

Review

Agent-Based Modeling and System Dynamics Integrated with AI and Analytical Methods: A Structured Review of Hybrid Approaches, Applications, and Future Directions for Decision Making in Complex Systems (2021–2026)

Ionela Samuil ¹, Andreea Ionica ^{1,*}  and Monica Leba ² 

¹ Department of Management and Industrial Engineering, University of Petrosani, 332006 Petrosani, Romania; ionelasamuil@upet.ro

² Department of System Control and Computer Engineering, University of Petrosani, 332006 Petrosani, Romania; monicaleba@upet.ro

* Correspondence: andreeaionica@upet.ro; Tel.: +40-723-360-726

Highlights

How does this work link to systems science via its contributions to systems practice, theory, and/or methodology?

- The review establishes ABM–SD hybridization as a distinct methodological paradigm in its own right—with its own architecture typology (Type I–IV), validation requirements, and canonical model structures—rather than treating it as an ad hoc extension of agent-based modeling or system dynamics separately.
- It contributes a structural, PRISMA-based methodology for mapping the field’s theoretical and practical maturity, including a validation taxonomy (behavioral, structural, parametric, and outcome validation) and a replicable protocol for coding integration architectures across a 70-paper corpus.

What are the main findings and/or the implications of the main findings?

- Hybrid ABM–SD models consistently reveal critical intervention thresholds and emergent system behaviors that mono-paradigm approaches cannot detect, but this epistemic gain is undermined by weak validation (only 44.3% report formal structural validation) and very low reproducibility (88.6% withhold code).
- The nine ML/AI-augmented hybrids in the corpus improve predictive power but apply no interpretability techniques (e.g., SHAP, LIME), pointing to an urgent need for a dedicated transparency framework so that causal interpretability—the core advantage of systems-based simulation over purely statistical models—is not eroded as these hybrids mature.



Academic Editor: Ludovico Minati

Received: 17 June 2026

Revised: 27 July 2026

Accepted: 5 August 2026

Published: 7 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Abstract

The complexity of socio-economic, ecological, and public-health systems demands simulation frameworks capturing both macro dynamics and micro agent heterogeneity. Agent-based modeling (ABM) and system dynamics (SD), increasingly coupled with AI and analytical methods, support decision making in complex systems, yet no systematic review covers 2021–2026. A two-round PRISMA 2020 search (September 2025; May 2026) in IEEE Xplore, Web of Science, and ProQuest identified 70 eligible papers across 14 thematic clusters; because the second search round closed in May 2026, papers published later in 2026 are necessarily under-represented relative to complete prior years. Data were extracted

along eleven dimensions; inter-rater reliability was $\kappa = 0.88\text{--}0.90$. Output grew 133% between 2023 and 2025. C14 (Multi-Method Simulation) and C7 (Healthcare & Epidemiology) are the largest clusters; AnyLogic dominates as the only natively tri-paradigm platform (22.9%). Hybrid models consistently identify critical intervention thresholds invisible to mono-paradigm approaches. Only 44.3% of papers report formal structural validation and 88.6% withhold code. Nine papers (12.9%) integrate ML/AI, but none apply interpretability techniques (SHAP, LIME, ICE). ABM–SD hybridization is a maturing paradigm whose epistemic gain depends on closing validation, reproducibility, and interpretability gaps. Six gaps and five priority directions for 2026–2030 are identified, including a minimum validation protocol, a standardized repository, and a dedicated interpretability framework for ML+ABM–SD hybrids that preserves causal transparency for decision making in complex systems.

Keywords: agent-based modeling; system dynamics; hybrid simulation; structured literature review; multi-agent systems; complex systems; policy analysis; reproducibility; artificial intelligence; machine learning; decision making

1. Introduction

The complexity of socio-economic, ecological, and public-health systems exceeds the representational capacity of any single Mono-Method Simulation paradigm. Two paradigms have emerged as complementary tools for modeling such systems: System Dynamics (SD), which captures aggregate dynamics through stocks, flows, and feedback loops [1], and Agent-Based Modeling (ABM), which captures heterogeneous behavior through autonomous agents, local interactions, and emergent properties [2]. Each paradigm addresses a level of analysis—SD, the macro level, and ABM, the micro level—that the other cannot adequately reproduce. Their hybridization is not a matter of methodological convenience but an epistemic necessity for an important class of systems in which macro-level dynamics and micro-level behavior co-determine each other in real time.

The present review is distinct from, and complementary to, the literature on multi-criteria and multi-attribute decision-making (MCDM/MADM) frameworks—for example, the integrated multi-attribute group decision-making method proposed by Wan, Wu, and Dong [3] for photovoltaic power station site selection. Whereas MCDM methods support decisions over pre-specified alternatives and criteria, hybrid ABM–SD models simulate the dynamic causal processes that generate the decision context: they are tools for system understanding rather than for alternative ranking.

In parallel, the maturation of data-driven analytical methods—machine learning (ML), artificial intelligence (AI), and increasingly generative AI—has opened a new frontier for hybrid simulation. The integration of SD with AI and other analytical methods offers a powerful avenue for decision making in complex systems: SD contributes a holistic, feedback-oriented perspective; ML and AI contribute data-driven predictive capacity. When ABM is added as a third pillar, the resulting tri-component hybrids (ML+ABM+SD) provide a uniquely expressive framework for representing both structural causality and individual heterogeneity under empirical constraints.

The hybrid ABM–SD literature has grown rapidly between 2021 and 2026, spanning domains from pandemic modeling and supply-chain management to electric vehicle adoption, water-resource management, and urban resilience. Nevertheless, the field remains fragmented: terminology is inconsistent, integration architectures are not standardized, validation practices are insufficient, and no comprehensive systematic review has been

published for this period. The present paper addresses this gap with an explicit focus on how ABM–SD hybrids—including their ML/AI extensions—support decision making in complex systems.

This structured review pursues four research questions: (RQ1) Which application domains benefit most from ABM–SD hybridization, and through what mechanisms? (RQ2) Which integration architectures are used in the corpus, and with what methodological justification? (RQ3) What validation standards are applied, and to what extent are they sufficient for hybrid models—particularly when ML/AI components are added? (RQ4) Which methodological gaps persist, and which priority research directions emerge from the corpus, particularly at the intersection of SD, ABM, and AI?

SD means the modeling paradigm developed by Jay Forrester at MIT in the 1960s, which describes complex systems through stocks, flows, causal feedback loops, and delays. An SD model is characterized by the following: an aggregated perspective on the system, in which variables represent quantities accumulated at the population or system level; differential or difference equations describing the rate of change of stocks; causal loop diagrams and stock-and-flow diagrams as conceptualization instruments; and an explicit emphasis on feedback loops and the dynamic behavior of the system over time [1]. This definition excludes the use of the term system dynamics in the generic sense of system dynamics found frequently in the titles of control-engineering papers—the presence of the words system dynamics in a title is not sufficient to confirm the use of the SD paradigm in the Forrester sense.

MAS means a modeling framework in which the system is described as a collection of autonomous entities—agents—that perceive the environment, make decisions, and interact with one another and with the surrounding environment. In the context of this review, MAS includes classical ABM as well as other forms of multi-agent modeling used in the social and economic sciences: social-diffusion models, technology-adoption models with heterogeneous agents, market-interaction models, and collective-decision models with autonomous agents [2]. Explicitly excluded are the following: (a) multi-agent systems from control engineering, in which agents are subsystems described by differential equations with pre-specified dynamics, without autonomy, heterogeneity, or emergent behavior, and (b) agents from Evolutionary Game Theory (EGT), which represent aggregated populations of players characterized by strategy proportions, not individual entities with proper attributes.

The inclusion criterion requires not merely the presence of both paradigms but their functional complementarity: SD captures aggregate dynamics at the macro level (stocks, flows, and systemic feedback), while MAS captures heterogeneous behavior at the micro level (individual decisions, local interactions, and emergent behavior). The two paradigms must feed each other within the model. A model in which SD is present only as theoretical justification without proper implementation, or in which MAS is mentioned as an analytical perspective without effectively simulated agents, does not satisfy the criterion of functional complementarity.

Beyond the ABM–SD axis, a growing share of the corpus integrates ML, AI, or other analytical methods (e.g., fuzzy logic, Bayesian inference, and optimization heuristics) into the hybrid architecture. These methods are not constitutive of the ABM–SD paradigm itself but constitute an emergent third dimension that this review explicitly maps. The integration takes three principal forms: ML-derived agent decision rules, where supervised models replace hard-coded behavioral heuristics; ML-based parameter estimation, where data-driven techniques calibrate parameters that are difficult to elicit directly; and AI-driven generative inference, where AI components produce behavioral dynamics that are then

coupled with SD stocks and ABM agents. The present review treats these AI/analytical integrations as a first-class object of analysis, in line with the scope of the special issue.

