
Oegstgeest
Alphen aan den Rijn
The Hague
Voorburg
Boskoop
Waddinxveen
Ypenburg
Monster
Gouda
Delft
s-Gravenzande
Berkel en Rodenrijs
De Lier
Maassluis
Schiedam
Rotterdam

Reading Between the Boxes

Using Scenario Discovery to Explore Tipping Points in the Behaviour of Human-Earth Systems

By

Gabriel H. Sher

in partial fulfilment of the requirements for the degree of

Master of Science
in Engineering & Policy Analysis

at the Delft University of Technology,
to be defended publicly on August 29, 2024 at 09:30 AM.

Thesis committee:	First Supervisor, Chair:	Prof. dr. ir. J. H. Kwakkel,	TU Delft
	Second Supervisor:	Prof. dr. T. Filatova,	TU Delft
	Advisor:	Dr. Alessandro Taberna	TU Delft

An electronic version of this thesis will be made available at http://repository.tudelft.nl/.

Cover page image taken from Surging Seas risk zone map (https://ss2.climatecentral.org) depicting the area around The Hague at present vs. under 5 feet of sea level rise.

“

For complex systems, locating boundaries between qualitatively similar regions in a space of alternatives can often be […] useful.

(Bankes, 2011, p. 597)

“

Nothing in the world is as soft, as weak as water; nothing else can wear away the hard, the strong, and remain unaltered.

(Lao Tzu, translated by Ursula K. Le Guin)

Executive Summary

Tipping points are an active and growing interest in both the scientific and political study of climate change: what are they, how can we identify them, and how can we avoid them (negative tipping points) or encourage them (positive tipping points). As climate change worsens, scientists and policy analysts have turned to computer models of complex, interconnected, human-Earth systems to help understand and address both its physical and social aspects. As this field has matured, so too has the complexity of both the models being developed and the questions being asked with them. Agent-based modeling (ABM) is one framework that has become popular for its ability to observe system-level behaviour without closed-form equations due to its encoding of heterogeneous individual-level behaviour.

Analyzing the data generated by ABMs is not straightforward, as they tend to have many input and output dimensions, most outputs are either temporally or spatially distributed (or both) and can be sensitive to stochastic effects. When applying a *exploratory modeling* or *deep uncertainty* lens—a philosophy that seeks to explore the effects of assumptions made in a model's development and parametrization, understanding more about the modeled system's behaviour as opposed to attempting to predict it—the complexity of this analysis grows further. However, this complexity should not discourage analysts from bringing existing *Decision-Making under Deep Uncertainty (DMDU)* methods to ABMs.

This study applies one such method (scenario discovery) to a complex ABM of household and firm climate adaptation in a coastal economy, attempting to uncover the existence of socio-environmental tipping points in the system. Based on a previously developed analogy connecting the output space of an ABM to the traditional notion of a physical phase diagram, Scenario Discovery is used to generate such a phase diagram and infer tipping points at the boundaries between distinct system states. Ultimately, a set of possible population-change tipping points are generated.

While this work demonstrates the fitness of scenario discovery as a tool for exploring the output spaces of ABMs and finding tipping points within them, it is very preliminary. The work should be repeated with several improvements. First, either the uncertain parameters varied in this study should be selected to be more policy-relevant, controllable system factors, or the system states and thus the tipping points should be expressed in terms of endogenous variables instead of input parameters. Second, this study demonstrates that the typical approach to processing stochastic replications in exploratory modeling—simply averaging all outcomes—is not fit for use with complex modeling like ABM. Despite the computational and cognitive load introduced by simultaneously handling both a wide uncertainty space and many stochastic replications, efforts must be made to ensure any dynamically distinct behaviour generated by the original model is not lost to averaging. Studies like this one that do not put in this effort risk enabling the extraction of incorrect political and policy lessons.

Acknowledgements

I would like to thank my supervisors Dr. ir. Jan Kwakkel & Dr. Tatiana Filatova for their support and guidance throughout the process of writing this Thesis. This project went through a few too many confused forms before settling where it is today—right where both Professors expected it to land. Your patience with me getting there is appreciated. Special thanks also to Dr. Alessandro Taberna, who, on top of developing the CRAB model and laying the groundwork for this research, provided me guidance at all stages of my work and helped me find my footing. Additional thank you to the ERC SCALAR project, supported by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Program (grant agreement number: 758014), which has funded the development of and research performed with the CRAB model.

I'd also like to thank my partner, Jacqueline, for being my sounding board throughout my research process. No matter the topic, whether it be tipping points, scenario discovery, the so-called Institution, or a crisis of confidence in my own abilities, you were there to listen. During the high-stress periods of my Thesis work, your support always came at the perfect moment. Thank you for helping me reach my goals this summer. LTFI.

Chris, Floris, and Angela, thank you for your close friendship over the last two years and in particular your willingness to listen over the last six months. I am very glad I was able to grow as an analyst, thinker, and a person alongside the three of you.

There are many others I've interacted with over the past months, both about my Thesis research and not so much, who have helped me on this brief but meaningful journey. Unfortunately, there are too many people to name, let alone remember. Rest assured that if I've spoken to you in any capacity since February 2024, you probably played a role in either my reflection on this report, or my (necessary) distraction from it.

Finally, I would like to thank my cat, Goose, for trying very hard to contribute to this document's word count. Unfortunately, none of his contributions made the final editttttttttttttttttttttt.

I do not know what the future holds beyond the next year or so, but I have many thoughts and dreams about how I can take the skills and experience I gained in the EPA program into my future. I look forward to it with excitement. I hope this is just my first of many step in trying to help improve the world around me, from my local community all the way up to the global one.

Contents

1 Introduction

As climate change worsens, scientists, economists, and other practitioners of knowledge—perhaps ironically—turn to computer models to understand the shifting dynamics of the Earth and our place in it: models of so-called coupled human-Earth systems or *socio-environmental systems* (SES) (e.g., Moallemi et al., 2022). These systems are large, often incomprehensibly complex, and usually riddled with interrelated components and feedback loops; their computational counterparts are similarly hard to grasp. Computer models of SES are intended to enable policymakers to make better-informed decisions in the face of their highly complex subjects (Castro et al., 2020; Filatova et al., 2013; Lippe et al., 2019).

Agent-based modeling (ABM) is one such paradigm for modeling SES. ABMs are unique in several ways, most notably their ability to model detailed, heterogeneous, behaviourally rich agents (people, organizations, governments, environmental hazards, etc.) (Filatova et al., 2013). Thus, ABMs are generally considered to more closely resemble reality than closed-form alternatives like computable general equilibrium models. However, this comes at the cost of additional challenges in design, execution, and output analysis (Lee et al., 2015; Filatova et al., 2013). ABMs are also often *stochastic* models, meaning there is an element of randomness in a model's behaviour; they must be run several times—even with the same structure and parameter settings—to evaluate the range of its possible outcomes.

Tipping points are a hot topic in the world of climate policy, which means they are a hot topic in the world of SES modeling. In brief, tipping points are the thresholds whose crossing are associated with abrupt regime shifts in a system's behaviour (Horan et al., 2011). The *Global Tipping Points Report* was released at last year's COP28 (Lenton et al., 2023). Included in it was not just a review of the known Earth-system tipping points and their impacts, but also discussion on the governance of tipping points, science-policy engagement on the topic, and the possibility of identifying positive economic and social tipping points that could accelerate the fight against climate change (Lenton et al., 2023).

The use and definition of the term *tipping point* is often disputed. Generally, it refers to a rapid (and often snowballing) change brought on by crossing a threshold; importantly, the change must be a large one, often to an entirely new state or dominant behaviour (van Nes et al., 2016). One research focus has been on identifying or anticipating tipping points in complex real-world systems (Scheffer et al., 2009; Scheffer et al., 2012). In 2016, Filatova et al. called for the pursuit of better statistical methods for searching for tipping points in complex SES using model output data, such as those gathered from ABMs.

An important concept in the field of model-based policy analysis is that of a *scenario*. The most common example might be the SSP/RCP scenario framework, which is used in almost all large-scale climate change studies (O'Neill et al., 2020). A scenario is a plausible future, drawing the link between an expected system behaviour and the conditions under which it occurs. There are numerous approaches to scenario development, with differing degrees of qualitative and quantitative influence (Wright et al., 2020). In essence, scenarios can be seen as an output of the

model analysis process and an input to the policymaking process: analysts equip policymakers with an understanding of the breadth of plausible futures (and their implications), and policymakers use this understanding to inform their decisions. It is important to remember, though, that scenarios can also be used as part of the model-based policy analysis process itself.

If a scenario can be conceived as a possible future (or set of futures), it must inherently be characterized by some unifying condition or behaviour. Perhaps a terrible thing happens, like the economy in a region collapses due to frequent flooding, and people are forced to move away. Perhaps a good thing happens, like a breakthrough in battery technology leading to rapid transition towards electrification across several sectors and emissions dropping faster than otherwise anticipated.

In this way we can see a connection between the concept of a scenario and the concept of a tipping point: a tipping point could be seen as the conditions at which a system transitions from one scenario to another. One's instinct might be to think of this in terms of time passing, that is, as we move forward in time, conditions change, and we transition (perhaps rapidly and abruptly, like when crossing a tipping point) to a new state. However, it is more helpful to think of these transitions in terms of the changing conditions themselves: the frequency of flooding has increased, or the quality of battery technology has increased.

Often, we do not have perfect knowledge of important quantities or conditions like these. Many quantities that characterize complex systems like SES are impossible to perfectly know or are subject to political or scientific disagreement. Even the structure of many SES can be similarly disputed. This unknowability is often termed *deep uncertainty*—situations where planners cannot perfectly know, precisely define, or agree upon how to characterize a system's structure, behaviour, or present and future states (Lempert, 2003; Walker et al., 2013). When translating complex systems into models, deep uncertainty manifests as a range of unknowns surrounding variables and structures critical to the model (Kwakkel & Haasnoot, 2019; Moallemi et al., 2020).

In 1993, the term *exploratory modeling* was introduced as a philosophy for modeling under such uncertainty (Bankes, 1993). Contrasted with traditional "consolidative" modeling where computational models are expected to predict how the real-world future unfolds, this is the use of models to explore the impact of the assumptions made in their construction, parametrization, and execution (Bankes, 1993). In the modern day, the field that has taken up Bankes's mantle of using models to address policy questions under such conditions is known as Decision Making under Deep Uncertainty (DMDU).

Thus, if scenarios map a set of conditions to an expected behaviour and SES are characterized by a deep uncertainty—an unknowability of their exact conditions—then we can conceive of a tipping point as the transition between possible scenarios, regardless of whether those are possible futures or possible presents. It is with this framing that this study will attempt to uncover tipping points in SES.

1.1 Research Gap

Just as scenarios can be used by policymakers in decision-making, so too can they be used by policy analysts for model-based policy evaluation. Elsawah et al. (2020) call for further research addressing the challenges of using scenarios as both the processes *and* the products of SES modeling. Amongst others, they identify two key challenges: the need to improve the visual communication of scenarios and the need to better bridge the gap between scenario development and decision-making. Linking the concept of tipping points to scenario analysis is one way to address the latter. As with all things in model-based policy analysis, finding effective ways to visually communicate results aids their impact, and thus visually communicating scenarios and tipping points together will be important.

One DMDU method that uses the concept of scenarios is *scenario discovery*, which exploits a model's uncertainty space (i.e., the range of parameter and structure options) to search for stand-out scenarios that may be of interest, and the conditions under which they occur (Bryant & Lempert, 2010; Jafino & Kwakkel, 2021; Moallemi et al., 2020). Given that this method explores a system and defines a set of key scenarios, it is natural to expect that it can also be used to explore the system's tipping points.

Furthermore, ABMs are often used in a manner that resembles what Bankes called consolidative modeling (Bankes, 1993). Though many ABM studies do consider uncertainty, often in a model's parameters, the full consideration of deep uncertainty is often missing from both model designs and the policy studies that use them. On the other hand, exploratory modeling approaches are often demonstrated on simpler deterministic models that lack the stochasticity and behavioural complexity of many ABMs. Thus, the two worlds do not meet as often as one would expect.

This thesis intends to contribute to the social goal of identifying tipping points in SES with the two technical fields of ABM and DMDU by bringing recent advancements in scenario discovery to a complex ABM of a real-world, policy-relevant system. Both ABM and exploratory modeling studies are often critiqued (compared to their alternatives) due to the computational resources required to carry them out. Thus, marrying the two paradigms comes with several computational challenges and trade-offs.

1.2 Research Question

The above research gap motivates the following main research question:

> (RQ) How can scenario discovery be used in complex ABMs to uncover information about tipping points in human-Earth systems?

This research question stands between two sub-fields of study. On the one hand, it must stand out from other recent advancements in scenario discovery methods. On the other, it must demonstrate that scenario discovery is both feasible in and adds a new element to ABM output analysis in a way other methods do not. This motivates the two sub-questions:

(SQ1) Which recent advancements in scenario discovery enable its use for exploring and discovering tipping points?

(SQ2) How can scenario discovery enable effective communication of tipping points?

(SQ3) How can scenario discovery be applied to the outputs of complex ABMs, which are often stochastic and can have many output variables, each with possible spatial and temporal dimensions?

These sub-questions address the main research question from opposite sides. The first is about surveying recent progress on scenario discovery methods from the field of DMDU and using them to develop a method for tipping point identification. The second follows up and, assuming tipping points can be discovered, looks at how they can be communicated. It is inherently a visual question: scenario discovery visualizations are often suited for a technical audience who understands the underlying techniques and algorithms, but to bridge Elsawah et al.'s scenario-decision making gap, these visualizations must be adapted to a more general audience (2020). The goal should be to create visuals that are clear enough for a non-technical decision-maker to look at and quickly enough grasp the conditions and/or implications of a tipping point.

The third sub-question is about surveying the challenges of using scenario discovery with ABM data, including the spatio-temporal dimensions of ABM output data and the stochastic nature of ABMs themselves. Stochasticity increases model complexity and requires running (and thus analysing data from) many more model runs to extract salient information. It is difficult to assess the stopping point for adding more replications to an ABM with chaotic, non-linear outputs, and even more so under deep uncertainty, where the effect of stochasticity might change across different portions of the uncertainty space.

1.3 Research Relevance

1.3.1. EPA Relevance

This research is highly relevant to the EPA curriculum. Both ABMs and DMDU methods are a core part of the modeling and simulation line of the curriculum, each having at least an entire module dedicated to their exploration. It is the uniquely prepared skillset of an EPA graduate to be able to understand and contribute to the literature in both fields. Methodologically, this thesis finds its home perfectly within TPM and EPA.

Methods are not the whole picture, though. EPA is about solving complex, many-actor problems and addressing global grand challenges. Uncovering tipping points in human-Earth systems is critical to understanding our ability to protect both ourselves and the environment as climate change and its impacts worsen. Tipping points like the ones studied in this paper are of critical importance to policymakers. Uncovering the conditions under which populations are expected to rapidly retreat from coastal regions, or under which people stay in coastal regions but intra-

regional inequality runs rampant, are key pieces of information that decision makers can use to protect those populations. To make the best policy decisions, we must understand as much as possible about the conditions and drivers of those possible futures.

1.4 Research Pipeline

Figure 1-1, below, breaks down the work done for this Thesis and indicates the structure of the document. In this section and the next one, initial research was performed to generate knowledge of the literature and identify the research gap to be addressed. Section 3 then covers the methods used in this research: first, a review of the CRAB model, the case study to which the scenario discovery method is applied; second, an explanation of the experimental design used in this study; and third, an explanation of the use of behaviour-based scenario discovery for identifying tipping points. Then, Section 4 covers analysis and results. Initially, results are inspected using Visual Analysis, in Section 4.1. Then Section 0 applied traditional sensitivity analysis to explore the input-output relationships in the model and demonstrate the common use of uncertainty analysis in ABM studies. Then, Section 4.3 validates that the experimental design used is fit to adequately address the research questions. Finally, Section 4.44.3 comprises the meat of this study, the full behaviour-based scenario discovery pipeline and its results. Section 5 then reflects on the impact of the results and the promise of the presented methods, noting especially how these methods can be improved in future work. Finally, Section 6 concludes.

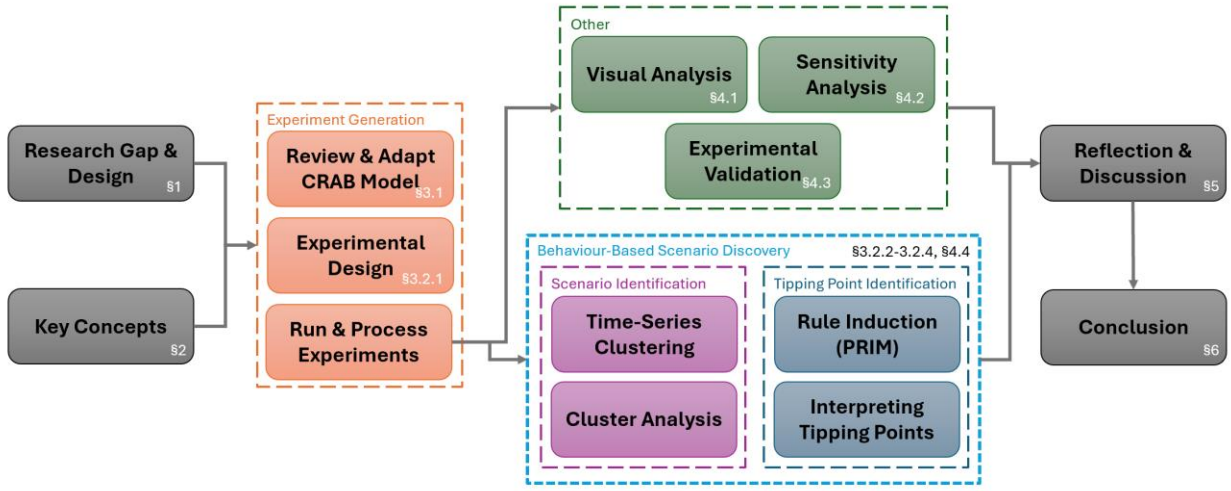


Figure 1-1: A visualization of the full research pipeline taken in this document. Attached to each box is the section of this report in which it is written about.

2 Key Concepts

2.1 Tipping Points and Regime Shifts in SES

The history of tipping points research focuses on tipping points in physical systems, like the melting of polar ice. For example, Lenton et al. (2008) identified a list of policy-relevant tipping elements in climate systems, including several major forests and ice sheets. However, several researchers have also discussed the topic in association with both purely social systems and SES.

Filatova et al. (2016) review approaches to using complex SES models to explore *regime shifts*—"significant, persistent changes" (p. 333) in a system, often due to a change in conditions enabling new dynamics in the system. They propose several criteria assessing how well a given modeling approach can capture these shifts and identify that many models do not consider these criteria, suggesting a need to develop capacity for considering regime shifts in modeling. Milkoreit et al. (2018) conduct a literature review to find and compare definitions of tipping points in SES scholarship. They ultimately propose 23 characteristics, including *external causes*, *multiple causes*, and *multiple stable states*. The characteristic of external causes is interesting, as some other research presents tipping points due to endogenous process and amplifying feedback (Dietz et al., 2021; Barnard et al., 2021; van Ginkel et al., 2020). This difference indicates the existence of two types of tipping point, one triggered by external forces and one in response to endogenous changes.

Otto et al. (2020) study positive tipping points, seeking to uncover "social tipping interventions" that represent shifts towards rapid reductions in net emissions. In this light, tipping points are seen as just moments of major shift, *not* strictly moments of rapid collapse. van Ginkel et al. (2020) similarly look at climate change-induced tipping points in socio-economic systems (*socio-economic tipping points* or SETPs) and identify 22 such tipping points that could be relevant to the policy arena in Europe. One such tipping point is migration induced by sea level rise (SLR). Based on expert opinion, they suggest that a transition to mass migration from coastal regions may be more likely to be triggered by a single extreme event than a gradual rise in SLR. They also comment on the complexity of such a system, noting that the adoption of private adaptation actions (i.e., household floodproofing) may cause the whole system to reconfigure and behave differently. They call for further research, including the development of dynamic models to study SETPs, a more tangible study of the impacts of crossing SETPs, and investigation into the role of mitigatory policy in dampening the likelihood and impacts of crossing SETPs.

Though an inherent feature of regime shifts and crossings of tipping points is their rapid onset, Scheffer et al. (2009) propose that there are generalizable early-warning signals when a system is nearing a tipping point. One such indicator is a significant increase in the autocorrelation of a critical outcome just prior to its tipping. Another study Scheffer cites suggests that rapid flickering between overall states or regimes often precedes tipping. Some of these signals, such as state flickering, were found across systems with little similarity. The authors point out that one need not have a good understanding of a system's mechanisms and dynamics to recognize these warning

signals. This suggests the possibility of algorithmically searching for these warning signals in model results.

Many of the above studies fall prone to the trap of thinking about tipping as occurring intertemporally in a single timeline: i.e., within a single model execution. van Ginkel et al.'s (2022) study on identifying climate change-induced SETPs lays this out most clearly, by characterizing model runs using a meta-metric indicating whether an SETP occurred: for each timeseries outcome of interest in the modelled system, evaluate whether, during a single model run, there was ever an abrupt change from one stable state (i.e., level) to another. First, this paper supposes that system state is a function of a single endogenous variable, rather than a combination of several. Second, its conception of tipping points, while intuitively accessible, is not the only framing.

An alternative way to conceive of tipping points in complex systems can be found in Gualdi et al. (2015). Here, the authors' use of phases and borrow the notion of phase diagrams, phase transitions, and critical points from physics. Rather than searching for tipping within a time-series describing a single model run, they look for tipping within the model's parameters. The example they provide defines two distinct system states—a "good" economy with low unemployment, and a "bad" economy with high unemployment. Figure 2-1(a), below, shows how the model transitions between states as a specific parameter varies. This framing of tipping as phase transition works well in a highly parametrized model, such as one designed for study under deep uncertainty.

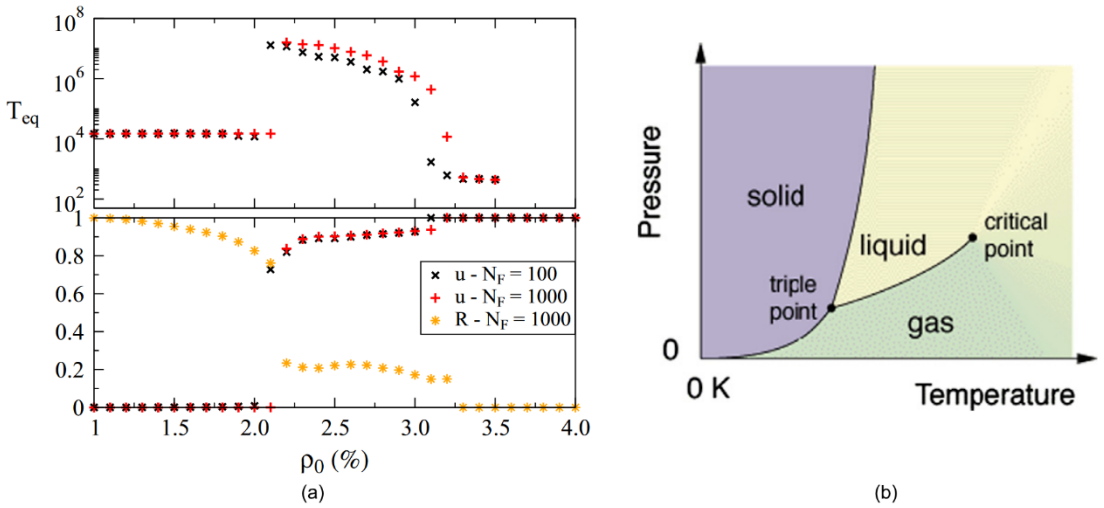


Figure 2-1: (a) Demonstration of a tipping point in a model's parameter space, showing unemployment (y-axis) as a function of the interest rate. (b) The physical concept of a phase diagram. Taken from Gualdi et al. (2015).

The analogy of a phase diagram is important, as it enables the linking of scenario thinking to tipping points. Using this analogy, a scenario or model state stands in for the physical notion of a phase: instead of water being a gas at a high temperature provided pressure is not too high, a a city's population might be in an acceptable state of steady growth if unemployment is low and local businesses are thriving. Gualdi et al.'s (2015) notion of phases in an ABM already closely mirrors that of a modelled scenario.

2.2 Scenario Discovery

Scenario Discovery describes an approach to exploratory modelling that seeks to identify the conditions under which specific outcomes of interest occur (Moallemi et al., 2020). Broadly speaking, there are three steps involved in scenario discovery (Steinmann et al., 2020):

(1) **Measurement**, i.e., collecting output data from a model-based experiment, running the model across enough samples to sufficiently cover its reasonable operating space. Samples here are usually points in the model's parameter space but can cross multiple candidate structures as well. In the case of stochastic models, model replications are usually aggregated to a single measure per sample, for example by averaging.
(2) **Identification,** i.e., using model output data to identify scenarios of interest. This is traditionally a binary classification—a scenario can either be of interest, or not (Kwakkel & Jaxa-Rozen, 2016). There is also *multiclass* scenario discovery, where scenarios are put into several groups according to their measured outcomes, and any number of those groups may be of interest to an analyst. In essence, this step compresses the full scope of outcomes measured for a given sample point into a single scalar value.
(3) **Rule Induction**, i.e., using algorithmic methods to associate each identified class from Step 2 with "rules" in terms of the model's parameter space. These are often expressed as ranges of values in some or all of the model's input variables ranges.

