From Model-Based Screening to Data-Driven Surrogates: A Multi-Stage Workflow for Exploring Stochastic Agent-Based Models

Paul Saves[1 [0000−0001−5889−2302]], Matthieu Mastio[1 [0009−0002−3486−3739]], Nicolas Verstaevel[1 [0000−0002−7879−6681]], and Benoit Gaudou[1 [0000−0002−9005−3004]]

IRIT, Université Toulouse Capitole, Toulouse, France. {first.last}@irit.fr

Abstract. The systematic exploration of Agent-Based Models (ABMs) is challenged by the curse of dimensionality and their inherent stochasticity. We present a multi-stage pipeline integrating the systematic design of experiments with machine learning surrogates. Using a predator-prey case study, our methodology is carried out in two steps. First, an automated model-based screening identifies dominant variables, assesses outcome variability, and segments the parameter space. Second, we train Machine Learning models to map the remaining nonlinear interaction effects. This approach automates the discovery of *unstable* regions where system outcomes are highly dependent on nonlinear interactions between many variables. Thus, this work provides modelers with a rigorous, hands-off framework for sensitivity analysis and policy testing, even when dealing with high-dimensional stochastic simulators.

Keywords: Agent-based models · Simulation · Machine learning · Uncertainty Quantification · Sensitivity Analysis

1 Introduction

Agent-based models are powerful tools for simulating complex systems based on autonomous agents that interact within an environment. These models enable the study of emergent phenomena when local interactions give rise to global behaviors that are often difficult to capture with traditional analytical methods [4]. In ecology and socio-environmental systems, ABMs are particularly appealing because they allow researchers to relax strong analytical assumptions (*e.g.*, well-mixed populations or perfect rationality) and explicitly represent space, stochasticity, and interaction networks [2]. Despite their flexibility, ABMs are often criticized for lacking systematic exploration protocols, making it difficult to assess the robustness, generality, and policy relevance of their results [15].

In this paper, we propose a general protocol for model exploration that systematically investigates the behavior of agent-based simulation models as a function of actionable parameters. Rather than focusing on empirical calibration or prediction, our objective is to support exploratory analysis and *what-if* reasoning. The proposed protocol is designed to be transferable across simulation

models for which external levers, such as regulatory, legal, or institutional constraints, can be meaningfully applied. To illustrate this protocol, we rely on a deliberately simple toy model inspired by predator–prey dynamics with renewable resources [7]. We aim to enhance the credibility of model predictions by systematically quantifying their sensitivity to parameter variability and intrinsic stochasticity. To support reproducibility, the open-source research pipeline, including the NetLogo simulation model [23,13], analysis scripts, datasets, and supplementary figures, is freely available in the online repository[1].

The remainder of this paper is organized as follows. Section 2 positions our work within the existing literature. Section 3 details the spatial predator-prey simulation case study. Section 4 introduces our proposed multi-stage workflow: Section 4.1 presents the experimental design and quantifies the simulation's intrinsic stochasticity; Section 4.2 performs a preliminary model-based screening using linear and tree-based methods; and Section 4.3 details the machine learning surrogate approach for nonlinear data-based uncertainty quantification. Finally, Section 5 sums up our contributions and discusses future perspectives.

2 Related works

Assessing the robustness and validity of Agent-Based Models (ABM) requires moving beyond single-trajectory simulations toward systematic global exploration. Still, the combination of high-dimensional parameter spaces, nonlinear dynamics, and intrinsic stochasticity makes this task computationally prohibitive [2]. In complex systems analysis, the primary goal of model exploration is to map the relationship between input parameters and output variability. While early approaches relied on local one-at-a-time methods, the field has shifted toward Global Sensitivity Analysis (GSA) to capture interactions over the full parameter space [9]. Variance-based methods are often considered the gold standard for GSA as they decompose the output variance into contributions from single parameters and their interactions. However, calculating these indices requires Monte-Carlo sampling schemes that are computationally expensive for slow-running ABMs. Furthermore, simpler screening methods may fail to capture the complex, non-monotonic responses typical of ecological or social simulations, potentially leading to misleading conclusions about parameter importance [20].

To tackle the computational cost associated with GSA, one can use surrogate model predictions for data augmentation. They generally consist of statistical or machine learning approximations trained on a limited set of ABM simulations to predict outcomes at a fraction of the cost [1]. To do so, Gaussian processes or random forests have historically been preferred for their built-in uncertainty estimates, but they struggle to scale beyond low-dimensional spaces. Consequently, more complex methods such as gradient boosting, as well as deep neural networks, have gained traction [1]. The latter may be less interpretable but offers the flexibility to approximate highly nonlinear response surfaces, interactions, and discontinuous regimes (*e.g.*, tipping points) inherent to multi-agent systems [10].

[1] https://github.com/ANR-MIMICO/MABS2026_Preys_Predators

A critical yet overlooked challenge in ABM exploration is the distinction between sources of uncertainty. Standard surrogate approaches often treat the simulator as a deterministic blackbox, ignoring the *aleatoric uncertainty* (intrinsic stochasticity) arising from random seeds in the ABM. Conversely, *epistemic uncertainty* arises from the surrogate's lack of training data in certain regions. Failing to separate these can lead to overconfident predictions in unstable regimes [8]. Furthermore, the black-box nature of complex machine learning surrogate models poses an interpretability challenge. In this context, conformal prediction has emerged as a critical framework for reliable uncertainty quantification. Unlike traditional Bayesian or error-variance methods, it provides distribution-free and finite-sample tools to construct rigorous prediction intervals that bound the surrogate's approximation error with a pre-specified confidence level [19]. To bridge the gap between accurate prediction and mechanistic understanding, post-hoc explainable techniques, such as Partial Dependence Plots (PDP) and Individual Conditional Expectation (ICE), are becoming essential [14]. These tools allow modelers to visualize not just which parameters matter, but how they influence the system (*e.g.*, identifying thresholds and phase transitions). In this paper, we propose a pipeline that unifies these elements: leveraging machine learning whenever necessary for nonlinear approximations while rigorously decomposing uncertainty to identify robust stability boundaries and critical regimes.