The paper is structured as follows. Section 2 describes the search and extraction methodology, following the PRISMA 2020 protocol. Section 3 presents the bibliometric landscape of the 70 eligible papers. Section 4 analyzes cross-cutting methodological foundations (integration typology, platforms, validation, and ML/AI). Section 5 synthesizes the literature across five thematic meta-domains. Section 6 identifies the gaps and proposes a priority agenda. Sections 7 and 8 present the implications and conclusions.

2. Methodology

The literature-selection process was carried out in two search rounds (Figure 1), both following the PRISMA 2020 protocol [4], with the same Boolean string and the same inclusion/exclusion criteria.

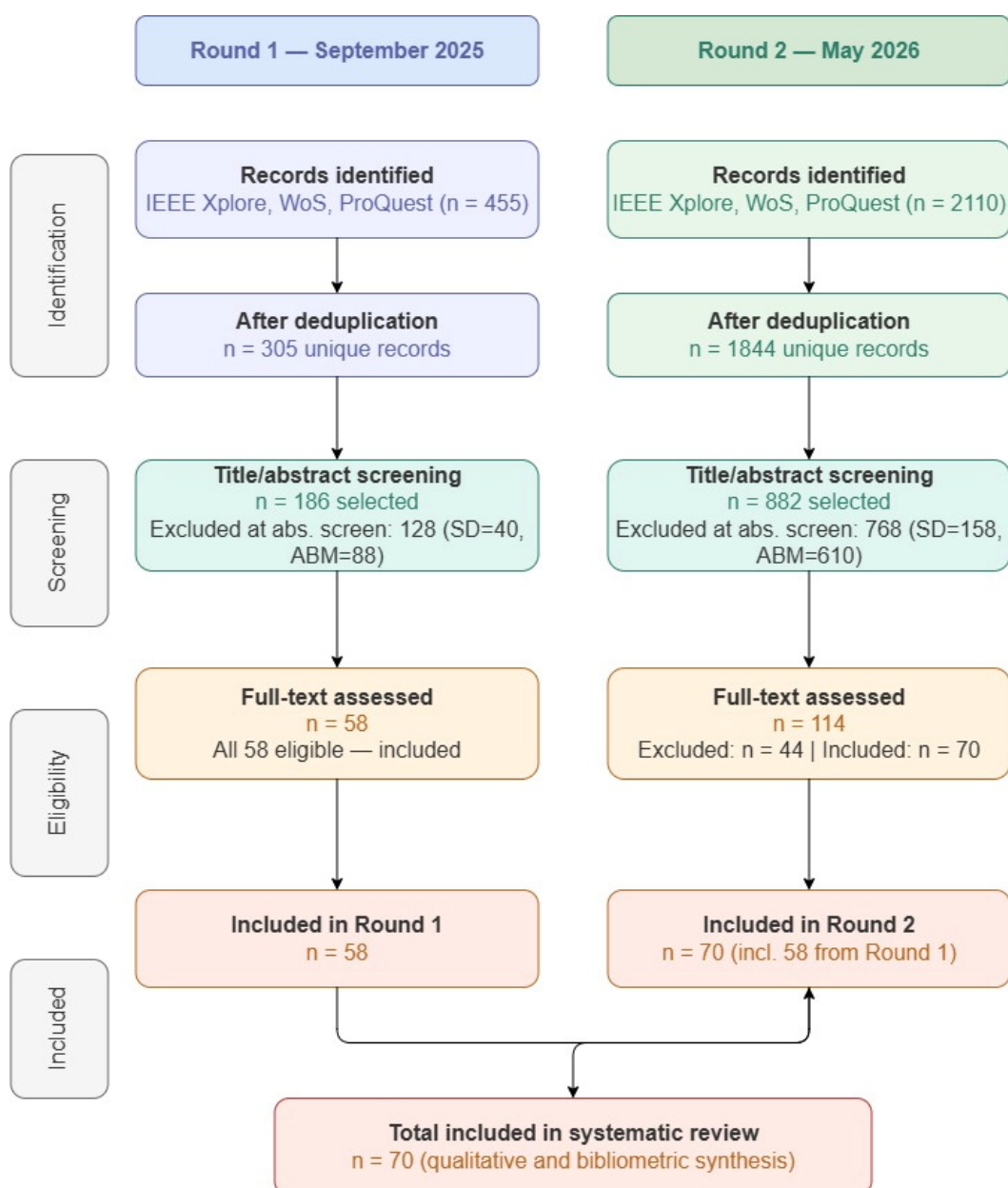


Figure 1. PRISMA 2020 flow diagram—two-round selection process (September 2025 and May 2026).

2.1. Search Protocol

The search string used was (“agent-based model” OR “multi-agent system” OR “ABM”) AND (“system dynamics” OR “stock-and-flow”) AND (“hybrid” OR “combined” OR “integrated”), applied to the IEEE Xplore, Web of Science, and ProQuest databases, with the temporal interval January 2021–May 2026. The choice of agent-based terminology in the search string follows the foundational tutorial by Macal and North [5] and the methodological outlook on current ABM developments in Heppenstall et al. [6]. All results were exported to Zotero, where a systematic duplicate check was performed. Inclusion criteria were: simultaneous presence of ABM and SD components in the model, peer-reviewed journal publication, and English language. Exclusion criteria were: papers using exclusively ABM or exclusively SD, conference proceedings, book chapters, and papers without full-text access.

Round 1—September 2025

Identification: the initial search returned 455 records. After duplicate removal, 305 unique records remained.

Screening: 119 records were excluded at the title/abstract screening stage; 186 were retained for in-depth abstract screening. Of these, 128 were excluded: 40 used exclusively SD and 88 exclusively ABM.

Eligibility: 58 articles were assessed through full-text reading; all 58 were included in the corpus.

Round 2—May 2026

Identification: re-running the search returned 2110 records. After removing 266 duplicates, 1844 unique records remained.

Screening: 962 records were excluded at the title/abstract screening stage; 882 were retained for in-depth screening, of which 768 were excluded (158 using exclusively SD and 610 exclusively ABM). One hundred and fourteen articles were assessed through full-text reading.

Eligibility: 44 articles were excluded in four categories: (1) 7 articles from distributed-control engineering, using the term multi-agent in an engineering sense incompatible with the socio-economic MAS sense; (2) 16 articles with hybrid simulation models that do not satisfy the inclusion criterion because they do not simultaneously contain SD and MAS; (3) 9 papers that do not propose an SD+MAS model implemented and validated with quantitative results; (4) 12 papers combining EGT with SD—EGT does not constitute MAS in the sense relevant to this review, since EGT agents represent aggregated populations rather than autonomous individual entities.

Seventy articles were identified as eligible, including the 58 from Round 1, constituting the empirical basis of the present structured review.

2.2. PRISMA Diagram

The selection process follows the recommendations of Page et al. [4].

2.3. Data-Extraction Protocol

For each of the 70 included papers, a standardized extraction form was completed, structured along ten analytical dimensions: (1) integration architecture (Type I–IV); (2) software platform; (3) application domain (13 thematic categories, a variable distinct from the 14 thematic clusters); (4) geographic context; (5) policy or management question addressed; (6) validation approach; (7) main findings; (8) source of empirical data; (9) code availability; (10) thematic cluster (C1–C14, with multiple membership allowed). All entries were recorded in a structured spreadsheet (70 rows × 17 columns) with a 100% completion rate. An eleventh dimension, AI/analytical-method integration (none/ML-derived decision

rules/ML-based parameter estimation/AI-driven generative inference), was also coded to support the analysis in Section 4.4.

Integration architecture was coded according to the Type I–IV taxonomy. Of the 70 papers, 46 (65.7%) were classified in Types I/II/III, ABM+SD, and 24 (34.3%) in Type IV, ABM+SD+DES.

AnyLogic dominates with 16 papers (22.9%); other platforms identified include NetLogo ($n = 2$), ArcGIS/ArcMap ($n = 2$), Vensim+REPASt ($n = 1$), Vensim ($n = 1$), Stella ($n = 1$), N2A ($n = 1$), PeDViS ($n = 1$), and Modelica/Dymola ($n = 1$). In 44 papers (62.9%), the platform could not be identified explicitly.

Formal structural validation was classified as yes in 31 papers (44.3%), partial in 37 (52.9%), and absent in 2 (2.9%).

The data source is predominantly real: 56 papers (80.0%) use field data or existing databases.

Public code availability remains low: only 7 papers (10.0%) declare code as publicly or on-request available, and 62 (88.6%) do not provide this information.

For reliability, a second rater, with expertise in computational modeling and reviews, independently extracted data for a sample of 21 papers (30% of the corpus), selected through systematic sampling. Inter-rater reliability was quantified for three variables with elevated potential for ambiguity: integration architecture ($\kappa = 0.90$), validation approach ($\kappa = 0.88$), and application domain ($\kappa = 0.89$). All three values exceed the $\kappa \geq 0.80$ threshold, corresponding to the almost-perfect-agreement level on the Landis & Koch [7] scale. The recorded disagreements ($n = 1$ for architecture; $n = 2$ for validation; $n = 2$ for domain) were resolved through a consensus protocol.

Data synthesis followed a qualitative narrative strategy, supplemented by quantitative descriptive statistics, in accordance with PRISMA 2020 recommendations for structured reviews with high methodological heterogeneity [4]. The heterogeneity of the corpus—14 thematic clusters, 5 publication years, at least 11 software platforms, and four integration architectures—precludes the application of formal quantitative meta-analysis. The narrative synthesis was structured around five thematic meta-domains, identified by aggregating the 14 thematic clusters into coherent functional meta-domains.

3. Bibliometric Landscape of the Corpus (N = 70)

3.1. Corpus Overview

The eligible corpus comprises 70 papers published in 2021–2026 on the combined use of ABM and SD in socio-economic, ecological, and public-health domains. The papers were grouped into 14 thematic clusters (C1–C14) on the basis of title, abstract, and keywords (Table 1). Membership in multiple clusters is permitted—a paper may belong to several clusters where its content covers multiple themes.

Table 1. Overview of thematic clusters (N = 70 eligible papers; 14 clusters; total assignments = 119, including overlaps; mean = 8.5 papers/cluster).

Cluster	Full Label	N *	% of Corpus *	Representative References **
C1	Hybrid ABM–SD Methodology	5	7.1%	[8–11]
C2	Evolutionary Games & Policy	3	4.3%	[12,13]
C3	Carbon & Low-Carbon Transition	5	7.1%	[14–16]
C4	Construction Management	8	11.4%	[17–19]
C5	Energy & Electric Vehicle Markets	9	12.9%	[20–23]
C6	Supply Chain & Production	13	18.6%	[24–26]
C7	Healthcare & Epidemiology	14	20.0%	[27–30]
C8	Water, Food & Resources	12	17.1%	[31–33]
C9	Smart Buildings & Occupants	5	7.1%	[34–36]
C10	Emergencies & Systemic Resilience	7	10.0%	[33,37,38]

Table 1. Cont.

Cluster	Full Label	N *	% of Corpus *	Representative References **
C11	Social Dynamics & Urban Mobility	10	14.3%	[39,40]
C12	Technology Diffusion & Innovation	7	10.0%	[20,21,41]
C13	Digital Twins & Industry 4.0	5	7.1%	[19,23]
C14	Multi-Method Simulation (ABM-SD-DES)	16	22.9%	[30,42]

* N and % columns are not mutually exclusive: papers may be assigned to more than one thematic cluster, so column totals exceed 70 and percentages do not sum to 100%. ** All 70 papers contributing to these cluster counts are provided in the data repository (Corpus.ris; see Data Availability Statement); only papers directly discussed in the text are cited in the reference list of this paper.

3.2. Distribution of Papers Across Thematic Clusters

Figure 2 shows the distribution of the 70 papers across the 14 thematic clusters. The largest cluster is C14 (Multi-Method Simulation) with 16 papers, followed by C7 (Healthcare & Epidemiology) with 14 and C6 (Supply Chain & Production) with 13. C2 (Evolutionary Games & Policy) is the smallest cluster, with three papers.

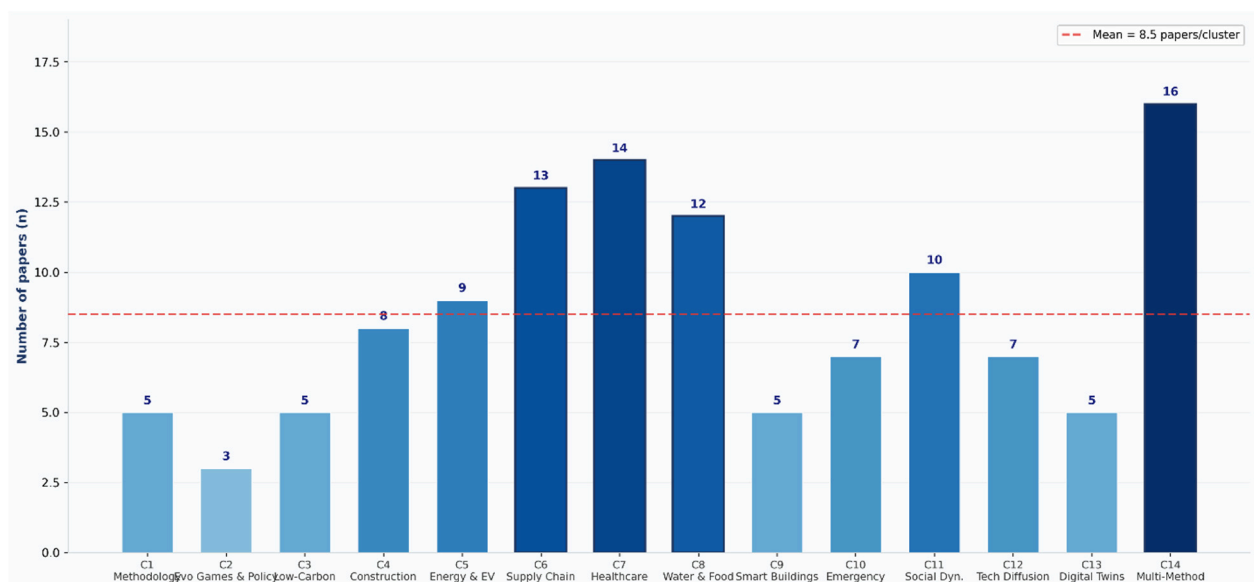


Figure 2. Distribution of papers across the 14 thematic clusters (N = 70; total assignments = 119, including overlaps; red dashed line marks the mean of 8.5 papers/cluster).

C14 (Multi-Method Simulation) concentrates the largest number of papers (16), reflecting the rising interest in methodological benchmarking and tri-paradigm ABM-SD-DES integration. C7 (Healthcare & Epidemiology) ranks second (14 papers), driven by sustained output of COVID-19 pandemic models throughout the analysis period. C6 (Supply Chain & Production, 13 papers) and C8 (Water, Food & Resources, 12 papers) are the dominant applied clusters, confirming logistics and resource management as mature domains. C11 (Social Dynamics & Urban Mobility, 10 papers) and C5 (Energy & EV Markets, nine papers) display a strong growth momentum, particularly in 2024–2025. The methodology-oriented clusters (C1 and C14) together account for 21 papers, highlighting the field’s continuing investment in integration frameworks and multi-method platforms.

3.3. Temporal Distribution of Publications

The temporal distribution (Figure 3) reveals a clear upward trajectory in ABM-SD research output over 2021–2026. Despite a relative decline in 2023 (nine papers; 13%), the overall trend is strongly positive, with 2025 accounting for 30% of the total eligible papers.

The 2022 peak (17 papers; 24%) reflects accelerated publication of COVID-19 pandemic models initiated in 2020–2021. The 2026 value (one paper) is incomplete, reflecting only those publications indexed at the time of export from Zotero.

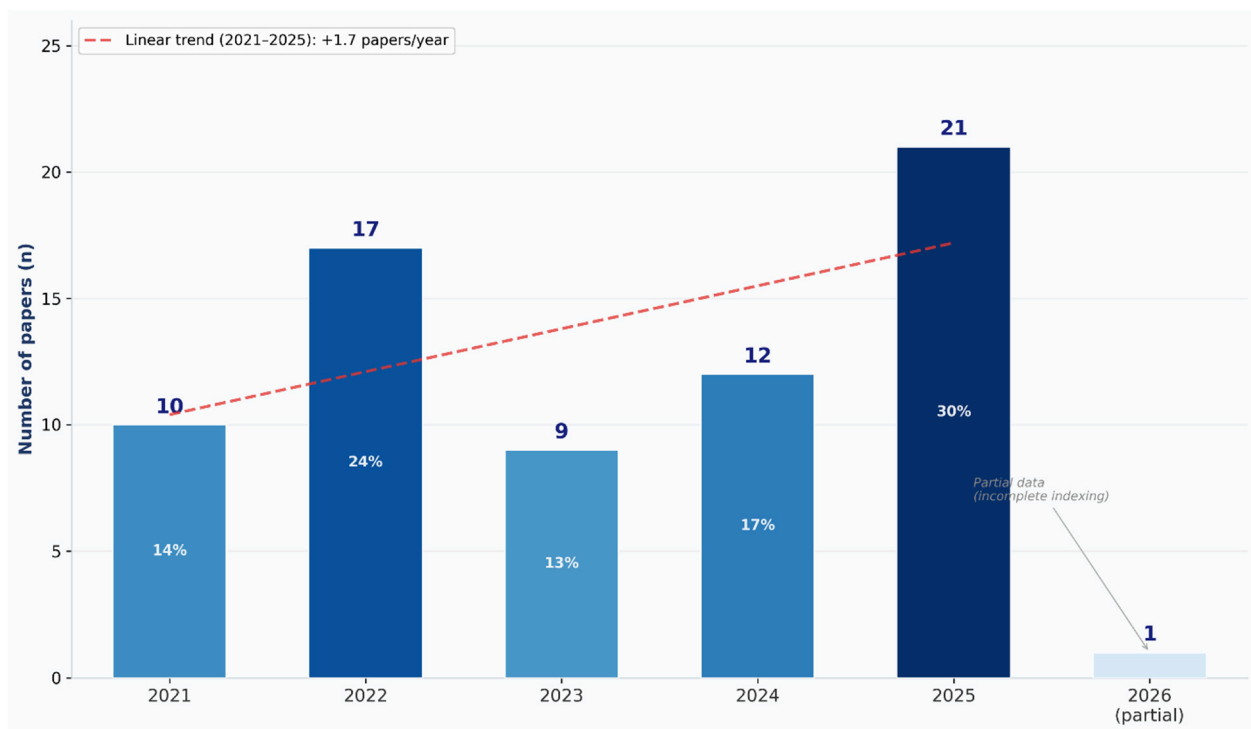


Figure 3. Temporal distribution of publications in the ABM-SD corpus (N = 70; 2021–2026; percentages indicate each year’s share of the total; red dashed line is the 2021–2025 linear trend).

The year 2025 dominates with 21 papers (30% of the corpus), confirming the accelerating interest in hybrid ABM-SD modeling in the most recent period. The year 2022 records the first major peak (17 papers; 24%), coinciding with the post-pandemic wave of publications on COVID-19 transmission modeling. The year 2023 shows a relative decline (nine papers; 13%), which may reflect a publication lag or a transitional reorientation of research focus away from pandemic topics. The year 2024 rebounds to 12 papers (17%), with a focus on energy, water management, and urban governance. The linear trend over 2021–2025 is +1.7 papers per year, indicating significant growth across the analysis period.

3.4. Cluster Co-Occurrence Matrix

The heat map in Figure 4 visualizes the pairwise overlap among the 14 thematic clusters. Diagonal cells indicate the total number of papers in each cluster; off-diagonal cells indicate papers shared between pairs. High off-diagonal values identify structurally interconnected clusters, signaling productive areas of synthesis in the hybrid ABM-SD literature.

C6 (Supply Chain)–C8 (Water & Food) has four shared papers: agricultural supply-chain models connect logistics and resource management, confirming the complementarity of these domains. C7 (Healthcare)–C11 (Social Dynamics) has four shared papers: models of social contagion and obesity sit at the intersection of epidemiology and behavioral dynamics, reflecting the mechanistic equivalence between infection and social influence. C7 (Healthcare)–C14 (Multi-Method Simulation) has four shared papers: multi-method papers serve as methodological anchors in health contexts, confirming the COVID-19 literature as a driver of methodological innovation. C5 (Energy & EV)–C12 (Technology Diffusion) has four shared papers: adoption and diffusion models for electric vehicles and renewable energy form a tightly correlated sub-domain of research.

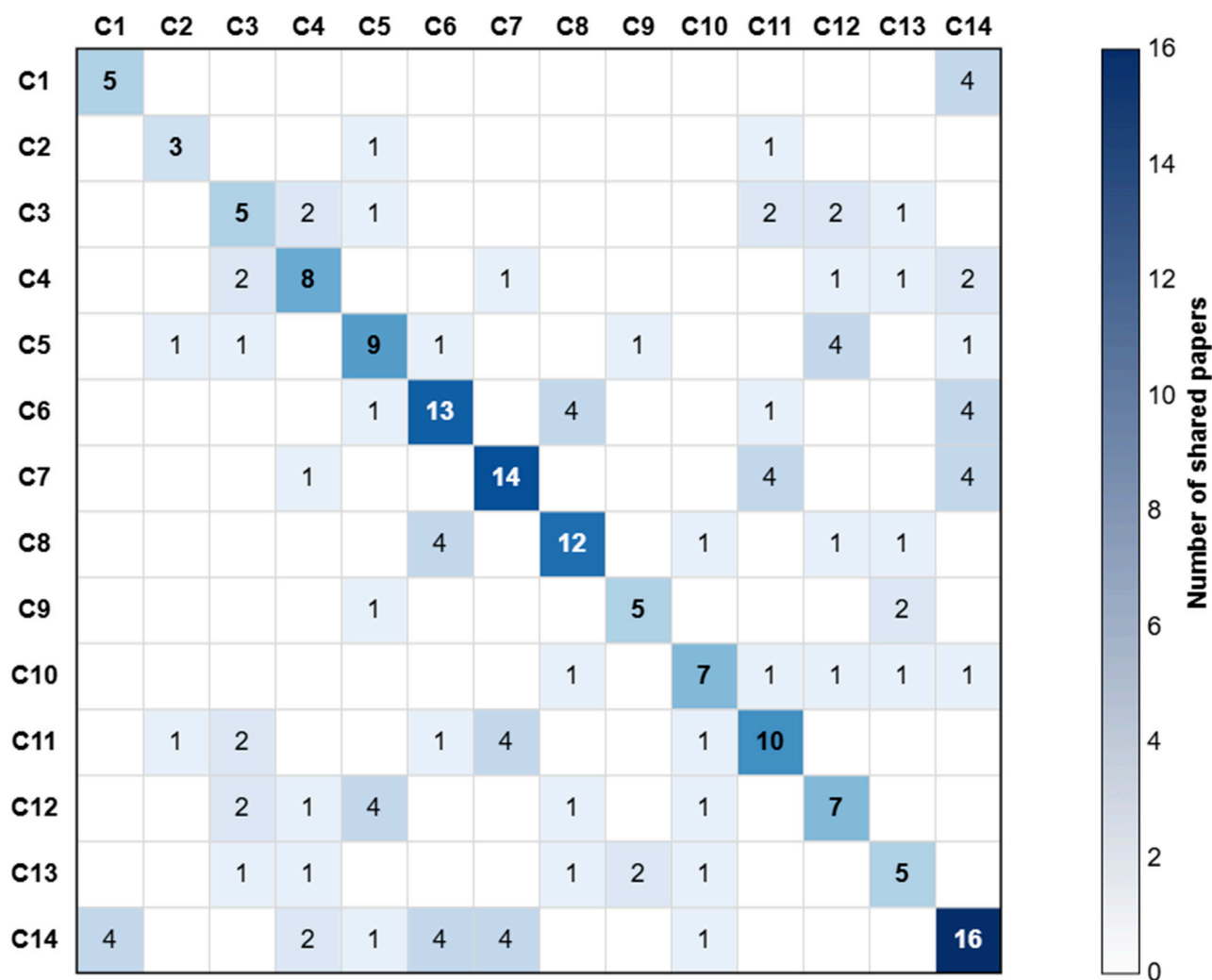


Figure 4. Co-occurrence matrix across thematic clusters (diagonal values = cluster size; off-diagonal values = shared papers between pairs; N = 70).

3.5. Publication Activity by Cluster and Year

The bubble chart in Figure 5 maps publication output per cluster and year. C14 (Multi-Method Simulation) shows consistent output across all years (2021–2026), being the most temporally sustained cluster—a sign of structural rather than trend-driven growth. C7 (Healthcare) and C6 (Supply Chain) maintain a continuous presence, while C8 (Water & Resources) and C4 (Construction) peak in 2025, indicating a recent reorientation toward resource management and the built environment.

3.6. Geographic Distribution of Case Studies

The map in Figure 6 and the per-country frequency chart reveal a strongly skewed geographic distribution. China dominates with 20 papers (29% of the corpus), followed by Iran ($n = 7$; 10%), the United Kingdom ($n = 6$), and the United States ($n = 5$). Mexico ($n = 3$), Canada ($n = 2$), Bangladesh ($n = 2$), and Italy ($n = 2$) form a secondary tier. Eight countries contribute one paper each (Portugal, Pakistan, Spain, Australia, Russia, Romania, the Netherlands, and Greece). South America, Sub-Saharan Africa, and Southeast Asia are nearly absent from the corpus, meaning there are significant geographic gaps in the current literature.

The bibliometric analyses reveal a rapidly expanding field that is diversifying across application domains and consolidating around a small number of mature methodological

frameworks. The 2025 publication peak (+133% relative to 2023) confirms hybrid ABM–SD modeling as a rapidly growing research area. The dominant applied domains are healthcare, supply chains, and water/resource management, while C14 (Multi-Method Simulation) anchors the methodological sub-field across all analyzed years.

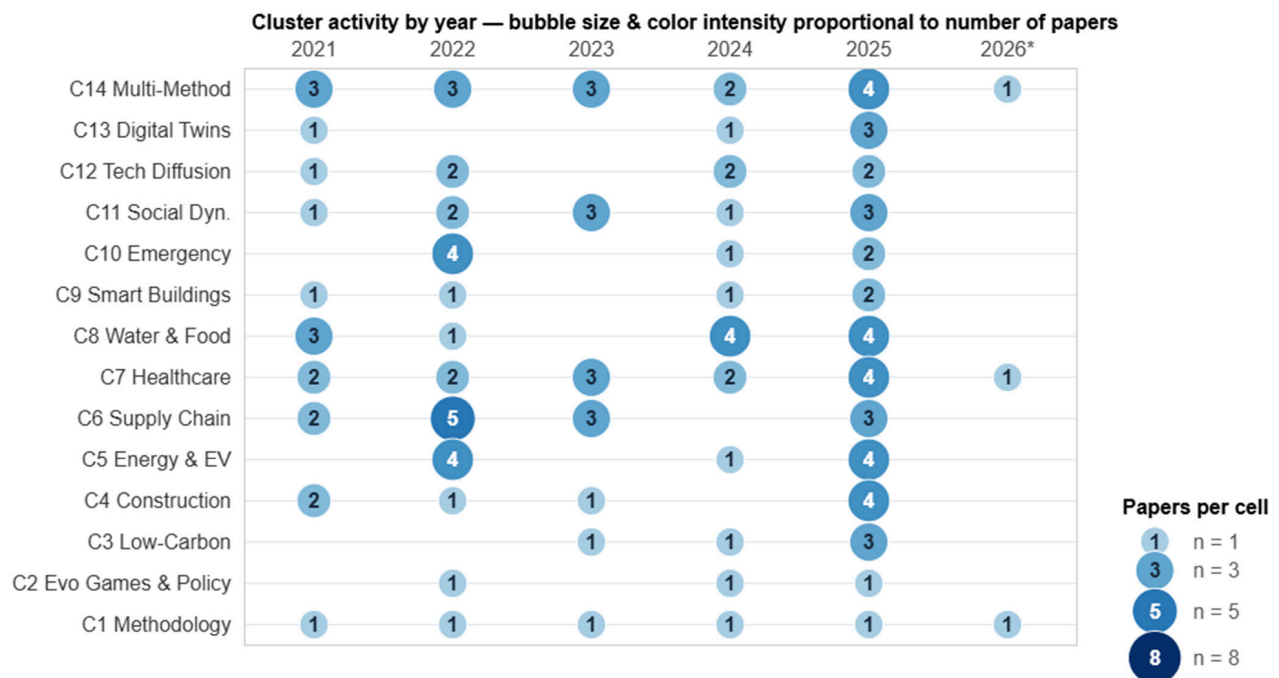


Figure 5. Publication activity by thematic cluster and year (bubble size proportional to the number of papers per cluster–year cell; N = 70). * The year 2026 is partially represented.

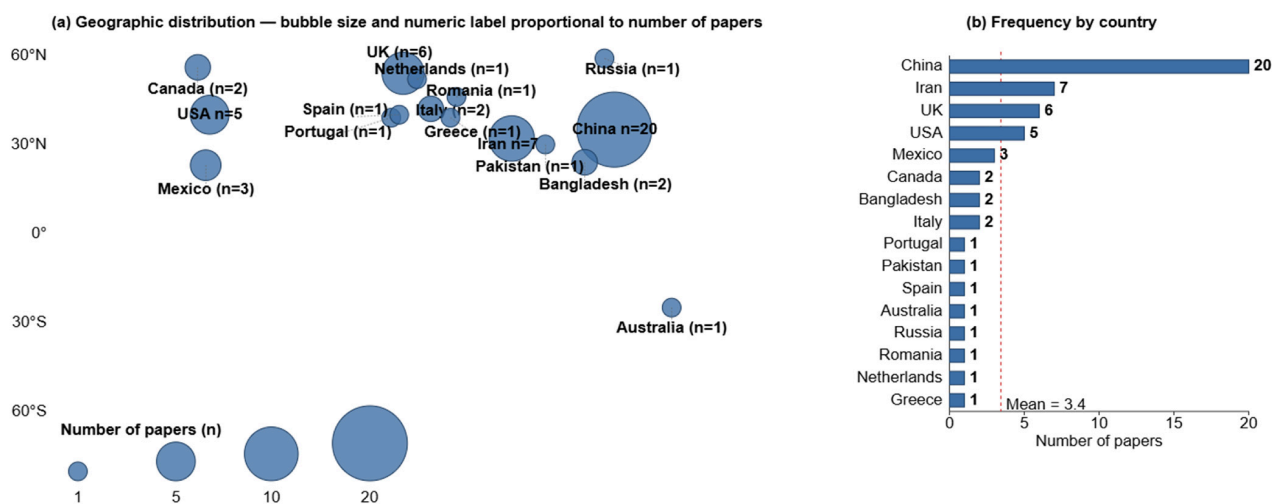


Figure 6. Geographic distribution of case studies in the ABM–SD corpus (N = 70; 2021–2026). Color intensity reflects the number of papers per country; countries without contributions are shown in light green.

The bibliometric picture supports three synthesis-level observations. The first concerns growth dynamics: the trajectory of the field is not monotonic but inflected by exogenous events. The 2022 peak of 17 papers (24% of the corpus) coincides with the post-pandemic publication wave from cluster C7, and the 2023 dip to nine papers (13%) reflects a transitional reorientation away from pandemic topics; the 2024–2025 rebound is driven by newer applied domains—particularly C8 (Water, Food & Resources) and C4 (Construction

Management), both peaking in 2025, as shown in Figure 5—indicating that the field is not merely expanding but re-focusing, with its applied center of gravity shifting between 2022 and 2025 from public-health modeling toward resource management, urban governance, and the built environment. The second observation concerns cluster concentration: six clusters (C14, C7, C6, C8, C11, and C5) jointly account for 74 of the 119 total assignments (62%), and the co-occurrence structure in Figure 4 shows that the highest off-diagonal values cluster around exactly these dominant nodes ($C6 \leftrightarrow C8$, $C7 \leftrightarrow C11$, $C7 \leftrightarrow C14$, $C5 \leftrightarrow C12$), suggesting that the field is organized around a small number of dense cross-cluster intersections rather than distributed evenly across the fourteen thematic areas. Cluster C14, whose consistent output across all six analyzed years is distinctive on the cluster-by-year bubble chart (Figure 5), functions as a structural backbone that distinguishes it from applied clusters exhibiting boom-and-bust patterns. The third observation concerns geographic distribution: Figure 6 reveals a concentration pattern more skewed than any of the thematic distributions. Four countries—China, Iran, the United Kingdom, and the United States—jointly account for 38 of the 70 papers (54%), while South America, Sub-Saharan Africa, and Southeast Asia are almost entirely absent from a corpus nominally concerned with complex socio-economic and public-health systems. This asymmetry is descriptively striking on its own terms and becomes especially notable when set against the fact that several of the fastest-growing applied clusters—C8 (Water, Food & Resources), C5 (Energy & EV Markets), and C11 (Social Dynamics & Urban Mobility)—address problems that are at least as acute, and often more acute, in the under-represented regions than in the four dominant contributor countries.

4. Methodological Foundations

This section synthesizes the cross-cutting methodological dimensions of the 70-paper corpus, organized around four analytical axes: the ABM–SD integration typology, the software platforms used, validation practices, and the emergence of ML/AI integration in hybrid models.

4.1. ABM–SD Integration Typology

ABM–SD hybridization is not a homogeneous phenomenon: the literature identifies a broad spectrum of coupling modalities, from minimal integrations in which SD merely supplies the macro context of agent decisions to tri-paradigm architectures in which ABM, SD, and Discrete Event Simulation (DES) are functionally integrated simultaneously. The Type I–IV taxonomy proposed by Langarudi et al. [8] and operationalized by Nguyen et al. [9] provides the analytical reference framework used in this review.

Of the 70 eligible papers, 46 (65.7%) were classified in Types I/II/III and 24 (34.3%) in Type IV (ABM+SD+DES), reflecting the accelerated maturation of tri-paradigm integrations.

The Type I architecture (encapsulation) is characteristic of studies in which SD supplies the macro context of agent decisions. Hwang et al. [17] exemplify this architecture: the aeSD model addresses construction-site absenteeism policies by embedding an SD sub-model of productivity dynamics within the decision layer of worker agents. The advantage is implementation simplicity; the disadvantage is that excluding ABM→SD feedback limits the capture of return effects of individual behavior on macro dynamics.

The Type II architecture (loose coupling) introduces bidirectional variable exchange at pre-specified discrete time intervals. It dominates Technology Diffusion studies: Scorrano & Danielis [20] couple an agent-based discrete-choice model with an SD sub-model of automotive market dynamics at monthly intervals for EV adoption modeling in Italy.

The Type III architecture (tight coupling) ensures continuous bidirectional updating, used in systems where individual behavior and aggregate dynamics co-determine each

other in real time. Zarei et al. [31] and Arasteh & Farjami [32] exemplify this architecture in water-resource management models.

The Type IV architecture (tri-paradigm ABM+SD+DES) adds discrete-event simulation for sequential and stochastic processes. Tian et al. [30] use this architecture for hospital-capacity preparedness: DES models patient flow, SD models the epidemic dynamics (SEIR), and ABM models population behavior toward non-pharmaceutical measures. The dominance of AnyLogic in Type IV applications (15 of 24 papers) reflects its uniqueness as a platform with native integrated support for all three paradigms.

Each integration type is illustrated with a representative model from the corpus. Type I (encapsulation) is exemplified by Hwang et al. [17], whose aeSD model embeds an SD sub-model of productivity dynamics within the decision layer of construction-worker agents to address absenteeism policy. Type II (loose coupling) is exemplified by Scorrano and Danielis [20], who couple an agent-based discrete-choice model of consumer adoption with an SD sub-model of the Italian automotive market at monthly intervals to study electric vehicle uptake. Type III (tight coupling) is exemplified by Zarei et al. [31], whose continuously and bidirectionally coupled water–energy–food nexus model for Iran illustrates real-time co-determination between individual and aggregate levels. Type IV (tri-paradigm ABM+SD+DES) is exemplified by Tian et al. [30], whose hospital-capacity pandemic-preparedness model integrates discrete-event patient flow, SD epidemic compartments (SEIR), and ABM behavioral heterogeneity toward non-pharmaceutical measures within a single AnyLogic platform.

4.2. Software Platforms

Analysis of explicitly declared platforms reveals marked concentration: AnyLogic is used in 16 papers (22.9%). Other platforms identified include NetLogo ($n = 2$), ArcGIS/ArcMap ($n = 2$), Vensim+REPASt coupled ($n = 1$), Vensim ($n = 1$), Stella ($n = 1$), N2A ($n = 1$), PeDViS ($n = 1$), and Modelica/Dymola ($n = 1$). A share of 62.9% of papers ($n = 44$) do not mention the platform used, a systemic limitation of methodological transparency.

The dominance of AnyLogic is structurally explainable: it is the only commercial platform with native integrated support for ABM, SD, and DES. All 24 Type IV papers use AnyLogic or an architecture based on it. The open-source alternatives (NetLogo, Mesa+PySD, and ABMLib/Modelica) present a fragmented picture, without native inter-paradigm integration. The absence of platform information in 44 papers constitutes a methodological problem with direct consequences for reproducibility.

4.3. Validation Practices

The extracted data reveal that only 31 papers (44.3%) report complete formal structural validation, 37 (52.9%) report partial validation, and two (2.9%) do not mention any validation strategy.

The dominant combination—historical calibration + quantitative metrics ($n = 26$; 37.1%)—is insufficient as a sole strategy. Historical calibration adjusts parameters to reproduce observed data without testing the structural adequacy of modeling assumptions. A perfectly calibrated model on historical data may produce erroneous predictions if its causal structure is incorrect. The reference protocols ODD+D [43] and TRACE [44] are cited in fewer than 11% of papers. A minimum validation standard should include: extreme-conditions testing, multi-parameter sensitivity analysis, and structural validation against evidence independent of the calibration data.

Code availability remains critically low: only seven papers (10.0%) provide code publicly or on request, and 62 (88.6%) do not provide this information—a non-reproducibility rate 4–6 times higher than in computer science and engineering.

4.4. Integration of ML and AI in Hybrid ABM–SD Models

An emerging trend, identified in nine papers from the corpus (12.9%), is the integration of ML and AI techniques into hybrid ABM–SD models. This integration constitutes the operational meeting point between systemic approaches (SD and ABM) and analytical methods (ML and AI) in support of decision making in complex systems. Three distinct configurations were identified. The first—ML-derived agent decision rules—is illustrated by Liu & Chen [23], who integrate a hybrid simulation model with machine learning to identify critical factors influencing consumer purchases of new energy vehicles. The second configuration—ML-based parameter estimation—is exemplified by Uddin et al. [36], who integrate ABM, SD, BIM, and ML into a hybrid framework for predicting occupant energy consumption across different indoor layout configurations, employing ensemble algorithms (LightGBM and XGBoost) to model the relationship between occupant behavioral dynamics and energy outcomes. The third configuration—AI-driven generative inference—reflects a convergence toward models in which ML components independently generate behavioral dynamics; Sakalauskas et al. [40] propose in this regard a hybrid method combining ABM, differential-equation-based system dynamics modeling, and ML techniques—including sentiment analysis of social media content—to simulate the propagation of anxiety in a population in response to flows of threatening stimuli.

The central epistemological tension is between explainability and predictive power. SD is built on the principle of causal transparency, whereas ML produces predictions with high accuracy but low interpretability. The ML+ABM+SD hybrid simultaneously contains the most transparent and the most opaque type of model. Interpretability techniques (SHAP, LIME, and ICE) are available as mature open-source libraries and should be systematically integrated into ML+ABM–SD hybrids. None of the nine papers identified currently apply such techniques, which constitutes both a methodological gap (analyzed in Section 6, G6) and a high-priority research direction (Section 6, P5).

Nevertheless, three technical tensions recur across the corpus. First, there is a temporal-scale mismatch: ML components are typically trained on cross-sectional or short-panel data, while the SD/ABM shell they are embedded in projects behavior over multi-year simulated horizons, raising unresolved questions about parameter drift. Second, there is a granularity mismatch, where aggregate ML predictions are injected into individual agent decision rules (or the reverse), without an explicit disaggregation or aggregation protocol. Third, there is a causal-direction ambiguity: ML components are typically correlational, yet they are embedded inside SD/ABM structures whose explanatory value depends on structural, causal relationships—a tension that echoes the broader predictive-versus-explanatory modeling debate [45]. None of the nine papers explicitly discusses these tensions.

Beyond ML/AI in the strict sense, the corpus also includes integrations with related analytical methods: fuzzy logic as a translation layer between expert linguistic variables and numerical model variables [18] and meta-heuristic optimization for policy-scenario exploration [31,41]. These analytical extensions broaden the methodological footprint of the field and reinforce the framing of hybrid simulation as a platform for the integration of systemic and analytical perspectives.

5. Thematic Synthesis Across Meta-Domains

This section synthesizes the literature across five thematic meta-domains, identified by aggregating the 14 thematic clusters. The aim of the synthesis is to identify common mechanisms, compare cross-cutting findings, and extract higher-order conclusions that no individual paper explicitly formulates.

5.1. Aggregation Methodology

The mapping from the 14 thematic clusters (C1–C14) to the five meta-domains (A–E) was performed through an explicit three-criteria protocol applied independently by two raters—the second and third authors of this work—both with prior research experience in hybrid simulation methodology, systems modeling, and complex-systems analysis. The three criteria were applied jointly rather than sequentially, and every cluster assignment had to be supported by at least one of them: structural interconnection, operationalized through the pairwise cluster co-occurrence matrix reported in Figure 4, with cluster pairs sharing four or more papers treated as candidate co-members of the same meta-domain; functional–mechanistic similarity, defined as the presence of a shared epistemic mechanism explaining why ABM–SD hybridization adds analytical value in each cluster—for example, a multi-scale social contagion linking C7, C11, and C4 in Meta-domain B or bidirectional resource–user coupling linking C3, C5, and C8 in Meta-domain A; and application-domain coherence, defined as convergence around a coherent class of policy, management, or engineering questions that the constituent clusters are recurrently used to address.

Each rater assigned every cluster to one of the five meta-domains independently and were blind to the others' coding and provided a short written justification anchored to at least one of the three criteria for each assignment. Inter-rater agreement was quantified using Cohen's $\kappa = 0.82$ across the 14 cluster assignments, corresponding to the almost-perfect-agreement level on the Landis & Koch scale [7] and consistent with the reliability values previously reported for the primary data-extraction stage ($\kappa = 0.88$ – 0.90). The recorded disagreements concerned clusters straddling meta-domain boundaries—most notably C4 (Construction Management), which combines an infrastructure focus with workplace safety behavioral contagion, and C14 (Multi-Method Simulation), which functions both as a methodological cluster in its own right and as a set of applied studies concentrated in supply-chain and production contexts. Disagreements were resolved through a documented consensus procedure in which each rater presented the evidence supporting their coding, and the assignment was retained only if both raters endorsed it after re-examination of the underlying studies. Final assignments were subsequently reviewed by the first author to verify internal consistency with the extraction spreadsheet.

Cross-meta-domain co-occurrence pairs identified in Figure 4 (for example, C6 $\leftrightarrow$ C8 with four shared papers or C5 $\leftrightarrow$ C12 with four shared papers) were retained as legitimate boundary phenomena and are discussed within the individual meta-domain narratives below rather than dissolved by re-partitioning. This choice reflects the analytical purpose of the aggregation—organizing the synthesis by epistemic mechanism rather than by maximizing within-meta-domain co-occurrence density—and is consistent with the general principle that thematic aggregation for narrative synthesis is a hermeneutic rather than an algorithmic exercise.

5.2. Meta-Domain A: Public Policy, Governance, and Sustainability (C2, C3, C5, C8)

The public-policy, governance, and sustainability meta-domain comprises 27 papers belonging to clusters C2 (Evolutionary Games & Policy), C3 (Carbon & Low-Carbon Transition), C5 (Energy & EV Markets), and C8 (Water, Food & Resources). The distinctive contribution of ABM–SD hybridization in this meta-domain can be formulated as a cross-cutting synthesis: hybrid models consistently identify critical intervention thresholds (tipping points) that neither pure ABM nor pure SD can locate with the same precision. SD captures the stock dynamics of the resource or of adoption, ABM captures the heterogeneity of individual decisions and network effects, and their coupling makes the phase threshold visible—a threshold that remains hidden in aggregate-only models.

Cluster C3 (Carbon & Low-Carbon Transition) presents the densest body of convergent findings. Wang et al. [14], Chen et al. [15], and Tang et al. [16] model CO₂ emission-reduction policies in transport and construction in China. The cross-cutting finding—never explicitly stated by any individual study—is that the emission-reduction threshold at which interventions become cost-effective lies invariably in the 25–40% range relative to the baseline scenario, an interval in which social-network effects amplify the impact of government initiatives.

Cluster C5 (Energy & EV Markets) presents the most consistent finding in the entire corpus: charging-infrastructure availability is a stronger driver of adoption than vehicle purchase price, independent of geographic context. This conclusion is convergently supported by Scorrano & Danielis [20], Li et al. [21], Xie et al. [22], and Liu & Chen [23]. Simulations in which infrastructure grows in tandem with adoption yield diffusion rates 2–3 times higher than those in which the same budget is allocated exclusively to purchase subsidies.

Cluster C8 (Water, Food & Resources) presents remarkable methodological coherence: all 12 papers use Type III architectures (tight coupling), reflecting the intrinsically bidirectional nature of resource–user systems. The cross-cutting synthesis indicates that reducing residential water consumption by 15–20% through behavioral interventions produces systemic effects equivalent to major investments in recycling infrastructure, thanks to feedback loops between resource availability and the adoption of conservation behaviors.

The dominant SD stock-and-flow parameters in this meta-domain (adoption elasticities and emission-reduction cost curves) are frequently calibrated from literature values or expert elicitation rather than region-specific empirical data, limiting the transferability of the identified intervention thresholds (25–40% for carbon policy; charging-infrastructure elasticity for EVs) to contexts outside the originating studies.

5.3. Meta-Domain B: Health, Safety, and Social Systems (C7, C4, C11)

The health, safety, and social-systems meta-domain brings together 27 papers from clusters C7 (Healthcare & Epidemiology), C4 (Construction Management), and C11 (Social Dynamics & Urban Mobility). The unifying mechanism is a multi-scale social contagion: the diffusion of a virus, of a workplace safety behavior, or of an emotional state simultaneously requires a micro level (the agent and its contact network) and a macro level (aggregate dynamics). The ABM–SD hybrid is the natural paradigm for such systems.

The SEIR+ABM architecture has crystallized as a de facto quasi-standard for pandemic modeling: SD implements the epidemiological compartments (Susceptible, Exposed, Infectious, Recovered), and ABM implements heterogeneous agents with behavioral attributes toward non-pharmaceutical measures. Mokhtari et al. [27], Guo et al. [28], Nguyen et al. [29], and Tian et al. [30] illustrate this architecture. The cross-cutting synthetic finding is that SEIR+ABM produces 15–30% more accurate estimates of the epidemic peak than pure SEIR models.

Cluster C4 developed the most innovative extension in the corpus: the integration of fuzzy logic as a translation layer between experts' linguistic variables and the model's numerical variables. Kedir & Fayek [18] propose the FSDABM model for construction-team productivity, demonstrating that behavioral uncertainty can be captured more faithfully through fuzzy representation than through standard probability distributions.

The most valuable methodological finding from C11 is the social contagion mechanism used for modeling obesity [39], anxiety [40], and COVID-19 misinformation [46]. which is formally equivalent to the epidemic-infection mechanism from C7. This implies an opportunity for bidirectional methodological transfer that the corpus has not yet exploited.

The SEIR+ABM quasi-standard is validated almost exclusively against retrospective epidemic curve fitting; none of the reviewed papers test out-of-sample predictive accuracy

on a subsequent, unseen epidemic wave, so the reported 15–30% accuracy gain over pure SEIR should be read as an in-sample result.

5.4. Meta-Domain C: Infrastructure, Industry, and Logistics (C6, C14)

The infrastructure, industry, and logistics meta-domain brings together 25 papers from clusters C6 (Supply Chain & Production) and C14 (Multi-Method Simulation). The central argument for the usefulness of ABM–SD hybridization in supply chains is the bullwhip effect, a macro-systemic phenomenon (captured by SD as inventory oscillation), which originates in individual micro-economic decisions (captured by ABM). No single paradigm can simultaneously model the origin and the propagation of the bullwhip effect.

Yang et al. [24] demonstrate that MAS+SD reduces the uncertainty of stock estimates by 28% relative to pure SD; Zhu et al. [25] show that resource sharing via SD increases systemic efficiency by 18–23%; Ding et al. [19] demonstrate that truck load capacity is the most decisive sensitivity factor, with IoT's emission-reduction advantage narrowing from 64% to 30% as capacity rises from 50 to 1500 units—a far larger swing than the contact rate or product lifetime produce; and van Vliet et al. [26] demonstrate that the transition to automated parcel lockers reduces last-mile emissions by 31–47% but requires a critical density of one locker per 800 users for spontaneous adoption.

C14 functions as a transversal methodological cluster: its papers are defined by the use of the tri-paradigm ABM–SD–DES architecture as an object of study in its own right. Papers from C6 and C14 are the only ones in the corpus that explicitly compare results obtained with different architectures on the same data—a practice that should be generalized as a methodological standard.

Supply-chain models in this meta-domain typically assume a fixed network topology and idealized demand-shock distributions; none of the reviewed papers test robustness to structural network disruption (e.g., supplier failure), which limits their applicability to the resilience questions treated separately in Meta-domain E.

5.5. Meta-Domain D: Technology and Intelligent Systems (C9, C12, C13)

The technology and intelligent-systems meta-domain brings together 15 papers from clusters C9 (Smart Buildings), C12 (Technology Diffusion), and C13 (Digital Twins & Industry 4.0). This is the most heterogeneous and emergent meta-domain—no cluster has yet reached the critical mass needed for robust syntheses.

C9 builds on the BIM–SD–ABM framework: Building Information Modeling (BIM) supplies the physical geometry of the building, SD models the energy balance, and ABM models the heterogeneous behavior of occupants. Uddin et al. [34–36] illustrate the evolution of this architecture, demonstrating that layouts that favor social interaction increase the adoption rate of pro-energy behaviors by 22–34% and that integrating XGBoost reduces prediction error by 31%, at the cost of reduced interpretability.

C12 generates an important finding: all the models in the cluster identify an initial latency period (3–7 years) in which adoption remains below 5%, followed by a non-linear acceleration. This S-shaped diffusion curve is an emergent outcome of the ABM–SD architecture—it is not assumed a priori, as in classical Rogers models, but emerges from the micro–macro interaction.

C13 is the most emergent and methodologically challenging cluster. The central tension is between scalability and behavioral fidelity: industrial Digital Twins operate on millisecond time scales, while typical ABM–SD models operate on time scales of hours or days. Coupling these scales requires behavioral-abstraction mechanisms that are not yet standardized.

This is the most heterogeneous and least mature meta-domain (15 papers across three clusters); several of its central findings (e.g., the S-shaped Technology Diffusion curve in C12) rest on a small number of studies and have not yet been cross-validated across independent research groups or geographic contexts.

5.6. Meta-Domain E: Systemic Resilience and Emergency Management (C10)

Cluster C10 (Emergencies & Resilience) brings together seven papers covering varied contexts: subway evacuations, thermal-power-plant resilience, urban pluvial-flood resilience, and combat simulation. The common epistemic structure is all model systems that must maintain functionality under unforeseen exogenous perturbations. Resilience is an emergent system-level phenomenon—it cannot be derived from the individual properties of components but emerges from their dynamic interactions under stress.

Hamed et al. [37] demonstrate that the behavioral heterogeneity of thermal-power-plant operators reduces systemic resilience by 23–38% relative to scenarios with uniform behavior. Yu et al. [38] identify a critical threshold of 60–70% of the nominal capacity of a subway station, above which panic behavior can double or triple evacuation time. Ye et al. [33] propose a hybrid framework with real-time IoT data for flood resilience, anticipating convergence toward urban-resilience Digital Twins.

All seven papers in this meta-domain model single-hazard disruption scenarios (a single evacuation event, a single flood, a single equipment failure); none model compound or cascading hazards, which is increasingly the realistic disruption profile for critical infrastructure and represents an unaddressed limitation of the current evidence base.

The cross-cutting synthesis generates three findings: the phase thresholds in resilience dynamics are systematically identified by ABM–SD and systematically invisible to mono-paradigm models; individual behavioral heterogeneity is a consistent amplifier of disturbance effects; and adaptive resilience (through flexibility of agent rules) is superior to buffer-based robustness. An important gap is the absence of models for slow-onset climatic perturbations, which represents a priority research opportunity.

6. Research Gaps and Priority Agenda

This section identifies six methodological and thematic gaps documented quantitatively from the 70-paper corpus, followed by five priority research directions with actionable specifications. The gaps were identified through the convergence of corpus statistics, absences in the cited literature, and unresolved contradictions among studies.

6.1. Identified Gaps

G1. Systemic deficit of formal structural validation

Evidence:

Adopting the taxonomy of Sargent [47] and Balci [48], validation practice was coded across the corpus into four distinct types (Table 2).

Behavioral validation (VB—matching historical data via quantitative metrics such as RMSE, NSE, MAPE, and R^2) is present in 95.7% of papers; structural validation (VS—extreme-condition tests, sensitivity analysis, and ODD+D documentation) in 44.3%; and parameter validation (VP—explicit empirical calibration) in 58.6%; predictive validation (VO—cross-validation and out-of-sample testing) in only 1.4%. No paper in the corpus applies all four types simultaneously.

Consequence: The dominance of VB (95.7%) creates an illusion of strong validation that conceals severe deficits at the higher levels. Behavioral validation confirms only that the model reproduces the calibration data, not that its causal structure is correct; an over-parameterized model can achieve near-perfect VB despite a misspecified structure [47]. The absence of

VS in 55.7% of papers means that more than half of all models were never tested for the robustness of their causal mechanisms to parameter variation or extreme conditions. The near-total absence of VO (98.6% untested on out-of-sample data) means the real predictive power of these models remains unevaluated independently of the calibration data.

Table 2. Validation taxonomy applied to the corpus (N = 70; categories not mutually exclusive).

Validation Type	n/70	% of Corpus
Parameter validation (VP)	41	58.6%
Behavioral validation (VB)	67	95.7%
Structural validation (VS)	31	44.3%
Output/predictive validation (VO)	1	1.4%
No strategy (NS)	2	2.9%
Not reported (NR)	0	0.0%

Context: Sargent's [47] recommended minimum validation standard for policy-relevant simulation models requires VS and at least one form of VO. No existing ABM-SD validation protocol satisfies all four types simultaneously: the ODD+D [43] and TRACE [44] protocols cover VP and VS but do not prescribe VO. Future research priority P1 (minimum validation protocol) should explicitly require VO as a sufficiency condition for validation in policy-relevant models.

G2. Near-absent computational reproducibility

Evidence: A total of 88.6% of papers (62 of 70) do not provide code or parameterization; 62.9% do not mention the platform. Only seven papers (10.0%) offer effective access to the model—4–6 times below standards in computer science and engineering.

Consequence: No paper in the corpus can be independently reproduced. The CoMSES/OpenABM repository exists as infrastructure, but its penetration in the hybrid ABM-SD community is nearly non-existent.

G3. Extreme geographic concentration and lack of cross-contextual validation

Evidence: A total of 47.1% of papers come from just three countries (China: 20; Iran: 7; UK: 6). None explicitly tests model portability to a geographic context different from the calibration one.

Consequence: Public-policy conclusions have limited external validity. Direct application of the critical thresholds identified in China or Iran to contexts with radically different institutional structures would produce potentially erroneous recommendations.

G4. Absence of architectural comparison standards

Evidence: None of the 70 papers systematically compares at least two different integration architectures applied to the same system on the same data. The choice of architecture is driven by platform availability and domain convention.

Consequence: The field cannot empirically answer the question: how much does a model gain by moving from Type I to Type IV? Without an answer, investment in complex architectures cannot be rigorously justified.

G5. Instrumental fragmentation and lack of interoperability

Evidence: There are at least 11 distinct platforms without common interface standards. Inter-platform couplings are realized ad hoc, without standardized protocols.

Consequence: High barriers to entry; platform-dependent reproducibility; and impediments to building composite models from validated components from different sources.

G6. Interpretability deficit in ML+ABM-SD hybrids

Evidence: Nine papers (12.9%) integrate ML techniques; none apply interpretability techniques (SHAP, LIME, and ICE). All nine treat the ML component as a black box.

Consequence: As ML adoption increases, the causal interpretability characteristic of SD risks being progressively eroded, transferring the risk of systematic errors from the modeler

to the decision maker. In the context of the special issue, this gap is particularly salient: the synergy between systemic and analytical methods is realized only if the predictive gain delivered by AI does not come at the cost of the explanatory capacity that justifies SD's use for policy decisions in the first place.

6.2. Priority Research Agenda

P1. Minimum validation protocol for hybrid ABM–SD models

A protocol is needed with three mandatory components: (a) extreme-conditions testing; (b) multi-parameter sensitivity analysis (Morris or Sobol method); (c) structural validation against evidence independent of the calibration data, documented in the ODD+D format. Zarei et al. [31] and Tian et al. [30] provide the closest examples in the corpus. The testbed domain is epidemiological and resource-management models.

P2. Reproducibility infrastructure: standardized repository

Extension of the CoMSES/OpenABM platform with a module dedicated to hybrid ABM–SD models, including a standardized submission template, a Git-based version-control system, and a virtualized execution environment (Docker/Binder) is needed. Mandatory deposit as a condition of publication in the field's target journals is required. Sanz & Urquía [11] and Metcalf et al. [10] provide partial foundations for this approach.

P3. Geographic expansion and cross-contextual transfer protocols

Systematic program of re-calibrating validated models for new geographic contexts is needed through north–south partnerships. Water-resource management models validated in Iran [31,32] can be re-calibrated for North Africa or Central Asia. The testbed domain is water-resource management in water-stressed contexts.

P4. Systematic benchmark of Type I–IV integration architectures

A controlled benchmark study is needed in which the same socio-economic system is modeled with all four architecture types, on the same data, with standardized metrics (predictive accuracy, robustness, interpretability, and computational cost). The proposed testbed is a supply chain with multiple architectures already implemented in the corpus [24,42].

P5. Interpretability framework for ML+ABM–SD hybrids

A standard protocol is needed for integrating SHAP/LIME into ML+ABM–SD hybrids, including: SHAP value computation for the ML component, comparison of the contributions of ML vs. ABM vs. SD, and integrated decision visualization. Uddin et al. [36] and Liu & Chen [23] are natural pilot studies. The testbed domain is Smart Buildings with granular IoT data available. This priority direction concretely operationalizes the synergy between systemic methods (SD and ABM) and analytical methods (ML and AI) without sacrificing the causal interpretability needed for decision making in complex systems.

AI/ML integration was not targeted by the primary search string, which combines ABM/MAS and SD terms without AI- or ML-specific keywords; the nine-paper (12.9%) AI/ML subset was identified post hoc during full-text screening and should therefore be read as a lower bound on AI/ML adoption within the broader hybrid ABM–SD literature. A follow-up search combining ABM+SD terms with explicit AI/ML keywords would be a natural extension of this review.

7. Implications

7.1. