P Van Der Sluijs, Marjolein BA Van Asselt, Peter Janssen, and Martin P Krayer von Krauss. Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support. *Integrated assessment*, 4(1):5–17, 2003.

[12] Daniel Wirtz and Wolfgang Nowak. The rocky road to extended simulation frameworks covering uncertainty, inversion, optimization and control. *Environmental Modelling & Software*, 93:180–192, 2017.