From Model Validation to Epistemic Assurance in Disaster Simulations

Rodrigo Costa*[1], Gemma Cremen[2], Sara Hamideh[3], Ali Nejat[4], and Elaina J. Sutley[5]

[1]Department of Systems Design Engineering, University of Waterloo, Waterloo, Ontario, Canada

[2]Department of Civil, Environmental and Geomatic Engineering, University College London, London, United Kingdom

[3]School of Marine and Atmospheric Sciences, Stony Brook University, Stony Brook, 11790, United States of America

[4]Department of Civil, Environmental, and Construction Engineering, Texas Tech University, Lubbock, TX, USA

[5]Department of Civil, Environmental and Architectural Engineering, University of Kansas, Lawrence, KS, USA

July 20, 2026

*Corresponding author: Rodrigo Costa, rodrigo.costa@uwaterloo.ca

Abstract

Disaster simulation models are increasingly used to support decisions about future hazards, climate adaptation, and post-disaster recovery, yet many of the socio-economic processes they represent cannot be directly observed or fully validated. Instead, modelers rely on proxies, inferred relationships, and simplified mechanisms to represent latent processes, creating the possibility that defensible representations produce different decision-relevant conclusions despite similar agreement with historical observations. We argue that confidence in a model implementation and confidence in the conclusions it supports are distinct epistemic questions. We propose an epistemic assurance framework for disaster simulation models that distinguishes proxy, structural, and outcome validity as complementary properties of the model while introducing epistemic robustness as a property of decision-relevant conclusions. The framework first establishes whether a model provides a defensible basis for inference and then evaluates whether important conclusions depend on particular representational choices. Together, these complementary assessments provide a structured basis for qualifying the scope of the model-based conclusions that can be responsibly defended and used to support decisions. Illustrative examples demonstrate how alternative representations can reveal fragile conclusions even when competing models exhibit similar agreement with historical observations. More broadly, the framework complements existing approaches to model validation and uncertainty analysis by reframing the evaluation of disaster simulations from asking whether a model has been validated to asking which conclusions can be responsibly supported, under what assumptions, and for which intended purposes.

1 Introduction

Consider the following scenario: A research team develops a high-resolution disaster recovery model using a rich dataset from the 1994 Northridge Earthquake. Household relocation, financing, contractor allocation, and infrastructure restoration are informed by years of detailed observations, yielding a model that reproduces the observed recovery with remarkable fidelity. The team is later asked to update input datasets and use the model to design housing recovery policies for a future major earthquake affecting the same region.

Should the research team assure decision makers that they can act on the model's recommendations simply because it reproduces a well-documented disaster recovery case?

At first glance, the answer appears straightforward. A model that reproduces one of the best-documented disaster recoveries ever observed would seem well-positioned to inform future policy. Yet reproducing the past and supporting decisions about an unobserved future are fundamentally different epistemic tasks. The first confirms that a model provides a defensible representation of historical system behavior; the second requires demonstrating that decision-relevant conclusions remain stable when applied to unobserved contexts driven by partially latent processes. This distinction motivates the central argument of this paper.

The scenario above is illustrative but increasingly reflects how disaster simulation models are designed. Rather than supporting only retrospective analyses, they are increasingly used to evaluate future hazards, climate adaptation strategies, and recovery interventions [e.g. 15, 17, 18]. In doing so, they must represent socio-economic processes that unfold over months to years, including housing instability, displacement, mental health impacts, inequality, and long-term community well-being [e.g., 26, 33, 37].

Many of these processes remain only partially observable even after major disasters, e.g., household relocation decisions, access to financing, workforce availability, adaptive behavior, and mental health outcomes [42, 48]. Model developers therefore rely on proxies to represent these latent processes. Here, a *proxy* is defined as a modeled variable, relationship, or mechanism used to represent a target process that cannot be directly observed at the spatial, temporal, or conceptual resolution required by the simulation. Such representations may be derived from empirical observations and theory, expert judgment, machine learning, spatial inference, or combinations thereof. We refer to models that rely on these representations as *proxy-driven* because multiple defensible representations of latent or partially observed processes exist, and the choice among them can affect important conclusions.

Proxy-driven modeling is an unavoidable consequence of reasoning about complex socio-economic systems using incomplete evidence. However, proxy-driven modeling introduces a distinct epistemic challenge. Limitations in data and theoretical understanding often permit multiple defensible representations of the same latent process. These representations may preserve different properties of the underlying process, produce different system dynamics, and ultimately support different decision-relevant conclusions. Agreement with observed outputs alone therefore does not guarantee that the mechanisms and relationships required for the intended inference have been adequately represented [41]. Instead, credibility depends on understanding how conclusions vary under unresolved representational assumptions and whether they remain stable across defensible alternative representations [7, 25]. For proxy-driven models, confidence is established not only by demonstrating that a model behaves plausibly, but by showing that the conclusions used to support decisions are robust to reasonable representational choices.

Several established approaches provide guidance for evaluating model credibility under uncertainty. Sensitivity analysis quantifies how outputs respond to parameter and assumption uncertainty within a fixed representation [e.g., 45, 46]; it tells the modeler how uncertain a result is, not which results they can stand behind. Exploratory Modelling and Robust Decision Making explore alternative structures and futures to support decisions that perform well across them [4, 28, 29, 53]; their object is the decision, and their primary user is the decision analyst. Pattern-Oriented Modeling evaluates whether representations reproduce characteristic patterns, a question of model evaluation [e.g., 22, 23]. The equifinality literature observes that many defensible structures may fit the same data equally well [7]. Each of these is diagnostic or decision-analytic. None addresses the modeler's distinct, prior question: given that these conditions hold, which of my own conclusions can I responsibly assure, under what assumptions, and which must I disclose as conditional or fragile?

In this paper, we propose an epistemic assurance framework for proxy-driven disaster simulations. We use epistemic assurance to mean the combined evidence, drawn from both establishing the model's validity and testing the robustness of its conclusions, that determines which of those conclusions can be responsibly defended and under what assumptions. The framework distinguishes proxy, structural, and outcome validity as complementary properties of the model, providing a structured basis for assessing whether a particular implementation constitutes a defensible basis for inference. Building upon this foundation, epistemic robustness evaluates whether decision-relevant conclusions depend on the choice of representation of influential latent processes. Together, these complementary assessments support the qualification of model-based conclusions by identifying which are robust, which remain conditional on unresolved assumptions, and which are too fragile to support their intended purpose.

Our framework builds on the concept of equifinality, which holds that multiple model structures can be equally consistent with the available data [6, 7]. Equifinality, however, characterizes a property of the model space rather than prescribing what a modeler should claim once that non-uniqueness is acknowledged. Our central claim supplies that missing step: the credibility of proxy-driven simulations should not be evaluated solely by whether a model reproduces observed system behavior, but established through a structured argument linking model representations, mechanisms, empirical evidence, and the decision-relevant conclusions they support.

We therefore argue that simulation models should be accompanied not only by evidence supporting their validity, but also by an explicit statement describing which conclusions they can responsibly support, under what assumptions, and for which intended purposes.

The contribution of this framework is therefore not a new method for exploring uncertainty, but a shift in object and purpose: from characterizing how uncertain a model is, or which decision is robust, to qualifying which conclusions the modeler can responsibly assure. Where existing approaches are largely analytic or decision-analytic, the framework proposed here is a disclosure discipline owned by the modeler.

In this context, the framework is intended as a tool for modelers themselves, not as an external validation gate. Its purpose is to help conscientious developers identify, before publication or use, which of their conclusions are sufficiently supported to defend and which depend on unresolved representational choices. In this sense, it is a safeguard against inadvertent overclaiming, a risk that grows as models become more detailed and are applied to higher-consequence decisions, rather than a procedure for adjudicating whether a model is 'valid.' Just as a structural engineer scopes and stamps an analysis to define the conditions under which it can be relied upon, the framework lets a modeler state the scope within which their conclusions can be professionally defended.

2 The Epistemic Challenge of High-Resolution Disaster Simulations

Modern disaster simulation models increasingly represent interactions among heterogeneous households, infrastructure systems, institutions, markets, and adaptive behaviors to evaluate the societal consequences of extreme events [e.g., 8, 15, 16, 37]. These models are frequently used to reason about future events, evaluate alternative interventions, and compare policies whose outcomes cannot be directly observed or empirically validated in advance [e.g., 34, 51]. Consequently, complete validation of every modeled process is often unattainable. The challenge is therefore not simply to determine whether a model reproduces historical observations, but to establish a sufficient basis to infer that the decision-relevant conclusions it supports are credible.

Many of the socio-economic processes represented in disaster simulations remain only partially observable. Household relocation, financing access, institutional response, infrastructure interdependencies, workforce availability, and adaptive behavior are not often measured at the spatial and temporal resolution required for direct simulation (i.e., they are underdetermined) [e.g., 43, 44]. Model developers therefore rely on proxies and inferred mechanisms to represent these latent processes. Importantly, multiple defensible representations may exist for the same process, each preserving different properties of the underlying phenomenon (i.e., they are equifinal). These alternative representations can produce similar aggregate behavior while implying different relationships among households, infrastructure systems, institutions, and resources. Models may therefore appear empirically similar yet support contradictory conclusions about dominant mechanisms or effective interventions because they represent latent processes differently.

Underdetermination and equifinality are known challenges in the study of complex systems and processes [6, 7], but are underexplored in disaster simulations where they lead to ambiguity and directly affect the credibility of the conclusions drawn from a model. To formalize this challenge, consider a disaster impact simulation represented as

$$Y = f(\mathbf{X}, \mathbf{A}, \mathbf{M}, \theta, \varepsilon), \tag{1}$$

where $\mathbf{X}$ denotes observed or latent system attributes (e.g., household income or building characteristics), $\mathbf{A}$ denotes relationships or assignments among entities (e.g., household-building occupancy), $\mathbf{M}$ denotes modeled mechanisms (e.g., financing, relocation, or contractor allocation processes), θ denotes model parameters, and ε represents stochastic variability. The boldface indicates $\mathbf{X}$, $\mathbf{A}$, and $\mathbf{M}$ are not scalars and may have complex structures.

In practice, many components of $(\mathbf{X}, \mathbf{A}, \mathbf{M})$ cannot be directly observed at the required level of detail. Instead, the implemented simulation operates on representations of these latent processes,

$$\widetilde{Y} = f(\widetilde{\mathbf{X}}, \widetilde{\mathbf{A}}, \widetilde{\mathbf{M}}, \theta, \varepsilon), \tag{2}$$

where $\widetilde{\mathbf{X}}$, $\widetilde{\mathbf{A}}$, and $\widetilde{\mathbf{M}}$ denote representations of latent variables, relationships, and mechanisms. For example, $\widetilde{\mathbf{X}}$ may represent household income inferred from aggregate census data, $\widetilde{\mathbf{A}}$ may represent inferred household–building assignments, and $\widetilde{\mathbf{M}}$ may represent simplified behavioral or institutional processes. Different representations may preserve different properties of the underlying process while remaining equally defensible given the available evidence.

This distinction has important implications for model evaluation. Conventional uncertainty analysis typically assumes that the chosen representation is fixed and examines how uncertainty in parameters (θ) or stochastic variability (ε) propagates through the model. Proxy-driven disaster simulations introduce an additional source of epistemic uncertainty: uncertainty regarding how latent variables, relationships, and mechanisms should be represented in the first place. As a result, we cannot establish that these models are defensible solely through uncertainty quantification within a single representation. It also requires evaluating whether the chosen representations provide a defensible basis for inference and whether the resulting decision-relevant conclusions do not significantly change when defensible representations are considered. This distinction motivates the framework developed in the following sections.

2.1 Alternative Representations of Latent Processes

The central epistemic challenge in proxy-driven modelling is that proxies are not unique. Different representations of the same latent process may preserve distinct properties relevant to different analyses. For example, one representation may accurately reproduce aggregate population distributions, while another may better preserve household-level relationships or spatial dependencies. Both may be defensible for particular purposes, yet support different inferences about system behavior or intervention effectiveness. The question is therefore not whether proxies should be used, but how confidence can be established in the conclusions they support.

Alternative representations arise whenever available evidence does not uniquely determine how a latent process should be modeled. Differences may stem from alternative data sources, different methods for constructing proxies from the same data, competing theoretical explanations, or different choices regarding the level of abstraction. For example, household income may be inferred from census microdata, mortgage information, or property values. Even when using the same underlying census data, modelers may construct income proxies through random sampling, marginal matching, or probabilistic inference. Similarly, when adapting the Northridge-based model to future Los Angeles earthquake, household relocation might be represented using empirical Northridge observations, survey-informed behavioral models, utility-based decision rules, or machine-learning approximations trained on more recent disasters.

Not all alternative representations are equally well supported. Some may be poorly supported by available evidence or inconsistent with established theory, whereas others may constitute reasonable alternative explanations of the same latent process. Rather than assuming that all alternatives are equally plausible, we view them as *candidate representations* whose suitability should be evaluated relative to the intended inference. The following section introduces three complementary forms of validity that provide a structured basis for determining whether a particular representation constitutes a defensible basis for inference.

Table 1 summarizes common reasons why multiple candidate representations arise in proxy-driven disaster simulations. These sources do not establish that a representation is valid. Rather, they explain why multiple representations may legitimately require evaluation.

Table 1: Common sources of representational ambiguity in proxy-driven disaster simulations. These sources explain why multiple candidate representations may exist, but do not by themselves establish that a representation is suitable for a particular analysis.

Source of ambiguity	Description	Illustrative example
Empirical limitations	Available observations are insufficient to uniquely determine a single representation.	Multiple statistical models provide comparable explanations of post-disaster relocation behavior.
Competing theories	Alternative theoretical perspectives imply different mechanisms linking the same inputs to the same outcomes.	Place-attachment and opportunity-based theories suggest different representations of relocation decisions.
Alternative modeling approaches	Different modeling paradigms represent the same process in different ways.	Agent-based and system dynamics models represent recovery through different mechanisms.
Expert-informed assumptions	Representations derived from contradictory expert knowledge where systematic observations are limited.	Contractor allocation rules derived from interviews with reconstruction professionals.
Scale and abstraction	Different levels of abstraction preserve different properties of the same process.	Household-level and neighborhood-level financing models emphasize different mechanisms while remaining consistent with available evidence.

Processes that are weakly constrained by empirical evidence or theory generally admit a wider range of candidate representations than processes governed by well-established physical principles. Consequently, the degree of representational ambiguity varies across model components. Later in the paper, we argue that the validity of candidate representations should be evaluated before assessing the robustness of the resulting conclusions.

3 A Layered Framework for Establishing Epistemic Assurance in Proxy-Driven Simulations

Assessing whether a proxy-driven disaster simulation is a defensible basis for inference requires demonstrating that a model reproduces expected outcomes and that model-driven conclusions remain defensible when important latent processes admit multiple plausible representations. We therefore propose a structured framework that distinguishes two complementary objects of evaluation: the *model* and the *decision-relevant conclusions* it supports. We assess whether a proxy-driven disaster simulation provides a defensible basis for inference through three complementary forms of validity: proxy, structural, and outcome. Building upon this foundation, epistemic robustness evaluates whether the conclusions drawn from that model remain stable across defensible representations of influential latent processes. Here, we call the combination of a validity and robustness evaluation an *epistemic assurance assessment*.

The epistemic assurance assessment is intended as a professional self-assessment undertaken by modelers seeking to determine which conclusions they can responsibly defend, rather than as an adversarial procedure capable of preventing strategic misuse. It therefore presumes a good-faith effort to identify relevant validity requirements, influential latent processes, and defensible alternative representations, including those that may weaken preferred conclusions. A modeler could formally satisfy the framework by defining the intended purpose, validity criteria, candidate representations, or conclusions in a strategically narrow manner. Such an exercise, however, would not constitute epistemic assurance as defined here, because its purpose is not to maximize the number of conclusions that can be defended, but to expose foreseeable dependencies before they undermine a decision or the modeler's professional credibility. The framework makes these judgments explicit and contestable, but their integrity ultimately depends on the modeler treating assurance as a safeguard against overclaiming rather than as a means of legitimizing claims already chosen.

Our framework builds on a long tradition of thinking about model validity in the simulation literature, most directly the work of Sargent and Barlas [e.g., 5, 47]. Sargent distinguishes verification (building the model right) from validation (building the right model) and emphasizes that validity is established relative to a model's intended purpose rather than in the abstract, decomposing it into conceptual model validity, operational validity, and data validity. Barlas draws a complementary distinction between structure validity, whether the model's internal relationships correspond to those of the real system, and behavior validity, whether the model reproduces observed behavior patterns. The three forms of validity we propose can be read as a refinement of these ideas for proxy-driven disaster simulation models. Proxy and structural validity concern what Barlas calls structure and what Sargent locates in conceptual model validity, addressing whether individual proxies and their dependencies defensibly represent the latent processes at issue. Similarly, outcome validity corresponds to behaviour and operational validity, asking whether model outputs reproduce empirically and theoretically expected patterns.

Where our account departs from both is in the setting it addresses. In the disaster simulations we consider, the structure that Barlas would have us validate is itself only partially observable, so structure validity cannot be established by comparison against a known system and must instead be argued from theory and preserved properties across defensible alternative representations. Epistemic robustness extends Barlas' modified-behavior prediction testing by shifting the object of evaluation from the validity of the model to the stability of the decision-relevant conclusions it supports, which is the property that ultimately governs whether those conclusions can be responsibly defended. Therefore, the terminology we introduce is not intended to claim wholly novel forms of validity or robustness, but to adapt and distinguish established concepts in a way that is analytically useful for proxy-driven disaster simulations.

Figure 1 summarizes the epistemic assurance workflow. The workflow begins by defining the intended decision context and the conclusions the model is expected to support. For the motivating Northridge example, the intended conclusion might be whether Los Angeles will experience a temporary housing shortage following a future major earthquake. It then establishes whether the model provides a defensible basis for inference through the three complementary forms of validity, and, where warranted, evaluates the robustness of the resulting conclusions to alternative representations. The outcome is not simply an assessment of model quality, but a transparent qualification of the scope of defensible use, identifying the decisions for which the available evidence provides sufficient epistemic assurance and those for which additional robustness assessment or caution is required. Table 2 summarizes the distinct roles of each component within this process.

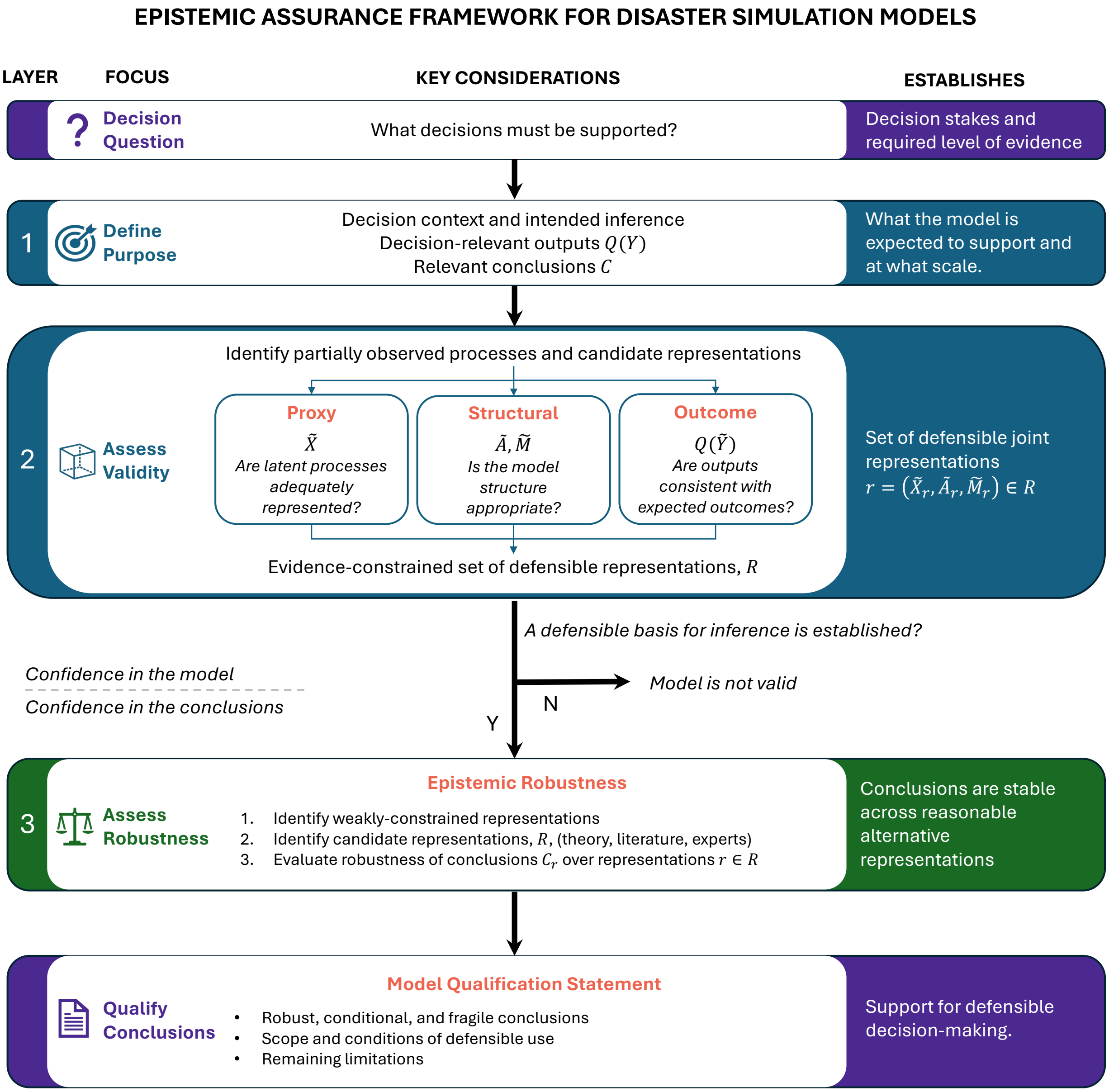

Figure 1: Epistemic assurance framework. Proxy, structural, and outcome validity evaluate whether different model implementations, i.e., a joint set of $\widetilde{\mathbf{X}}_r$, $\widetilde{\mathbf{A}}_r$, and $\widetilde{\mathbf{M}}_r$ are defensible bases for inference. Epistemic robustness builds on this foundation by evaluating whether conclusions depend on a particular representation of latent processes.

The principle that conclusions surviving across alternative formulations warrant greater confidence has a long lineage in the philosophy of science, where it is termed robustness analysis [30, 54, 55]. That literature explains why robust results are more trustworthy. Our contribution is to operationalize this principle for proxy-driven disaster

Table 2: Summary of the proposed framework.

Concept	Primary object of evaluation	Provides evidence for	Core question
Proxy validity	$\widetilde{\mathbf{X}}$	The model	Are the individual processes represented appropriately?
Structural validity	$\widetilde{\mathbf{A}}, \widetilde{\mathbf{M}}$	The model	Are the interactions among those processes represented appropriately?
Outcome validity	$Q(\widetilde{\mathbf{Y}})$	The model	Do model outputs reproduce empirically and theoretically expected patterns at the relevant scale?
Epistemic robustness	C_r across $r \in R$	Decision-relevant conclusions	Do decision-relevant conclusions change across defensible representations of latent processes?

simulations: specifying how the set of defensible representations is assembled (e.g., through systematic literature review), restricting it to representations that satisfy the validity requirements, and converting the result into a disclosed qualification of decision-relevant conclusions. Whereas robustness analysis asks whether a theoretical result is an artifact of modeling idealizations, our concern is whether the conclusions a model supports remain defensible across the representations, and how that determination is communicated.

Epistemic assurance claims inevitably require additional effort, including identifying alternative representations, implementing competing formulations, and evaluating their implications. However, using a model beyond what its supporting evidence justifies also carries a cost, increasing the risk that decision-makers place confidence in conclusions that depend on unresolved representational assumptions. The purpose of the proposed framework is not to maximize epistemic assurance but to establish a burden of evidence commensurate with the model's intended use. Just as engineering analyses require different levels of verification depending on the consequences of failure, the level of epistemic assurance required from a simulation should depend on both the degree of representational ambiguity and the consequences of the decisions it is intended to inform. Models dominated by well-constrained physical processes may require little additional robustness assessment, whereas models relying heavily on weakly constrained socio-economic processes or supporting high-consequence decisions may warrant a more extensive evaluation of defensible representations.

3.1 Proxy Validity

Proxy validity provides evidence that the variables and attributes used in a model constitute defensible representations of the latent processes they are intended to capture. Because many socio-economic disaster processes cannot be observed directly at the required spatial, temporal, or conceptual resolution, the objective is not to reproduce the underlying process exactly, but to preserve the properties required for the intended analysis.

Let $\mathbf{X}$ denote a target variable or attributes and $\widetilde{\mathbf{X}}$ its representation within the model. Let P denote the set of properties required for the intended inference, such as marginal distributions or rank ordering. A representation valid for a given purpose can be assessed by $\widetilde{\mathbf{X}} \sim_P \mathbf{X}$, where $\sim_P$ denotes similarity with respect to the properties in P. For example, a synthetic household-income variable may reproduce tract-level income distributions while failing to preserve relationships with tenure or housing characteristics. Such a representation may be adequate for estimating aggregate losses, yet inappropriate for evaluating distributional impacts or recovery inequality.

The proxy validity of a representation $\widetilde{\mathbf{X}_r}$ is, therefore, purpose-dependent. It provides evidence that the $\widetilde{\mathbf{X}_r}$ preserves the characteristics required for the intended conclusions, but does not, by itself, establish that the interactions among model components or the resulting conclusions are reliable.

3.2 Structural Validity

Structural validity provides evidence that the relationships, dependencies, and mechanisms connecting model components constitute a defensible representation of the system being studied. Whereas proxy validity evaluates individual variables, structural validity evaluates how they interact to produce system behavior.

In the context of Eq. 2, structural validity depends primarily on the representation of relationships and mechanisms through $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{M}}$. A model may contain realistic households, buildings, and infrastructure while still producing misleading conclusions if the dependencies among these entities are distorted. For example, a household–building assignment that weakens the relationship between income, tenure, and housing costs may substantially alter estimates of recovery inequality even when the marginal distributions of each variable remain realistic.

Structural validity provides evidence that the relationships, $\widetilde{\mathbf{A}}$, and mechanisms, $\widetilde{\mathbf{M}}$, connecting model components constitute a defensible representation of the system for the intended inference. Because these relationships and mechanisms are often only partially observable, structural validity does not generally establish that the particular representations $\widetilde{\mathbf{A}}_r$ and $\widetilde{\mathbf{M}}_r$ correspond uniquely to the true system structure. Instead, it evaluates whether they are supported by available empirical evidence, theory, domain knowledge, and established constraints, and whether they generate dependencies and system behavior consistent with what is known. When multiple structures satisfy these requirements, and available evidence cannot justify preferring one over the others, they remain defensible alternatives and should be retained in the set of defensible model representations for robustness assessment.

3.3 Outcome Validity

Outcome validity provides evidence that the behavior of the model is consistent with available empirical evidence and theoretical understanding. Rather than evaluating individual variables or mechanisms, outcome validity evaluates whether the model, under a given representation of $\widetilde{\mathbf{X}}$, $\widetilde{\mathbf{A}}$, and $\widetilde{\mathbf{M}}$, produces patterns that are plausible at the scale relevant to the intended application.

Let $Q(\widetilde{Y}(r))$, where $r = (\widetilde{\mathbf{X}}_r, \widetilde{\mathbf{A}}_r, \widetilde{\mathbf{M}}_r)$, denote a decision-relevant quantity derived from model outputs, such as median recovery time, infrastructure restoration, displacement rates, or inequality across population groups. Outcome validity requires that $Q(\widetilde{Y}(r))$ reproduce patterns observed in comparable disasters or remain consistent with established theoretical expectations.

Because outcome validity is often the most visible aspect of model evaluation, it is also the easiest to overinterpret. Similar aggregate patterns may emerge from substantially different underlying representations and mechanisms. Consequently, agreement with observed outcomes alone does not establish that the model provides a reliable basis for inference. Outcome validity should therefore be interpreted together with proxy and structural validity when assessing a model.

Outcome validity does not require setting-specific empirical data. Model outputs can be assessed against available empirical evidence, where it exists, as well as against well-established theoretical understanding and accumulated disaster experience. A model predicting that Los Angeles would fully recover from a major earthquake within two weeks would violate outcome validity, even in the absence of specific recovery data, because the prediction is inconsistent with robust expectations about post-disaster recovery timescales. Outcome validity, therefore, is applicable to transferred and prospective applications, where it is established against expected patterns and plausibility bounds rather than direct observation of the target.

Expectation-based outcome validity, however, is a coarse screen. It can rule out grossly implausible behavior but cannot discriminate among defensible representations that all produce outputs consistent with available evidence and expectations while supporting different decision-relevant conclusions. Outcome validity, whether established from data or from well-grounded expectations, is therefore necessary but not sufficient, and must be complemented by epistemic robustness, which evaluates whether the conclusion itself is stable across those equally plausible representations.

3.4 Epistemic Robustness

Whereas proxy, structural, and outcome validity establish that a model is a defensible basis for inference, epistemic robustness builds confidence in the decision-relevant conclusions supported by that model. We define epistemic robustness as the extent to which decision-relevant conclusions persist across defensible representations of uncertain or partially observed model components. Epistemic robustness is therefore a property of conclusions rather than models.

This distinction is fundamental. A model may provide a defensible representation of the system and satisfy all three forms of validity, yet still deliver fragile conclusions if those conclusions depend strongly on a particular representation of a weakly constrained latent process. Conversely, conclusions that remain stable across multiple defensible representations provide greater epistemic assurance because they depend less on any individual modeling assumption.

Only representations that satisfy the validity requirements appropriate for the intended analysis should be considered when evaluating epistemic robustness. Robustness is therefore assessed across alternative *defensible* representations rather than arbitrary model variations. Let R denote the set of defensible model representations, where each $r \in R$ satisfies the proxy, structural, and outcome validity requirements appropriate for the intended analysis. The objective is to gather sufficient evidence to determine whether the intended conclusions can be responsibly defended. Consider $Q(\widetilde{Y}(r))$, where $\widetilde{Y}(r)$ is the model output produced under representation r, and $Q(\cdot)$ extracts the decision-relevant quantity. Different conclusions C_r drawn from the same model may exhibit different levels of epistemic assurance if $Q(\widetilde{Y}(r))$ is inconsistent across different r. Robustness assessment, therefore, may need to be evaluated across different sets $\{C_r : r \in R\}$.

Note that model-supported conclusions C_r must be drawn from the statistics of $Q(\widetilde{Y}(r))$, rather than point estimates. That is, a large ensemble of stochastic realizations of $Q(\widetilde{Y}(r))$ must be run and summarized by its location and scale, so that comparisons across representations are between distributions rather than single outcomes. Fragility thus refers to differences across representations that are evident at the level of these distributions, not to within-representation stochastic variability.

3.5 Qualification of Decision-relevant Conclusions

The purpose of robustness assessment is not to determine whether a model is universally robust, but to characterize the degree to which individual conclusions can be supported for their intended use. For example, a model may consistently rank alternative recovery policies across defensible representations while producing substantially different estimates of absolute recovery time. In this case, the policy ranking may be considered robust, whereas conclusions about recovery duration should be interpreted with caution.

The outcome of a robustness assessment is therefore a model qualification statement that describes the model's purpose and the degree to which individual conclusions can be supported, rather than a binary judgment of the model itself. A conclusion may be considered *robust* when it remains sufficiently stable across the defensible representations examined that the intended inference or decision is unaffected. It may be considered *conditional* when representational uncertainty influences the conclusion, but not enough to undermine its intended use. In such cases, the conclusion remains useful provided that its limitations are clearly communicated. Finally, a conclusion is considered *fragile* when defensible representations alter the intended inference or decision, and the available evidence is insufficient to justify preferring one representation over the others. Fragile conclusions do not imply that the model is invalid. Rather, they indicate that the current state of knowledge is insufficient to support that particular conclusion for the intended purpose.

Table 3 summarizes several forms of conclusion-specific robustness that may be relevant for different decision contexts. Rather than prescribing universal quantitative thresholds, the table identifies the types of evidence required to qualify common forms of decision-relevant conclusions. Different applications require different forms of robustness

Model Qualification Statement (template)

Intended purpose and decision context. The decision the model is used to support, and the scale(s) at which conclusions are drawn.

Decision-relevant conclusions. The specific conclusions C the model is intended to support.

Validity basis. Brief evidence for proxy, structural, and outcome validity appropriate to the intended analysis.

Influential latent processes and candidate representations. The weakly constrained processes examined and the defensible alternative representations R considered, with their source (e.g., systematic literature review).

Robustness findings. Each conclusion classified as *robust*, *conditional*, or *fragile* across the representations examined.

Scope of defensible use. The decisions the model may, and may not, be used to support.

Remaining limitations and recommended caution. Outstanding dependencies and the evidence needed to resolve them.

Figure 2: Template for a Model Qualification Statement. Each field corresponds to a component of the epistemic assurance framework.

depending on the intended inference and the decisions the model is expected to inform. For example, policy evaluation may primarily require ranking robustness, whereas planning analyses may depend more heavily on mechanism or interpretive robustness. The appropriate level of epistemic assurance should therefore be determined by the conclusions being supported rather than by the model as a whole.

Table 3: Illustrative robustness criteria for qualifying decision-relevant conclusions. The required evidence depends on the intended inference rather than on universal quantitative thresholds.

Conclusion type	Robustness question	Evidence supporting robustness
Metric	Are the estimated magnitudes sufficiently stable for the intended use?	Estimated values remain within a range that does not alter the intended inference or decision.
Ranking	Does the preferred ordering of policies, locations, groups, or scenarios remain unchanged?	The preferred ranking is preserved across the defensible representations examined.
Mechanism	Are the dominant drivers and dependencies preserved?	The same principal mechanisms are identified across the defensible representations examined.
Interpretation	Does the broader qualitative interpretation remain unchanged?	The principal interpretation remains unchanged despite different defensible representations.

4 Revisiting the Opening Question and Establishing Epistemic Assurance for Decision Support

We now revisit the motivating question posed in the introduction:

> *Should the research team assure decision makers that they can act on the Northridge-based model's recommendations simply because it reproduces a well-documented disaster recovery case?*

The proposed framework suggests that the answer is neither an automatic yes nor no. Rather, the Northridge-based model should first be evaluated to determine whether it provides a defensible basis for the intended inference and whether its decision-relevant conclusions remain robust across defensible representations. The following hypothetical example illustrates how the proposed workflow could be applied in practice. The following scenario serves as a step-by-step conceptual walkthrough of the framework. The subsequent section provides empirical demonstrations using other disasters.

Suppose that emergency planners in Los Angeles wish to evaluate whether the region would experience a shortage of temporary housing following a future major earthquake. The decision question is therefore:

> *Does Los Angeles require additional temporary housing capacity following a future major earthquake?*

Following the proposed framework, the first step is to define the intended purpose of the analysis. The decision context is a present-day earthquake near Los Angeles. The decision-relevant model outputs, $Q(\widetilde{Y}(r))$, include the number of displaced households, the number of undamaged dwellings that could provide temporary accommodation, and the duration of unmet temporary housing demand. These outputs support the principal conclusion, C_r, that Los Angeles is (or is not) expected to experience a temporary housing shortage following the event.

The second step is to establish a defensible basis for inference through the three complementary forms of validity. The modeler first identifies influential latent processes that remain weakly constrained by empirical evidence, e.g., contractor allocation and household relocation behavior. For each process, the literature is systematically reviewed to identify proposed alternative representations and to evaluate the empirical, theoretical, or practical justification supporting each. Appendix A illustrates this procedure by compiling alternative representations reported in the disaster recovery literature for several common latent processes.

Validity checks are then conducted to identify *defensible* alternative representations that capture the principal ways the process has been conceptualized. Proxy validity considers whether the variables and attributes in the original model remain appropriate for Los Angeles. Structural validity assesses whether the model mechanisms developed for the 1994 Northridge Earthquake adequately represent modern counterparts. Outcome validity assesses whether the adapted model reproduces observations or theory relevant to the intended purpose. In the Los Angeles application, outcome validity must be assessed not only against observations from the 1994 Northridge Earthquake but against established expectations about plausible recovery patterns and magnitudes for an event of this kind. Alternative joint representations $r = (\widetilde{\mathbf{X}}_r, \widetilde{\mathbf{A}}_r, \widetilde{\mathbf{M}}_r)$ that remain defensible in light of the validity assessment are collected into the set R, which provides the basis for the next step.

The third step evaluates epistemic robustness of conclusions across R. The objective of this exercise is not to assess the robustness of the model itself, but to determine whether the model's recommendations remain stable across these alternative representations. The model is run multiple times for each joint defensible representation $r = (\widetilde{\mathbf{X}}_r, \widetilde{\mathbf{A}}_r, \widetilde{\mathbf{M}}_r) \in R$, and the statistics of $Q(\widetilde{Y}(r))$ are collected. The level of robustness of the decision-relevant conclusions C_r is assessed as the degree to which different r support different C_r. For example, the number of displaced households may prove robust because estimates remain sufficiently consistent to support planning decisions. Estimates of available temporary accommodation may be conditional if they depend on assumptions regarding household sharing behavior, while the estimated duration of temporary housing shortages may prove fragile if different defensible representations produce materially different recovery timelines and available evidence cannot justify preferring one representation over another.

The resulting Model Qualification Statement communicates the scope of the model's defensible use rather than a single measure of model quality. For example, the modeler may conclude that the model provides robust evidence that Los Angeles is likely to experience a temporary housing shortage following a future earthquake, while simultaneously stating that estimates of the duration of that shortage remain conditional on assumptions. Conclusions regarding neighborhood-level recovery timing may remain fragile and, therefore, should not be used to support planning decisions without additional evidence.

As anticipated in the introduction, it is worth distinguishing two distinct sources of uncertainty in this illustrative application: contextual drift (e.g., shifts in demographics, housing stock, or building codes between 1994 and the present) and representational ambiguity (e.g., competing, defensible representations of latent household behaviors). While updating baseline datasets addresses contextual drift, the proposed epistemic assurance framework targets representational ambiguity. It demonstrates that even if contextual drift were completely eliminated by holding system attributes identical, alternative defensible representations of latent socio-economic processes can still produce fragile policy conclusions.

5 Using the Epistemic Assurance Framework to Identify Fragile Conclusions in Existing Disaster Simulations

The following examples illustrate how the framework can be applied to different classes of proxy-driven disaster simulations. Rather than providing complete model evaluations, they demonstrate how validity arguments and epistemic robustness are used in an epistemic assurance assessment, which highlights fragility in the models' decision-relevant conclusions.

5.1 Aggregate Agreement Can Mask Divergent Local Recovery Dynamics

The first example applies the proposed framework to the *RAAbIT* housing recovery simulation model [15]. Rather than evaluating the model itself, the objective is to examine whether its decision-relevant conclusions remain stable across defensible representations of an influential latent process. In this example, the latent process is contractor allocation, for which multiple plausible representations exist despite limited empirical evidence.

RAAbIT simulates post-fire housing recovery following the 2017 Tubbs Fire in Santa Rosa, California. Two alternative contractor-allocation representations are considered. The first preferentially allocates contractors within spatially concentrated rebuilding areas using an urban-density heuristic, whereas the second allocates contractors sequentially using a first-come, first-served (FCFS) process without explicitly preserving spatial reconstruction patterns. Both representations are theoretically plausible, produce realistic regional recovery trajectories, and satisfy the proxy, structural, and outcome validity requirements appropriate for the intended analysis. The proposed framework therefore treats both as defensible representations suitable for robustness assessment.

Figure 3 summarizes the resulting simulations. At the aggregate level, both representations reproduce similar rebuilding trajectories and provide comparable agreement with observed regional recovery curves (Figures 3.a-b). If the intended conclusion were limited to estimating regional rebuilding trends, the framework would therefore identify this conclusion as robust because the alternative representations produce essentially the same inference.

A different picture emerges when examining building- and neighborhood-scale recovery. The urban-density heuristic reproduces observed reconstruction timing better (Figures 3.c-d), and the two contractor-allocation mechanisms produce materially different spatial recovery dynamics despite similar aggregate behavior. This difference becomes even more apparent in Figures 3.e–g, which show that alternative representations generate spatially structured differences in estimated recovery duration and identify neighborhoods whose predicted recovery timing depends strongly on the contractor-allocation mechanism adopted. In this hindcast, observed building-level reconstruction timing provides stronger support for the urban-density representation. However, such evidence is available only because the event has already occurred. In a prospective application, comparable outcome data would not exist, and the available pre-event evidence may be insufficient to exclude FCFS as a defensible representation.

Viewed through the proposed framework, these results reveal that different decision-relevant conclusions exhibit different levels of epistemic assurance. Aggregate recovery trajectories remain robust because they are largely invariant across the alternative contractor-allocation representations examined. In contrast, neighborhood-level recovery timing and intervention prioritization are fragile because they depend on the particular contractor-allocation represen-

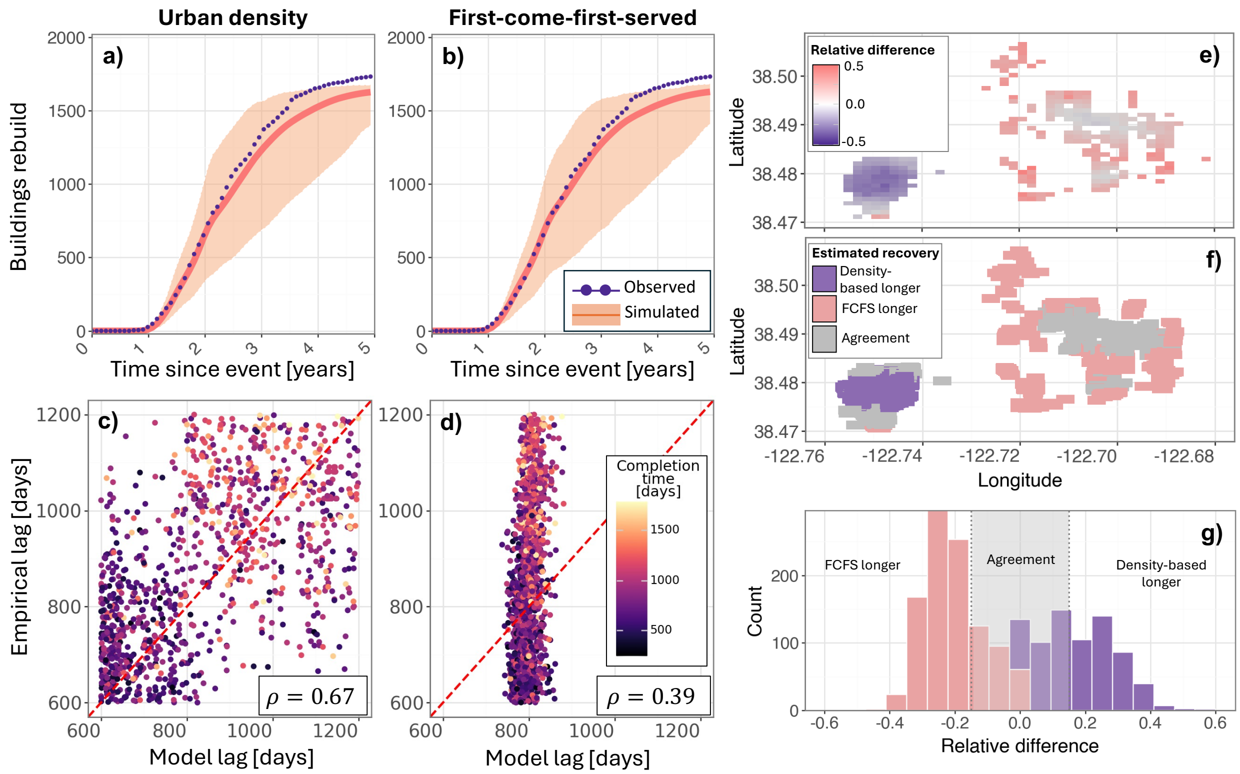

Figure 3: Comparison of two alternative contractor-allocation representations in the RAAbIT housing recovery model using empirical rebuilding data from Santa Rosa, California, following the 2017 Tubbs Fire. Panels (a–b) compare aggregate rebuilding trajectories under urban-density and first-come-first-served (FCFS) contractor allocation. Panels (c–d) compare simulated and observed building-level reconstruction timing. Panel (e) shows the normalized spatial difference between the two representations, panel (f) identifies locations where estimated recovery durations agree or differ substantially, and panel (g) summarizes the distribution of spatial differences.

tation adopted, and the available evidence does not justify preferring one representation over the other. Consequently, although the model provides strong support for regional recovery assessment, its use for neighborhood-level planning or prioritization should be accompanied by explicit qualification regarding the influence of contractor-allocation assumptions. These findings should be reported as in Fig. 4 using the Model Qualification Statement introduced in Section 3.5.

Model Qualification Statement: RAAbIT 2017 Tubbs Fire housing recovery

Intended purpose. Estimate post-fire housing recovery in Santa Rosa, CA, to support recovery assessment and planning at regional and neighborhood scales.

Decision-relevant conclusions. (C1) Regional rebuilding trajectory and pace of recovery; (C2) neighborhood-level recovery timing and prioritization of intervention areas.

Validity basis. *Proxy*: housing stock, damage, and rebuilding inferred from parcel and permit records for the affected area. *Structural*: contractor-mediated rebuilding dependencies represented. *Outcome*: simulated regional rebuilding reproduces observed recovery curves (Fig. 3a-b); both representations are retained as defensible for prospective use, although the hindcast indicates that the urban-density heuristic better reproduces observed local recovery timing.

Influential latent process examined. Contractor allocation, weakly constrained by empirical evidence. Two defensible representations drawn from the literature: an urban-density heuristic and a first-come, first-served (FCFS) process.

Robustness findings. *C1 (regional trajectory):* **robust** — essentially invariant across both representations. *C2 (neighborhood-level timing and prioritization):* **fragile** — the two representations produce materially different spatial recovery dynamics (Fig. 3c-g), and available evidence does not justify preferring one representation over the other.

Scope of defensible use. The model provides strong support for regional recovery assessment. It should not be used for neighborhood-level recovery timing or intervention prioritization without additional evidence constraining the contractor-allocation mechanism.

Remaining limitations. Sub-regional conclusions are fragile and depend on the contractor-allocation representation; resolving this fragility requires empirical data on how contractors are allocated after disasters.

Figure 4: Model Qualification Statement for the RAAbIT example. The model supports the regional conclusion (C1) while explicitly declining to vouch for the neighborhood-level conclusion (C2), protecting both the decision and the modeler's credibility.

5.2 Parameter Sensitivity Can Depend on Model Representation

The second example illustrates that conclusions drawn from conventional sensitivity analyses may themselves depend on higher-level representational assumptions. The example revisits the RecovUS agent-based recovery model developed by Moradi and Nejat [37], focusing on the influence of flood insurance penetration under two alternative representations of insurance eligibility. The objective is to examine whether conclusions regarding the importance of insurance penetration remain robust across defensible representations of the institutional context.

Both analyses vary insurance penetration from 10% to 100%, but differ in how they represent insurance eligibility. In Figure 5.a, insurance is available only to properties located within FEMA-designated high-risk flood zones, reflecting the contemporary implementation of the National Flood Insurance Program (NFIP). In Figure 5.b, insurance eligibility is extended to all households, representing a plausible alternative policy in which all homeowners are eligible for flood insurance compensation. Both representations are defensible for their respective policy contexts, but they imply fundamentally different institutional boundaries governing access to financial recovery resources.

Under the NFIP representation (Figure 5.a), insurance penetration appears to have little influence on community recovery. Across the full range of penetration values, recovery trajectories differ by only approximately one percentage point after 24 months because only a small fraction of damaged properties are eligible for insurance. Under the alternative representation (Figure 5.b), however, the identical penetration parameter becomes highly influential. Differences between the lowest and highest penetration scenarios reach approximately 22 percentage points after six months and remain close to 10 percentage points after two years.

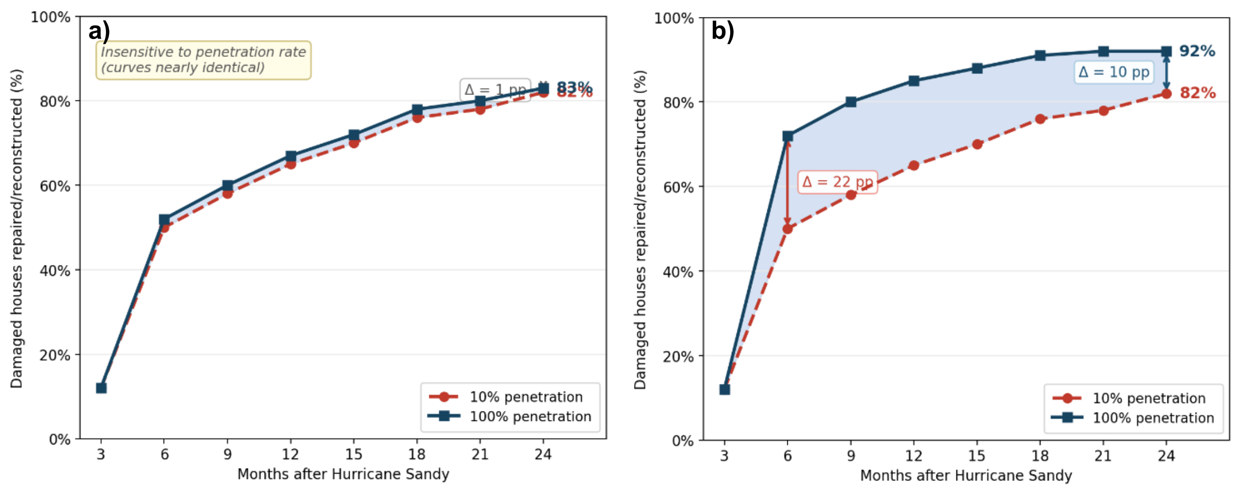

Figure 5: Community recovery trajectories under alternative representations of flood insurance eligibility in the RecovUS agent-based model (adapted from Moradi and Nejat [37]). Panel (a) restricts insurance eligibility to FEMA high-risk flood zones, whereas panel (b) extends eligibility to all households. In both cases, insurance penetration varies from 10% to 100%.

Viewed through the proposed framework, the decision-relevant conclusion is not the numerical value of the penetration parameter itself, but whether increasing insurance penetration is expected to accelerate community recovery. This conclusion is fragile due to its dependency on the representation of the insurance system. A conventional sensitivity analysis performed under the original NFIP representation correctly characterizes parameter uncertainty within that representation, yet would likely conclude that insurance penetration has little influence on recovery. An equally defensible representation of institutional eligibility leads to the opposite conclusion.

This example, therefore, demonstrates that parameter sensitivity is not an intrinsic property of a model but may itself depend on higher-level representational assumptions. Consequently, conclusions regarding the effectiveness of insurance-based recovery policies should be qualified by the institutional assumptions under which they were derived. More broadly, the framework identifies the representation of insurance eligibility (not the insurance penetration parameter itself) as the weakly constrained process responsible for the fragility of the resulting policy conclusions. Consequently, the Model Qualification Statement for this example should be as shown in Fig. 6.

Model Qualification Statement: RecovUS flood insurance penetration

Intended purpose. Assess whether increasing flood insurance penetration accelerates community recovery, to inform insurance-based recovery policy.

Decision-relevant conclusion. (C1) The effect of increasing insurance penetration on the pace of community recovery.

Validity basis. *Proxy*: household damage, financing, and recovery decisions represented as in the RecovUS agent-based model [37]. *Structural*: insurance compensation linked to household recovery through eligibility and penetration. *Outcome*: recovery trajectories remain consistent with expected behavior under each institutional representation (Fig. 5a–b); both eligibility representations are defensible for their respective policy contexts.

Influential latent process examined. The representation of insurance *eligibility* — the institutional boundary governing access to financial recovery resources — rather than the insurance-penetration parameter itself. Two defensible representations: eligibility restricted to FEMA high-risk flood zones (current NFIP implementation) and eligibility extended to all households (a plausible alternative policy).

Robustness findings. *C1:* **fragile** on the eligibility representation. Under the NFIP representation, penetration has little influence on recovery (recovery differs by $\sim$1 percentage point at 24 months, Fig. 5a). Under universal eligibility, the identical parameter is highly influential ($\sim$22 percentage points at 6 months, $\sim$10 at 24 months, Fig. 5b). The two defensible representations support opposite conclusions, and available evidence does not resolve which institutional context applies.

Scope of defensible use. Conclusions about the effectiveness of insurance-based recovery policy must be stated together with the institutional eligibility regime under which they were derived. They cannot be reported as a general property of the model or of insurance penetration.

Remaining limitations. Parameter sensitivity here is not intrinsic to the model but depends on a higher-level representational assumption. The weakly constrained process is the eligibility representation; resolving the fragility requires specifying the institutional and policy context the analysis is intended to inform.

Figure 6: Model Qualification Statement for the RecovUS example. The effect of insurance penetration is defensible only when qualified by the institutional eligibility assumption under which it was derived; reported without that qualification, it would mislead.

6 Discussion

The examples in this paper share a common lesson: conclusions drawn from the same model, validated to the same standard, can carry very different levels of assurance. This motivates the paper's principal contribution - separating confidence in a model implementation from confidence in the decision-relevant conclusions it supports. Validation is therefore not the endpoint of model development. Validity establishes that a model is a defensible basis for inference, and epistemic robustness asks whether a given conclusion remains stable across the defensible representations of the latent processes it depends on. Read together, they allow a developer to state not whether a model is 'validated,' but which of its conclusions can be responsibly defended, for what purpose, and under what assumptions.

This perspective reframes the role of simulation models in disaster risk assessment. Models should not be viewed as generators of universally valid predictions, but as decision-support tools whose conclusions require different levels of evidence depending on their intended use. Two models exhibiting similar agreement with historical observations may provide substantially different levels of epistemic assurance if one supports conclusions that remain stable across defensible representations while the other depends critically on unresolved modeling assumptions. The objective, therefore, shifts from demonstrating that a model is correct to establishing the scope within which its conclusions are professionally defensible.

This perspective also changes how representational fidelity should be interpreted. Increasing spatial resolution, behavioral realism, or institutional detail does not automatically increase confidence in a model's conclusions. Greater fidelity often introduces additional latent processes whose behaviors are only weakly constrained by empirical evidence, thereby expanding the space of defensible representations that should be considered. Consequently, representational fidelity and epistemic assurance should be viewed as complementary rather than synonymous objectives.

The proposed framework likewise complements rather than replaces existing approaches to uncertainty analysis. Sensitivity analysis, uncertainty propagation, exploratory modeling, and robust decision making remain essential tools for understanding model behavior and supporting decisions under uncertainty. Table 4 situates the proposed framework among established approaches for reasoning about model credibility under uncertainty. Rather than ordering these approaches by sophistication or scope, it distinguishes them by what they evaluate and whom they are primarily intended to serve. Sensitivity analysis, Exploratory Modelling, Robust Decision Making, Pattern-Oriented Modeling, and the equifinality literature each illuminate a different facet of model behavior or decision performance, and each remains essential for the purposes it was designed to serve. What none of them provides is a means for the modeler to determine, and then disclose, which of their own conclusions can be responsibly assured. Epistemic assurance occupies that role: it takes the outputs of these methods as inputs and shifts the object of evaluation from the model or the decision to the claim, and the user from an external or decision-side analyst to the modeler accounting for their own work.

Approach	Primary question	Primary user / object
Sensitivity analysis	How do outputs respond to parameter uncertainty?	Modeler / model behavior
Exploratory Modelling	What outcomes are possible across structures and futures?	Analyst / outcome space
Robust Decision Making	Which decision performs well across plausible futures?	Decision analyst / the decision
Pattern-Oriented Modeling	Do representations reproduce characteristic patterns?	Modeler / model credibility
Equifinality / GLUE	Do many defensible models fit equally well?	Modeler / diagnosis of non-uniqueness
Epistemic assurance (this work)	Which conclusions can the modeler responsibly assure, and what must be disclosed?	Modeler / the claim

Table 4: Positioning of the proposed framework relative to established approaches for reasoning about model credibility under uncertainty, distinguished by the question each addresses and its primary user or object of evaluation. Existing approaches are diagnostic or decision-analytic; epistemic assurance shifts the object to the modeler's own conclusions and to what must be disclosed.

6.1 Implications for Modeling Practice

The primary beneficiary of this framework is the model developer. Rather than defending a model against a standard of complete validation, a developer can proactively state the purpose of the analysis, the evidence supporting its key representations, the robustness of its conclusions across defensible alternatives, and the resulting scope of defensible use. This protects both the credibility of the science and the decisions that rely on it, by ensuring claims do not outrun their evidentiary basis.

For reviewers, the framework shifts attention away from evaluating individual assumptions in isolation toward assessing whether the evidence presented is sufficient to support the claimed conclusions. This perspective reduces unrealistic expectations of complete model validation while encouraging greater transparency regarding the assumptions that most influence decision-relevant inferences. It also provides a common language for distinguishing between robust, conditional, and fragile conclusions without requiring universal quantitative thresholds.

Beyond the evaluation of individual studies, the framework may also help identify recurring sources of representational fragility across the disaster modeling literature. If important conclusions consistently depend on particular proxy choices or weakly constrained mechanisms, these become priorities for future empirical research, benchmark datasets, model intercomparison exercises, and methodological development. In this sense, epistemic robustness serves not only as an evaluation tool but also as a mechanism for directing scientific effort toward reducing the uncertainties that most limit confidence in decision-relevant conclusions.

6.2 Limitations and Future Directions

The proposed framework is intentionally descriptive rather than prescriptive. It does not prescribe universal thresholds for validity, robustness, or epistemic assurance because the level of evidence required depends on the model's intended purpose, the degree of representational ambiguity, and the consequences of the decisions it supports. Future work could develop application-specific guidance describing appropriate levels of epistemic assurance for different classes of disaster simulation and decision contexts.

Another limitation concerns the construction of the set of defensible representations R. In the examples presented, R is assembled primarily through a systematic review of the literature, which limits it to representations already published in the field. Three consequences follow. First, the review inherits the biases of the published record, including publication bias and the influence of dominant modeling paradigms. Second, genuinely novel target contexts may have no published representation of a relevant process. Third, the assurance the framework provides is scope-dependent: a conclusion shown to be robust is robust across the defensible representations known to the field at the time of analysis, not across all conceivable representations.

We regard this as a bound to be disclosed rather than a flaw to be hidden. It mirrors established engineering practice, in which analyses are checked against a codified set of cases rather than every conceivable one, and it is consistent with the framework's purpose of qualifying, rather than guaranteeing, the conclusions a model can support. Where the literature is sparse, R can be supplemented through expert elicitation, structured analogs from related events, and participatory or stakeholder input, while maintaining transparency about the assumptions these sources introduce. Finally, the absence of a credible representation for an influential process is itself informative: it identifies a priority for empirical research, benchmark development, and model intercomparison.

Although the examples presented here focus on post-disaster housing recovery, the proposed framework is intended to apply broadly to proxy-driven disaster simulations, including infrastructure restoration, evacuation, climate adaptation planning, public health, and other applications involving latent human and institutional processes. Evaluating the framework across these domains will help determine whether different application areas require distinct forms of validity evidence, robustness assessment, or model qualification.

Finally, the framework intentionally remains agnostic regarding the specific methods used to construct and explore the space of defensible representations. Developing systematic approaches for identifying influential latent processes, constructing representative alternative formulations, and determining when sufficient epistemic assurance has been achieved represents an important direction for future research.

7 Conclusion

Disaster simulations are increasingly expected to support decisions about future hazards, adaptation strategies, recovery processes, and other socio-technical systems whose behavior cannot be fully observed or empirically validated in advance. As these models become more detailed, they necessarily rely on proxies and inferred representations of latent processes that remain only partially constrained by empirical evidence. For this class of models, validation alone cannot establish whether their decision-relevant conclusions are sufficiently supported for the purposes they are intended to inform.

This paper proposes an epistemic assurance framework for proxy-driven disaster simulations that distinguishes between confidence in a model implementation and confidence in the conclusions it supports. Proxy, structural, and outcome validity establish that a model provides a defensible basis for inference. Building upon this foundation, epistemic robustness evaluates whether decision-relevant conclusions depend on a particular representation of influential latent processes. Together, these complementary assessments provide a structured basis for determining what conclusions can be responsibly supported, under what assumptions, and for which intended purposes.

Rather than viewing robustness assessment as an additional burden on model developers, the proposed framework reframes it as a means of establishing and communicating the scope of the model's defensible use. Just as engineers routinely communicate the assumptions and limitations associated with physical analyses, developers of socio-technical simulation models should be able to identify which conclusions are robust, which remain conditional on unresolved assumptions, and which are too fragile to support the intended decision. In doing so, the framework shifts the emphasis from demonstrating that a model is universally "validated" to providing an explicit epistemic assurance case for its appropriate use.

Ultimately, the goal of proxy-driven disaster simulations is not to eliminate uncertainty or produce perfectly validated predictions of inherently uncertain futures, but to support better decisions despite persistent uncertainty. We argue that this objective is best achieved not only by asking whether a model has been validated, but also by establishing whether the conclusions that inform decisions can be responsibly defended. In this sense, the proposed framework provides a pathway to move from model validation to epistemically assured decision support.

Appendix

A Examples of Alternative Representations of Latent Process in Disaster Recovery Simulation

Table 5 illustrates the process of identifying *alternative* representations for influential latent processes in socio-economic disaster simulations. The representations listed are not intended to be exhaustive nor prescriptive. Rather, they demonstrate how the literature can be systematically reviewed to identify competing formulations that may be appropriate for robustness assessment. The set of *defensible* representations can be obtained from this list as empirical evidence accumulates, competing approaches are evaluated across applications, and stronger theoretical and empirical understanding emerges regarding the properties and dependencies that must be preserved for different classes of decision-relevant conclusions.

Table 5: Literature-derived examples of alternative representation uncertainty in computational housing and household recovery models. The table summarizes abstractions that have appeared in the literature for latent processes that are incompletely observed and not uniquely specified by accepted theory. A validity assessment must evaluate whether they are defensible representations.

Latent recovery process	Reasons for ambiguity	Alternative representations synthesized from the literature
Financing	Funding pathways combine insurance, grants, loans, savings, personal resources, public assistance, and other formal and informal sources whose timing, eligibility, adequacy, and procurement are only partially observable.	• *Financing assumed immediately available or absorbed into generic recovery time* [3, 10, 31]. • *Condition- or delay-based financing availability*, where funds become available after income, insurance, eligibility, or delay conditions are satisfied [14, 37, 52]. • *Process-based financing acquisition*, where multiple funding sources are represented through probabilities of receiving funds, amounts received, source interactions, and disbursement times [2]. • *Type–tenure-specific financing*, where available sources and recovery pathways differ for owner-occupied, renter-occupied, single-family, and multi-family housing [36].
Repair initiation	The start of repair is governed by partially observed administrative, financial, engineering, and logistical conditions. Models must decide whether repairs can begin immediately, only after prerequisites are met, or while some prerequisites are addressed in parallel.	• *Immediate repair initiation*, where repair begins once damage and repair time are known and no explicit pre-repair process is represented [31, 32]. • *Prerequisite-based repair initiation*, where repair waits for impeding factors such as inspection, engineering assessment, financing, permitting, contractor mobilization, or long-lead components [1, 3, 8, 11, 14, 24, 52]. • *Partially parallel prerequisite completion*, where some prerequisites proceed concurrently and initiation depends on the maximum or combination of delays rather than a strictly serial chain [2, 8, 14].

Latent recovery process	Reasons for ambiguity	Alternative representations synthesized from the literature
Household recovery decision-making	Household preferences, expectations, constraints, risk perceptions, social ties, and relocation intentions are not directly observable and may change during recovery. Models therefore require behavioral assumptions about whether households repair, rebuild, wait, relocate, sell, or return.	• *Rule- or threshold-based decisions*, where households act when specified conditions, such as sufficient financing or adequate community recovery, are met [14, 37]. • *Statistical recovery propensity*, where reconstruction or recovery probability is inferred from observed household, neighborhood, damage, and spatial attributes [39]. • *Utility-, objective-, or strategy-based decisions*, where households or stakeholders evaluate alternatives according to specified objectives, expected benefits, or equilibrium strategies [10, 20, 38]. • *Neighbor- or context-dependent decisions*, where household decisions respond to perceived signals from neighbors, community assets, infrastructure, employment access, or other recovery conditions [21, 37, 38]. • *Type–tenure-specific decision-maker*, where the relevant recovery actor differs across owner-occupied, renter-occupied, single-family, and multi-family housing [36].

Latent recovery process	Reasons for ambiguity	Alternative representations synthesized from the literature
Construction resource allocation	Contractor labor, materials, equipment, inspectors, and repair crews may become scarce after large disasters, but household-level access to these resources is rarely observed. Once scarcity is represented, the model must assume how limited resources are allocated among competing recovery needs.	• *No competition for construction resources*, where resources are assumed sufficient and do not constrain the recovery trajectory [10, 31]. • *Scarcity with first-come-first-served allocation*, where households or buildings receive available services in order of arrival or readiness [14, 24]. • *Aggregate capacity constraint*, where recovery is slowed by finite labor, contractor, material, or repair capacity without detailed household-level allocation [1, 13, 19, 27, 51]. • *Policy- or priority-modified allocation*, where public intervention, resource redistribution, prioritization, or targeted programs modify who receives scarce resources first [2, 40].
Building recovery progression	The physical state of a damaged housing unit is only partially observable over time, and models differ in whether recovery is represented as repair completion, functional restoration, occupancy safety, or transitions across discrete damage/functionality states.	• *Repair-time progression*, where damaged buildings recover after a sampled or calculated repair duration [14, 52]. • *Engineering-functionality progression*, where recovery is represented through re-occupancy, functional recovery, or partial/full functionality states [1, 9, 10, 11, 35]. • *Discrete-state stochastic progression*, where buildings transition among functionality states using stochastic state-transition models [31, 32]. • *Portfolio aggregation*, where individual building recovery states are aggregated to characterize neighborhood or community building-stock functionality [10, 31, 32, 52].

Latent recovery process	Reasons for ambiguity	Alternative representations synthesized from the literature
Household occupancy transitions	Where households live while permanent housing recovers is not fully observed, and temporary arrangements may include emergency shelter, temporary shelter, temporary housing, rental substitution, out-of-community relocation, or return to permanent housing.	• *Stage-based occupancy progression*, where households move through emergency shelter, temporary shelter, temporary housing, and permanent housing stages [8, 49, 50]. • *Social-vulnerability-conditioned transitions*, where transition probabilities or durations vary with vulnerability, damage, or household characteristics [8, 50]. • *Temporary-housing demand and supply*, where displaced households seek limited temporary accommodation while permanent housing recovers [51]. • *Resident–contractor temporary-housing competition*, where out-of-town reconstruction workers also demand temporary housing and affect displaced households' options [51].
Institutional assistance and unmet need	Public assistance depends on administrative definitions, application behavior, eligibility, data availability, and agency-specific rules. The true distribution of unmet need is not directly observable shortly after disasters.	• *Program eligibility representation*, where assistance is assigned according to program-specific eligibility, caps, and loss definitions [2, 36]. • *Unmet-need estimation*, where remaining need is computed as losses minus expected assistance from insurance, FEMA, SBA, HUD, or other sources [12, 36]. • *Approval-rate/amount modeling*, where expected assistance is estimated probabilistically from applicant, hazard, state, ownership, income, insurance, and residence-type attributes [12].

Data and code availability

NA

Competing interests

NA

References

[1] Mohammad Aghababaei and Maria Koliou. Community resilience assessment via agent-based modeling approach. *Computer-Aided Civil and Infrastructure Engineering*, 38(7):920–939, 2023.

[2] Irene Alisjahbana, Ana Moura-Cook, Rodrigo Costa, and Anne Kiremidjian. An agent-based financing model for post-earthquake housing recovery: Quantifying recovery inequalities across income groups. *Earthquake Spectra*, 38(2):1254–1282, 2022.

[3] Ibrahim Almufti and Michael Willford. Redi™ rating system: Resilience-based earthquake design initiative for the next generation of buildings. Technical report, Arup Co, 2013.

[4] Steve Bankes. Exploratory modeling for policy analysis. *Operations research*, 41(3):435–449, 1993.

[5] Yaman Barlas. Formal aspects of model validity and validation in system dynamics. *System Dynamics Review: The Journal of the System Dynamics Society*, 12(3):183–210, 1996.

[6] Keith Beven. A manifesto for the equifinality thesis. *Journal of hydrology*, 320(1-2):18–36, 2006.

[7] Keith Beven and Jim Freer. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the glue methodology. *Journal of hydrology*, 249(1-4):11–29, 2001.

[8] Abdullah M Braik, Xu Han, and Maria Koliou. A framework for resilience analysis and equitable recovery in tornado-impacted communities using agent-based modeling and computer vision-based damage assessment. *International Journal of Disaster Risk Reduction*, 121:105427, 2025.

[9] Henry V Burton, Gregory Deierlein, David Lallemant, and Yogendra Singh. Measuring the impact of enhanced building performance on the seismic resilience of a residential community. *Earthquake Spectra*, 33(4):1347–1367, 2017.

[10] Henry V Burton, Scott B Miles, and Hua Kang. Integrating performance-based engineering and urban simulation to model post-earthquake housing recovery. *Earthquake Spectra*, 34(4):1763–1785, 2018.

[11] Dustin T Cook, Abbie B Liel, Curt B Haselton, and Maria Koliou. A framework for operationalizing the assessment of post-earthquake functional recovery of buildings. *Earthquake Spectra*, 38(3):1972–2007, 2022.

[12] Rodrigo Costa and Jack W Baker. A methodology to estimate postdisaster unmet housing needs using limited data: Application to the 2017 california wildfires. *Risk Analysis*, 44(4):850–867, 2024.

[13] Rodrigo Costa and Terje Haukaas. The effect of resource constraints on housing recovery simulations. *International journal of disaster risk reduction*, 55:102071, 2021.

[14] Rodrigo Costa, Terje Haukaas, and Stephanie E Chang. Agent-based model for post-earthquake housing recovery. *Earthquake Spectra*, 37(1):46–72, 2021.

[15] Rodrigo Costa, Ali Nejat, and Sara Hamideh. Post-wildfire housing recovery simulation via an agent-based model. *Scientific Reports*, 15(1), 2026.

[16] Gemma Cremen, Carmine Galasso, and John McCloskey. Modelling and quantifying tomorrow's risks from natural hazards. *Science of the Total Environment*, 817:152552, 2022.

[17] Gemma Cremen, Carmine Galasso, and John McCloskey. A simulation-based framework for earthquake risk-informed and people-centered decision making on future urban planning. *Earth's Future*, 10(1):e2021EF002388, 2022.

[18] Rachel A Davidson, Linda K Nozick, Jamie Kruse, Joseph E Trainor, Meghan Millea, Ian Sue Wing, Dahui Liu, Caroline Williams, and Jingya Wang. Stakeholder-based tool for the analysis of regional risk. *Natural Hazards Review*, 26(3):04025031, 2025.

[19] Rafael Diaz, Sameer Kumar, and Joshua Behr. Housing recovery in the aftermath of a catastrophe: material resources perspective. *Computers & Industrial Engineering*, 81:130–139, 2015.

[20] Mohamed S Eid and Islam H El-adaway. Integrating the social vulnerability of host communities and the objective functions of associated stakeholders during disaster recovery processes using agent-based modeling. *Journal of Computing in Civil Engineering*, 31(5):04017030, 2017.

[21] Saman Ghaffarian, Debraj Roy, Tatiana Filatova, and Norman Kerle. Agent-based modelling of post-disaster recovery with remote sensing data. *International journal of disaster risk reduction*, 60:102285, 2021.

[22] Volker Grimm, Uta Berger, Donald L. DeAngelis, J. Gary Polhill, Jarl Giske, and Steven F. Railsback. The ODD protocol: A review and first update. *Ecological Modelling*, 221(23):2760–2768, 2010.

[23] Volker Grimm, Eloy Revilla, Uta Berger, Florian Jeltsch, Wolf M Mooij, Steven F Railsback, Hans-Hermann Thulke, Jacob Weiner, Thorsten Wiegand, and Donald L DeAngelis. Pattern-oriented modeling of agent-based complex systems: lessons from ecology. *science*, 310(5750):987–991, 2005.

[24] Derek Huling and Scott B Miles. Simulating disaster recovery as discrete event processes using python. In *2015 IEEE Global Humanitarian Technology Conference (GHTC)*, pages 248–253. IEEE, 2015.

[25] Reto Knutti, Reinhard Furrer, Claudia Tebaldi, Jan Cermak, and Gerald A Meehl. Challenges in combining projections from multiple climate models. *Journal of Climate*, 23(10):2739–2758, 2010.

[26] Maria Koliou, John W van de Lindt, Therese P McAllister, Bruce R Ellingwood, Maria Dillard, and Harvey Cutler. State of the research in community resilience: Progress and challenges. *Sustainable and resilient infrastructure*, 5(3):131–151, 2020.

[27] Sameer Kumar, Rafael Diaz, Joshua G Behr, and Ange-Lionel Toba. Modeling the effects of labor on housing reconstruction: A system perspective. *International Journal of Disaster Risk Reduction*, 12:154–162, 2015.

[28] Jan H Kwakkel and Erik Pruyt. Exploratory modeling and analysis, an approach for model-based foresight under deep uncertainty. *Technological Forecasting and Social Change*, 80(3):419–431, 2013.

[29] Robert J Lempert. Robust decision making (rdm). In *Decision making under deep uncertainty: From theory to practice*, pages 23–51. Springer, 2019.

[30] Richard Levins. The strategy of model building in population biology. *American scientist*, 54(4):421–431, 1966.

[31] Peihui Lin and Naiyu Wang. Stochastic post-disaster functionality recovery of community building portfolios i: Modeling. *Structural Safety*, 69:96–105, 2017.

[32] Peihui Lin and Naiyu Wang. Stochastic post-disaster functionality recovery of community building portfolios ii: Application. *Structural Safety*, 69:106–117, 2017.

[33] Ram Krishna Mazumder and Elaina J Sutley. Modeling the influence of social vulnerability and financial resources on housing functionality restoration. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 11(4):04025065, 2025.

[34] Carlos Mesta, Luis Ceferino, Gemma Cremen, and Carmine Galasso. Investigating post-earthquake hospital transportation for casualties through agent-based modeling. *Earthquake Spectra*, 41(3):2028–2059, 2025.

[35] MW Mieler and J Mitrani-Reiser. Review of the state of the art in assessing earthquake-induced loss of functionality in buildings. *Journal of Structural Engineering*, 144(3):04017218, 2018.

[36] Emily Mongold, Rodrigo Costa, Ádám Zsarnóczay, and Jack W Baker. Modeling post-disaster recovery: Accounting for rental and multi-family housing. *Earthquake Spectra*, 40(2):1353–1375, 2024.

[37] Saeed Moradi and Ali Nejat. RecovUS: An agent-based model of post-disaster household recovery. *Journal of Artificial Societies and Social Simulation*, 23(4):13, 2020.

[38] Ali Nejat and Ivan Damnjanovic. Agent-based modeling of behavioral housing recovery following disasters. *Computer-Aided Civil and Infrastructure Engineering*, 27(10):748–763, 2012.

[39] Ali Nejat, Roxana J Javid, Souparno Ghosh, and Saeed Moradi. A spatially explicit model of postdisaster housing recovery. *Computer-Aided Civil and Infrastructure Engineering*, 35(2):150–161, 2020.

[40] E. Opabola and Carmine Galasso. A probabilistic framework for post-disaster recovery modeling of buildings and electric power networks in developing countries. *Reliability Engineering & System Safety*, 2023.

[41] Naomi Oreskes, Kristin Shrader-Frechette, and Kenneth Belitz. Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, 263(5147):641–646, 1994.

[42] Sangung Park, Takahiro Yabe, and Satish Ukkusuri. Post-disaster recovery policy assessment of urban socio-physical systems. *Computers, Environment and Urban Systems*, 114:102184, 2024.

[43] Nicole Paul, Carmine Galasso, and Jack Baker. Household displacement and return in disasters: A review. *Natural Hazards Review*, 25(1):03123006, 2024.

[44] Varun Ramachandran, Suzanna K Long, Tom Shoberg, Steven M Corns, and Héctor J Carlo. Post-disaster supply chain interdependent critical infrastructure system restoration: A review of data necessary and available for modeling. *Data Science Journal*, 15:1–1, 2016.

[45] Andrea Saltelli, Gabriele Bammer, Isabelle Bruno, Erica Charters, Monica Di Fiore, Emmanuel Didier, Wendy Nelson Espeland, John Kay, Samuele Lo Piano, Deborah Mayo, et al. Five ways to ensure that models serve society: a manifesto. *Nature*, 582(7813):482–484, 2020.

[46] Andrea Saltelli, Marco Ratto, Terry Andres, Francesca Campolongo, Jessica Cariboni, Debora Gatelli, Michaela Saisana, and Stefano Tarantola. *Global sensitivity analysis: the primer*. John Wiley & Sons, 2008.

[47] Robert G Sargent. Verification and validation of simulation models. *Journal of simulation*, 7(1):12–24, 2013.

[48] S. Sharif, Peyman Habibi Moshfegh, and H. Kashani. Simulation modeling of operation and coordination of agencies involved in post-disaster response and recovery. *Reliability Engineering & System Safety*, 235:109219, 2023.

[49] Elaina J Sutley and Sara Hamideh. An interdisciplinary system dynamics model for post-disaster housing recovery. *Sustainable and Resilient Infrastructure*, 3(3):109–127, 2018.

[50] Elaina J Sutley and Sara Hamideh. Postdisaster housing stages: A markov chain approach to model sequences and duration based on social vulnerability. *Risk Analysis*, 40(12):2675–2695, 2020.

[51] Chenbo Wang, Rodrigo Costa, and Jack W Baker. Simulating post-disaster temporary housing needs for displaced households and out-of-town contractors. *Earthquake Spectra*, 38(4):2922–2940, 2022.

[52] Wanting Lisa Wang and John W van de Lindt. Quantitative modeling of residential building disaster recovery and effects of pre-and post-event policies. *International Journal of Disaster Risk Reduction*, 59:102259, 2021.

[53] Christopher P Weaver, Robert J Lempert, Casey Brown, John A Hall, David Revell, and Daniel Sarewitz. Improving the contribution of climate model information to decision making: the value and demands of robust decision frameworks. *Wiley Interdisciplinary Reviews: Climate Change*, 4(1):39–60, 2013.

[54] Michael Weisberg. Robustness analysis. *Philosophy of science*, 73(5):730–742, 2006.

[55] William C Wimsatt. Robustness, reliability, and overdetermination (1981). In *Characterizing the robustness of science: After the practice turn in philosophy of science*, pages 61–87. Springer, 2012.