Analysts can then combine model outcomes covered by each identified class (Step 2) with their associated rules (Step 3) to create ("discover") simplified, representative scenarios for use in policy analysis and decision making (Wang et al., 2013). In some cases, further translation of the numerical results (i.e., parameter ranges associated with outcome levels) to qualitative narratives could be considered a fourth step of scenario discovery (Lamontagne et al., 2018). By narrowing the full experiment set into a smaller set of qualitatively distinct scenarios, this brings the results of an exploratory modeling experiment more in line with the traditional qualitative process of scenario development (Carlsen et al., 2016; Greeven et al., 2016).

There are several decisions analysts can make in each of the three steps that differentiate scenario discovery approaches. Bryant and Lempert's (2010) original scenario discovery paper measured a single scalar-value outcome for each experiment and denoted scenarios of concern using a simple threshold—the 90^{th} percentile of the outcome across their experimental design. This threshold approach is the traditional approach to binary class identification. As they discuss, some policy environments lend themselves well to such a threshold, for example due to a budget constraint. The choice of identification scheme—i.e., what makes an outcome *of interest*—has great influence on the results and meaning of an scenario discovery study. Finally, they use the Patient Rule-Induction Mechanism (PRIM) to algorithmically determine a "rule" defining which portions of the parameter space are associated with the scenario of interest. PRIM produces such a rule in the form of a many-dimensional box; for example, Parameter A might need to be in the lower half of its range for a scenario of interest to occur, while Parameters B and C can take on any value. In this case, the discovered box is a 3D cube comprising half the original parameter space. They propose two performance metrics for such a box: coverage, which is the proportion of cases of interest that are indeed in accordance with the rule (i.e., fall within the box); and density, which is the proportion of cases in accordance with the rule that are indeed of interest. A

density of 80% implies a 20% false positive rate, while a coverage of 80% implies a 20% false negative rate. Coverage and density typically come at a trade-off; an optimal rule tries to maximize both. The induced rule can be interpreted as the scenario in which the outcome of interest is most likely to occur.

There have been two primary improvements on Bryant and Lempert's method that are relevant to this work. First, several studies have achieved multiclass scenario discovery by clustering model runs according to similarities in their outcomes (Wang et al., 2013; Steinmann et al., 2020; Jafino & Kwakkel, 2021). This can help deepen analysis by differentiating several distinct outcomes of interest. For analysts, it shifts the responsibility of identification from an *a priori* measure like a threshold, which can be arbitrary (e.g. Bryant and Lempert's choice of deeming the 90th percentile cost "unacceptably high (2010, p. 40)), to a statistical algorithm. The latter imposes the additional burden of analysing and explaining the outcomes associated with each cluster (Jafino & Kwakkel, 2021). Jafino & Kwakkel (2021) also explore how the choice of clustering algorithm affects results, deciding on the K-means method for their study, and explore the use of an algorithm that enables concurrent clustering and rule induction.

For multiclass scenario discovery, several papers have discussed the performance metrics of *input-* and *output-space separability* (Jafino & Kwakkel, 2021; Steinmann et al., 2020). These refer, respectively, to the ability to distinguish between the parameter constraints of the induced rules for each cluster, and between the clusters themselves. Input-space separability can be evaluated by the number of input samples that fall within the input region associated with another cluster (Steinmann et al., 2020). Output-space separability can be quantified by the ratio of within-cluster distances to between-cluster distances (Jafino & Kwakkel, 2021). A set of classifications and their associated rules (or scenarios) perform well when both types of separability are maximized.

Second, (Steinmann et al., 2020) use scenario discovery methods to study the *behaviour dynamics* of a model outcome. That is, by clustering model runs according to a timeseries output, they switch from the notion of an *outcome of interest* to a *behaviour of interest*. As such, the induced rules (and thus scenarios) that result from the scenario discovery process are a collection of input parameter ranges each associated with certain classes of model behaviour.

Other improvements to Bryant & Lempert's original work include the exploration of alternative rule induction algorithms. Several authors, including (Jafino & Kwakkel, 2021), use some variation on Classification and Regression Trees (CART) as a means of inducing rules and defining scenarios. Other authors rely on visual tools like dimensional stacking (Suzuki et al., 2015). Indeed, there is an entire field of data science beyond the world of DMDU that studies rule induction algorithms (Grzymala-Busse, 2023).

3 Methods

3.1 CRAB Model

To demonstrate scenario discovery as a method for discovering, explaining, and communicating tippint points in complex SES, this study will use the CRAB model as a case study. The CRAB model, introduced by Taberna et al. across a series of papers (Taberna et al., 2021; Taberna et al., 2023; Taberna et al., 2023), captures a detailed regional economy with three economic sectors—capital goods, consumption goods, and services—and uses survey data to characterize household adaptation behaviours. It was developed as part of the European Research Council's SCALAR project, which seeks to bring information about micro-level behaviours from social science to macro-level climate-economy simulations (ERC Scalar, 2020). Figure 3-1, adapted from (Taberna et al., 2023), illustrates the model's structure.

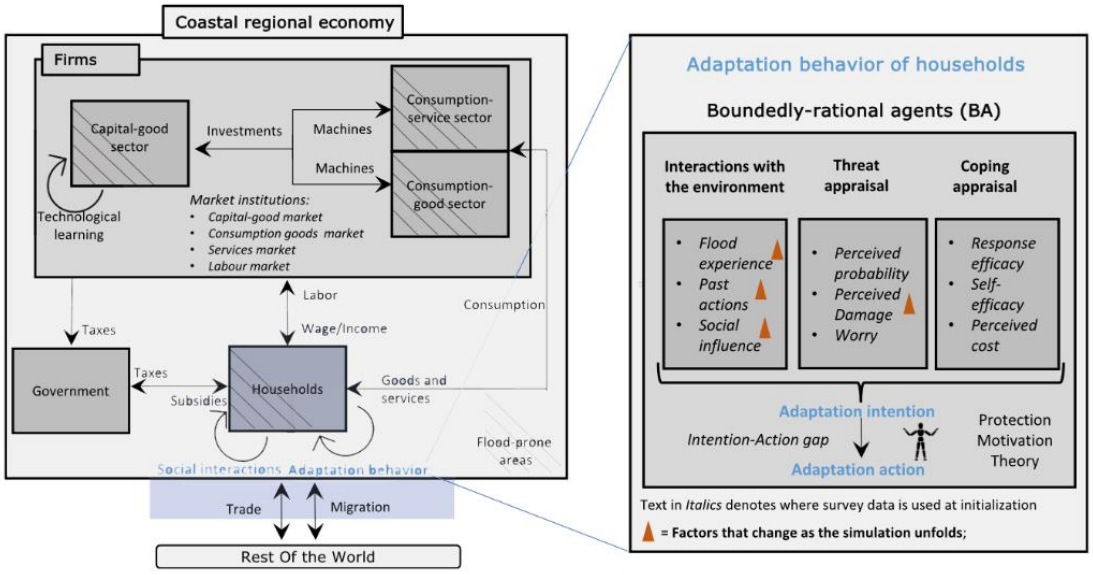


Figure 3-1: Visual summary of the CRAB model. Taken from Taberna et al. (2023)

The left-side box in Figure 3-1 explains the economic activity present in the CRAB model. Capital-good firms sell machines to consumption good and service firms. Over time, they also invest some of their revenue into R&D, producing better technology that firms then buy when they need to replenish, upgrade, or expand their capital goods stock. Households supply labour to all three types of firms, earning wages accordingly, and consume goods and services from those firms. Households and firms both pay taxes to the government, which are used to subsidize household flood adaptation investments.

Firms can dry-proof their facilities, choosing to do so based on a purely rational, discounted expected utility calculation. Households can choose to dry-proof, wet-proof, and/or elevate their homes, with their likelihood to do so determined according to their risk perception, worry, and

preferences between each intervention type. These traits are assigned according to distributions gathered from a survey of households in a real-world coastal region, though some change over the course of a model run. Households also interact with and learn from their neighbours, adopting some of their traits or being more likely to invest in adaptation if those in their social network also do so. The box on the right-hand side in Figure 3-1 explains these behavioural traits.

The synthetic, heterogeneous population of households and firms are varyingly exposed to the risk of floods, emulating the effects of a spatial distribution of flood risk without incorporating an explicitly spatial element in the model. When a flood occurs, damages are calculated using calibrated data from a real-world region, and firms and households must commit savings towards repairs either by limiting their consumption or redirecting funds intended to be saved for investment in adaptation.

For this initial experiment, it is assumed that one flood of a fixed intensity will occur at some point in the model's execution, with variable timing. Thus, one could consider the underlying system to be that of a *flood-exposed* coastal region, not just one with rising flood *risk*.

Finally, households can migrate in and out of the coastal region. This is done mostly according to changes in the region's wages and employment prospects and is not tied to households' behavioural traits. Firms enter the market when there is sufficient opportunity and exit when they go bankrupt. Thus, CRAB is effectively able to model *coastal retreat*, i.e., adapting to rising flood risk by choosing to leave a coastal area (Haasnoot et al., 2021).

3.2 Exploratory Modeling

All code used for the modelling and analysis in this report is available here and archived here (Sher, 2024a).

3.2.1. Experimental Design

To assess the relationship between model inputs or modelling choices on the outcomes of a model, exploratory modeling studies seek by nature to cover a wide range of model formulations. The differences between these formulations can be structural (a fundamental difference in the assumed behaviour of the underlying system), parametric (a difference in the level of a parameter used to calibrate or characterize the model, often considered an input to the model), or stochastic (all else held equal, the difference in outcomes is purely due to random effects inside the model's behaviour). Thus, analysts and exploratory modelers must design experiments that effectively cover the range of uncertainties they wish to explore.

The Exploratory Modeling & Analysis (EMA) Workbench provides a convenient environment in Python to generate intelligent experimental designs and iterate through many runs of a model (Kwakkel, 2013). The Workbench uses the *XLRM* framework for defining model structure: a model is defined by the e**X**ogenous factors, policy **L**evers, and **R**elationships within the system it models, as well as the **M**easures of system performance. **X** and **L** could be considered model parameters: these are inputs that influence how the model performs. **R** could be considered model structure. **M** refers to the measured variables that serve as indicators for the outcome of a model run. Any

model programmed in Python can be used with the Workbench if you explicitly identify **X**, **L**, and **M** (as **R** is the programmed model itself). For this study, **X** and **L** are taken together—it is of less importance whether the studied uncertainties are exogenous factors to be monitors or are levers controllable by meta-actors in the system (e.g., the coastal region's government).

The CRAB model was adapted for this study to expose several assumed values as input parameters. Table 3-1 describes these parameters and the ranges across which they were explored. The selection of relevant input parameters was informed by several factors, including prior observations of parameters influential to the model's calibration as well as the identification of assumed or arbitrarily selected values in the existing codebase and in Taberna's (2021) paper introducing the model.

Table 3-1: CRAB model input parameters and their ranges used in this study

Name	Description	Range
Debt-to-sales Ratio	The amount of debt a firm can take on, as a proportion of their total sales	0.8-5.0 [dmnl]
Wage Sensitivity (Productivity)	The sensitivity of wages to firm productivity in the previous timestep	0.0-1.0 [dmnl]
Initial Markup	The markup (pricing) rate for all firms at timestep 0	0.05-0.50 [dmnl]
Capital-Output Ratio	The size of the capital stock a firm requires as a ratio of its output	0.2-0.6 [dmnl]
Emigration Minimum Unemployment	The minimum unemployment level required for people to emigrate from the coastal region	2-8 [% empl.]
Migration Unemployment Bounds Range	The difference between *Emigration minimum unemployment* and the maximum unemployment level under which immigration to the region can occur	10-25 [% empl.]
DEU Discount Factor	The discount factor firms use in their discounted expected utility calculations for investing in flood adaptation	0.8-1.0 [dmnl]
Flood Timing	The time after start of model execution at which the fixed-intensity flood occurs	30-80 [quarters]

The parameters in Table 3-1 are sampled using a Latin Hypercube sampler, which is an alternative to random (Monte Carlo) sampling that enables effective uncertainty analysis using smaller sample sizes (Helton & Davis, 2003). For this study, 2000 sample points were taken across the parameter space. This number was selected to minimize compute time for an initial run while still providing sufficient coverage of the space. The model was run for 40 replications at each of the 2000 sample points, using consistent seeding values for the random processes across the sample points to ensure comparability. The number of replications for this experiment was chosen somewhat arbitrarily according to computational limits afforded me. Its impact will be evaluated later in this report, and future experiments should have more considered replication numbers.

Table 3-2 defines the outcomes that were measured as part of the model's execution. Each outcome of interest is a timeseries, i.e., it is an endogenous system variable measured over the course of the model run. Many variables were measured, but for the sake of simplicity only Household Population, GDP, and the Gini Coefficient are carried forward for further analysis. In line with Milkoreit et al. (2018) and van Ginkel et al. (2020), these simple outcomes are the most meaningful ones in which to look for tipping in the economic system of a flood-exposed coastal region.

Table 3-2: Four primary CRAB model outcomes measured in this study.

Name	Description
Household Population	The total number of households in the coastal region, as a time-series
GDP	The total productive output of the coastal region, as a time-series
Median Wage	The median wage across employed households in the coastal region, as a time-series
Gini Coefficient	A measure of income and wealth inequality across households in the coastal region, as a time-series

A more thorough description of the CRAB model, its behaviours, and its possible parameters can be found in Taberna et al. (2021).

The model is run with a time horizon of 120 quarters, or 30 years. The first five years are used as burn-in time, i.e., the data from this period is removed after model execution. This is because the model uses this period as a sort of calibration period. Also, not all effects are active during this period, as migration only turns on at timestep 20. Data is collected at each timestep—each quarter—though for some analysis, it may be resampled to every four timesteps (emulating yearly data).

The high-performance computing cluster DelftBlue was used to carry out the experiments using the CRAB model (Delft High Performance Computing Centre, 2024). A single model run takes about two minutes to execute, so the 80000 model realizations in this experiment used roughly 2600 hours of compute time, which ran across 320 cores for roughly 8 hours each. After experimentation, model runs were processed as follows:

- Remove the first 20 timesteps from model execution, representing 5 years in real-world time. This acts as a *burn-in* time, as it is the period before all model functionality is online (notably, migration is not enabled until timestep 20).
- The outcomes for each sample point were calculated by taking the mean across the outcomes from each individual replication. To study the effects of stochasticity on model performance, the variance and confidence intervals of each outcome at each sample point were also measured.

3.2.2. Visual Analysis

Once the model runs have been executed and thus the model outcomes measured for each sample point, it is time to explore the outcomes as they differ across the parameter space. For

this purpose, an interactive, browser-based dashboard was developed for on-demand generation of time-series plots, enabling the highlighting or exclusion of certain portions of the parameter space. A screenshot of the controls available in this dashboard is included below, as Figure 3-2.

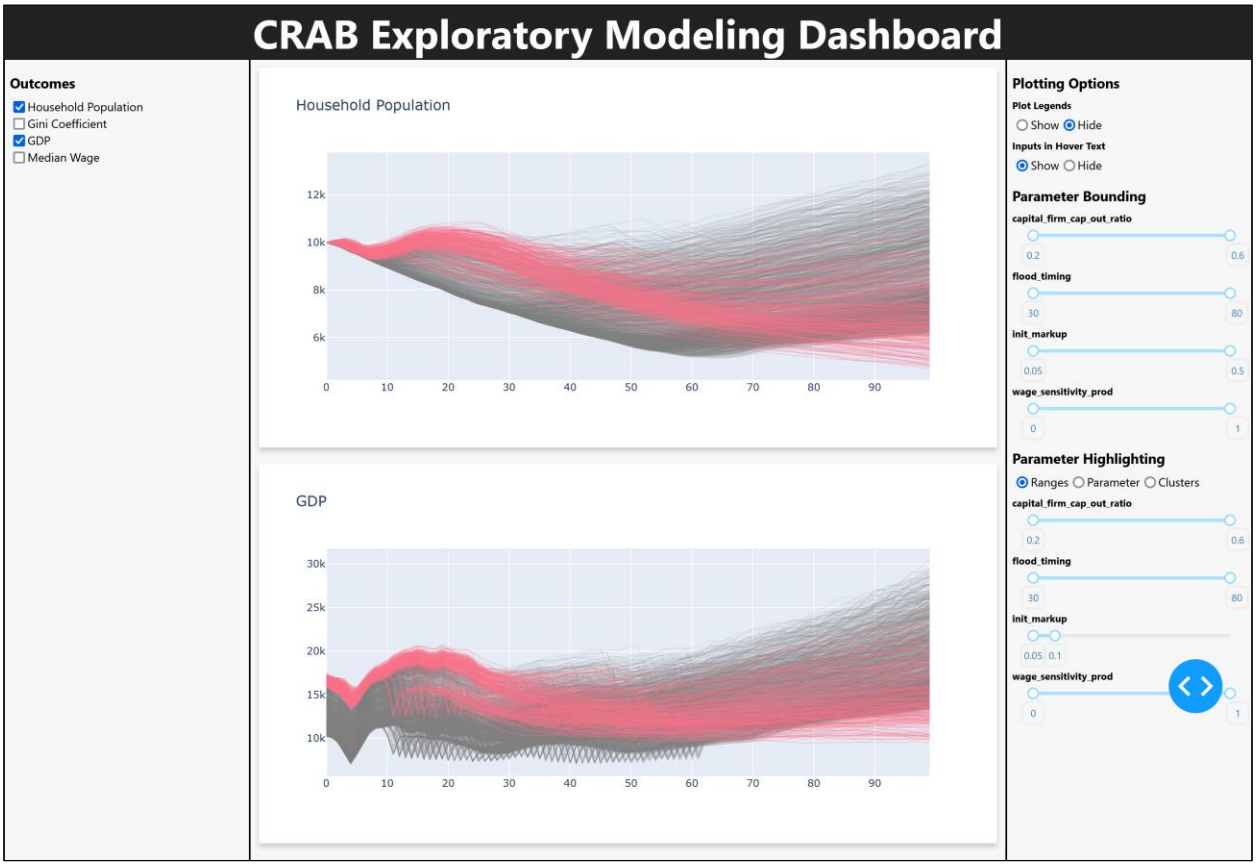


Figure 3-2: Screenshot of CRAB dashboard developed in Dash. The left-hand column allows users to select which outcomes to plot. The right-hand column allows users to select which samples to exclude or highlight within the plot.

The dashboard was developed for visual analysis and open exploration of the output space, allowing me and other analysts to explore how changes in the parameters influence the outcomes before moving to statistical methods in further steps. It was built in Python using Plotly's *Dash* package, which is free and open source. Dash provides a useful interface for developing both an HTML page and dynamic graphing objects using just Python code. This interface was then connected to the model output data generated by running the CRAB model using the EMA Workbench. It is implemented such that it should be agnostic to any model data generated with the Workbench: it infers the input and output variables and their ranges and uses those to expose dynamic controls to the user. Beyond the default range-highlighting mode, the dashboard can apply a sequential colour palette to depicted model runs according to a specific input variable of choice. If there are any clustering variables represented in the model data tables, the dashboard can also group timeseries outputs according to those clusters.

The primary use case of the dashboard is to look at the portions of the output space associated with specific areas of the input space, on demand. One way this could be (and was) used for this study is isolating the top and bottom ten percentile cases of samples, in each parameter one at a

time. This enables visualizing the effect of extreme input values on model outcome. The dashboard enables further exploration of interaction effects, by isolating small portions of the input space associated with specific ranges of many input variables at once. Finally, the dashboard enables a quick view of the entire output space at once, as a user can see the full range of outcomes as well as select and study any arbitrary individual model run. The dashboard is broadly useful for initial exploratory analysis of complex time-series model output, though at present is implemented exclusively for experiments run using the EMA Workbench (Kwakkel, 2013). The source code is available here and archived here (Sher, 2024b).

3.2.3. Sensitivity Analysis

Sensitivity analyses are a class of model analysis method that evaluate the influence each model input—parameters, in this case—on each model output. In essence, they calculate the degree to which input variable contributes to variance in an output variable (Saltelli et al., 2019). Sensitivity analysis is typically applied to scalar outcomes. To bring sensitivity analysis to ABMs, whose outputs typically have a temporal dimension, several authors have suggested computing sensitivity at each time step of the model, creating a time-series measure of input variable sensitivity (Magliocca et al., 2018; Ligmann-Zielinska & Sun, 2010).

A brief sensitivity analysis of the CRAB model is carried out using the feature scoring package built into the EMA Workbench. The package measures sensitivity as the proportion of an outcome's variability that is attributable to a specific input. For any given outcome, the sum of the sensitivity indices for each input must be 1. The package's methods are applied to the model outcomes once per simulation-year (every 4 timesteps) to assess input variable importance as a time-series.

Sensitivity analysis serves two purposes in this study. First, it acts as the first statistical indicator of which input variables might be most responsible for the divergent dynamics of the outcomes. Much like visual analysis, this helps the analyst gain an intuitive understanding of the model's behaviour and relationships before moving forward. Second, it can be used as a validation of the number of input samples in the experimental design. The time-series sensitivity for each input-output pair can be computed using an increasing number of samples, and then results can be plotted to observe whether the sensitivities converge as the number of used samples approaches the total number of measured samples.

3.2.4. Behaviour-Based Scenario Discovery

Steinmann et al. (2020) introduced behaviour-based scenario discovery as a way of expanding the application of scenario discovery to models with timeseries outcomes, framing scenarios of interest based on the dynamics they are associated with, rather than end states. As described in Section 2.2, scenario discovery is broken down into three main steps: measurement, identification, and rule induction. The Measurement process has already been described in Section 3.2.1. The further two steps constitute the scenario discovery-specific analyses used in this study.

3.2.4.1. *Time-Series Clustering*

First, runs are grouped based on behavioural similarity in each key outcome. Steinmann et al. (2020) perform univariate clustering, meaning each outcome variable is associated with its own set of clusters. This makes the clusters relatively easy to interpret, validate, and understand, as visualizing each cluster against the whole set of time series should highlight their distinct features. Multivariate time series clustering is more complex to compute, visualize, and interpret, but several methods exist (Zhou & Chan, 2014; Singhal & Seborg, 2006; Li & Liu, 2021). Using multivariate time series enables clusters to represent richer and more meaningful model states, but analysts must spend more time studying the clusters and ensuring that the clusters represent qualitatively distinct, real-world phenomena. This relies a bit on *a priori* knowledge of the system being modelled. To simplify analysis, this study will rely on univariate clustering.

Steinmann (2018) compared several similarity metrics for use in time-series clustering of complex systems models. They selected Complexity-Invariant Distance (CID) (Batista et al., 2014), a measure that uses the Euclidean distance between same-time values in each pair of time-series weighed according to the time-series' complexity. The latter step is taken to ensure that complex, dynamic time-series are not incorrectly considered far away from other time-series even when their dynamics are similar. In a different study, Jafino & Kwakkel (2021) compared several clustering algorithms for use in scenario discovery. Their conclusions found K-means clustering performs better than the alternatives, for the purposes. K-means clustering based on CID is already implemented in the EMA Workbench and will be used for this study.

There are several ways to select K (the number of clusters) when performing K-means clustering. This study emulates Jafino & Kwakkel (2021) in using the *elbow method*. This method performs K-means clustering for increasing K values until a set of clusters is found such that moving to the next K up and performing clustering again does not sufficiently improve the clusters' performance. Clustering performance here is measured by *explained variance*, which is a measure of the degree to which a set of clusters fully explains the variance in the dataset. The explained variance of a set of clusters of size *K* can be measured as:

$$EV_K = 1 - \frac{\sum_{k=1}^{K} SSE_k}{SSE_{all}}$$

where SSE_k is the sum of the squared errors of the members of cluster k, and SSE_{all} is the sum of the squared errors for the entire dataset (Jafino & Kwakkel, 2021).

Once a K value is selected, the original experimental results are expanded to save the cluster with which each sample was associated. Finally, representative samples are selected from each cluster. This is done by calculating the centroid of each cluster and selecting the sample that minimizes the mean-squared-error between it and all other runs in the cluster.

3.2.4.2. *Rule Induction*

Rule induction is performed using the PRIM method as it is implemented in the EMA Workbench. The independent variables used in the PRIM process are the model parameters described Table 3-1, i.e., the dimensions which PRIM looks to restrict. The dependent variable is the cluster to

which a given input sample belongs. This version of PRIM requires the dependent variable to be a binary variable, thus PRIM is run separately for each cluster, and cases are marked as in or out of that cluster. Ultimately, a rule will be generated to explain each cluster of model outputs, expressed in terms of the model's input parameters—loosely, a scenario.

PRIM's primary output is a sequence of candidate *rules*, each represented by a set of restricted ranges within the model's input parameters. These restrictions can be visualized as a bounding box within the multidimensional input space, so PRIM rules are often called "boxes." As described in Section 2.2, a PRIM box can be evaluated according to its coverage (how many of the cases of interest fall inside the box) and density (how many of the cases inside the box are of interest). Each box found by PRIM is more restrictive than the last and usually trades off some performance in terms of coverage for added performance in terms of density. Thus, we must decide which boxes to select such that we maximize coverage and density. As a rule of thumb, we will try to select boxes that maintain a coverage above 80%: the rule needs to be able to explain at least 80% of the cases that fall within a given cluster. However, we will manually study the coverage-density trade-off of each sequence of PRIM boxes to ensure an adequate choice is made.

Each selected box from PRIM defines a portion of the parameter space associated with a qualitatively distinct behaviour: together, these make a scenario. If the model and experiment have been designed correctly, the set of scenarios for each cluster provides an analyst with the full set of behaviour scenarios possible in this system (at least in terms of the outcome(s) of interest). Amalgamating the rules for each of these scenarios forms a sort of function that maps the model's parameters to an expected dynamic system behaviour.

Ideally, only a small number of input dimensions are restricted across the amalgamated set of rules. This aids both comprehensibility and visualization.

3.2.4.3. *Identifying Tipping in the Parameter Space*

In essence, we have first defined the system's phases (by performing clustering on its outcome dynamics) and then identified which parameters are key predictors of that phase, and how so (via rule induction). While this is where traditional scenario discovery may end, this study seeks to then find tipping points, i.e., points at which the system tips from one expected behaviour to another. In other terms, we are looking for the conditions under which a realization of a system stops belonging to one scenario and begins to belong to another.

With this framing, the use of the phase diagram becomes clear. Plotting the mapping function described in Section 3.2.4.2 would produce the system's phase diagram. Then, we can look at the boundaries between phases to find points in the model's parameter space where the system changes phase: a tipping point. However, just as with traditional scenario discovery, the job is not finished without further analysis. It is unhelpful to generate a phase diagram and suggest that that defines all relevant tipping points in a system. Practically, phases could be overlapping (i.e., there could be bad *input space separability* between the phases (Jafino & Kwakkel, 2021)) or there could be large gaps in the phase diagram, preventing the easy interpretation of tipping points. Additionally, one must relate the results back to the real-world system under study and attempt

the causes and implications of the tipping point, as well as how we might look out for it in the real world.

4 Results & Analysis

This section will use the CRAB model to demonstrate the methodology for discovering tipping points described above in Section 3. For this demonstration, the output variable *Household Population* will be used as the focus of study. This variable serves as an indicator of the overall performance of a coastal region. As coastal denizens increasingly consider the option of retreat (i.e., migrating away from coastal regions), it is important to understand the conditions under which coastal regions are still able to maintain or grow their populations, and the conditions under which their populations collapse. The latter is important to know so that governments can prepare for coastal retreat by investing in support for migrants as well as ensuring services are provided for all citizens, even as populations shrink. Similar analysis applied to the *GDP* and *Gini Coefficient* (inequality) outcomes can be found in the Appendix.

Part of the reason for this single-variable focus is that the clustering performed in this process relied on a univariate clustering algorithm. As discussed in Section 3.2.4.1, multivariate clustering is a powerful explanatory tool, but it is both computationally more expensive and requires more effort to translate into tangible scenarios. Future versions of this research should be carried out using multivariate clustering.

Alongside the results for the main set of model runs, we will present the results of a set of simulations run at the same set of input samples but with a modified model that produces no flooding. This might provide an additional way to isolate the effect of the flood itself and explain the results of the model runs.

4.1 Visual Analysis

For an initial exploration of potential tipping in CRAB's parameter space, the processed model data was loaded into the interactive, browser-based dashboard developed for open exploration of timeseries data, as described in Section 3.2.2. Figure 4-1 shows the 2000 timeseries outcomes plotted together with different highlighting rules, and Figure 4-2 shows the same outcomes for the no-flood run.

It is clear from this initial analysis that some parameters—in this case, the initial markup rate—have a stronger influence on the dynamics of the population outcome than others. The top 10 percentile of cases by initial markup comprise a group of model realizations with highly similar dynamics. It is expected that this parameter might be predictive of a cluster in this *Household Population* outcome.

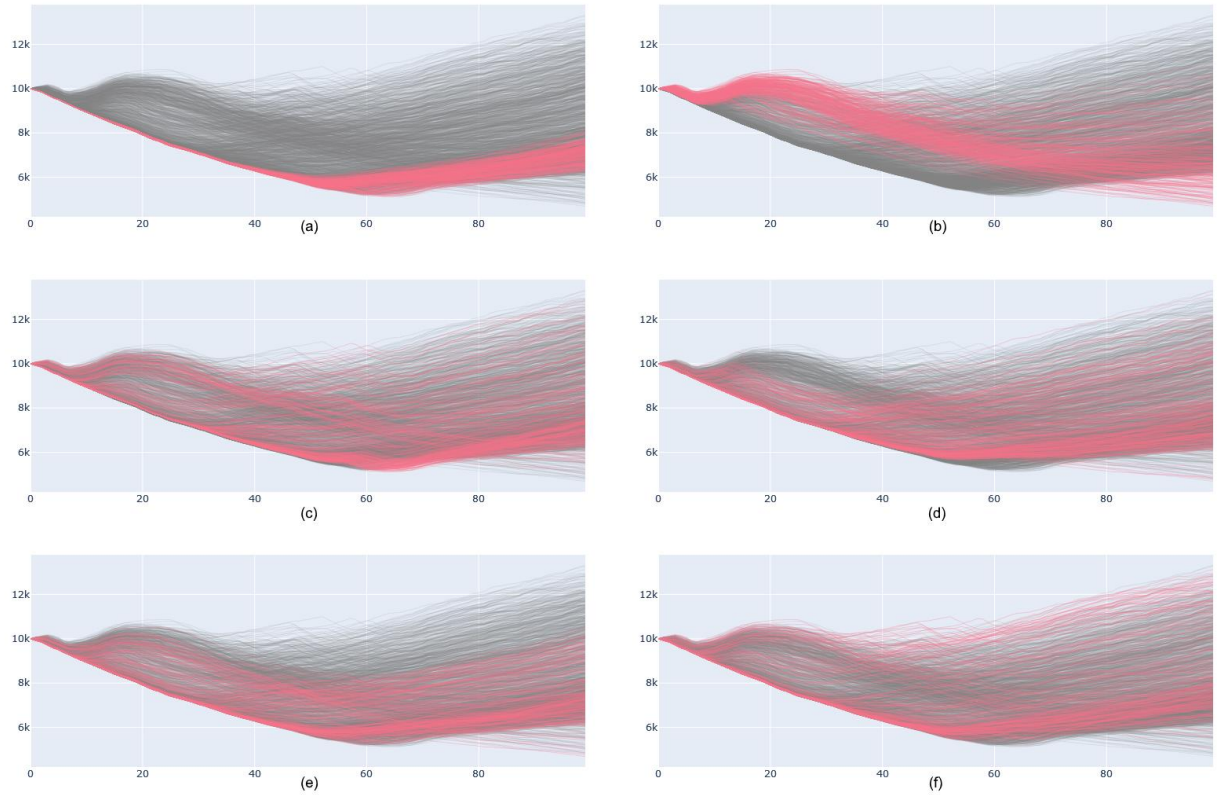


Figure 4-1: 2000 realizations of the Household Population outcome of the CRAB model. The highlighting in (a) and (b) reflect the top and bottom 10 percentile cases in terms of the initial makup parameter, respectively. (c) and (d) are highlighted according to the flood timing parameter. (e) and (f) are highlighted according to the capital-output ratio.

We can study the model runs generated at the same sample points but without flooding to get a sense of the role of the flood on the performance of the model:

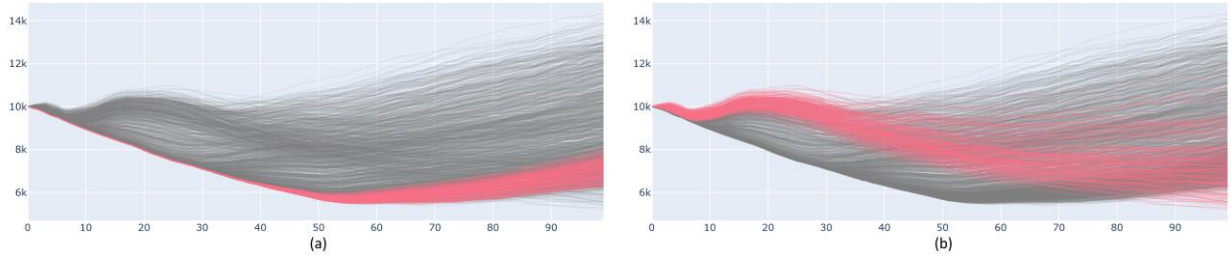


Figure 4-2: 2000 realizations of the Household Population outcome of the CRAB model, run with no flooding. (a) reflects the top and (b) reflects the bottom 10 percentile cases in terms of the initial markup parameter.

Initially, the lack of flooding seems to minorly decrease the dispersion of model results (seen most notably around timestep 20), implying that flooding does have at least some effect on the model behaviour. In Figure 4-1 and Figure 4-2, the flood can first occur at timestep 10 (per Table 3-1, flooding starts at timestep 30, but a 20-step burn-in time has been accounted for before creating these plots).

4.2 Sensitivity Analysis

The subplots in Figure 4-3 shows the timeseries sensitivity of the outcomes named on the y-axis to each parameter.

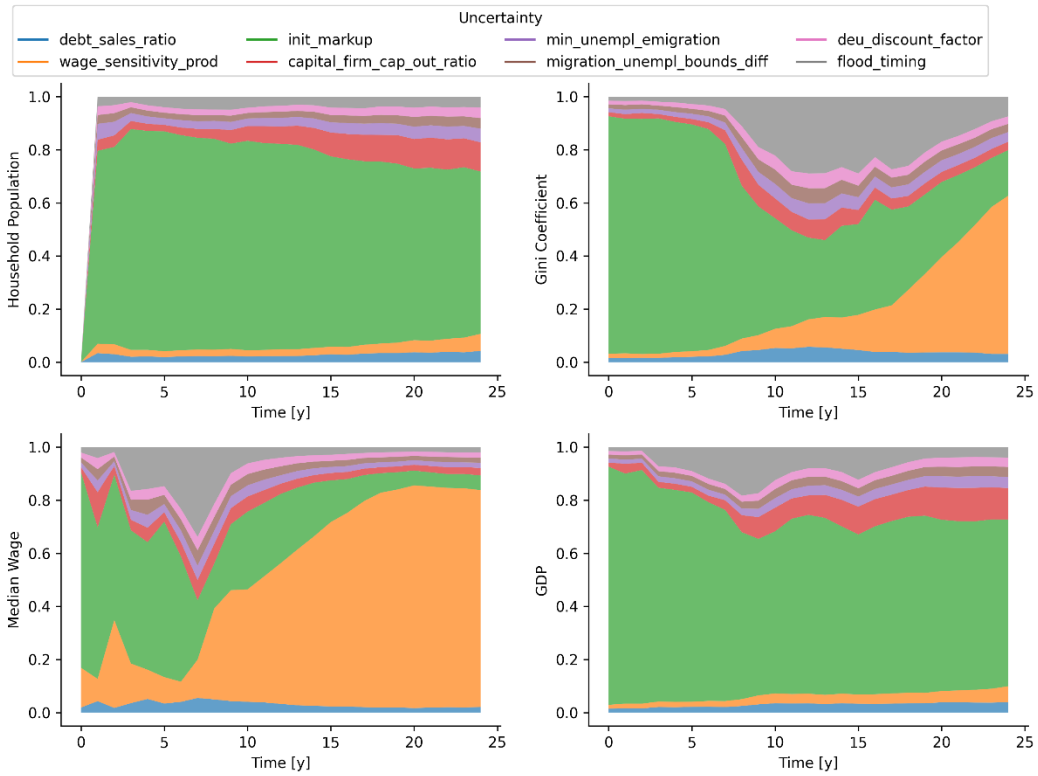


Figure 4-3: Time-series sensitivity of each key outcome to each input parameter in the CRAB model.

Again, the *Initial markup* parameter dominates. Its importance is most clear in the *Household Population* and *GDP* outcomes, where it remains significantly more important than other factors across the time horizon. For *Gini Coefficient* and *Median Wage*, however, its importance wanes as the model run advances. This makes sense, as it is a parameter that sets the initial value of a variable which then changes endogenously. In both cases, the *Sensitivity of wages to productivity* parameter overtakes it in importance as time goes on. This also makes sense, as this parameter dictates how wages are reset at each timestep, and thus its influence compounds as model time goes on. The *Gini Coefficient* is partially calculated based on the distribution of wages across the population, so it is expected that both it and *Median Wage* would be sensitive to a parameter that informs wages.

It is notable that in at least two of the outcomes, the sensitivity to each parameter changes drastically over time. This implies that the parameters predictive of a variable's level *at a given timestep* may change over time. This does not, however, address which parameters are predictive of an outcome's dynamics or behaviour across the *entire* time-series. For that, we must turn to behaviour-based scenario discovery.

According to sensitivity analysis, the model behaves relatively similarly when flooding is taken out, per Figure 4-4, below. A key difference is illustrated by the factors with high influence on inequality (Gini Coefficient, top right) in the later stages of the time horizon. In the flooded version of the CRAB model, *Flood timing* is an important factor to inequality, so in the unflooded version, its influence on the outcome is distributed across the other factors, notably as lingering importance for the *Initial markup* parameter. However, the dominant control behaviour remains: for most of the model run, the *Initial markup* parameter is most important to three of the four key outcomes. Only the *Sensitivity of wages to productivity* factor begins to become meaningfully influential near the end of the model run, and in that, only for two of the four outcomes.

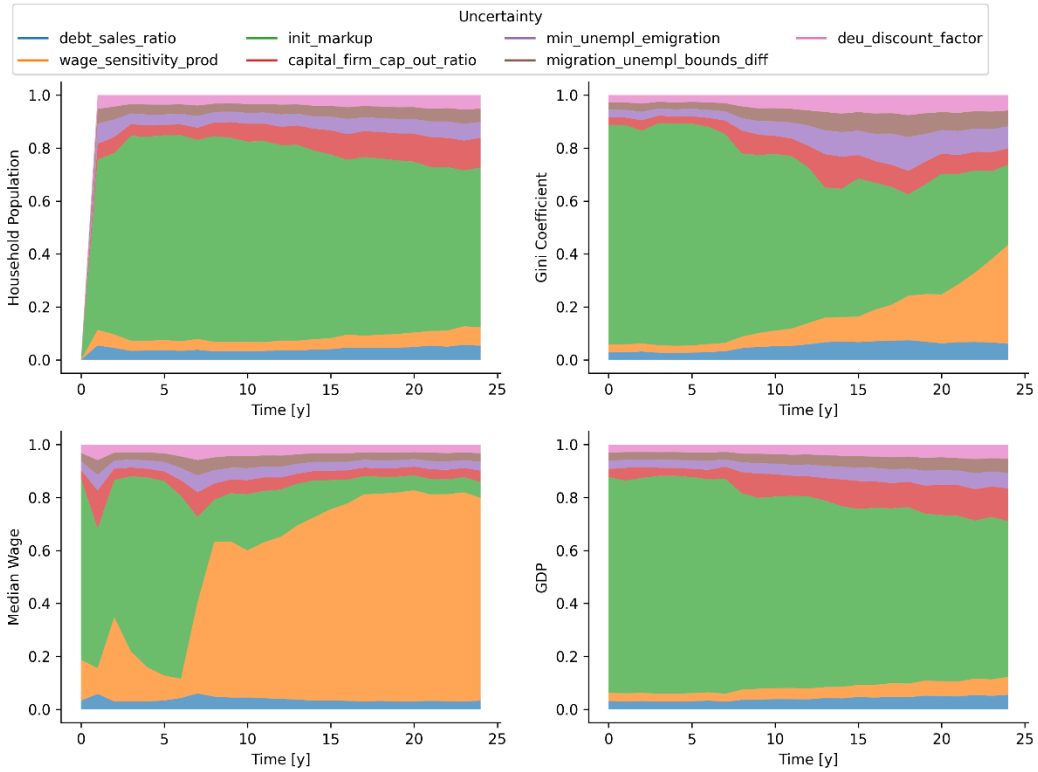


Figure 4-4: Time-series sensitivity of each key outcome to each input parameter in a version of the CRAB model without flooding.

4.3 Validation of Experimental Setup

Before we can use the above results to make claims about the relationships between uncertainties and outcomes in the CRAB model, we must validate that the experiment considers a sufficient portion of the model's uncertainty space.

4.3.1. Selection of sample size

Sensitivity analysis already provides a measure through which we can evaluate the strength of an individual input's influence on a model outcome. As more samples are used (and thus more of the model's parameter space is covered), these sensitivities should converge towards their true values. Thus, we can observe how input sensitivity changes as more samples are added to

validate the final number of samples used. Figure 4-5 presents such a plot. Each subplot shows the time-series sensitivity of *Household Population* to the input parameter named on the y-axis. The lines represent the time-series sensitivities using sample sizes increasing by a step of 100, with larger samples shown using brighter lines. This method exploits the feature of Latin Hypercube sampling where each $N+1^{th}$ sample fittingly expands the space covered by the previous *N* samples as if the size had been N+1 from the start.

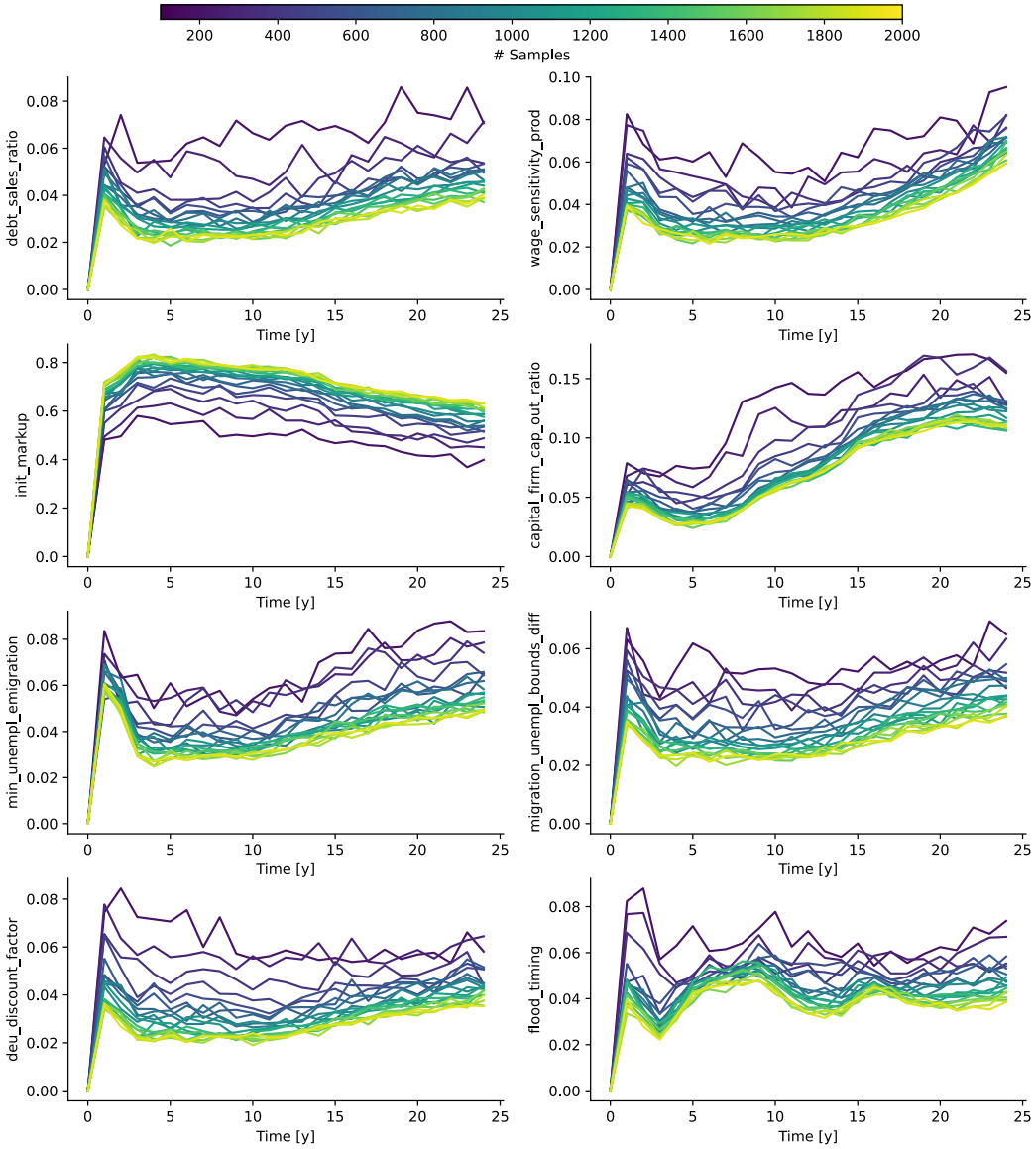


Figure 4-5: Convergence of input sensitivities as the number of samples increases. The y-axis denotes the importance of the named parameter to the outcome of interest, Household Population. Note the lack of shared y-axis.

This method is in part adapted from the tendency in ABM literature to select the number of experiment replications where output variance coefficients begin to converge (Lorscheid et al., 2012). Rather than looking at the variance across a set of replications, we look at the sensitivity across a set of sample points. Since the model is expected to behave differently at different

portions of the parameter space, sensitivity serves as a better basis than variance for an information convergence check as sample size increases.

As is visible, sensitivity begins to converge at around 1000-1200 samples; the sensitivity time-series for N_{samp}=2000 are not so different from those with N_{samp}=1000. Interestingly, only for *init_markup*—the most explanatory variable—does sensitivity increase as sample size increases. This suggests that, at small sample sizes, the effect of important variables cannot be as clearly distinguished, and some of the outcome variability that should be attributed to *init_markup* is instead distributed across the other parameters.

4.3.2. Selection of number of replications

For this study, the input space is the cross between the CRAB model's parameter space and its stochastic variability. Though the main analysis of this study combines replications into one realization per sample point, and thus does not treat stochastic uncertainty as an explanatory variable, it is still important to ensure enough replications are used to cover the range of model behaviours across stochastic replications.

One way to measure the size of the effect of stochastic uncertainty on combined model performance is via the time-series variance of each key metric. Variance is a statistical measure of a variable's deviation from its mean. High variance means that, at a given sample point, the model's performance is highly sensitive to stochastic uncertainty. Low variance means that, at a given sample point, the model behaves relatively similarly across all model replications. As the number of replications grows and thus the range of distinct model behaviours is increasingly covered, variance should converge. Figure 4-6, below, shows how variance in *Household Population* converges as the number of replications increases. The subplots show variance convergence at six different sample points which display qualitatively distinct dynamics in the *Household Population* variable. (In fact, they are the points identified as being representative samples for each of the clusters of that variable's dynamics, which will be identified in Section Number).

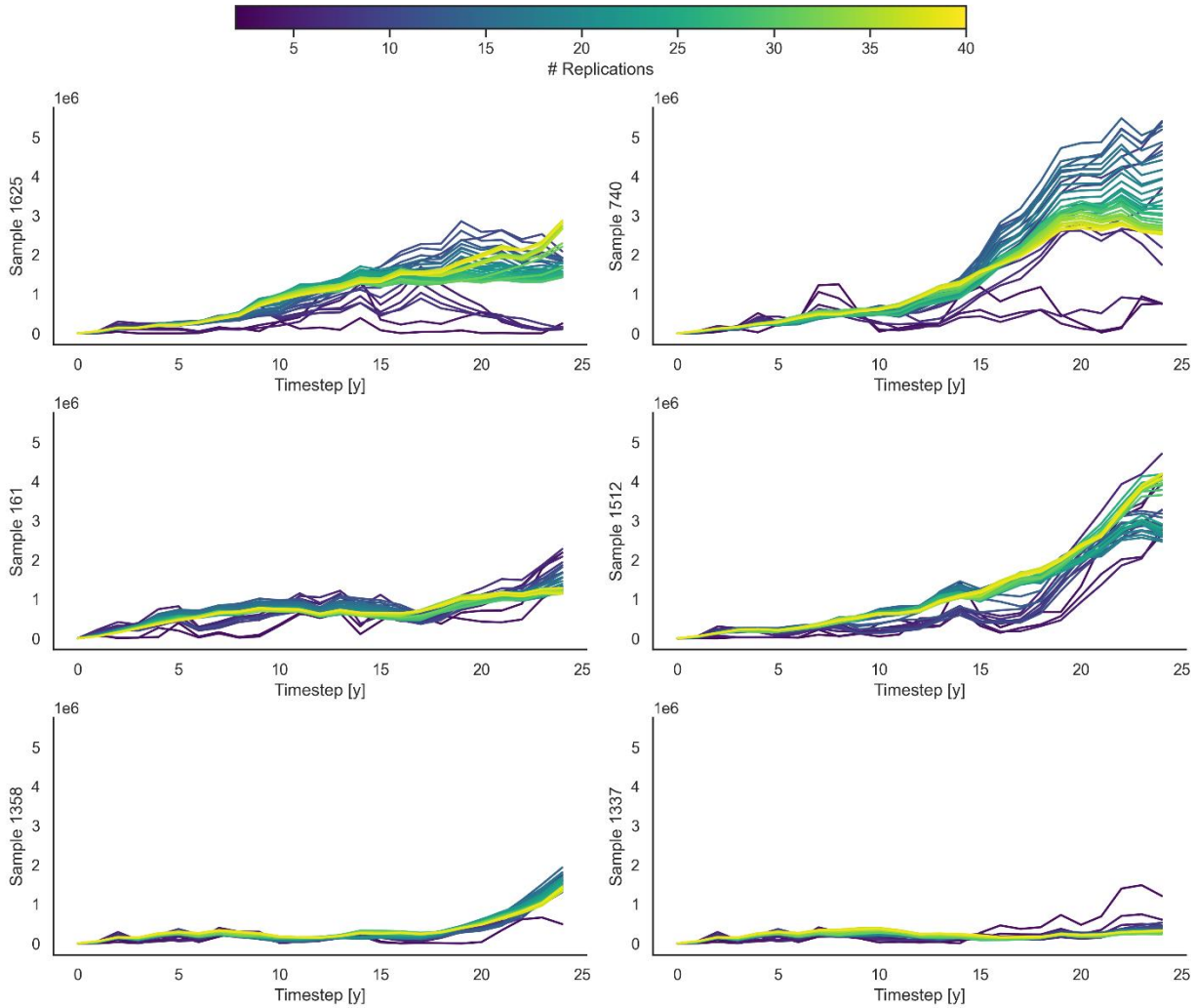


Figure 4-6: Convergence of time-series variance in Household Population outcome across an increasing number of replications. Variances are calculated at six different input samples, selected to be representative of each discovered cluster (scenario).

Two things are of note. First, the six subplots are, for the most part, visually distinct from one another. This implies that time-series variance differs in distinct portions of the model parameter space: i.e., the range of stochastic effects changes as the input parameters change. For example, Sample 1337 (bottom right) has a much lower and much flatter variance than Sample 740 (top right). Second, there are occasionally sudden jumps between subsequent lines in a given subplot. This is perhaps most visible in the plot for Sample 1625 (top left), at the end of the time horizon, as the lines for N_{reps} goes from ~30 to 40. This suggests individual replications were introduced that are radically different from the prior ones and thus have a large influence on overall variance. Given the abruptness with which this occurs, it is hard to assess whether the choice of N_{reps}=40 is enough. It is also unclear what would be a sufficient stopping rule using this analysis.

4.3.3. On choosing to combine replications

To explore more closely the breadth of stochastic realizations in the CRAB model at a given sample point, Figure 4-7 presents individual model runs alongside the processed mean as a visual indicator of information lost.

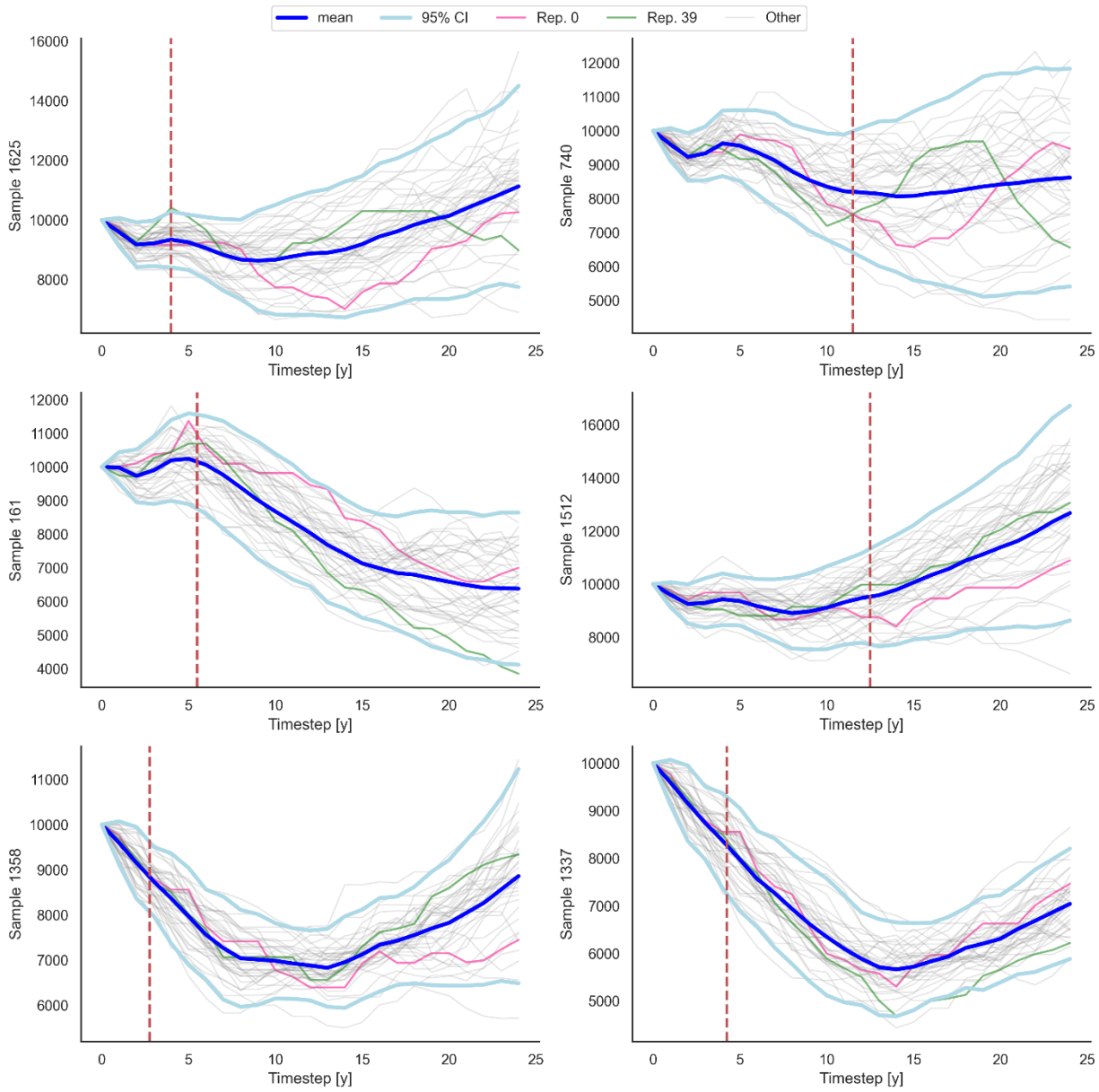


Figure 4-7: Aggregated dynamics of Household Population (mean and 95% confidence intervals, shown in blue, taken across 40 replications) atop dynamics of individual model replications, plotted separately for each of six sample points. Two arbitrary samples are highlighted in pink and green to help demonstrate distinct dynamics. For visual acuity, time-series have been resampled intertemporally (one value per year, instead of per quarter). The vertical red line indicates the timing of the flood for the relevant input sample.

It is clear from Figure 4-6 and Figure 4-7 that variance in the outcomes of the CRAB model can be relatively high, and that there are clearly distinct stochastic realizations of the model, even at

the same sample point (for example, Samples 1625 and 740 show wildly different dynamics between the two highlighted replications, whereas for the other four sample points, the two highlighted replications show more similar dynamics that just differ in level). Thus, some dynamic information is lost by taking the mean and processing replications together. In tangible terms, by using the average of stochastic replications, we could be looking at outcome behaviours that would never occur in reality, as they are the smoothed result of averaging several distinct behaviours together. The decision to combine replications is thus largely informed by the trade-off of realism and computational resources. Further discussion on the implications of this decision and ways to address the trade-off can be found in Section 5.2.

4.4 Behaviour-Based Scenario Discovery

4.4.1. Time-Series Clustering

Clustering was performed by first calculating the CID between the *Household Population* timeseries at each sample point and then applying k-means clustering to the sets of distances. To select an adequate number of clusters, one can use the elbow method, as discussed in 3.2.4.1. We apply this clustering for a range of K-values (total number of clusters) and select the one after which subsequent increases in K provide small or diminishing improvements in terms of explanatory value.

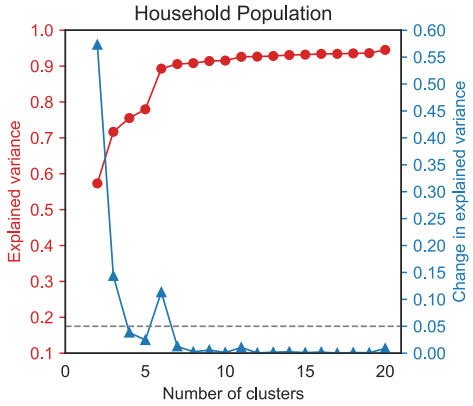


Figure 4-8: Explained variance and changed in explained variance gained by increasing K, for the Household Population outcome. The grey dotted line represents a 5% change in explained variance, the threshold that is used to determine an optimal K.

Figure 4-8 plots the explained variance (a measure of a set of clusters' explanatory value) in red, alongside the change in explained variance in blue. A traditional application of the elbow method might select $K = 3$, as the gain in explanatory value from moving to $K = 4$ falls below the threshold of 5% (this is the same value used in Jafino & Kwakkel (2021), but is somewhat arbitrary: it is selected to be a small value). However, it is clear from this plot that a move to $K = 6$ also produces a pronounced increase in explained value. Thus, let us look at both sets of clusters ($K = 3, K = 6$) and select one of the two.

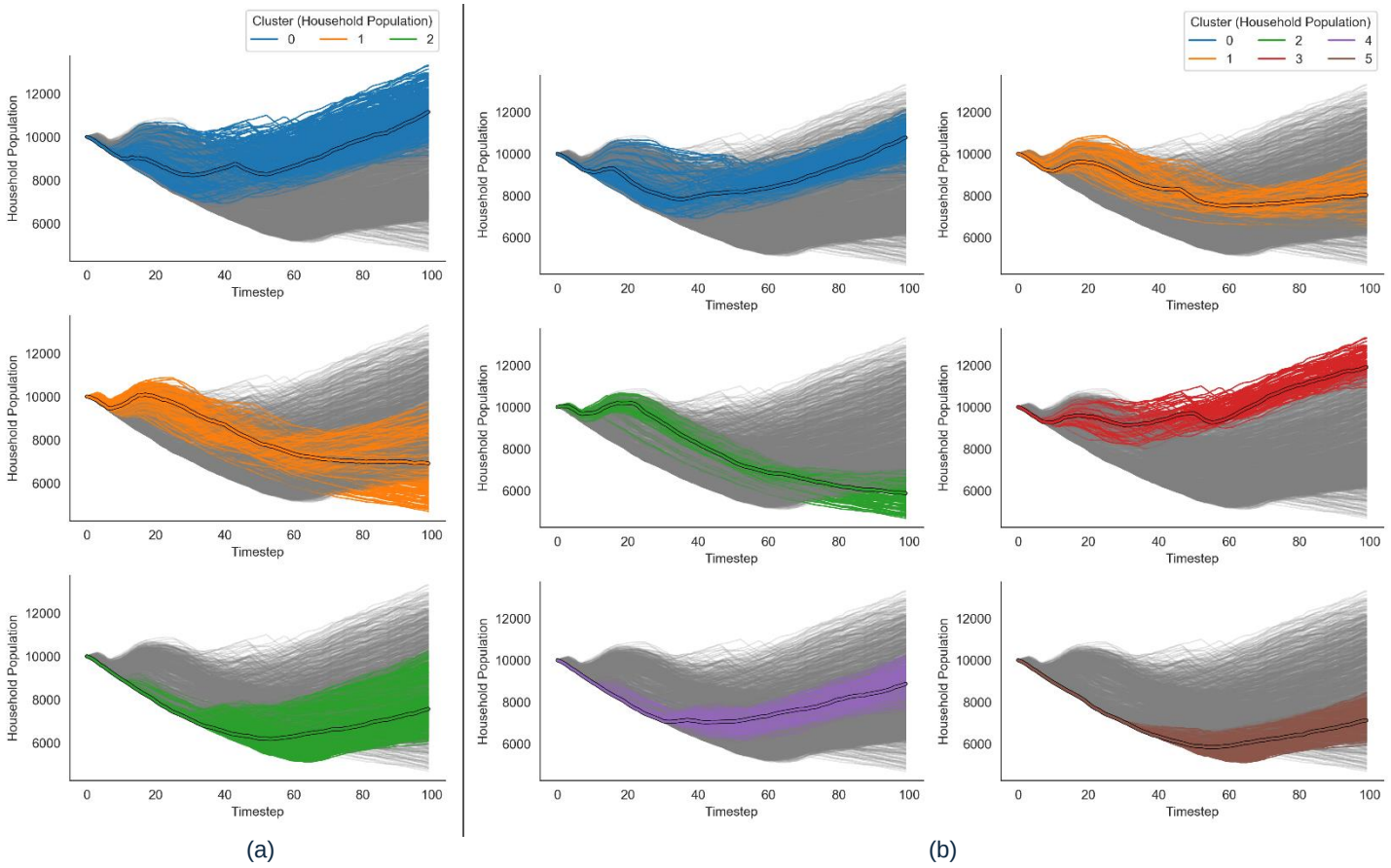


Figure 4-9: 2000 realizations of the CRAB model, visualized according to cluster in Household Population. (a) depicts the clusters chosen with $K = 3$, (b) depicts the clusters chosen with $K = 6$. Representative samples are highlighted with a black border.

Figure 4-9 shows the time-series model realizations clustered in the *Household Population* outcome for both $K = 3$ and $K = 6$. Each cluster is highlighted in its own plot, though all 2000 model realizations are captured in every plot. Representative samples—model runs typical of their cluster—are shown highlighted with a black border. It appears that each cluster from the $K = 3$ set splits into two clusters in the $K = 6$ set, which may explain why $K = 6$ had a pronounced increase in explained variance relative to $K = 4, K = 5$. The differences follow:

- The first cluster in Figure 4-9(a) seems to be represented in the blue and red clusters in Figure 4-9(b). These clusters depict the cases with high population and somewhat steady population growth. In the $K = 6$ set, the red cluster represents the highest-population model runs, while runs in the blue cluster tend to see a fall to about 6000-8000 households in the early timesteps before turning to growth. The red cluster sees population start high and stay high. The rate of growth for both clusters is relatively high. Most runs in the blue cluster and all runs in the red cluster end with a population greater than the starting point, meaning more households moved to the region than left. Especially within the blue cluster, there is some amount of dissimilarity between the model runs in the early timesteps. Generally, both clusters are characterized by growth.

- The second cluster in Figure 4-9(a) is split between the orange and green clusters in Figure 4-9(b). These clusters depict the cases with steady population decline. In the $K = 6$ set, we see more differentiation between cases that eventually reach a steady state with a somewhat depressed population (orange) and cases that continue to decline throughout the model run (green). The green cluster demonstrates a small amount of levelling off at the end of the time horizon, mirroring the orange cluster but much later and at a much lower population level. Notably, the cases in the green cluster have the highest density of high growth in the very early timesteps, though cases in the blue, orange, and red clusters also demonstrate this phenomenon. There is a bit of dissimilarity within the orange cluster, as some model runs look as though they begin to grow by the end of the time horizon. While the representative sample appears to level off and remain steady, the divergent behaviours in this cluster should not be ignored if a scenario were to be developed using this cluster.
- The third cluster in Figure 4-9(a) is split between the purple and brown clusters in Figure 4-9(b). These clusters depict the cases that have a very pronounced collapse in population at the start of the model run, and until roughly timestep 30 there is very little difference between runs in these clusters. However, runs in the purple cluster begin to recover earlier (around timestep 30), whereas runs in the brown cluster tend to take until about timestep 50 to begin to recover. After these recovery points, both clusters take on slow-but-steady growth.

It is important to note that in these timeseries, timestep 10 is the first moment at which a flood can occur in the model. This coincides with several notable dynamics, especially where model runs in the blue, orange, green, and red clusters have some uptick in population before continuing onto their dominant dynamic, and where those in the blue, orange, red, and purple clusters have some divergence before mostly coming back together. In the early timesteps of the latter set, the orange, red, and purple clusters tend to have shared dynamics (but divergent levels) in the early timesteps, whereas the blue cluster shows some divergent dynamics.

The biggest gain in information from moving to $K = 6$ is the differentiation between the orange and green cases. When lumped together for $K = 3$, the fact that some cases continue to decline while others level off and maybe even grow is lost. Additionally, the temporal separation between the recovery points in the purple and brown clusters is important. Thus, we will move forward with the $K = 6$ set. Each cluster will be named as follows:

- Blue: Fall-Growth
- Orange: Fall-Level Off
- Green: Growth-Collapse
- Red: Steady Growth
- Purple: Collapse-Recovery @ 30
- Brown: Collapse-Recovery @ 50

Each model run will be marked as belonging to one of these six clusters, serving as the foundation for the subsequent multi-class scenario discovery. An analyst could here decide to drop one or several clusters from the remaining steps of analysis. For example:

- The Collapse-Recovery clusters might be considered dynamically similar enough that only one may be of great importance to study (perhaps the brown cluster, which includes the more extreme subset of the model runs showing early population collapse).
- The red and possibly blue clusters might be considered "good" cases, as they both end with high population and somewhat steady growth (presumably towards a levelling-off point eventually). An analyst may therefore decide these are not worth the close study afforded to the other clusters.

However, this study is interested in tipping points, and therefore is looking at the boundaries between scenarios rather than scenarios themselves. To plot a "phase diagram" of the CRAB model, it is important to understand the portions of the parameter space associated with *all* possible outcome dynamics. Thus, all six clusters will be carried forward.

4.4.1.1. *Comparison with No-Flood Case*

We can also study the results of the same experiment applied to the floodless version of the CRAB model. In this case, a hard application of the 5% change in explained variance threshold selects $K = 4$. There is still in explained variance at $K = 7$, however, it is less than 5%. Six scenarios is already asking a lot of both the analyst and the policymaker to keep track of, as one has to juggle both their distinct behaviours and the conditions under which each scenario occurs. Thus, especially since its gain in explained variance is smaller than the flooded-model counterpart, we will select $K = 4$.

Per Figure 4-10, four distinct scenarios exist. The blue cluster is representative of the *Steady Growth* cluster from the flooded model, where population only sees some minor wobbling at the start of the time horizon before settling into a growth pattern. The orange cluster is representative of the *Growth-Collapse* cluster from the flooded model, though there are some cases represented that might be representative of the *Fall-Level Off* cluster. The green cluster is highly reminiscent of the *Fall-Growth* cluster from the other version, and at the bottom of the cluster some runs that might have been classified as *Collapse-Recovery @ 30* are included. Finally, the red cluster constitutes this version's *Collapse-Recovery* cluster, with sharp population decline and eventual slow recovery.

The no-flood version of the model produces relatively similar results to that of the flooded model. Thus, further analysis will be done just on the flooded version of the model, as that is the CRAB model's original intent. It will thus be a surprising result if the *Flood timing* parameter is shown to be an important one in the scenario discovery process.

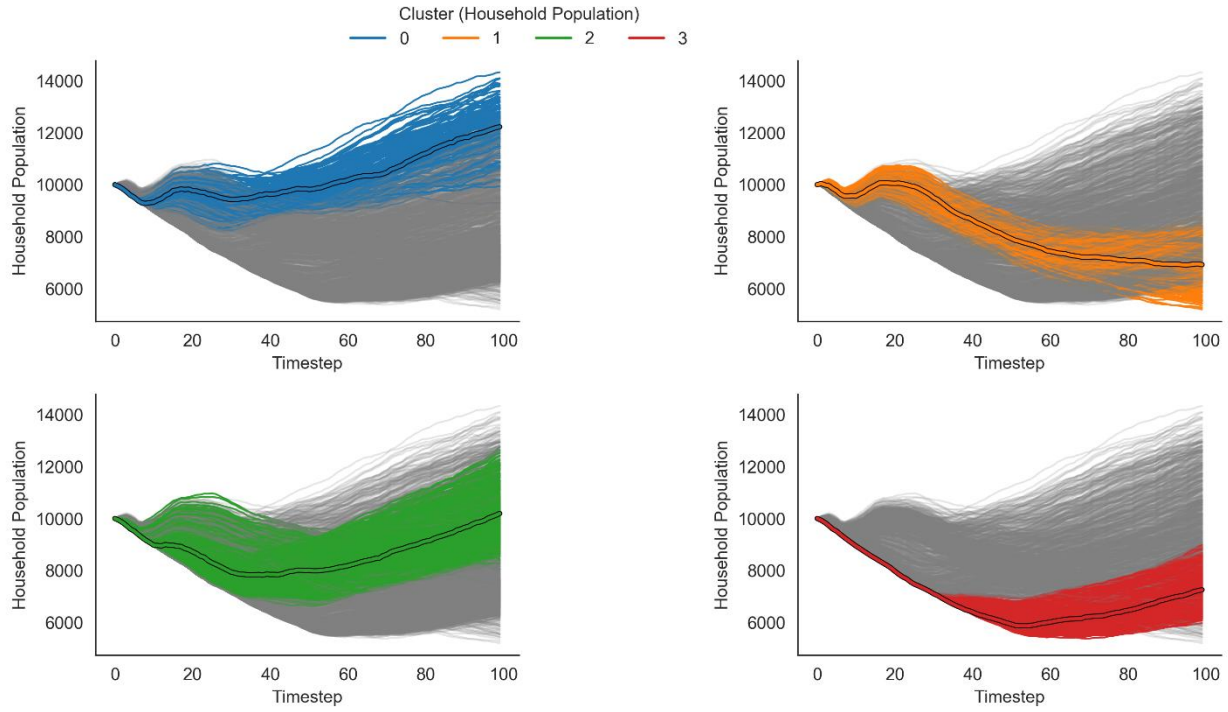
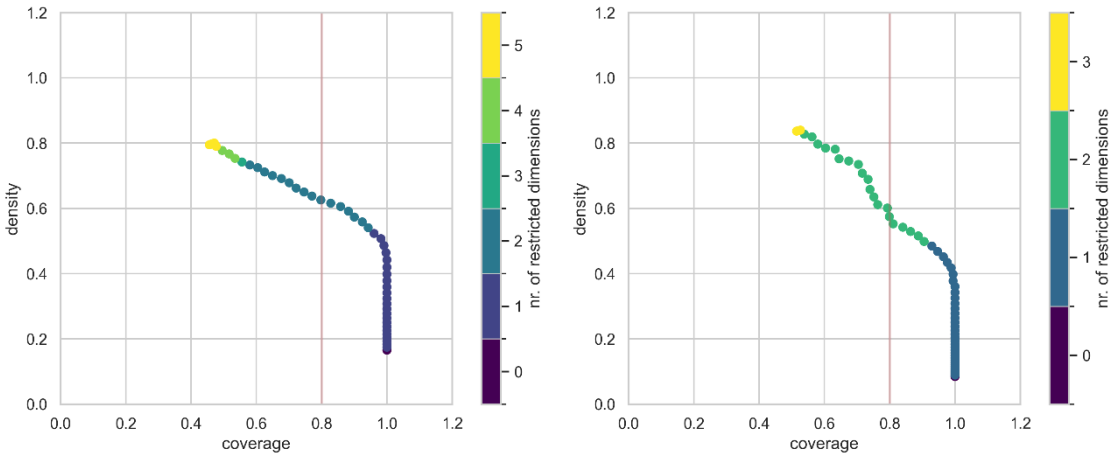


Figure 4-10: 2000 realizations of the CRAB model without flooding, visualized according to cluster in Household Population. $K = 4$.

4.4.2. Rule Induction

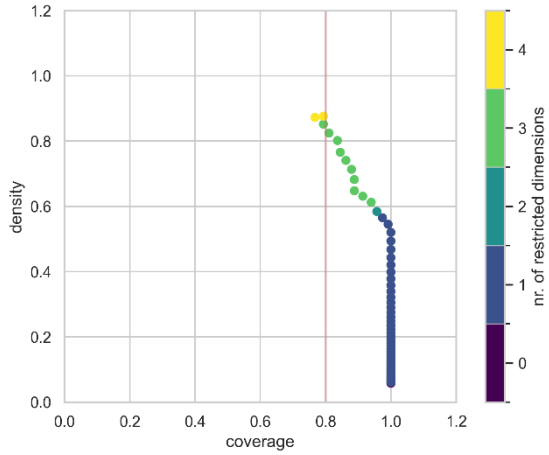
Figure 4-11 shows the coverage-density trade-offs generated by applying the PRIM algorithm to the each of the time-series clusters of the household population outcome. Each dot in the coverage-density trade-off plot is a candidate box that the PRIM algorithm has identified as a possible rule in its input space, the model parameters. PRIM starts with the unrestricted space—the right-most dot, coloured purple in all six plots. Then it additively applies restrictions in the input space that produce the largest gain in density with the smallest loss in coverage. It proceeds this way until it finds a box with 100% density, if possible—i.e., all samples within the box are of the correct cluster of interest. The vertical red line in these plots is added for the convenience of indicating 80% coverage, which will be used as a rule of thumb for a minimum-acceptable coverage for this study: the discovered rule needs to be able to explain at least 80% of the cases that fall within a given cluster.

A decision must be made regarding *which* of these candidate bounding boxes to select as the scenario-defining rule for each behavioural cluster. As the rules associated with each box become stricter, boxes become denser (i.e., more cases that fall within the box belong to their associated cluster), but they also lose coverage (i.e., fewer of the total instances of their cluster fall within the box). For some of the clusters, the selection is clear. For the *Collapse-Recovery @ 50* cluster (f), there exists a box with 100% density that has greater than 80% coverage, so this box is selected.

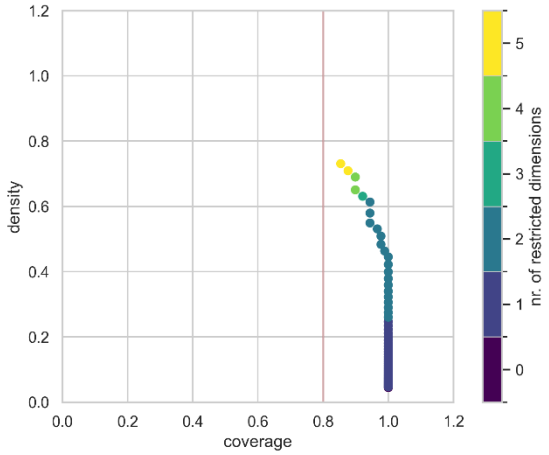


(a): Fall-Growth

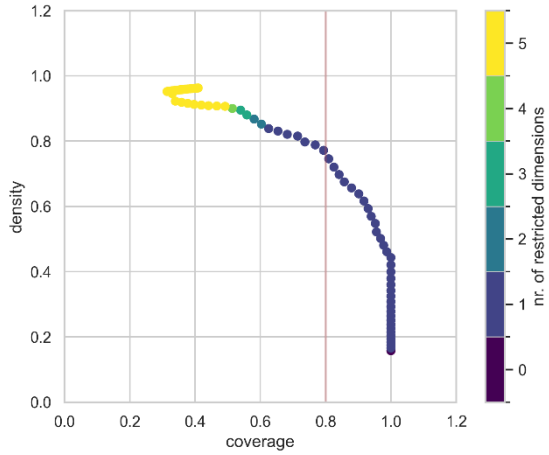
(b): Fall-Level Off



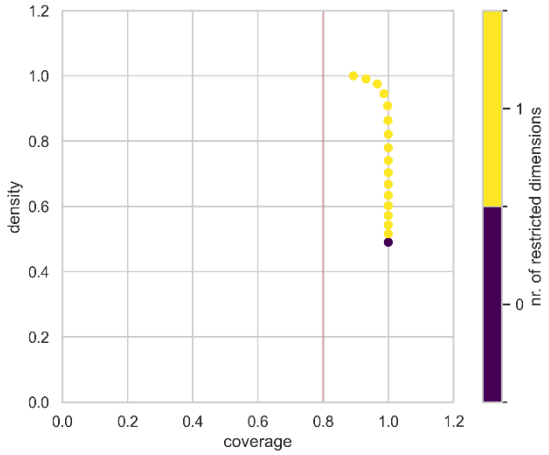
(c): Growth-Collapse



(d): Steady Growth



(e): Collapse-Recovery @ 30



(f): Collapse-Recovery @ 50

Figure 4-11: Coverage-density trade-off plots resulting from applying PRIM to each cluster.

The simple choices stop there, though. The *Steady Growth* cluster (d) has no box with less than 80% coverage, so the maximum-density box could be selected. However, this box restrictions in *five* of the eight input dimensions. To assist in comprehensibility of the selected scenarios and thus the tipping points between them, we should try to limit the number of dimensions restricted by any given rule. The upper-most green box restricts only 4 dimensions for a very marginal loss in density. The upper-most cyan box restricts only 2 dimensions and has a density just over 60%, compared to roughly 75% of the most-restrictive box. Either of these could be selected: we should look more closely at the induced rules to decide.

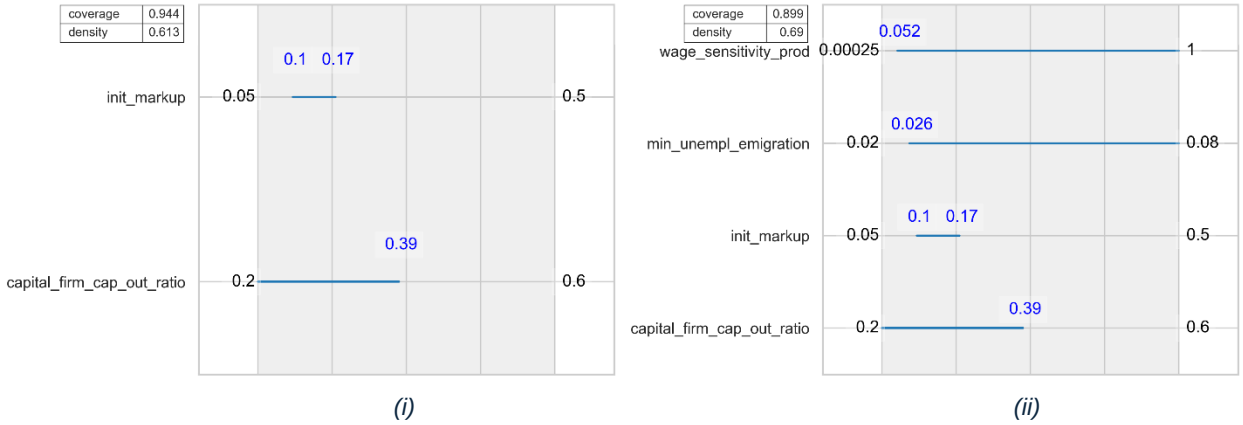


Figure 4-12: Two candidate bounding boxes for inducing a rule to describe the Steady Growth *cluster.*

Figure 4-12 visualizes the rules associated with each box. While (ii) shows that adding restrictions in two more dimensions leads to an 8-percentage point gain in density for just a 5-point loss in coverage, it is also notable that the restrictions in these dimensions are very slight. Thus, while it might be mathematically true that this restriction is beneficial, it is harder to ultimately convert into a meaningful scenario narrative. For this reason, we will select the box described in (i).

The *Growth-Collapse* cluster (c) gives a somewhat simpler choice. The four highest-density boxes all have relatively similar density and coverage, and all hover around 80% coverage. Thus, we can comfortably select one of the boxes that only restricts three dimensions.

The *Collapse-Recovery @ 30* cluster (e) has a candidate bounding box just under 80% density right at both the 80% coverage line and the elbow of its tradeoff curve, so that box will be selected.

The first two clusters present more of a quandary. Around the 80% coverage threshold, both have boxes with only 60% density, meaning that 40% of the cases that fall within these boxes belong to different clusters. The *Fall-Growth* cluster (a) demonstrates a mostly linear trade-off between density and coverage beyond a certain point, so there are no outsize gains to be made by breaking the 80% coverage threshold. Therefore, the maximum-density box just to the right of that threshold will be selected. The *Fall-Level Off* cluster (b), however, does show sudden jumps in density as boxes become more restrictive. Density grows much faster than coverage falls as coverage processes downwards to ~70% or just below it. Thus, one could argue that a ~75% density-70% coverage box might be a better choice than one with 60% density and 80% coverage.

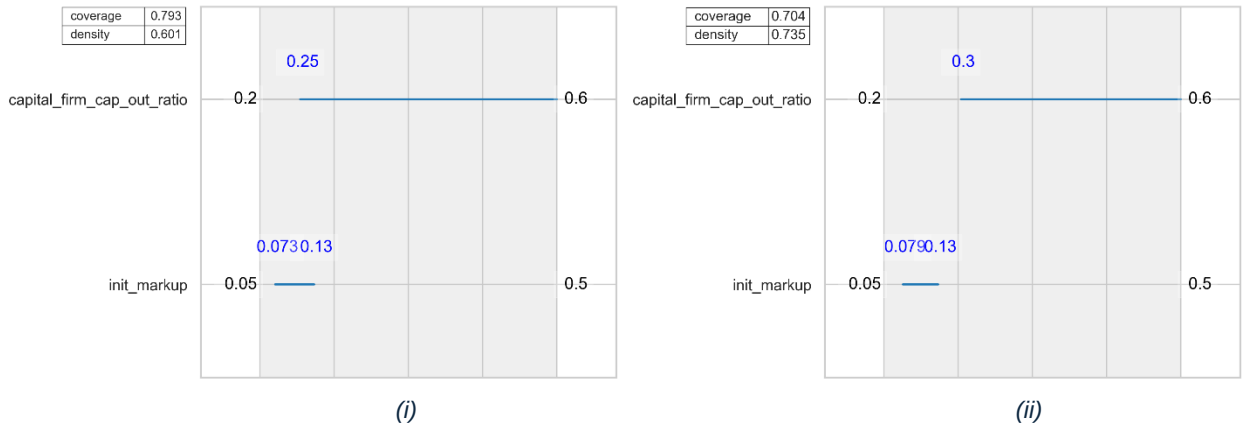


Figure 4-13: Two candidate bounding boxes for inducing a rule to describe the Fall-Level Off *cluster.*

Figure 4-13 shows the comparison of the two candidate boxes. The restrictions are very similar, and one can see how narrative scenarios developed from these quantitative boundaries could end up erasing the difference. Thus, the choice is not of great concern. For this analysis, we will move forward with box (ii), as it is a more restrictive box and might aid in the separation of the input space across the six clusters.

Table 4-1, below, summarizes the rules captured by the six selected bounding boxes.

Table 4-1: Induced rules from applying PRIM to each cluster of Household Population dynamics.

	Init. Markup		**Cap-Out Ratio**		**Min. Unempl.**		**Flood Timing**	
	min	max	min	max	min	max	min	max
Fall-Growth	0.0999	0.2307	0.2196	0.5234	-	-	-	-
Fall-Level Off	0.0786	0.1322	0.3038	0.5999	-	-	-	-
Growth-Collapse	0.5007	0.0787	-	-	0.0200	0.0786	35.5	80.0
Steady Growth	0.1032	0.1677	0.2000	0.3901	-	-	-	-
Collapse-Recovery @ 30	0.2058	0.2826	-	-	-	-	-	-
Collapse-Recovery @ 50	0.3031	0.5000	-	-	-	-	-	-

Only four of the eight input parameters appear as restricted dimensions across any of the six clusters. In fact, two of the restricted parameters are only restricted in one cluster (*Growth-Collapse*). Referring back to Table 3-1, these two dimensions have almost their full range represented here: *Emigration Minimum Unemployment* has a range of 2-8% (0.02 to 0.08) and *Flood Timing* has a range of 30 to 80. Again, while these restrictions might be numerically important to the discovered boxes, they probably do more to complicate the development of scenarios than they do to serve those scenarios. Further, we must remember that the goal of this process is to study the boundaries between the discovered scenarios as tipping points. For both reasons, it serves the qualitative results of this study to ignore those two restrictions and focus just on the two most critical dimensions from Table 4-1: *Initial Markup* and *Capital-Output Ratio*.

4.4.3. Identifying Tipping in Parameters

Though following this process will not always result in just two restricted dimensions, it does serve visualization of the resultant phase diagram. While Table 4-1 provides the collected restrictions as a table, it will be helpful to visualize them in the two-dimensional input (sub)space involving the two restricted dimensions.

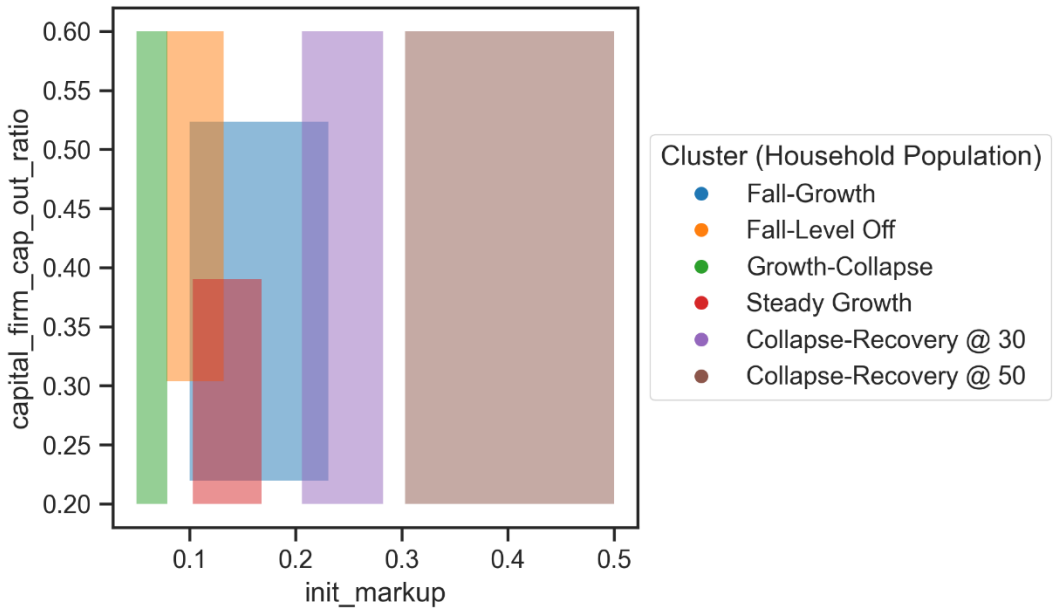


Figure 4-14: Induced rules for each cluster in Household Population, plotted as boxes in the two primary restricted dimensions of the parameter space.

In essence, the job is done here, as this constitutes a phase diagram of the CRAB model. It is important to note this is just *one* phase diagram for the model: one where each phase is defined by a qualitatively distinct dynamic in the *Household Population* output variable.

Several things stand out. First, there exists a clear, one-dimensional tipping point between the purple and brown dynamics (the two *Collapse-Recovery*) dynamics. If you consider those distinct enough dynamics, then there is a tipping point at roughly an Initial Markup of 0.3. The first four clusters have more interesting phase transitions. *Growth-Collapse* clearly exists at the very low end of the Initial Markup dimension and crossing beyond 0.08 exits that phase. In cases where Capital-Output Ratio is above 0.30, the phase transitions to *Fall-Level Off*. In cases where Capital-Output Ratio is below 0.39, the phase is near a transition boundary towards *Steady Growth*. The overlap between these two boxes reveals a limitation of this method, as it is hard to say exactly what dynamic occurs within that subspace. The *Fall-Growth* phase adds further complication here, as it overlaps with most of the *Steady-Growth* phase and some of the *Fall-Level Off* and *Collapse-Recovery @ 30* phases. However, the overlap with the latter cleanly constitutes a tipping point itself: at roughly an Initial Markup of 0.21-0.23, there exists a tipping point towards a dynamic of early, sharp collapse.

Table 4-2 summarizes the tipping points that can be extracted from Figure 4-14.

Table 4-2: Descriptive population tipping points discovered in the parameter space of the CRAB model.

Identifier	Clusters	Rule(s)		Consequence(s)
HP-TP1-a	*Growth-Collapse* → *Fall-Level Off*	Initial markup crosses above ~0.08-0.10.	Capital-output ratio is above 0.30.	Initial population dynamics switch from growth to slight collapse, but end population levels off.
HP-TP1-b	*Growth-Collapse* → *Steady Growth*		Capital-output ratio is below 0.39.	Population continues to grow throughout time horizon.
HP-TP1-c	*Growth-Collapse* → *Fall-Growth*		Capital-output ratio is between 0.22 and 0.52.	Initial population dynamics switch from growth to slight collapse, but end population later grows.
HP-TP2	*(Several)* → *Collapse-Recovery @ 30*	Initial markup crosses above ~0.21.		Initial dynamics see sharp collapse. After roughly 8 years, population begins slow recovery.
HP-TP3	*Collapse-Recovery @ 30* → *Collapse-Recovery @ 50*	Initial markup crosses above ~0.30.		Initial dynamics remain same (sharp collapse), but recovery only begins after 12 years.

In this summary, **HP-TP1-(a, b, c)** are noted as part of the same family of tipping points, as they all require a similar change in one parameter (initial markup) and therefore system behaviour "after" the tipping point is dependent on the level of the second parameter, even though behaviour "before" the tipping point is not.

5 Discussion & Limitations

5.1 Extracting Policy-Relevant Lessons

There are two ways that policymakers could use the results of this study, or another study following this methodology. First, if the parameters underpinning the discovered scenarios (and thus the tipping points between them) are controllable (i.e., can be influenced by policy or other means), policymakers can reflect on the consequences of tipping points to identify particularly desirable or undesirable ones. If there are existing policy levers that are known to drive the parameters in a particular direction, combining these levers with effective monitoring can help drive a system either towards a positive tipping point or away from a harmful one. For an example in the CRAB model, if a negative tipping point was discovered at the high end of the debt-to-sales ratio parameter, governments could impose limitations on firms regarding the amount of debt they take on.

Second, if parameters are not influenceable or describe exogenous effects, investing in monitoring of the parameters' real-world parallels is important. Understanding the real-world indicators that correspond to certain levels of our model parameters allows us to identify where in the phase space the system currently lies, and where it might be going. If data showed that a system was moving towards a negative tipping point, for example, policy efforts could be directed towards mitigating the consequences of crossing that tipping point.

Unfortunately, the CRAB model parameters used in this study are mostly too abstract (uninfluenceable and unmonitorable) to be connected to policy recommendations. Curiously, the most important parameter in this study—*Initial Markup*—is quite tangible, but it is hard to know exactly what markup rate firms use and it is hard to influence via policy outside price controls. The model parameter itself also reflects only the initial conditions of an endogenous variable, thus it is unclear how to adequately monitor the real-world analog of this parameter. Thus, this study serves mostly as a methodological testing ground, rather than a policy study. To get meaningful, policy-relevant results regarding tipping points in a model's parameter space, the model's parameters themselves must be constructed in a way that is policy-relevant.

What we can take from the results is that, according to the CRAB model, pricing is an important predictor of system-wide success. This likely results from the economic bias encoded into CRAB's underlying Keynes & Schumpeterian economic model: when firms charge more early in a model run, they grow and become more stable, which also grows the region's economy and ensures it is more able to withstand the economic disruption brought by a major flood (Taberna et al., 2021).

5.1.1. Parameter- vs. Variable-Based Tipping Points

This study looked at tipping in model parameters. That is, the model's input parameters were chosen as the independent variables in which scenario rules were induced. An alternative approach is to consider variable-based tipping. In model speak, variable-based tipping points are those expressed in terms of a model's endogenous variables as they evolve throughout the model

run. Accordingly, the phase transitions would then also be expressed in terms of the levels of endogenous variables. Meta-variables like the variance or rate of change of certain endogenous variables could also be calculated and used as independent variables for this purpose. The difference between parameter- and variable-based tipping points mirrors the difference between externally and endogenously caused tipping points discussed in Section 2.1.

To integrate this variable-based approach into this study, endogenous variables such as population, wage, or GDP would replace the model parameters as the independent variables used by the PRIM algorithm. Thus, PRIM's results would be expressed as a set of rules restricting the endogenous variable space. These variables would have to be sampled at a certain timestep, such as at the timestep when the simulated flood occurs.

This would enable several improvements. First, there is a clearer analogy between this approach and the framing of a phase diagram. A simple physical phase diagram usually expresses a material's physical state (its phase) in terms of conditions like temperature and pressure: these conditions are much more like endogenous state variables than exogenous parameter inputs.

Similarly, second, it might be easier to develop policies that address or attempt to drive endogenous state variables than those that drive exogenous inputs: as such, a study of variable-based tipping might lend itself better to answering questions with real-world policy impact. Beyond just driving system variables, understanding variable-based tipping points could serve the development of early-warning signals for tipping or other undesirable model outcomes. If an endogenous variable at an early or otherwise identifiable timestep is highly predictive of a particular model end-state or behaviour, then monitoring that variable in the real world becomes very important for anticipating and pre-empting system behaviour.

Some scholars might argue that this study has not done anything at all about tipping points. The time-series visible in Figure 4-1 and Figure 4-7 do not display, within a given or representative model run, a rapid change of state like those described in much tipping point literature. However, the tipping points discovered in this study do display such features as abrupt change and distinct system states, just in the model's parameter space, rather than its temporal domain. While this differs from van Ginkel et al.'s (2022) stepwise approach to recognizing tipping points, it is useful to policymakers in very similar ways.

5.1.2. Other CRAB Input Parameters

One intent of the CRAB model was to bring a behavioural economics perspective to the study of coastal climate adaptation (Taberna et al., 2023). As such, within the model, a synthetic population of agents is generated that are representative of a real-world coastal region, with behavioural characteristics like *Worry* and *Flood experience* (see Figure 3-1). This representative sample is generated according to a statistical distributed deduced from surveys performed by members of the CRAB model's initial development team (Noll et al., 2022).

The more traditional, rational-economics variables that were selected as control parameters for this study. Behavioural population traits were left out of the parametric sampling in part because of the nature of the statistics gathered from the survey. If we were to take, say, the mean of one

of these treats as a parameter, the nature of the exploratory modeling approach would require us to vary it across a reasonable range of values. However, co-linearities between the empirically measured traits mean that others would have to change accordingly. If more than one such trait was included in the parameter space, then arbitrarily varying both parameters would create unrealistic populations and may sacrifice the realism gained by using behavioural data in the first place (Taberna et al., 2023). However, it may still have been interesting to attempt this, knowing the limitations of the method, to compare the relative influence of the behavioural and rational parameters on model dynamics.

5.2 Potential Methodological Improvements

5.2.1. Stochastic Replications

Based on the discontinuities in Figure 4-6 and the variety of dynamics in Figure 4-7, one's instinct might be to keep each replication as a separate model realization when using a model as complex as the CRAB model, and then perform the above scenario discovery pipeline on the full dataset. Unfortunately, this is a computationally intensive task. Even the relatively small experiment performed in this study produced 80000 distinct model realizations (replications $\times$ input samples). To perform clustering on such a dataset using CID requires at least 50 GB of working program memory.

There are several workarounds, all of which constitute areas for improvement of this study. For one, a more intelligent means of storing and accessing the distances measured during the initial clustering steps could be developed, to minimize program memory required. Tools like PyArrow could enable such computation with limited active memory use (pyarrow 17.0.0, 2024).

Alternatively, one should look for ways to reduce the total number of realizations without losing information. One possible method would be to perform clustering *within* each sample point to find groups of distinct behaviours across the stochastic uncertainty space. If one were to find that a clustering with $K = 3$ explains much of the variance within the 40 replications at a given sample point, then one could carry forward a representative replication for each of those clusters and drop the remaining 37 replications, effectively decimating the size of the dataset without losing too much information.

It is not just important that losing the distinct dynamics covered by different stochastic replications means that fewer total possible model dynamics are covered. What is perhaps most critical is that, by representing a model run (sample point) as the average of its replications, one might create brand new dynamics that are not represented in any of the individual model realizations. Especially when variance changes over the parameter space (as it does, per Figure 4-6), the effects of averaging on the overall measured dynamics could lead to mistaken conclusions. Averaging replications is a somewhat standard approach in the DMDU field and is the standard implementation of the `process_replications()` method included in the EMA Workbench (Kwakkel, 2013), but perhaps is not suitable when applied to complex applications like ABMs.

5.2.2. Rule Induction

As indicated in the review of the literature (2.2), there is an entire subfield of data science devoted to developing and improving rule induction methods. The methods used in DMDU (PRIM, CART, and some others) are relatively primitive by comparison. From this study, several drawbacks of PRIM become clear. First, PRIM is restricted to defining rules orthogonally: one restricted range for each restricted parameter. This is why PRIM's rules are referred to as "boxes," because their edges in the parameter space are strictly straight. Traditional physical phase diagrams such as the one depicted in Figure 2-1(b) are not restricted to "boxy" phases, and in fact such a limitation would severely hinder their utility. Figure 5-1, below, depicts a toy phase diagram with a simple separation between solid, liquid, and gas. Superimposed on the diagram are boxes, loosely representative of what it might look like if PRIM had been used to identify the rules for each phase from a set of observations. Even in this simple diagram, it is clear to see how orthogonal rules limit both accuracy and comprehensibility.

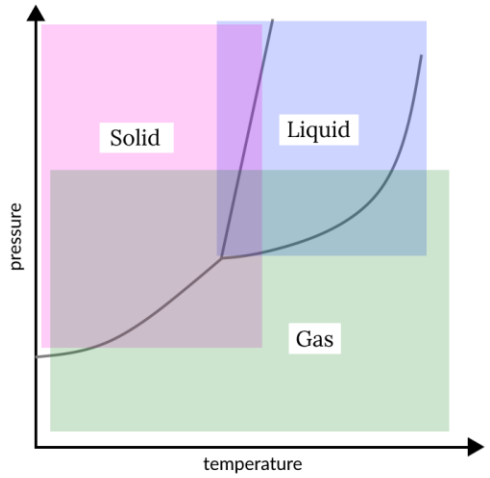


Figure 5-1: Toy phase diagram with square phase "boxes" superimposed

To evaluate the impact of "boxy" rules on the CRAB model's phase diagram, we can plot all sample points in the input space and colour each point according to its cluster: essentially, we can attempt to visually identify rules ourselves instead of relying on an algorithmic method like PRIM. With many dimensions (such as the eight in this study), this is very difficult. However, having run PRIM, we have discovered that only two dimensions are critical in defining the rules that separate our six model states. Thus, we can do this using a single two-dimensional scatter plot with the axes *init_markup* and *capital_firm_cap_out_ratio*.

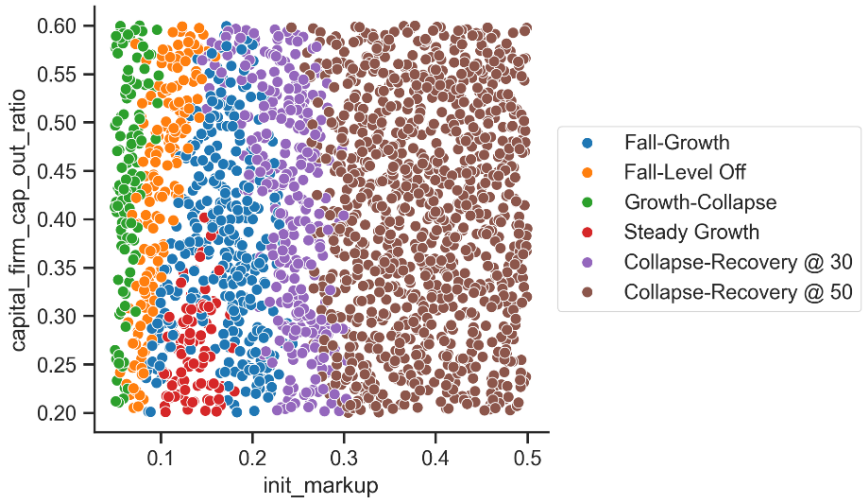


Figure 5-2: Scatterplot of 2000 CRAB model realizations, coloured according to cluster in Household Population outcome.

Indeed, Figure 5-2 demonstrates better input space separability than the phase diagram generated using the PRIM boxes (Figure 4-14). Figure 5-3, below, combines the two plots together, with smaller dots to enable cross-comparison. Only the brown *Collapse-Recovery @ 50* cluster (phase, system state, etc.) has a boxy shape. The green, orange, and red clusters all have triangular shapes, the purple roughly quadrilateral, and the blue cluster a very irregular shape enclosing the red one. This nuance better illustrates the boundaries between the phases and thus could lead to more meaningful tipping points than those summarized in Table 4-2, especially clarifying the transitions surrounding the *Fall-Growth* cluster. From this plot, it is clear to see that the *Steady Growth* cluster might be a special case of the conditions that usually enable the *Fall-Growth* cluster. If the parameters were more meaningful or controllable (see Sections 5.1.1 and 5.1.2), this could be an important conclusion to drive policy development.

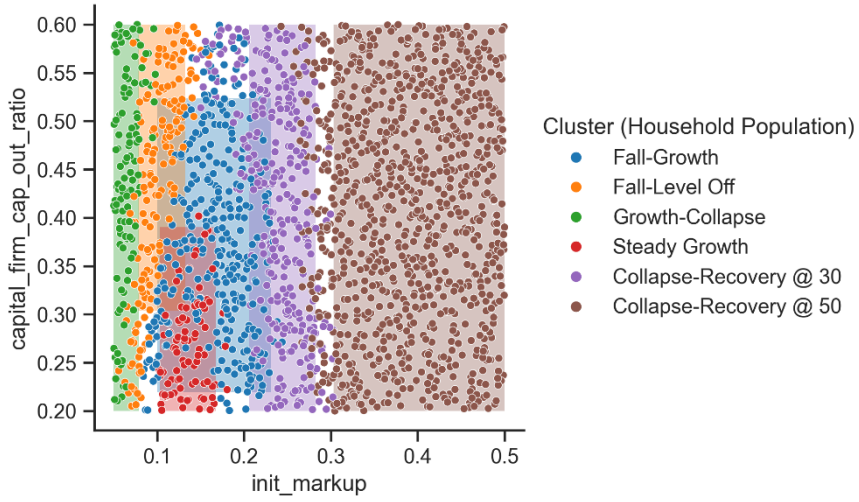


Figure 5-3: A replication of Figure 5-2: Scatterplot of 2000 CRAB model realizations, coloured according to cluster in Household Population outcome.Figure 5-2 (scatterplot of clustered realizations in most-critical dimensions, as suggested by PRIM) with the rules/boxes discovered by PRIM plotted behind for visual comparison.

With more information comes more communication. More precise (non-box) rules might also be harder to communicate. In two or three dimensions, where the rules are still easily visualizable in a plot, this loss is probably not felt as much. However, when there are more than three restricted dimensions and rules must be communicated via text or a combination of *several* 2-D plots, it may be easier to ground policy development in orthogonal rules: it is very easy to understand one restricted range per important variable, but it is harder to keep in mind if those ranges change dependent on the level of several other variables.

As a secondary issue, PRIM requires the manual selection of a candidate box for each case of interest. This is good, as it enables a policy analyst to personally weigh the trade-off between coverage, density, and information load (number of restricted dimensions), but the implications of this decision are not always explored in scenario discovery studies. Further, there are other factors beyond these three numerical measures that suggest whether a rule is good or not. For example with multiclass scenario discovery, input space separability is an important performance metric that enables clear communication of the discovered scenarios and the boundaries between them. A concurrent clustering and rule induction algorithm like the one used in Jafino & Kwakkel (2021) addresses this problem by inferring differentiating rules at the same time it makes the split in the output space. Alternatively, sequential methods like PRIM could be extended to perform several binary rule inductions in parallel and suggest a set of boxes that meet certain coverage and density requirements while minimizing overlap.

5.2.3. Clustering Algorithms

Finally, the choice of clustering algorithm is an important one that has downstream effects on the credibility of the identified clusters. Steinmann (2018) reviewed several distance metrics underpinning the use of clustering in behaviour-based scenario discovery and ultimately selected CID, which was the primary motivation for doing the same in this study. However, Steinmann used a simpler, deterministic model. In discussing the complexity of analyzing (spatio-)temporal outcomes of ABMs, Lee et al. (2015) suggest that Dynamic Time Warping (DTW) is a better method for evaluating the difference between two complex time-series as it is able to identify similar dynamics that are offset along the time horizon. Given that CRAB is a highly stochastic model and in this study is focused on identifying boundaries between system states (which are defined more by dynamics than by the exact moment in time when those dynamics occur), it is possible that DTW would have been a better distance metric to use for this study.

5.3 Future Work

There are several avenues that this Thesis opens up for future work, which are summarized as follows:

- The methodology covered in this study should be repeated, with improvements or as-is, on a model with input parameters that are more directly reflective of policy levers or things that can be influenced by policy levers. Ideally, this could be done in conjunction with a participatory model design process, such that the resultant model is itself informed by policymakers. The goal of such a study could be to demonstrate whether policymakers can make effective use of the tipping points that this method identifies.

- Instead of input parameters, the approach could be modified to use endogenous variables as the independent variables of rule induction (at specific time slices, or, if a mathematical method can be suggested, full time-series themselves). This could be repeated on a model like the CRAB model, as the endogenous variables tracked within the CRAB model might be more useful indicators for policymakers to track as tipping point warning signs, or targets for them to direct policy efforts towards.
- The methodology followed in this study could also be repeated but using multivariate clustering when defining the scenarios. This would help make scenarios more meaningful and broadly relevant. Instead of just a scenario of population growth or collapse, we could study more complex scenarios, like one characterized by all three of population growth alongside inequality growth and high governmental debt.
- Methods for keeping individual model realizations separate without averaging stochastic replications should be explored. Some suggestions are made in Section 5.2.1, such as using another, smaller round of time-series clustering to identify and extract dynamically distinct replications.
- The implications of various distance metrics underpinning time-series clustering should be investigated in the context of a complex, stochastic ABM like the CRAB model. This effectively repeats (Steinmann, 2018) with application to ABMs.
- Alternative rule induction methods (other than PRIM) could be explored. A method like the one described in (Jafino & Kwakkel, 2021) could enable concurrent splitting of the input and output space, which would lead to a more clearly separated “phase” space and better enable the phase space visualization of tipping points. Alternatively, PRIM could be extended to produce non-orthogonal rules or to handle rule induction for several scenarios in parallel.

5.4 A Note on Bringing DMDU to ABMs

Ultimately, this study was in part motivated by the need to improve or expand methods for the analysis of the outputs of ABMs. The last decade has seen plenty of work pushing this field forward (Lee et al., 2015; Magliocca et al., 2018; Ligmann-Zielinska et al., 2020). Meanwhile, while DMDU scholars have been creating and improving methods for analyzing large sets of model outputs across uncertainty ranges, there have been computational challenges in bringing such methods to complex models such as ABMs (Moallemi et al., 2020; Helgeson et al., 2022). This study has attempted to bridge this gap by using a standard DMDU analysis tool (scenario discovery) on a large, complex ABM to answer the type of question that a policy researcher may seek to answer using an ABM. In a sense, this study is attempting to show that the time has come to reject the perceived challenge of using DMDU methods in complex models.

Two things have been considered challenges in the marriage of these two sub-fields of computational policy support. First, ABMs tend to have longer runtimes than simpler models, especially when developed in Python, a standard language for scientific computation. This limits the number of model runs that can be used in an analysis of a model’s output space. However, pushing the limits of the student-level account’s access to TU Delft’s high-power computing platform still afforded me 10000 model runs per hour (Delft High Performance Computing Centre, 2024), which certainly enables a wide study of structural and parametric uncertainty. There is

ongoing work in the DMDU community to improve algorithms, sampling methods, and embedded support for parallel processing and high-performance computing such that the methods can be brought more frequently and successfully to studies using complex models (Moallemi et al., 2020). Computational supports like the EMA Workbench (Kwakkel, 2013) aid the application of DMDU methods to complex models with minimal adaptation in a model's codebase.

Second, it is not always clear how, when, or why to use DMDU methods in an ABM. DMDU methods are just one way to explore uncertainty in a model. There is already a growing literature on the application of sensitivity analysis in ABMs that lies outside the DMDU literature (Ligmann-Zielinska et al., 2020). Researchers may study the ABM literature and find that sensitivity analysis on its own is a sufficient tool for exploring uncertainty in an ABM. Thus, it is imperative that DMDU scholars make clear that their methods can be used in complex modeling applications and can answer complex questions like the one presented in this study. Furthermore, it is important that DMDU literature discusses its methods in such a way that is easy to adopt in future work, especially when publishing open-source code. Again, the use of standardized tools like the EMA Workbench can help here.

Thus, I believe that it is wrong to frame bringing DMDU to complex models such as ABMs as a *challenge*, and that doing so likely makes it appear more challenging than it is to researchers new to either part of the field. The supports are now in place to make standard both demonstrating new DMDU methods in the context of ABMs and using DMDU methods to explore or stress-test the boundaries of new ABMs.

6 Conclusion

This study explored the use of scenario discovery to search tipping points in complex coupled human-Earth systems using ABMs. Using the framing of phase transitions and phase diagrams, this research demonstrated this from start to finish, including illustrating an example of how tipping points can be visually communicated using the results of scenario discovery.

There were three research questions posed in Section 1.2 of this document. The first sub-question aimed to identify which recent advancements in scenario discovery enable its use for exploring and discovering tipping points. The study found that behavior-based scenario discovery (which brings time-series clustering and multi-class scenario discovery together) is suitable for this purpose. However, this study lacked in being able to make its result particularly useful or salient to policymakers. This could be improved by using more relevant input parameters, such as those that can be monitored by governments or driven by policy, or by using endogenous variables as the scenario-defining conditions instead.

The second sub-question about communicating and visualizing tipping points was addressed by building on an existing analogy of the phase diagram as a way of capturing distinct states of an ABM. This study suggests that scenario discovery, which breaks a model's output behaviours a set of qualitatively distinct scenarios of interest and associates portions of its input space with each scenario, can be used to construct such a phase diagram. Then, this serves as a convenient visualization that enables the intuitive interpretation of tipping points in a model (or the system it purports to represent).

Finally, the third sub-question addressed the challenges of applying scenario discovery to the output of ABMs. This is in line with other literature on the challenges of output analysis in ABMs, since they are often complex, have many dimensions (plus, outputs are often reported across space and/or time), and require many stochastic replications. The research confirmed that while ABMs introduce additional layers of complexity, particularly due to their stochastic nature, using scenario discovery to explore and simplify the relationships between their inputs and outputs is still feasible. More work should be done to explore the opportunity and impact of applying such a method without averaging replications together. In particular, the role of averaging should be studied in such a way that can identify whether it leads to potentially inferring incorrect policy lessons. In line with this question, the experimental design was validated, confidently showing the sample size was sufficient, though perhaps less definitely less so for the number of replications.

Tipping points are not going away as either a scientific or political tool important for framing the social, economic, and environmental impacts of climate change. It is thus important that scientists, analysts, advocates, and legislators are well-informed about the concept and have access to accurate information about tipping points across any number of human-Earth systems. This research builds on existing literature in exploratory modeling to propose one way of identifying and communicating such tipping points, expanding our methodological toolkit. Hopefully, this research can help computational policy analysis play a role in reversing climate change instead of advancing it.

Appendix A: Analysis Pipeline for Other Outcome Variables

This appendix will replicate the core of the analysis pipeline—in particular, Sections 4.3 (validation) and 4.4 (behaviour-based scenario discovery for tipping point identification) for two alternative outcome variables: GDP and the Gini Coefficient. Per Table 3-1, bothTable 3-2: Four primary CRAB model outcomes measured in this study. were chosen as convenient, accessible measures of their respected phenomena (economic activity and inequality, respectively). The Median Wage variable has been left out for brevity and due to a lack of interesting interactions (at least those that stand out from the other three).

GDP

Validation of Experimental Setup

The three plots from Section 4.3 are replicated here for the GDP variable.

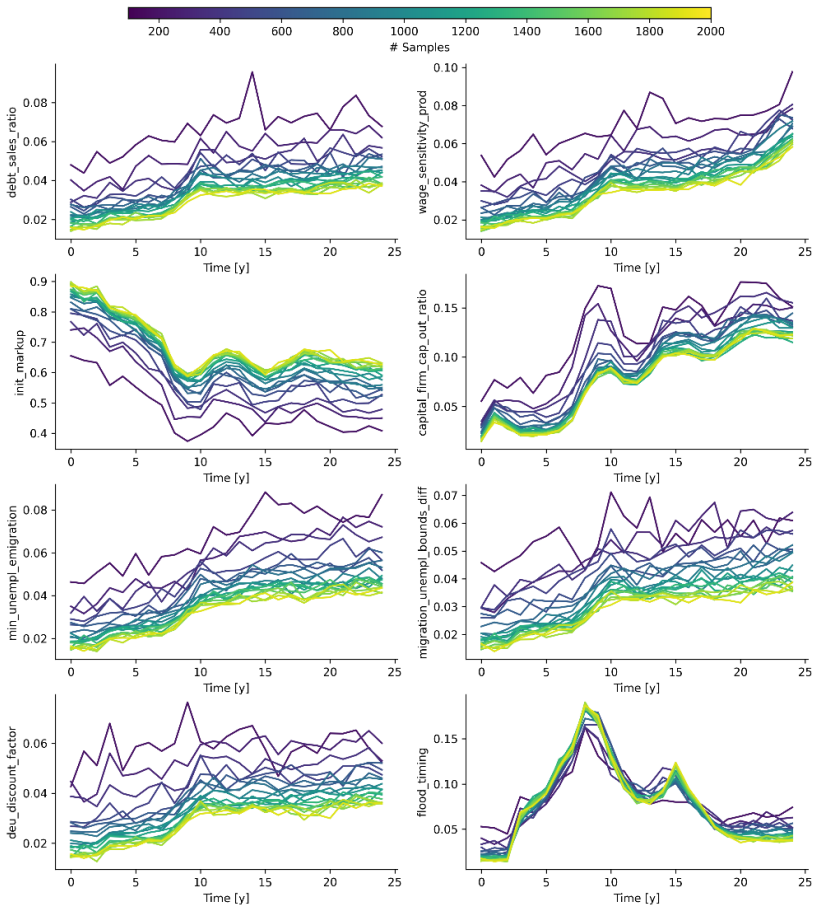


Figure A-1: Convergence of input sensitivities as the number of samples increases. The y-axis denotes the importance of the named parameter to the GDP outcome. Note the lack of shared y-axis.

Figure A-1, above, shows how sensitivity time-series (measuring how the importance of each input variable to the GDP output variable changes over time) converge as the number of model samples considered in the experimental run is increased. Results here largely mimic those in Figure 4-5 from the main text. For all but two variables, sensitivity coefficients are overestimated with a lower number of samples. The distance between subsequent sensitivity time-series is decreasing as N increases and seems quite small by the step from $N = 1900$ to $N = 2000$ (the full sample size). *Initial markup*, which is still found to be the most important parameter, has its importance underestimated at low sample sizes before converging on a higher number. Interestingly, the *Flood timing* parameter has a relatively consistent importance across all sample sizes.

Figure A-2, below, instead shows the convergence of time-series *variance* as replications increase from 2 to 40. The sets of variances are measured at four different points in the model's sample space, at four sample points that denote representative samples of the clusters of behaviours in this variable (Figure A-4).

There are several things to comment on. Unlike for *Household Population*, there are not many notable jumps in the shape of the progressive variance curves except at very low numbers of replications (where they are to be expected). This might suggest that stochasticity has less effect on the GDP outcome: there are fewer stochastically distinct behaviours generated at each sample point. The variance that *is* measured could be due to differences in GDP *level* between realizations, rather than its behaviour. A fair portion of the random number calls in the CRAB model occur during migration, which could explain why stochasticity has a more drastic effect on the behaviour of the *Household Population* outcome.

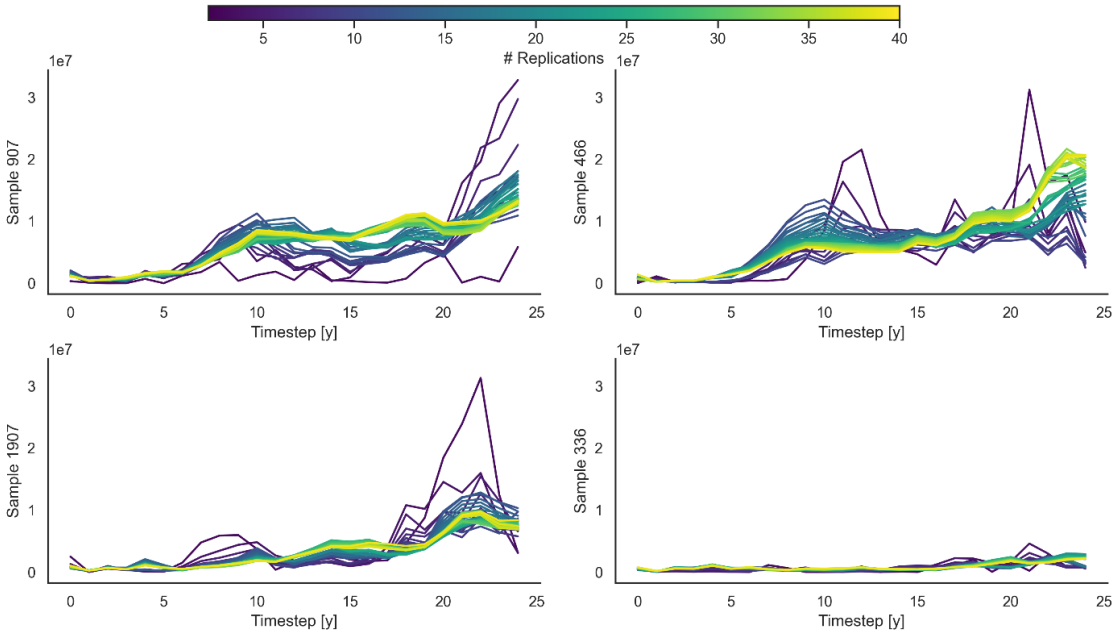


Figure A-2: Convergence of time-series variance in the GDP outcome across an increasing number of replications. Variances are shown for four sample points, each representative of a different scenario (clustered by GDP dynamics).

Samples 907, 466, and 1907 (top-left, top-right, and bottom-left of Figure A-2) show relatively similar time-series variances as replications near $N_{rep} = 40$, with the middle of the three differing the most (especially at the end of the time horizon) and taking the longest to converge to this curve shape. Sample 336 (bottom-right) shows *very* low variance compared to the other three. This is perhaps due to the phenomenon captured by its associated cluster having a more dominant effect than stochasticity at this part of the parameter space. However, the fact that variance itself can vary this much across the parameter space is further evidence that researchers studying uncertainty in models of complex SES should be careful about averaging replications: the effect of this averaging can be drastically different across the uncertainty space.

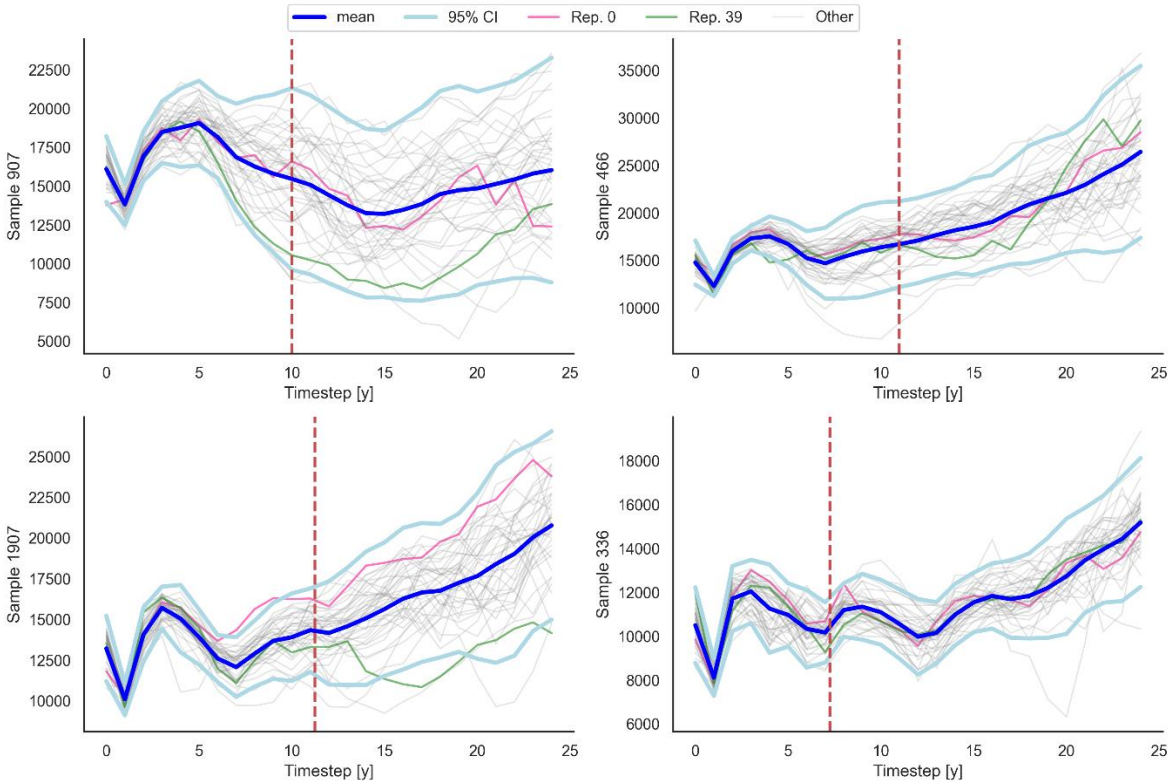


Figure A-3: Aggregated dynamics of GDP atop individual replications, plotted separately for each of four sample points. Two arbitrary samples are highlighted in pink and green to help demonstrate distinct dynamics. Time-series have been resampled to annual measurements. The vertical red line indicates the timing of the flood for the relevant input sample.

Finally, Figure A-3 shows the dynamics of the individual replications at each of the representative sample points before and after they are averaged together. Indeed, Sample 336 (bottom-right) has the least variance in both behaviour and level. In Samples 907 and 466 (top row), while there appear to be some distinct behaviours, their distinctions seem to perhaps be more exaggerations of the same dynamics: growth, slight (to large) fall, recovery. Sample 1907 is less clear. Some replications appear to fall or steady out shortly after the flood, while the dominant behaviour at the sample is one of medium growth both before and after it. This suggests that looking at the variances alone—even as they converge with a growing number of replications—does not tell the whole story.

Behaviour-Based Scenario Discovery

This section repeats Section 4.4 with application to the GDP outcome.

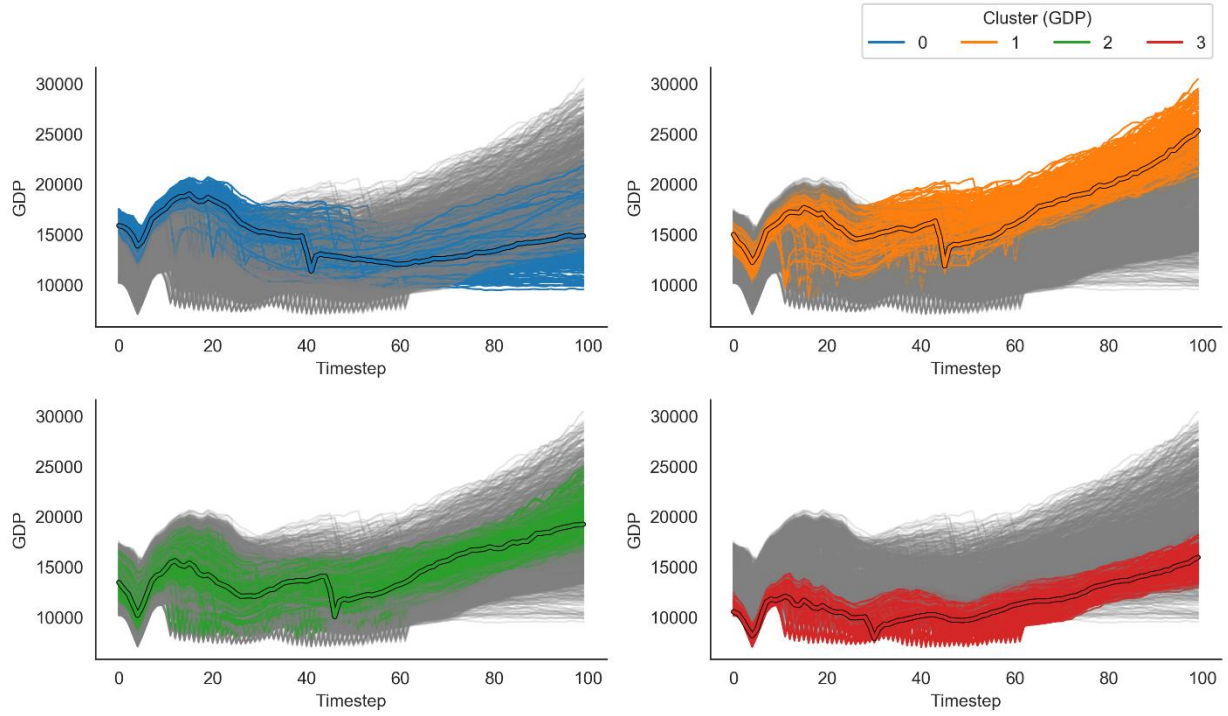


Figure A-4: 2000 realizations of the CRAB model, clustered ($K = 4$) according to behaviour of the GDP outcome.

Figure A-4 shows the four clusters of GDP time-series. The GDP variable is notably more erratic than Household Population. Each representative sample (highlighted with a black outline) shows a slight collapse at a midpoint timestep, which corresponds with the *Flood timing* parameter for that sample. Thus, flooding has a more direct and immediate effect on GDP than on population: this is to be expected.

The orange and green clusters share similar dynamics of growth, with the orange cluster exhibiting faster growth and higher GDP, even (especially) after the flood. In fact, according to the representative scenarios, the orange and green clusters show near-identical behaviour leading up to their floods, and only diverge in terms of flood recovery. The blue cluster demonstrates a wider distribution of post-flood outcomes, though has a relatively similar pre-flood pattern. Cases in the blue cluster vary from continued collapse up to a subtle rise back to and just past pre-flood peaks. However, the blue cluster clear comprises model runs where flooding severely hindered economic activity and growth, and in some case led to decline. Finally, the red cluster shows tight behaviour throughout. Its behaviour before flooding appears very distinct from the other three, implying that the economy in this scenario is already struggling even without flooding. The spikes at the bottom of the graph are due to different sample points having different flood timings, but the clustering algorithm is still able to correctly group these as dynamically similar model runs. GDP does eventually begin recovery in all runs found in this cluster, though very slowly.

We can name the clusters as follows:

- Blue: Post-Flood Stagnation/Collapse
- Orange: Rapid Post-Flood Growth
- Green: Adequate Post-Flood Growth
- Red: Weak Economy

As always, these names are imperfect, so the clusters themselves should be studied to remember the dynamics they display. Most notably, the blue cluster's name implies that all cases stagnate or collapse. Some cases in fact grow. However, that behaviour is not the dominant or unique behaviour of the blue cluster, especially compared to the orange and green ones. Furthermore, while the red cluster is so-named due to its starting state, in fact all cases in this cluster have a higher and faster-growing GDP than some cases in the blue cluster. Thus, its name is descriptive of the economy's starting state, but not necessarily the full time-series dynamic.

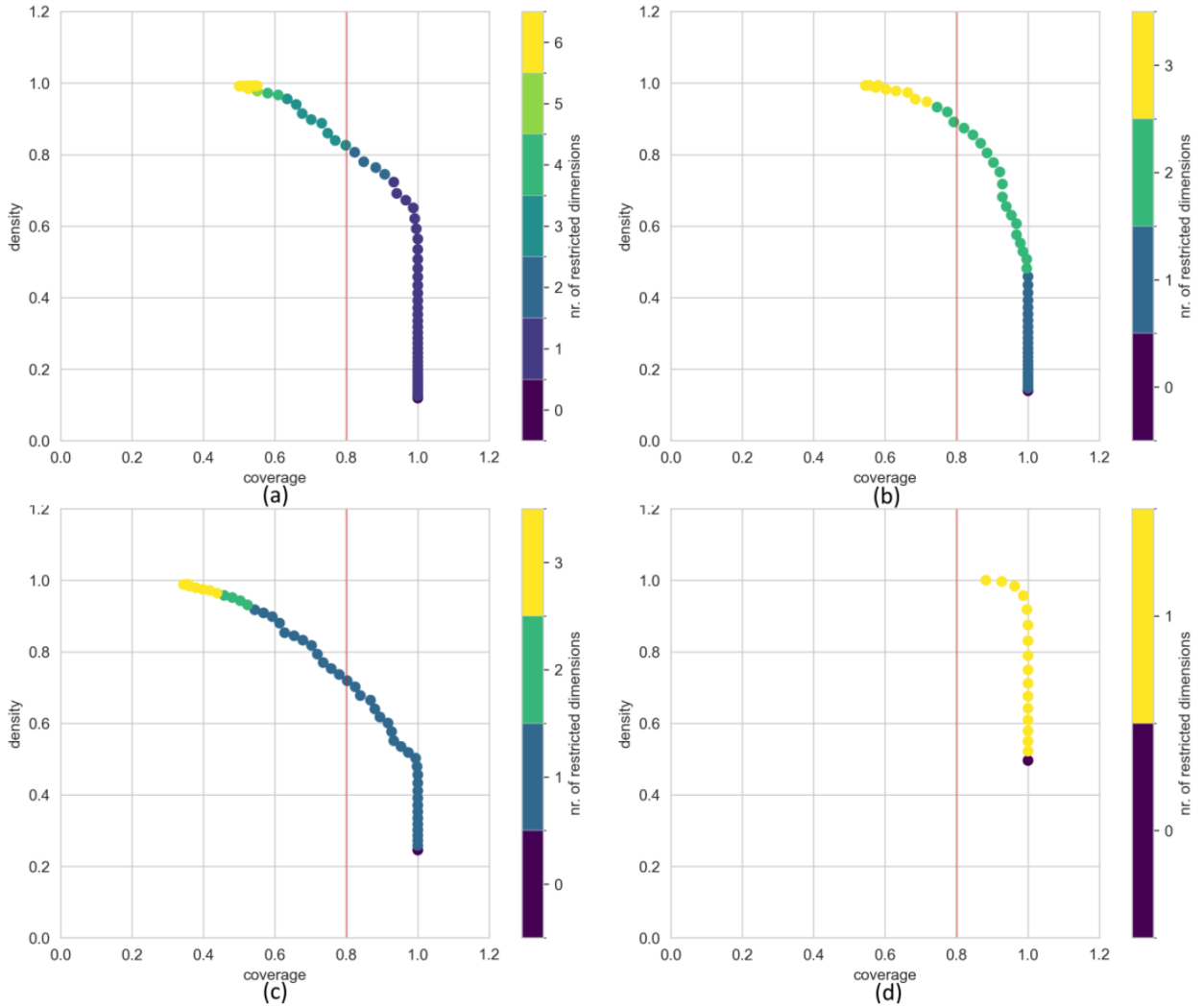


Figure A-5: Coverage-density trade-off curves from the PRIM algorithm applied to the clusters of GDP dynamics. (a): Post-Flood Stagnation/Collapse (blue cluster); (b): Rapid Post-Flood Growth (orange cluster); (c): Adequate Post-Flood Growth (green cluster); (d): Weak Economy (red cluster).

For brevity, the coverage-density tradeoffs for each rule induction (Figure A-5) are presented as they were in Section 4.4.2 but are not closely analysed. The process for selecting an ideal box from each closely mirrors that which was done for the *Household Population* outcome. Unlike *Household Population*, these decisions are mostly straightforward. The combined rules determined from the selected boxes are available in Table A-1:

Table A-1: Induced rules from applying PRIM to each cluster of GDP dynamics.

	Init. Markup		Cap-Out Ratio		Flood Timing	
	min	max	min	max	min	max
Post-Flood Stagnation/Collapse	0.0501	0.1155	-	-	39.5	80.0
Rapid Post-Flood Growth	0.0941	0.1957	0.2000	0.4277	-	-
Adequate Post-Flood Growth	0.1596	0.2828	-	-	-	-
Weak Economy	0.3031	0.5000	-	-	-	-

Again, *Initial Markup* dominates as the predictive variable. Two other variables appear once, both with restrictions on one end of their range. From these rules, the scenarios can be plotted in two dimensionally reduced spaces: *Initial Markup* against each of the other two variables. An alternative would be to visualize the resultant space in 3D, though due to the minimal restrictions in the two secondary dimensions, there might be too much overlap to make this visualization communicative.

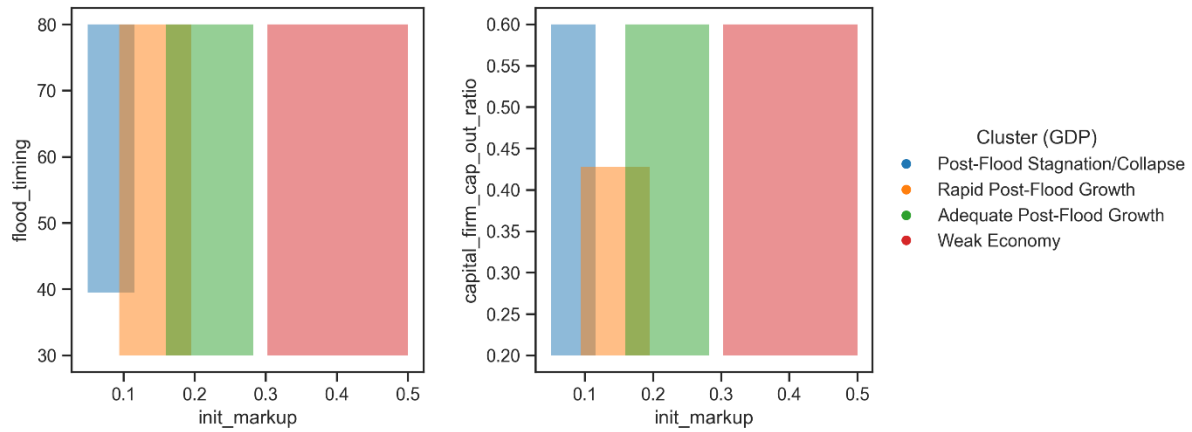


Figure A-6: GDP behaviour scenarios plotted in the dimensions in which their rules have restraints.

From this view of the system, tipping points in GDP come mostly in the *Initial markup* parameter. Broadly speaking, the "good" scenarios are the orange and green ones. Thus, we can deduce that having a very low or relatively high initial markup leads to unwelcome outcome behaviours. The same criticism of the salience of these results can be applied here as it was for the *Household Population* results: the parameters used in this study lead to a slightly obtuse reading of tipping points from this process, despite the method showing clear potential for their identification. One could take *Initial markup* as a proxy for markup rates and restrictive pricing in general. With this lens, these results imply that a coastal economy is most likely to recover well (economically speaking, at least) from a flooding event when it is affordable for individuals (prices are not set too high) while businesses are still able to turn some profit and keep resources "for a rainy day."

Gini Coefficient

Validation of Experimental Setup

The three plots from Section 4.3 are replicated here for the Gini Coefficient variable.

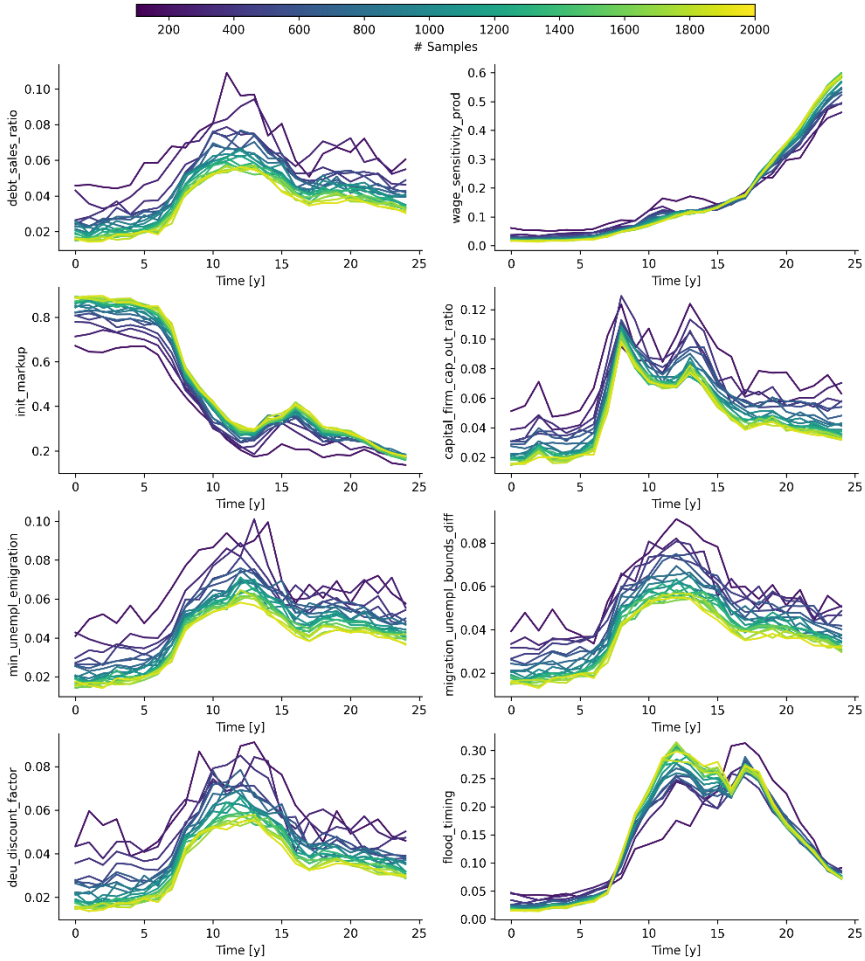


Figure A-7: Convergence of input sensitivities as the number of samples increases. The y-axis denotes the importance of the named parameter to the Gini Coefficient outcome. Note the lack of shared y-axis.

Figure A-7 (convergence of time-series sensitivities for the Gini Coefficient) shows very similar results to Figure A-1, except the *Sensitivity of wages to productivity* variable both becomes quite important to the model and has relatively consistent importance as more samples are added. This behaviour is consistent with behaviours seen elsewhere in the report. Interestingly, in the early timesteps where it is less important, convergence trends downwards, whereas when it is important at the end of the horizon, it trends upward like the *Initial markup* sensitivity.

Figure A-8 shows some jumps in time-series variance as replication count increases, more like the analog graph for *Household Population* than the one for *GDP.* The representative sample of the top-right cluster has low and relatively flat variance across its time-series. In contrast, the sample in the top-left cluster not only shows high variance, but the shape of the variance curve changes radically as certain replications are introduced, such as what appears to be the 37th and

39th replications in light green and almost-yellow, respectively. Not only is the across-replication variance thus highly dependent on input parameters, also the likelihood of stochastic replications being behaviourally distinct (and not just distinct in level) might vary as well.

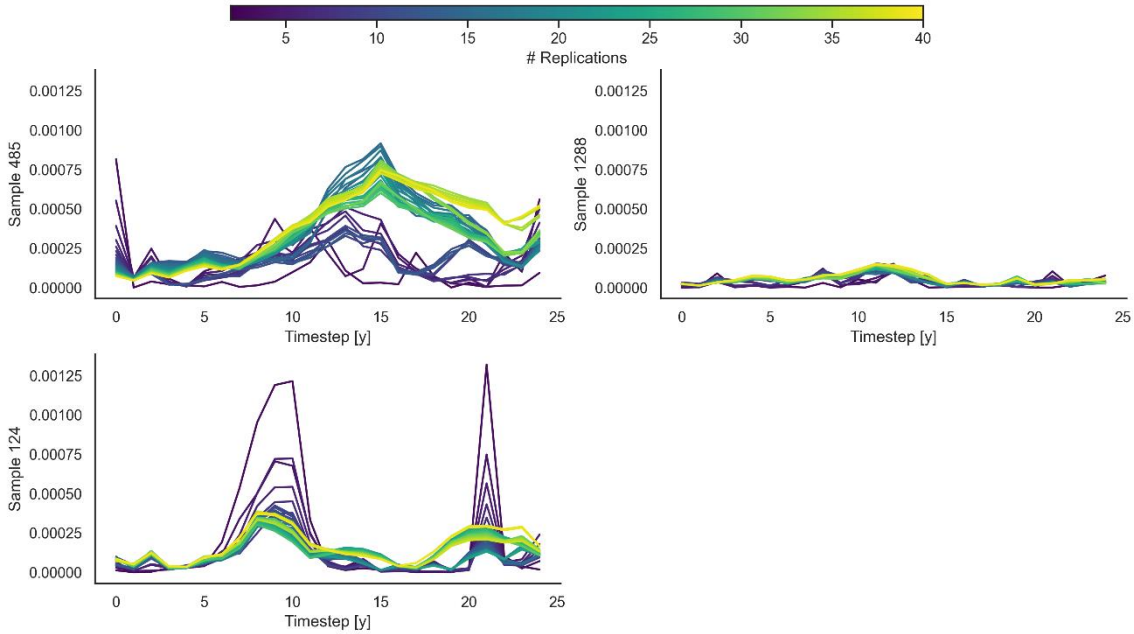


Figure A-8: Convergence of time-series variance in the Gini Coefficient outcome across an increasing number of replications. Variances are shown for four sample points, each representative of a different scenario (clustered by Gini Coefficient dynamics).

Indeed, Figure A-9 seems to agree. The left two plots show signs of some very distinct model behaviours, including short-term but sharp periods of high inequality visible in Sample 124 (bottom-left). Meanwhile, Sample 1288 appears quite consistent both in level and in dynamics.

Behaviour-Based Scenario Discovery

This section repeats Section 4.4 with application to the Gini Coefficient outcome. Figure A-10 shows the three clusters of Gini Coefficient dynamics. They are somewhat distinct, but in a way that is different from the previous two output variables studied. The orange and green clusters end with similar dynamics, but differ roughly in both levels and dynamics at the early stages of the model run: cases in the orange cluster start with high inequality, whereas cases in the green cluster less so. Cases in the blue cluster, however, start and end with a distinct dynamic. They very densely occupy the low end of inequality at the start of model monitoring, but rise to be the high-inequality cases by the end. Knowing what we know about what factors are influential on the model, it is already possible to suspect that *Initial markup* may play a large role in associating cases with this cluster. Low initial markup might lead to very accessible prices, good earnings and savings, and thus low inequality in the early stages, but if it leads to weak businesses without the capital capacity to weather a flood, then it could cause the highly disparate, unequal end states seen in this cluster.

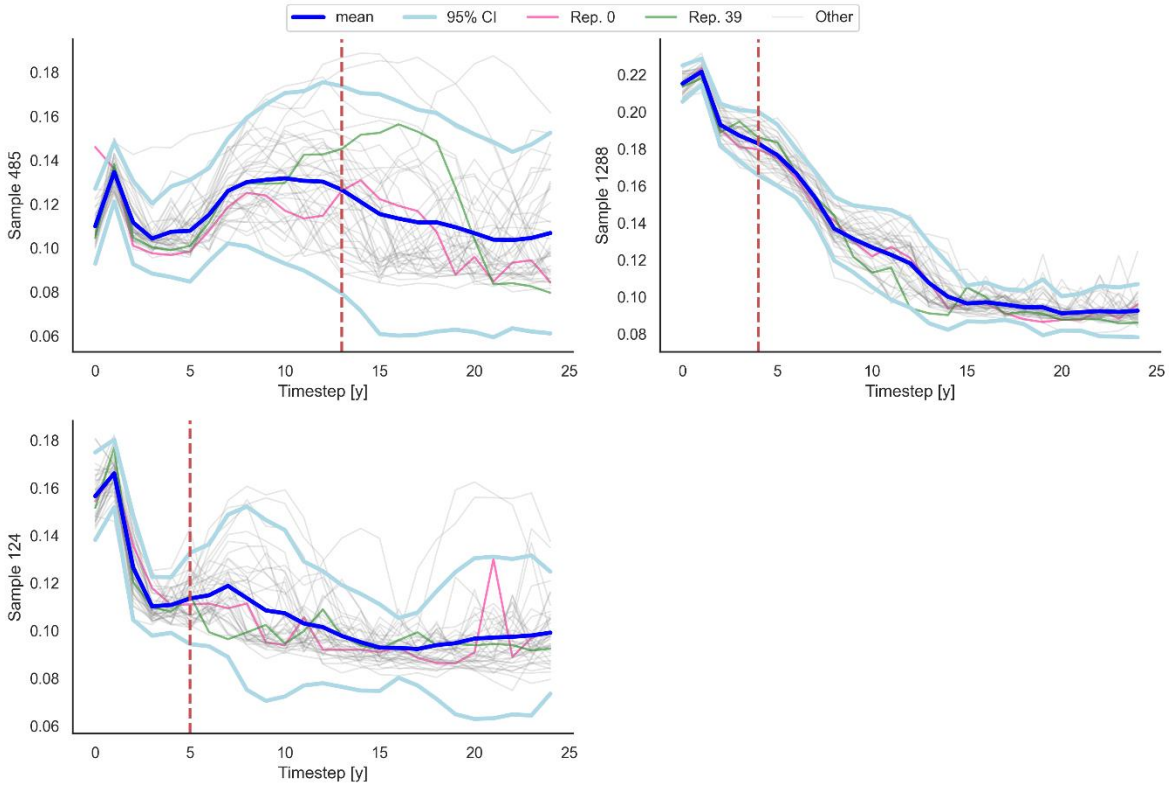


Figure A-9: Aggregated Gini Coefficient dynamics atop individual replications, plotted for each of four sample points. Two samples are highlighted in pink and green to help demonstrate distinct dynamics. Time-series have been resampled. The vertical red line indicates the timing of the flood for the relevant input sample.

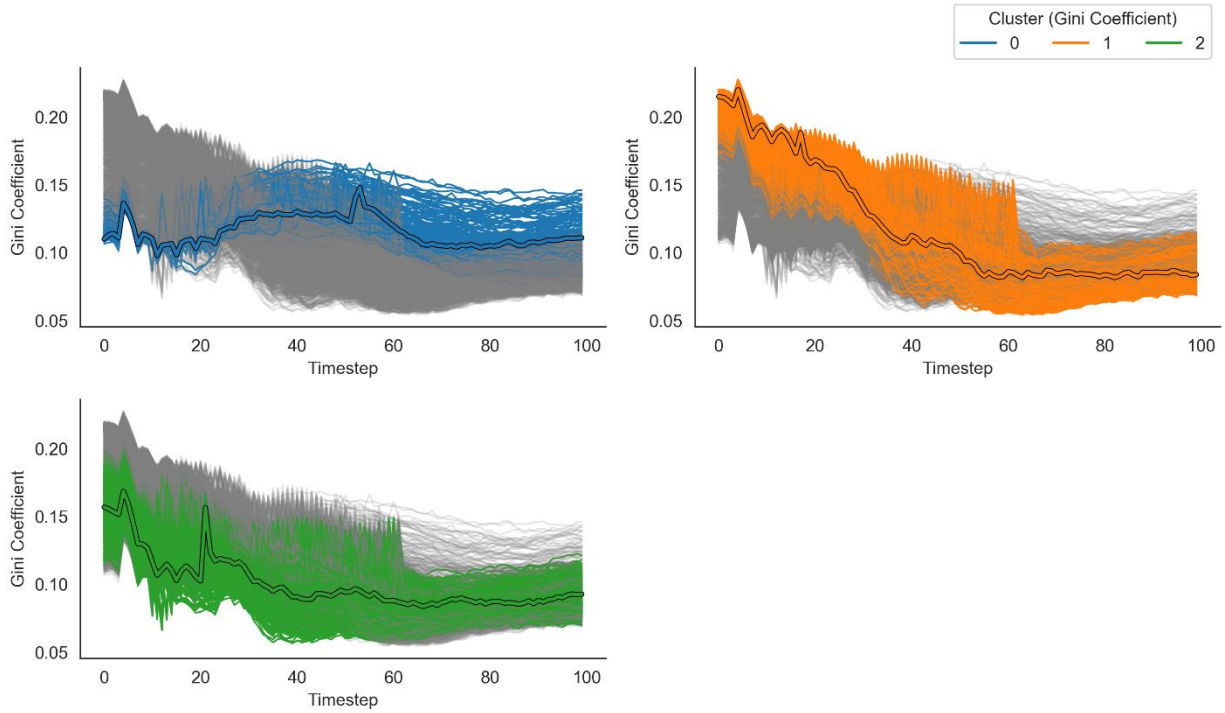


Figure A-10: 2000 realizations of the CRAB model, clustered ($K = 3$) according to dynamics in the Gini Coefficient.

We can name the clusters as follows:

- Blue: Low Initial-High Final
- Orange: High Initial-Low Final
- Green: Low Throughout

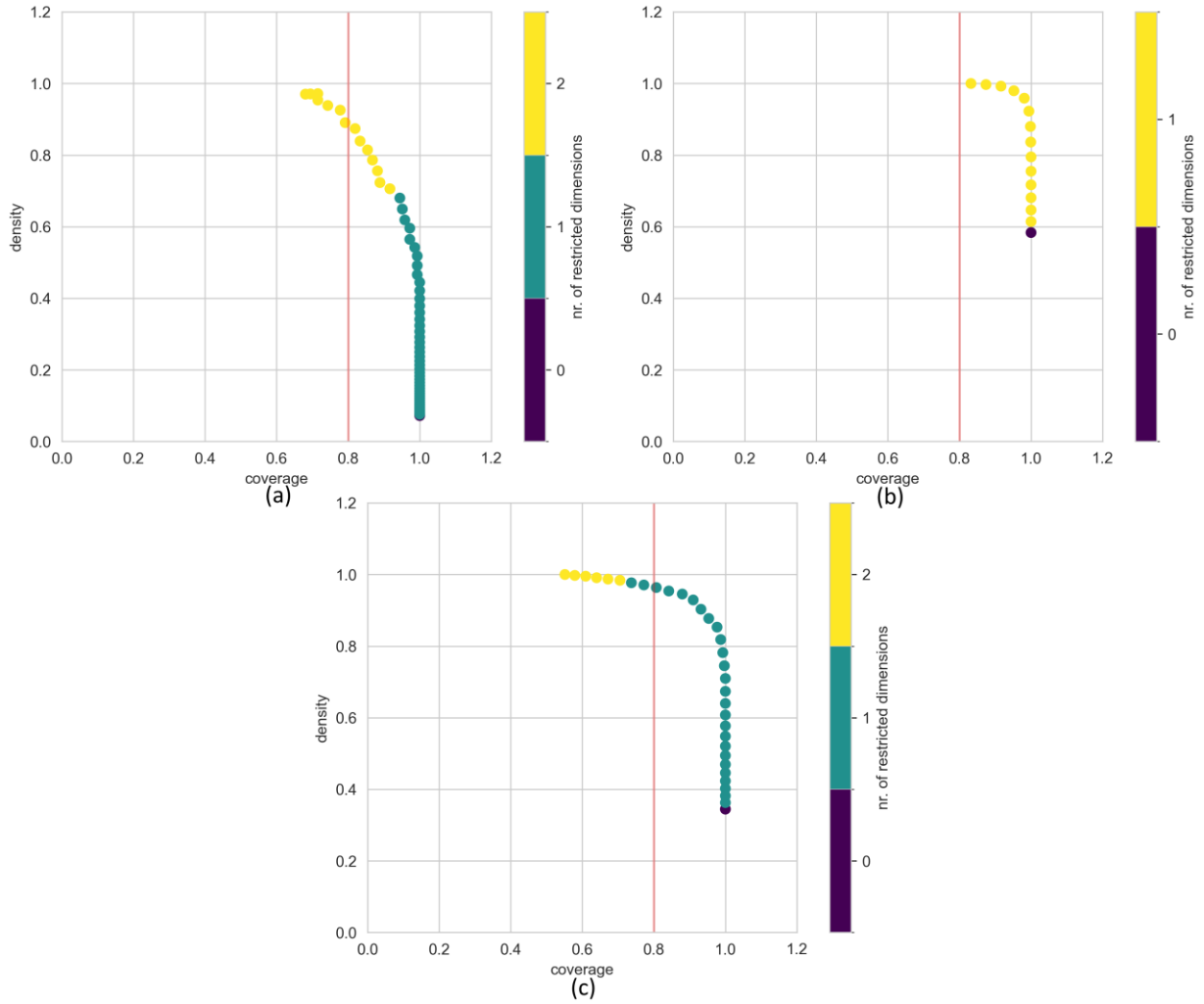


Figure A-11: Coverage-density trade-off curves from applying PRIM to clusters based on Gini dynamics. (a): Low Initial-High Final Inequality (blue cluster); (b) High Initial-Low Final (orange cluster); (c): Low Throughout (green cluster).

Again, the coverage-density trade-offs from Figure A-11 present relatively straightforward decisions, with a clear elbow (a, c), high-density cases with sufficient coverage (all three), and generally few restricted dimensions. The extracted rules follow in : Induced rules from applying PRIM to each cluster of Gini Coefficient dynamics.Table A-2. Again, *Initial markup* is highly dominant, and while *Flood timing* makes an appearance, just a tiny portion of its range is left out in the rule, and thus can be ignored: it adds more cognitive burden to a policymaker to keep track of that than the information lost by losing 2-3 quarters of timing accuracy, especially since flood timing is an inherently stochastic effective.

Table A-2: Induced rules from applying PRIM to each cluster of Gini Coefficient dynamics.

	Init. Markup		Flood Timing	
	min	max	min	max
Low Initial-High Final	0.0501	0.0817	32.5	80.0
High Initial-Low Final	0.2816	0.500	-	-
Low Throughout	0.0957	0.2256	-	-

As *Initial Markup* is the only highly relevant dimension, the tipping points in this variable could be visualized in one dimension, i.e., along a number line. For ease of creation and full information, it is still presented in Figure A-12 as a 2D plot with *Flood timing* on the y-axis. However, if explaining the identifiable tipping points to a client or colleague (one who is technical or otherwise), such a small or tangential restriction in a scenario definition could detract from the overall message.

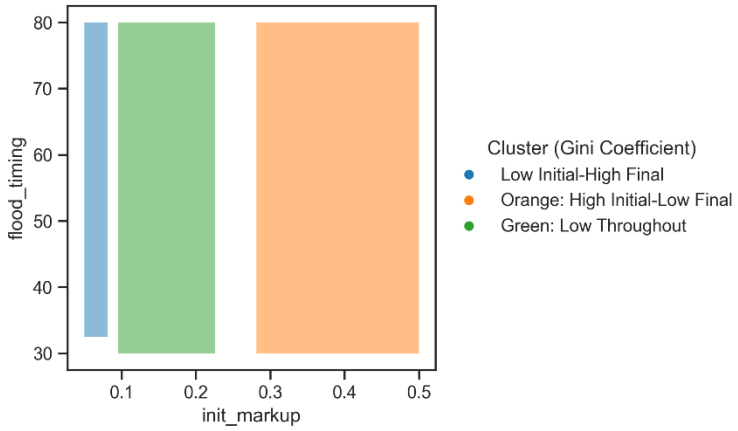


Figure A-12: Gini Coefficient dynamics scenarios plotted in the dimensions in which their rules have restraints.

It is clear the parameter-based tipping points occur in the *Initial markup* variable almost exclusively. This might be surprising, due to the high levels of importance that *Sensitivity of wages to productivity* has in the later model timesteps. Perhaps, this illustrates a weakness in this method: if the role of each parameter variable changes over the time horizon, behaviour-based scenario discovery might bias towards linking together samples that produce similar *early* dynamics, thus ignoring the role of variables that are potentially critical later. For this reason, the methods described in this study should not replace sensitivity analysis or any other method of assessing which factors are important (to a model's behaviour, its tipping points, etc.), but rather complement them.

Appendix B: References

Bankes, S. (1993). Exploratory Modeling for Policy Analysis. *Operations Research, 41*(3), 435-449.

Bankes, S. (2011). The Use of Complexity. In P. Allen, S. Maguire, & B. McKelvey (Eds.), *The SAGE Handbook of Complex and Management* (pp. 570-589).

Barnard, P. L., Dugan, J. E., Page, H. M., Wood, N. J., Finzi Hart, J. A., Cayan, D. R., Erikson, L. H., Hubbard, D. M., Myers, M. R., Melack, J. M., & Iacobelis, S. F. (2021). Multiple climate change-driven tipping points for coastal systems. *Scientific Reports, 11*(2021), 15560. https://doi.org/https://doi.org/10.1038/s41598-021-94942-7

Batista, G., Keogh, E., Tataw, O. M., & de Souza, V. (2014). CID: an efficient complexity-invariant distance for time series. *Data Mining and Knowledge Discovery, 28*(2014), 634-669. https://doi.org/https://doi.org/10.1007/s10618-013-0312-3

Bryant, B. P., & Lempert, R. J. (2010). Thinking inside the box: A participatory, computer-assisted approach to scenario discovery. *Technological Forecasting and Social Change, 77*(1), 34-49.

Carlsen, H., Lempert, R., Wikman-Svahn, P., & Schweizer, V. (2016). Choosing small sets of policy-relevant scenarios by combining. *Environmental Modelling & Software, 84*(2016), 155-164. https://doi.org/http://dx.doi.org/10.1016/j.envsoft.2016.06.011

Castro, J., Drews, S., Exadaktylos, F., Foramitti, J., Klein, F., Konc, T., Savin, I., & van den Bergh, J. (2020). A review of agent-based modeling of climate-energy policy. *WIREs Climate Change, 11*(4), e647. https://doi.org/https://doi.org/10.1002/wcc.647

Delft High Performance Computing Centre. (2024). DelftBlue Supercomputer (Phase 2). https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase2

Developers, A. A. (2024). *pyarrow 17.0.0*. pypi. https://pypi.org/project/pyarrow/

Dietz, S., Rising, J., Stoerk, T., & Wagner, G. (2021). Economic impacts of tipping points in the climate system. *PNAS Sustainability Science, 118*(34). https://doi.org/Economic impacts of tipping points in the climate system

Elsawah, S., Filatova, T., Jakeman, A. J., Kettner, A. J., Zellner, M. L., Athanasiadis, I. N., Hamilton, S. H., Axtell, R. L., Brown, D. G., Gilligan, J. M., Janssen, M. A., Robinson, D. T., Rozenberg, J., Ullah, I. I., & Lade, S. J. (2020). Eight grand challenges in socio-environmental systems modeling. *Socio-Environmental Systems Modeling, 2*, 16226. https://doi.org/https://doi.org/10.18174/sesmo.2020a16226

Elsawah, S., Hamilton, S. H., Jakeman, A. J., Rothman, D., Schweizer, V., Trutnevyte, E., Carlsen, H., Drakes, C., Frame, B., Fu, B., Guivarch, C., Haasnoot, M., Kemp-Benedict, E., Kok, K., Kosow, H., Ryan, M., Hedwig, & van Delden, H. (2020). Scenario processes for socio-environmental systems analysis of futures: A review of recent efforts and a salient research agenda for supporting decision making. *Science of the Total Environment, 729*(2020), 138393. https://doi.org/https://doi.org/10.1016/j.scitotenv.2020.138393

ERC Scalar. (2020). Center for Social Complexity of Climate Change. http://www.sc3.center/project/erc-scalar/

Filatova, T., Polhill, J. G., & van Ewijk, S. (2016). Regime shifts in coupled socio-environmental systems: Review of modelling challenges and approaches. *Environmental Modelling & Software, 75*, 333-347. https://doi.org/https://doi.org/10.1016/j.envsoft.2015.04.003

Filatova, T., Verburg, P. H., Parker, D. C., & Stannard, C. A. (2013). Spatial agent-based models for socio-ecological systems: Challenges and prospects. *Environmental Modeling & Software, 45*, 1-7. https://doi.org/https://doi.org/10.1016/j.envsoft.2013.03.017

Greeven, S., Kraan, O., Chappin, E. J., & Kwakkel, J. H. (2016). The Emergence of Climate Change Mitigation Action by Society: An Agent-Based Scenario Discovery Study Download PDF. *Journal of Artificial Societies and Social Simulation, 19*(3).

Grzymala-Busse, J. W. (2023). Rule Induction. In L. Rokach, O. Maimon, & E. Shmueli (Eds.), *Machine Learning for Data Science Handbook* (pp. 55-74).

Gualdi, S., Tarzia, M., Zamponi, F., & Bouchaud, J.-P. (2015). Tipping points in macroeconomic agent-based models. *Journal of Economic Dynamics & Control, 50*, 29-61. https://doi.org/https://doi.org/10.1016/j.jedc.2014.08.003

Haasnoot, M., Lawrence, J., & Magnan, A. K. (2021). Pathways to coastal retreat. *Science, 372*, 1287-1290. https://doi.org/https://doi.org/10.1126/science.abi6594

Helfmann, L., Heitzig, J., Koltai, P., Kurths, J., & Schutte, C. (2021). Statistical analysis of tipping pathways in agent-based models. *The European Physical Journal Special Topics, 230*, 3249-3271. https://doi.org/https://doi.org/10.1140/epjs/s11734-021-00191-0

Helgeson, C., Srikrishnan, V., Keller, K., & Tuana, N. (2022). Why Simpler Computer Simulation Models Can Be Epistemically Better for Informing Decisions. *Philosophy of Science, 88*(2), 213-233. https://doi.org/https://doi.org/10.1086/711501

Helton, J. C., & Davis, F. J. (2003). Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliability Engineering & System Safety, 81*(1), 23-69. https://doi.org/https://doi.org/10.1016/S0951-8320(03)00058-9

Horan, R. D., Fenichel, E. P., Drury, K. L., & Lodge, D. M. (2011). Managing ecological thresholds in coupled environmental–human systems. *Proceedings of the National Academy of Sciences of the United States of America, 108*(18), 7333-7338.

Jafino, B. A., & Kwakkel, J. H. (2021). A novel concurrent approach for multiclass scenario discovery using Multivariate Regression Trees: Exploring spatial inequality patterns in the Vietnam Mekong Delta under uncertainty. *Environmental Modelling & Software, 145*, 105177.

Kaaronen, R. O., & Streikovskii, N. (2020). Cultural Evolution of Sustainable Behaviours: Pro-enviromental Tipping Points in an Agent-Based Model. *One Earth, 2*, 85-97. https://doi.org/https://doi.org/10.1016/j.oneear.2020.01.003

Kwakkel, J. H. (2013). Exploratory Modeling and Analysis Workbench. https://github.com/quaquel/EMAworkbench

Kwakkel, J. H., & Haasnoot, M. (2019). Supporting DMDU: A Taxonomy of Approaches and Tools. In V. A. Marchau, W. E. Walker, P. J. Bloemen, & S. W. Popper, *Decision Making under Deep Uncertainty: From Theory to Practice* (pp. 355-374). Springer Cham.

Kwakkel, J. H., & Jaxa-Rozen, M. (2016). Improving scenario discovery for handling heterogeneous uncertainties and multinomial classified outcomes☆. *Environmental Modelling & Software, 79*, 311-321.

Lamontagne, J. R., Reed, P. M., Link, R., Calvin, K. V., Clarke, L. E., & Edmonds, J. A. (2018). Large ensemble analytic framework for consequence-driven discovery of climate change scenarios. *Earth's Future, 6*(3), 488-504. https://doi.org/https://doi.org/10.1002/2017EF000701

Lee, J.-S., Filatova, T., Ligmann-Zielinska, A., Hassani-Mahmooei, B., Stonedahl, F., Lorscheid, I., Voinov, A., Polhill, G., Sun, Z., & Parker, D. C. (2015). The Complexities of Agent-Based Modeling Output Analysis. *Journal of Artificial Societies and Social Simulation , 18*(4). https://doi.org/https://doi.org/10.18564/jasss.2897%0A

Lempert, R. J. (2003). *Shaping the Next One Hundred Years: New Methods for Quantitative, Long-Term Policy Analysis.* Rand Corporation.

Lenton, T. M., Held, H., Kriegler, E., & Schellnhuber, H. J. (2008). Tipping elements in the Earth's climate system. *Proceedings of the National Academy of the Sciences of the United States of America, 105*(6), 1786-1793.

Lenton, T. M., Laybourn, L., McKay, D. I., Loriani, S., Abrams, J. F., Lade, S. J., Donges, J. F., Milkoreit, M., Powell, T., Smith, S. R., Zimm, C., Bailey, E., Buxton, J. E., Dyke, J. G., & Ghadiali, A. (2023). *Global Tipping Points Report 2023.* University of Exeter, Exeter, UK.

Li, H., & Liu, Z. (2021). Multivariate time series clustering based on complex network. *Pattern Recognition, 115*, 107919. https://doi.org/https://doi.org/10.1016/j.patcog.2021.107919

Ligmann-Zielinska, A., & Sun, L. (2010). Applying time-dependent variance-based global sensitivity analysis to represent the dynamics of an agent-based model of land use change. *International Journal of Gerographical Information Sicence, 24*(12), 1829-1850.

Ligmann-Zielinska, A., Siebers, P.-O., Magliocca, N., Parker, D., Grimm, V., Du, E., Cenek, M., Radchuk, V., Arbab, N. M., Li, S., Berger, U., Paudel, R., Robinson, D. T., Jankowski, P., An, L., & Ye, X. (2020). 'One Size Does Not Fit All': A Roadmap of Purpose-Driven Mixed-Method Pathways for Sensitivity Analysis of Agent-Based Models. *Journal of Artificial Societies and Social Simulation, 23*(1). https://doi.org/https://www.jasss.org/23/1/6.html

Lippe, M., Bithell, M., Gotts, N., & Natali, D. (2019). Using agent-based modelling to simulate social-ecoological systems across scales. *GeoInformatica, 23*(2019), 269-298. https://doi.org/https://doi.org/10.1007/s10707-018-00337-8

Lorscheid, I., Heine, B.-O., & Meyer, M. (2012). Opening the 'black box'of simulations: increased transparency and effective communication through the systematic design of experiments. *Computational & Mathematical Organization Theory, 18*(2012), 22-62. https://doi.org/https://doi.org/10.1007/s10588-011-9097-3

Magliocca, N., McConnel, V., & Walls, M. (2018). Integrating Global Sensitivity Approaches to Deconstructing Spatial and Temporal Sensitivities of Complex Spatial Agent-Based Models. *Journal of Artificial Societiesi & Social Simulation, 21*(1). https://doi.org/https://doi.org/10.18564/jasss.362

Milkoreit, M., Hodbod, J., Baggio, J., Benessaiah, K., Calderón-Contreras, R., Donges, J. F., Mathias, Jean-Denis, Rocha, J. C., Schoon, M., & Werners, S. E. (2018). Defining tipping points for social-ecological systems scholarship—an interdisciplinary literature review. *Environmental Research Letters, 13*(3), 033005.

Moallemi, E. A., Gao, L., Eker, S., & Bryan, B. A. (2022). Diversifying models for analysing global change scenarios and sustainability pathways. *Global Sustainability, 5*(7). https://doi.org/https://doi.org/10.1017/sus.2022.7

Moallemi, E. A., Kwakkel, J. H., de Haan, F. J., & Bryan, B. A. (2020). Exploratory modeling for analyzing coupled human-natural systems under uncertainty. *Global Environmental Change, 65*, 102186.

Noll, B., Filatova, T., Need, A., & Taberna, A. (2022). Contextualizing cross-national patterns in household climate change adaptation. *Nature Climate Change, 12*(2022), 30-35. https://doi.org/https://doi.org/10.1038/s41558-021-01222-3

O'Neill, B. C., Carter, T. R., Ebi, K., Harrison, P. A., Kemp-Benedict, E., Kok, K. , Kriegler, E., Preston, B. L., Riahi, K., Sillmann, J., van Ruijven, B. J., van Vuuren, D., Carlisle, D., Conde, C., Fuglestvedt, J., Green, C., Hasegawa, T., Leininger, J., Monteith, S., & Pichs-Madruga, R. (2020). Achievements and needs for the climate change scenario framework. *Nature Climate CHange, 10*(2020), 1074-1084. https://doi.org/https://doi.org/10.1038/s41558-020-00952-0

Otto, I. M., Donges, J. F., Cremades, R., Bhowmik, A., Hewitt, R. J., Lucht, W., Rockstrom, J., Allerberger, F., McCaffrey, M., Doe, S. S., Lenferna, A., Moran, N., van Vuuren, D., & Schellnhuber, H. J. (2020). Social tipping dynamics for stabilizing Earth's climate by 2050. *Proceedings of the National Academy of Sciences, 117*(5), 2354-2365. https://doi.org/https://doi.org/10.1073/pnas.1900577117

Saltelli, A., Aleksankina, K., Becker, W., Fennel, P., Ferretti, F., Holst, N., Li, S., & Wu, Q. (2019). Why so many published sensitivity analyses are false: A systematic review of sensitivity analysis practices. *Environmental Modelling & Software, 114*(2019), 29-39. https://doi.org/https://doi.org/10.1016/j.envsoft.2019.01.012

Scheffer, M., Bascompte, J., & Brock, W. A. (2009). Early-warning signals for critical transitions. *Nature, 461*, 53-59. https://doi.org/https://doi.org/10.1038/nature08227

Scheffer, M., Carpenter, S. R., Lenton, T. M., Bascompte, J., Brock, W., Dakos, V., van de Koppel, J., van de Leemput, I. A., Levin, S. A., van Nes, E. H., Pascual, M., & Vandermeer, J. (2012). Anticipating Critical Transitions. *Science, 338*(6105), 344-348. https://doi.org/https://doi.org/10.1126/science.1225244

Sher, G. (2024a). Exploratory Modeling with the CRAB Model. https://zenodo.org/doi/10.5281/zenodo.13365139

Sher, G. (2024b). EMA Open Exploration Dashboard. https://zenodo.org/doi/10.5281/zenodo.13364651

Singhal, A., & Seborg, D. E. (2006). Clustering multivariate time-series data. *Journal of chemometrics, 19*(8), 427-438. https://doi.org/https://doi.org/10.1002/cem.945

Steinmann, P. (2018). *Behavior-Based Scenario Discovery.* Delft University of Technology [Thesis].

Steinmann, P., Auping, W. L., & Kwakkel, J. H. (2020). Behavior-based scenario discovery using time series clustering. *Technological Forecasting and Social Change, 156*, 120052.

Suzuki, S., Stern, D., & Manzocchi, T. (2015). Using Association Rule Mining and High-Dimensional Visualization to Explore the Impact of Geological Features on Dynamic Flow Behavior. *SPE Annual Technical Conference and Exhibition.* Houston, Texas. https://doi.org/https://doi.org/10.2118/174774-MS

Taberna, A., Filatova, T., Hadjimichael, A., Noll, & Brayton. (2023). Uncertainty in boundedly rational household adaptation to environmental shocks. *Proceedings of the National Academy of Sciences of the United States of America, 120*(44).

Taberna, A., Filatova, T., Hochrainer-Stigler, S., Nikolic, I., & Noll, B. (2023). Economic implications of autonomous adaptation of firms and households in a resource-rich coastal city. *Scientific Reports, 13*(1).

Taberna, A., Filatova, T., Roventini, A., & Francesco, L. (2021). Coping with increasing tides: technological change, agglomeration dynamics and climate hazards in an agent-based evolutionary model. *LEM Working Paper Series, 2021*(44).

van Ginkel, K. C., Botzen, W. J., Haasnoot, M., Bachner, G., Steininger, K. W., Hinkel, J., Watkiss, P., Boere, E., Jeuken, A., & de Murieta, E. S. (2020). Climate change induced socio-economic tipping points: review and stakeholder consultation for policy relevant research. *Environmental Research Letters, 15*(2), 023001.

van Ginkel, K. C., Haasnoot, M., & Botzen, W. J. (2022). A stepwise approach for identifying climate change induced socio-economic tipping points. *Climate Risk Management, 37*(2022), 100445. https://doi.org/https://doi.org/10.1016/j.crm.2022.100445

van Nes, E. H., Arani, B. M., Staal, A., van der Bolt, B., Flores, B. M., Bathiany, S., & Scheffer, M. (2016). What Do You Mean, 'Tipping Point'? *Trends in Ecology & Evolution, 32*(12), 902-904. https://doi.org/https://doi.org/10.1016/j.tree.2016.09.011

Walker, W. E., Lempert, R. J., & Kwakkel, J. H. (2013). Deep Uncertainty. In S. I. Gass, & M. C. Fu (Eds.), *Encyclopedia of Operations Research and Management Science (3rd edition)* (pp. 395-402).

Wang, P., Gerst, M. D., & Borsuk, M. E. (2013). Exploring Energy and Economic Futures Using Agent-Based Modeling and Scenario Discovery. In H. Qudrat-Ullah (Ed.), *Energy Policy Modeling in the 21st Century* (pp. 251-269). https://doi.org/https://doi.org/10.1007/978-1-4614-8606-0_13

Wright, D., Stahl, B., & Hatzakis, T. (2020). Policy scenarios as an instrument for policymakers. *Technological Forecasting and Social Change, 154*(2020), 119972. https://doi.org/https://doi.org/10.1016/j.techfore.2020.119972

Zhou, P.-Y., & Chan, K. C. (2014). A Model-Based Multivariate Time Series Clustering Algorithm. *Trends and Applications in Knowledge Discovery and Data Mining*, (pp. 805-817). Tainan. https://doi.org/10.1007/978-3-319-13186-3_72