3 Model Description

3.1 Baseline Model

We extend a classical predator–prey toy model [13] by introducing three spatially explicit mechanisms that drive complex system dynamics. We opted not to use the unmodified baseline model because its homogeneous resource distribution yields highly linear, simplistic dynamics that do not require complex exploration and can be modeled with an equation-based surrogate [22]. Two types of agents interact in this model: herbivores (bandicoots) gain energy by consuming the renewable resource, while predators (foxes) gain energy by consuming herbivores. The simulation is implemented in Netlogo[2] [2,23] and is stopped after 1000 timesteps (t). The environment is defined as a square lattice of 60×60 discrete patches. The simulation environment is defined as a toroidal grid to eliminate edge effects and avoid boundary handling.

Each patch contains a renewable resource (grass) that serves as the primary energy source for herbivorous agents. Resource availability on patch i at time t is represented by a continuous variable $R_i(t) \in [0, R_{\max}]$. Grass growth follows the regenerative linear process as $R_i(t+1) = \min\left(R_{\max},\ R_i(t) + g\right)$, where g is the intrinsic growth rate of the resource. When grass is consumed, $R_i(t)$ is reduced proportionally to the number of herbivores on the patch and their intake. Each agent a of species s is characterized by its position $\mathbf{x}_a(t)$, its energy $E_a(t)$, its age $A_a(t)$ and its maximum lifespan $A_s^{\max}$. Native herbivores feed on the grass

[2] http://ccl.northwestern.edu/netlogo/

in their current patch. When resources are abundant, bandicoots consume the vegetation and leave behind a patch of barren soil. Similarly, predators feed on native herbivores occupying the same patch, instantly removing a prey. In both cases, the successful agent gains energy according to $E_a(t+1) = E_a(t) + \alpha_s$, where α_s is the energy gain parameter specific to the species s. Energy decreases by 1 every tick and increases through successful feeding. Agents die if their energy becomes negative ($E_a(t) < 0$) or if their age exceeds the species-specific maximum lifespan ($A_a(t) > A_s^{\max}$).

Reproduction is asexual and energy-dependent. An agent has 50% chance of reproducing if $E_a(t) \geq E_s^{\text{rep}}$ and $A_a(t) \geq A_s^{\text{rep}}$, where E_s^{rep} and A_s^{rep} are species-specific thresholds. The number of offspring is drawn from a bounded discrete distribution, but since each child receives E_s^{rep} energy from its parent, the maximum number of offspring is therefore proportional to the parent's energy. Offsprings spawn one stride distance away from their parents.

3.2 Spatial extensions

Unlike classical predator–prey models assuming homogeneous resource availability [13], grass is distributed non-uniformly across space. Initial grass patches are generated around a limited number of spatial centers, with an exponentially decreasing patch fertility with respect to the nearest center distance: $\mathbb{P}(i \text{ is fertile}) \propto \exp(-kd_i)$, where d_i denotes the distance from patch i to the closest grass cluster center and k controls the spatial decay rate. This mechanism produces clustered resource landscapes, introducing spatial heterogeneity and localized competition.

Also, compared to the dummy implementation, agent movements are not purely stochastic when food sources or prey are locally available. Instead, agents exhibit directed movement toward relevant targets within their perceptual range to improve the spatialization of the model and the interactions between the agents and their environments. Each agent of species s is endowed with a perception radius r_s, defining a local neighborhood $\mathcal{N}_a$ within which relevant environmental information can be detected. Herbivorous agents move toward the nearest patch containing available grass within $\mathcal{N}_a$. Predatory agents move toward the nearest detectable prey individual within $\mathcal{N}_a$. When no suitable target is detected within the perception neighborhood, agents default to a random walk. Additionally, to keep the toy model's dimensionality manageable, bandicoots are modeled with perceptual capacities strictly limited to foraging, rather than complex predator-evasion mechanics.

Finally, we introduce a spatial risk component through the implementation of localized hunting zones. A subset of the fertile patches is randomly designated as regulated hunting areas. In these locations, both native herbivores and predators are subject to an additional stochastic death probability, $P_{death,s}$, which is proportional to the established hunting quotas for species s. This mechanism simulates anthropogenic pressure and introduces a non-biological source of mortality that varies across the landscape.

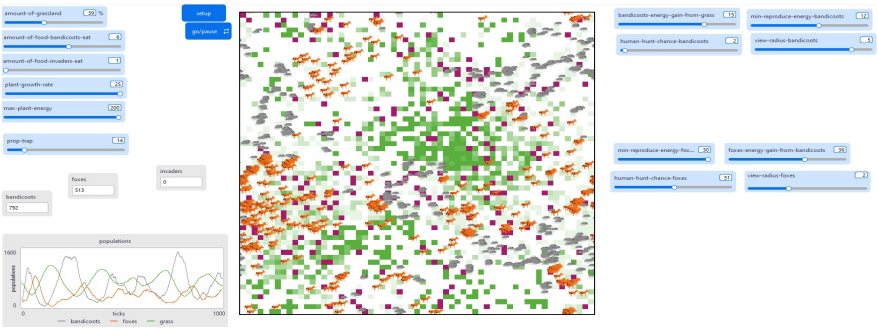


Fig. 1: A dynamic simulation for a given set of inputs.

The resulting graphical NetLogo interface is given in Fig. 1. The parallel execution of these simulations was orchestrated using the OpenMOLE platform, taking approximately 3 seconds per run by leveraging the 16 cores of an Intel Core Ultra 9 185H vPro processor [16].

4 Multi-Stage Exploration Pipeline

4.1 Multi-stage exploration workflow

The primary challenge in exploring stochastic ABMs lies in the computational cost of resolving high-dimensional interactions while accounting for aleatoric noise. To address this, we propose a hierarchical "zoom-in" pipeline, illustrated in Fig. 2, designed to bridge the gap between global statistical trends and local mechanical realities.

The methodology transitions from a coarse *model-based screening* to a fine-grained *data-driven analysis*. Initially, we assess internal stochasticity using linear and simple global screening to identify dominant drivers and interpretable extraction rules methods to segment stability regions locally. We then deepen this analysis by training, if needed, one or many machine learning surrogate models to capture nonlinear dynamics. This second stage facilitates a rigorous dual assessment: variance-based global sensitivity analysis ranks interactions [9], while local explainability tools map the precise functional forms and tipping points governing ecosystem resilience [11]. This paper illustrates a two-stage workflow, but depending on the underlying model at hand, more steps may be required, or the first step could be sufficient.

To systematically explore the model's behavior and the interactions between spatial heterogeneity and agent behavior, we identify 13 continuous variables and one categorical seed variable. These parameters characterize environmental constraints, prey (bandicoot) metabolism, and predator (fox) efficiency. The experimental variables and their respective ranges are summarized in Tab. 1. We assumed a uniform distribution across the 13-dimensional input space, E denotes energy units, and t denotes ticks.

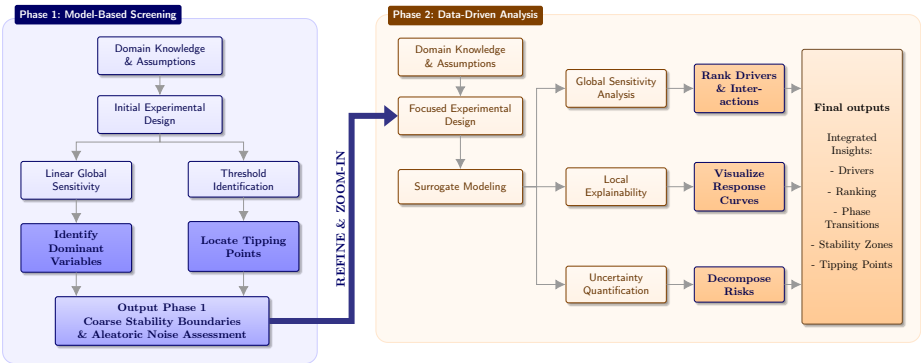


Fig. 2: Multi-stage exploration workflow.

To illustrate our methodology, we focus on population subsistence, $y \in \{\texttt{extinction}, \texttt{prey_survival}, \texttt{coexistence}\}$, as the primary ordinal output of the model [17]. For computational simplicity, we encode this output numerically as $y \in \{0, 0.5, 1\}$. Although assigning 0.5 to the intermediate state is an approximation and sophisticated methods exist to learn nonlinear warping operators for ordinal data [17], we adopt this linear encoding as a convenient proxy for stability ordering. This simplification does not impact the classification-based surrogate analyses performed later in the study. While our pipeline treats the underlying ABM as a blackbox, meaning no mechanistic model code needs to be modified, domain knowledge is strictly used to define the relevant input boundaries and select the output metrics of interest. Furthermore, while this case study focuses on continuous inputs, the pipeline natively supports categorical variables through standard one-hot encoding techniques before surrogate training.

4.2 Model-based Analysis

Following our workflow, we analyze the raw simulation data to identify global trends and assess the impact of stochasticity. This step informs the subsequent

Table 1: Model parameters, symbols, and experimental ranges.

Category	Parameter	Symbol	Range / Units
Environment	Grassland proportion on the map	Gr	$[10.0, 100.0]$ (%)
	Proportion of hunting zone among grassland	PH	$[0.0, 100.0]$ (%)
Plants	Maximum Plant Energy	PM	$[50.0, 250.0]$ (E)
	Plant Growth Rate	PR	$[5.0, 25.0]$ (E/t)
	Grass energy intake of eating bandicoots	BF	$[2.0, 10.0]$ (E)
Bandicoots	Energy Gain when eating	BG	$[1.0, 20.0]$ (E)
	Energy reserves required for reproduction	BR	$[8.0, 20.0]$ (E)
	Hunting quota on hunting zone	BH	$[0.0, 100.0]$ (%)
	View Radius to search for grass	BV	$[0.0, 3.0]$ (cells)
Foxes	Energy Gain when eating	FG	$[10.0, 50.0]$ (E)
	Energy reserves required for reproduction	FR	$[12.0, 30.0]$ (E)
	Hunting quota on hunting zones	FH	$[0.0, 100.0]$ (%)
	View Radius for hunting bandicoots	FV	$[0.0, 3.0]$ (cells)
Replications	Random Seed	S	$\{1, 2, 3, 4, 5\}$

data-driven surrogate model by highlighting the limitations of linear assumptions. We first start with a global analysis; then we dive into more local insights.

Initial Experimental Design and Data Analysis. To ensure efficient and space-filling sparse exploration, we employed Latin Hypercube Sampling (LHS) to generate $N = 650$ unique parameter configurations (50 points times the number of variables). To account for model stochasticity, each configuration was executed across $n = 5$ independent replications using distinct random seeds, resulting in a total of **3,250 simulation runs**.

In the data, the outcomes are significantly unbalanced because 60% of the data correspond to a total extinction, 27% correspond to a prey survival, and only 13% of the data correspond to a sustainable coexistence between the two species. Therefore, our first analysis will focus on understanding globally the response across the whole search space to identify dominant variables. Then, a second analysis will focus on understanding the drivers of coexistence with a decision tree to locate tipping points.

We quantify the inherent stochasticity of the simulation to establish a theoretical performance benchmark. Since the model output y is a random variable conditional on the seed S, no deterministic predictor can achieve 100% accuracy. We explicitly measured this *Aleatoric Uncertainty* to distinguish between surrogate modeling error and irreducible simulation noise (Bayes error rate). To determine the upper bound of predictability, we constructed a theoretical oracle predictor consisting of the Bayes optimal classifier [21]. For each unique parameter configuration X_i, we aggregated the five replicates $Y_i = \{y_{i,1}, \ldots, y_{i,5}\}$ and computed the group median $\tilde{y}_i$. We effectively asked: *"If a model perfectly learned the central tendency of the system, how often would it correctly predict the individual simulation runs?"* In our model, the theoretical maximum accuracy (Acc_{max}) is defined as the proportion of individual runs that align exactly with their group median: $Acc_{max} = \frac{1}{N \times n} \sum_{i=1}^{N} \sum_{j=1}^{n} \mathbb{I}(y_{i,j} = \tilde{y}_i)$. The analysis reveals a theoretical accuracy limit of 95.3%. This implies that only 4.7% of the variance is purely aleatoric (noise) and cannot be explained by the input parameters.

Leveraging ANOVA for Linear Global Sensitivity Analysis. To quantify the contribution of each parameter to the variance of the population outcome (y), we conducted a simple ANOVA on a generalized linear model fitted by ordinary least squares. The ANOVA used Type II sums of squares to estimate the proportion of variance explained by each factor while controlling for other main effects [9]. Note that we could have opted for other methods, such as one-at-a-time Morris elementary effects [9]. Analyzing the results reveals a clear hierarchy of linear drivers. The Proportion of Hunting Zones (PH) is the dominant factor, explaining approximately 25.7% of the total variance. This is followed by the Bandicoot Hunt Chance (BH, 9%) and the Bandicoot Energy Gain when eating (BG, 7E). Within the linear approximation, these results suggest that anthropogenic pressure contributes more to variability in outcomes than most individual metabolic parameters. The Seed (S) factor explains negligible vari-

ance, implying that internal stochasticity does not bias global trends. A χ^2 test across seeds ($\chi^2 = 1.25, p > 0.99$) further confirms that population regimes are independent of random initializations. However, the linear model achieves an R^2 of only 43%, leaving a large residual variance (56.5%). This substantial unexplained variance indicates that nonlinear interactions and higher-order effects play a major role in system dynamics and motivates the use of nonlinear surrogate methods in subsequent analysis. While the linear analysis highlights a strong sensitivity to anthropogenic parameters, the significant residual variance confirms that linear methods are insufficient to capture the model's full complexity. This motivates the subsequent use of nonlinear surrogate modeling to understand these interactions better.

Leveraging a Decision Tree for Threshold Identification. To resolve the nonlinearities identified by the ANOVA residuals, we trained a regression tree to identify specific parameter thresholds that trigger regime shifts [3], yielding the decision tree of Fig. 3. This allows for studying the identified local phenomena locally. The tree identifies a primary tipping point at $PH \approx 31\%$. When hunting zone density exceeds this threshold, the system stability relies heavily on BH. If $BH > 18\%$, the outcome converges toward total extinction (mean value of 8%). Conversely, in low hunting zone environments, metabolic factors become critical due to the fact that there are fewer hunting zones. In that case, a low energy gain when bandicoots are eating, $BG < 4$, also leads to the total extinction of both species (7% on average).

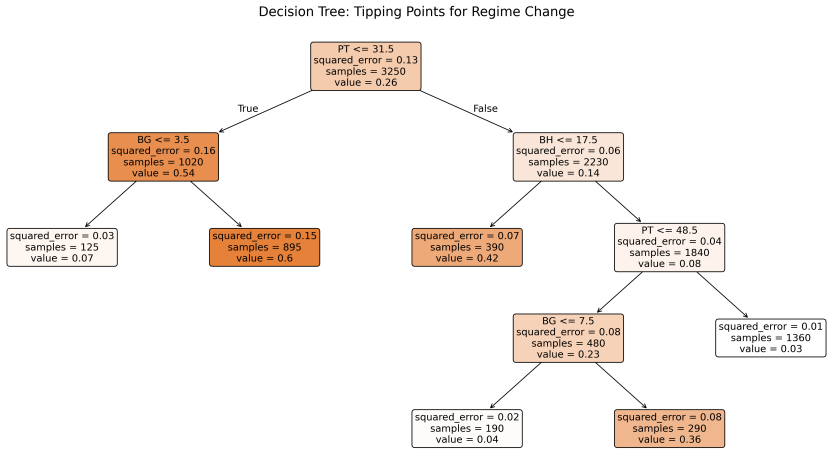


Fig. 3: **Model-Based Screening (Phase 1).** A Decision Tree (CART) trained on the initial dataset. It identifies the primary anthropogenic tipping point: a deterministic shift toward extinction when the Proportion of Hunting Zones (PH) exceeds $\approx 31\%$.

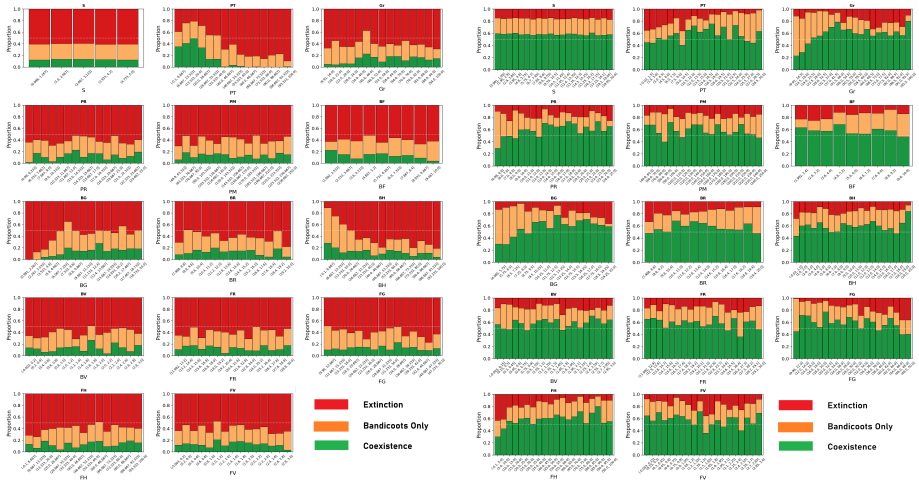
4.3 Data-based analysis

Refinement of the Search Space and Focused Experimental Design. To balance computational feasibility with statistical rigor, we determined the required number of replications, or number of seeds (n), using a sequential procedure based on the asymptotic normality of the sample mean. Relying on the central limit theorem, we assume that for a sufficiently large number of replications, the sampling distribution of the mean stability score approximates a Gaussian distribution, regardless of the underlying population distribution [12]. We targeted a confidence level of 95% ($Z_{\alpha/2} \approx 1.96$) with a desired margin of error of $\epsilon = \pm 0.1$. The optimal sample size n^* for a given parameter configuration is estimated as $n^* = \left(\frac{Z_{\alpha/2} \cdot \hat{\sigma}}{\epsilon}\right)^2$, where $\hat{\sigma}$ is the sample standard deviation estimated from previous runs. Our analysis of the pilot data revealed that the average required sample size across the input space is $n^* = 20.58$. Consequently, we fixed the experimental design at $n = 20$ replicates. This choice satisfies the statistical requirements for the majority of the parameter space.

Parts of the parameter space yield trivial extinction outcomes. Therefore the subsequent search space exploration will restrict the operational range of PH to $[0, 30]$%, BH to $[0, 20]$% and BG to $[5, 20]$, to focus computational resources on the transition zone where complex coexistence or bandicoots survival dynamics occur, especially since the ANOVA indicates that nonlinearities and interactions effects explains most of the variations in the simulation outcomes. Therefore, this first analysis successfully identified the search space regions leading to non-coexistence, and our second analysis can effectively focus on understanding the drivers of coexistence. Figure 4 shows variable-wise distributions for the initial and refined analyses, despite interaction effects playing a role in the comparison. Building on our preliminary findings, we sampled a new LHS of 650 points, each replicated 20 times for a total of 13,000 data points. The new class imbalance is of 16% extinction, 26% prey survival, and 58% coexistence. The theoretical accuracy limit is now 89.6%, implying a minimum of around 10% randomness in the results [8].

To capture nonlinear phase transitions, we train an MLP classifier (2×128 neurons, ReLU) using $\mathcal{L}^2$ regularization to prevent overfitting the irreducible noise [6]. Validated via Stratified Group 10-Fold Cross-Validation, the surrogate achieves 80.1% accuracy. Error analysis indicates that misclassifications primarily occur at the metastable boundary between coexistence and predator extinction. While the surrogate classifier is trained on three discrete states, the subsequent sensitivity and interpretability analyses focus specifically on the predicted probability of the coexistence regime, $P(y = \texttt{coexistence})$. This continuous metric will serve as a proxy for ecosystem resilience, allowing us to quantify how parameters drive the system toward or away from a steady state. Unlike discrete class labels, the probability score captures subtle gradients in stability, revealing regions where the ecosystem is technically stable but highly vulnerable to stochastic collapse versus regions of a more robust stability.

To quantify the computational efficiency gained by this multi-stage workflow, we compared the execution times of the native ABM against the trained surro-



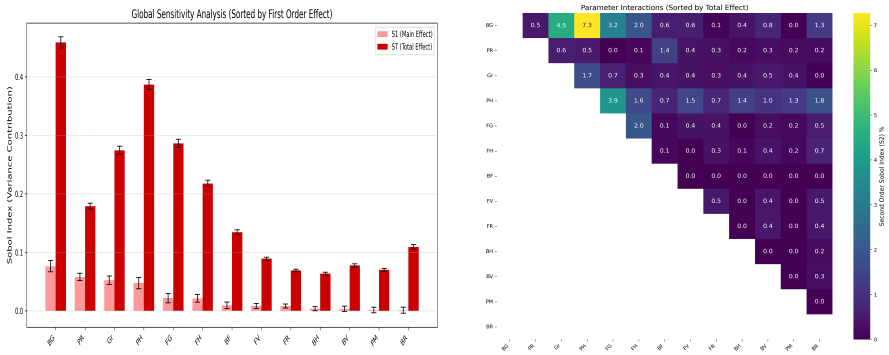
(a) Initial Exploratory Batch ($N = 3,250$) (b) Refined Focused Batch ($N = 13,000$)

Fig. 4: **Global Regime Dynamics.** Comparison of simulation outcome distributions between the initial broad exploration and the refined sampling strategy.

gates. While a single NetLogo simulation takes roughly 20 seconds natively, a single surrogate inference requires only 2 microseconds—a speedup factor of approximately 10 million (or 1 million with parallelization). Exploiting the trained surrogate models allowed us to evaluate approximately 3 million extra configurations. This exhaustive exploration of the behavioral space would be computationally prohibitive, relying solely on the underlying simulator.

Leveraging Sobol' indices for nonlinear global sensitivity analysis. To quantify parameter influence, we contrast a linear ANOVA on ordinal regimes, performed on the new dataset, with a variance-based Sobol' analysis derived from an MLP surrogate (Tab. 2). Note that while the specific selection of surrogate architectures and GSA methods is not the primary focus of this study, the rationale and workflow for such choices are detailed in our previous work [18]. Sobol' indices decompose output variance to capture nonlinearities: the First-Order Index (S_1) measures a parameter's pure contribution, while the Total-Order Index (S_T) accounts for its full effect, including all interactions. The two methods provide sharply different views of the input–output relationship: the linear ANOVA performs poorly (residual variance 87%), indicating that non-additive interactions and threshold effects beyond the reach of linear statistics largely govern system dynamics.

The comparison reveals a sharp disparity between linear and nonlinear sensitivity assessments. The linear ANOVA model exhibits poor explanatory power (residual variance $> 85\%$), suggesting a system dominated by top-down control



(a) Sobol' Indices (S_1 vs S_T) (b) Interaction Heatmap

Fig. 5: **Global Sensitivity Analysis (Phase 2).** (a) Comparison of First-Order and Total-Order indices reveals that variance is driven by interactions rather than single parameters. (b) The heatmap highlights the strong coupling between Bandicoot Energy Gain (BG) and Grassland Availability (Gr).

where fox-related traits are the primary drivers. Under this linear assumption, only the direct effects of predation emerge above the noise, while all other parameters appear statistically insignificant ($< 1\%$ explained variance).

In contrast, the Sobol' indices, given by Fig. 5, identify bandicoot-related parameters and environmental constraints as the dominant drivers. While predation is the proximal cause of mortality, coexistence is regulated bottom-up: stability depends on the prey's metabolic capacity and resource availability to buffer against stochastic extinction. The system nonlinearity is further characterized by the important gap between First-Order (S_1) and Total Order (S_T) effects. The low sum of first-order indices indicates that 69% of the variance arises from higher-order parameter interactions, confirming that regime shifts are driven by complex synergies rather than isolated parameter effects. Every variable plays a role when considering interactions; S_T varies from 46% to 6.4%. Note that the second-order Sobol' indices account for 55% of the variance, and almost exclusively in interaction with the top 6 variables given in Tab. 2. Therefore, only 17% of the variance comes from the interactions between 3 variables or more. Notably, the amount of food bandicoots eat (BF) is the only exception; it does not interact at order 2 but interacts strongly at high orders. This variable looks insignificant at orders 1 and 2, but plays an important role as a connecting factor between grassland-related variables and bandicoot-related variables.

Leveraging PDP and ICE for Threshold Identification. To interpret the nonlinear effects highlighted by the Sobol' analysis, we computed Partial Dependence Plots (PDPs) and Individual Conditional Expectation (ICE) curves from the MLP surrogate's predicted coexistence probability: PDPs show the marginal mean response, while ICE curves display individual *ceteris paribus* deviations [5].

Table 2: Comparison of Sensitivity Metrics: ANOVA vs. Sobol' indices (S_1, S_T).

Parameter	**ANOVA (%)**	**Sobol' S_1**	**Sobol' S_T**
Bandicoot Energy Gain (BG)	0.09	**7.6**	**45.9**
Prop. Hunting (PH)	3.00	4.7	38.7
Fox Energy Gain (FG)	2.86	2.2	28.6
Grassland % (Gr)	0.04	5.2	27.4
Fox Hunting (FH)	**4.11**	2.1	21.8
Plant Growth Rate (PR)	0.85	5.8	17.8
Model Fit	$R^2 = 12.9\%$	Surrogate Accuracy = 80.1%	

Those curves are plotted in Fig. 6 that displays the partial dependence as a function of BG values to illustrate the existence of interaction effects. We then quantify uncertainty, using conformal Prediction implemented through a dual Random Forest scheme [19]. This approach disentangles the epistemic component $\sigma_{\text{epistemic}}$ (model uncertainty) from the aleatoric component $\sigma_{\text{aleatoric}}$ (intrinsic stochasticity). The components were combined, based on the additive nature of assumed independent variances, as $\sigma_{\text{total}} = \sqrt{\sigma^2_{\text{aleatoric}} + \sigma^2_{\text{epistemic}}}$ [8]. Consequently, high values of σ_{total} identify critical phase-transition zones. Those uncertainty curves are plotted in Fig. 7, which displays the estimated uncertainty as a function of BG values to illustrate the existence of interaction effects.

We focus on the dynamics for the six most important variables given in Tab. 2, revealing the following critical thresholds and nonlinear interactions.

• *Metabolic Thresholds* (Bandicoots and Foxes energy gains BG, FG): BG exerts a critical influence at low values. We observe a stabilization threshold between 10E and 15E; beyond this range, both the system dynamics and the associated uncertainty plateau. FG exhibits minimal interaction with BG. It shows a non-monotonic effect on coexistence, peaking at $FG \approx 25$E before slowly declining. While overall uncertainty is stable, extreme values increase stochasticity. Specifically, the combination of low FG and low BG creates a highly unstable system, whereas excessive FG leads to over-predation, driving the system directly from coexistence to total extinction.

• *Anthropogenic Paradox* (Probability of hunting and foxes quotas PH, FH): PH promotes stability in robust metabolic regimes (high BG), stabilizing the system after a threshold of 10–15. However, under low metabolic conditions (low BG), increasing PH reduces coexistence probability. This suggests that hunting is viable only in resilient ecosystems; in fragile environments, it accelerates collapse. A tipping point appears at $PH \approx 15\%$, marked by a sharp drop in uncertainty around $PH \approx 12\%$. Similarly, FH favors coexistence up to a quota of $FH \approx 35\%$. Like PH, predator control is effective only when bandicoots are metabolically efficient (high BG). In these resilient environments, sufficient hunting minimizes uncertainty.

• *Spatial Resource Dynamics* (Amount of Grassland and Plant growth rate Gr, PR): The interaction between Gr and BG explains 7% of the total variance. While low values for both lead to deterministic extinction, high BG accelerates convergence to coexistence as the grassland increases. Conversely, at low BG,

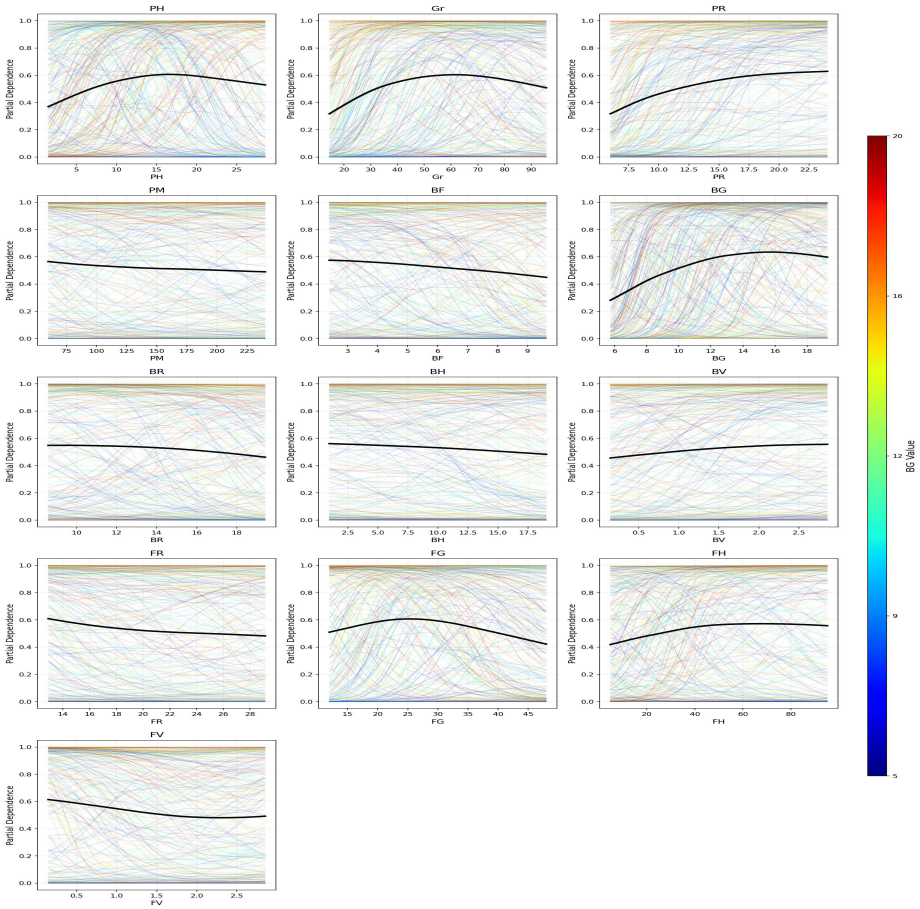


Fig. 6: **Nonlinear Response and Mechanisms.** Partial Dependence Plots (black lines) overlaid with Individual Conditional Expectation curves (colored lines). The coloring by BG reveals a "Metabolic Trap": increasing resources (Gr) only promotes coexistence if metabolic efficiency (BG) is sufficiently high (red).

excessive grassland favors a prey-only regime. The uncertainty landscape reveals a **regime shift** at $Gr \approx 28\%$: below this threshold, low BG results in low uncertainty (deterministic extinction); above it, low BG yields high uncertainty (outcome depends on stochastic fox survival), while high BG yields low uncertainty (robust coexistence). PR operates independently of BG. Higher growth rates correlate with increased stability and coexistence. A sharp drop in uncertainty occurs at a tipping point of $PR = 9\text{E}/\text{t}$; below this, the system becomes highly sensitive to the stochastic spatial distribution of resources.

Quantifying both *Total Uncertainty* and *Partial dependence* is a methodological contribution that allows modelers to precisely identify tipping point regions

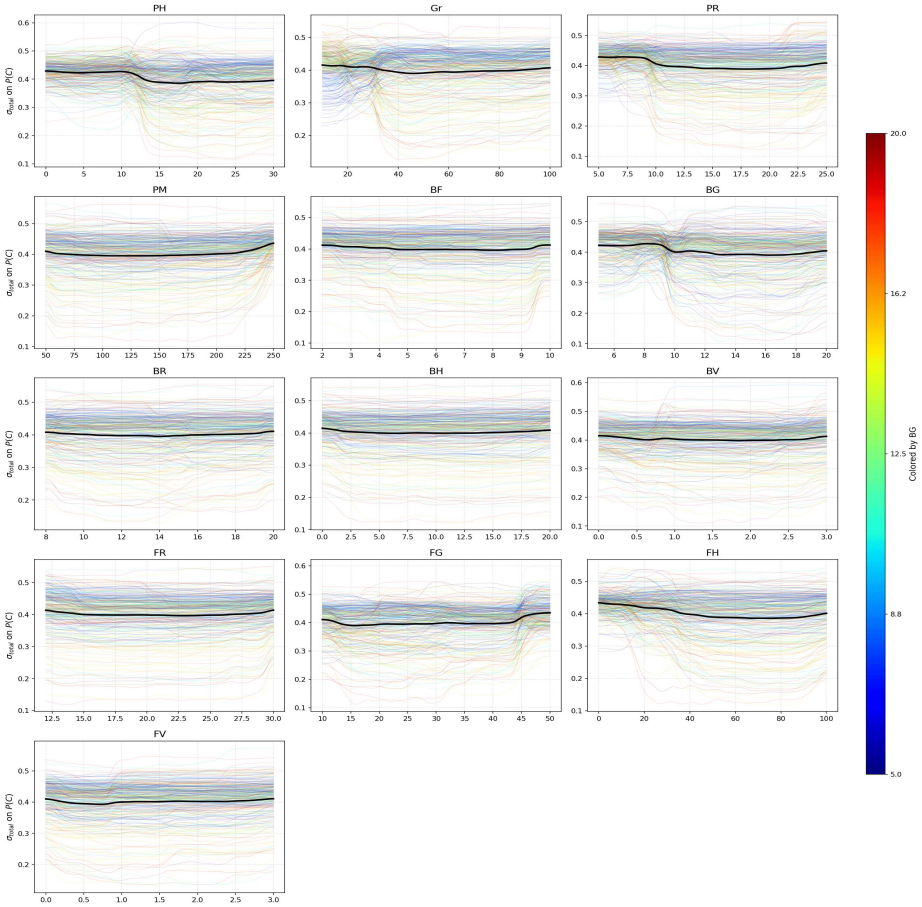


Fig. 7: **Phase Transition Zones.** A map of the Total Euclidean Uncertainty ($\sigma_{total} = \sqrt{\sigma^2_{aleatoric} + \sigma^2_{epistemic}}$) regarding the probability of coexistence. Peaks in uncertainty identify the precise location of tipping points where the ecosystem is structurally unstable.

where the ecosystem is structurally unstable, providing a rigorous framework to distinguish between deterministic regime shifts and irreducible stochastic risks.

5 Conclusion and Perspectives

In this work, we introduced a multi-stage, data-driven pipeline for the automated exploration of stochastic ABMs. By bridging the gap between classical Design of Experiments and Machine Learning surrogates, we addressed the dual challenge of high dimensionality and inherent stochasticity often found in complex simulations. Our methodology, validated on a spatially explicit predator-prey

simulation, demonstrates that linear methods, while useful for initial screening, fail to capture the critical nonlinear metabolic interactions and threshold effects that govern ecosystem resilience. While we used a discrete ordinal metric for end states to simplify our classification boundaries, future applications can easily adapt this pipeline to regress on nuanced, continuous time-series metrics (*e.g.*, fluctuating population counts). Furthermore, our framework specifically retains intrinsic model randomness rather than enforcing fixed seeds, as evaluating this irreducible aleatoric uncertainty is critical for accurately mapping real-world ecosystem vulnerabilities.

Future work focuses on three main axes. First, we will develop an active learning loop, where the uncertainty maps generated in the final stage are used as acquisition functions to iteratively refine the input space exploration with minimal additional simulations. Second, the differentiability of the trained surrogate model opens the way for gradient-based policy optimization, allowing for the automated discovery of robust management strategies in more complex socio-environmental digital twins. Finally, we advocate for a shift from manual, point-based calibration toward an automated, global characterization of the model's behavioral space. We recognize that the choice of a specific GSA method or surrogate architecture (e.g., ANOVA vs. Sobol, Random Forest vs. MLP) remains highly dependent on the model's structure and the problem at hand. Consequently, the ultimate goal is to connect this framework with recommendation tools within an *AutoXAI* pipeline [18]. This will automate the selection of the most appropriate analytical tools for a given problem, removing the bias of manual selection and standardizing the exploration of complex systems.

Acknowledgements

The research presented in this paper has been performed in the framework of the MIMICO research project funded by the Agence Nationale de la Recherche (ANR) n^o ANR-24-CE23-0380. This work was supported by the MUTTEC project (France 2030 – TIRIS, contract 23-AAP-TIRIS-01-062).

References

1. Angione, C., Silverman, E., Yaneske, E.: Using machine learning as a surrogate model for agent-based simulations. PLOS One **17** (2022)
2. Banos, A., Caillou, P., Gaudou, B., Marilleau, N.: Agent-based model exploration. In: Agent-based Spatial Simulation with NetLogo, pp. 125–181. Elsevier (2015)
3. Chen, J.J., Tsai, C.A., Moon, H., Ahn, H., Young, J.J., Chen, C.H.: Decision threshold adjustment in class prediction. SAR and QSAR in Environmental Research **17**(3), 337–352 (2006)
4. De Bosscher, B., Ziabari, S.S.M., Sharpanskykh, A.: Towards a better understanding of agent-based airport terminal operations using surrogate modeling. In: MABS (2023)

5. Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E.: Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. journal of Computational and Graphical Statistics **24**(1), 44–65 (2015)
6. Goodfellow, I.: Deep learning. MIT press (2016)
7. Grimm, V., Railsback, S.F.: Individual-based modeling and ecology. In: Individual-based modeling and ecology. Princeton university press (2013)
8. Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Machine learning **110**(3), 457–506 (2021)
9. Iooss, B., Lemaître, P.: A review on global sensitivity analysis methods. Uncertainty management in simulation-optimization of complex systems: algorithms and applications pp. 101–122 (2015)
10. Lamperti, F., Mandel, A., Napoletano, M., Sapio, A., Roventini, A., Balint, T., Khorenzhenko, I.: Towards agent-based integrated assessment models: examples, challenges, and future developments. Reg. Env. Change **19**(3), 747–762 (2019)
11. Mastio, M., Saves, P., Gaudou, B., Verstaevel, N.: Adaptive agents in spatial double-auction markets: Modeling the emergence of industrial symbiosis. In: Proceedings of AAMAS 2026. vol. 2026, pp. 1–10. IFAAMAS (2026)
12. Montgomery, D.C.: Design and analysis of experiments. John wiley & sons (2017)
13. Novak, M., Wilensky, U.: NetLogo Bug Hunt Predators and Invasive Species (2011)
14. Palar, P.S., Saves, P., Robani, M.D., Verstaevel, N., Garouani, M., Aligon, J., Shimoyama, K., Morlier, J., Gaudou, B.: Interpretable and explainable surrogate modeling for simulations: A state-of-the-art survey and perspectives on explainable ai for decision-making. Archives of Computational Methods in Engineering pp. 1–55 (2026)
15. Railsback, S.F., Grimm, V.: Agent-based and individual-based modeling: a practical introduction. Princeton university press (2019)
16. Reuillon, R., Leclaire, M., Rey-Coyrehourcq, S.: Openmole, a workflow engine specifically tailored for the distributed exploration of simulation models. Future Generation Computer Systems **29**(8), 1981–1990 (2013)
17. Saves, P., Hallé-Hannan, E., Bussemaker, J., Diouane, Y., Bartoli, N.: Modeling hierarchical spaces: A review and unified framework for surrogate-based architecture design. Structural and Multidisciplinary Optimization (2026)
18. Saves, P., Palar, P.S., Robani, M.D., Verstaevel, N., Garouani, M., Aligon, J., Gaudou, B., Shimoyama, K., Morlier, J.: Surrogate modeling and explainable artificial intelligence for complex systems: A workflow for automated simulation exploration. arXiv preprint arXiv:2510.16742;2025 (2025)
19. Shafer, G., Vovk, V.: A tutorial on conformal prediction. Journal of Machine Learning Research **9**(3) (2008)
20. Thiele, J.C., Kurth, W., Grimm, V.: Facilitating parameter estimation and sensitivity analysis of agent-based models: A cookbook using NetLogo and R. Journal of Artificial Societies and Social Simulation **17**(3), 11 (2014)
21. Tumer, K., Ghosh, J.: Estimating the bayes error rate through classifier combining. In: Proc. of ICPR '96. vol. 2, pp. 695–699. IEEE (1996)
22. Van Dyke Parunak, H., Savit, R., Riolo, R.L.: Agent-based modeling vs. equation-based modeling: A case study and users' guide. In: International workshop on multi-agent systems and agent-based simulation. pp. 10–25. Springer (1998)
23. Wilensky, U.: Netlogo: Center for connected learning and computer-based modeling. Northwestern University, Evanston, IL **4952** (1999)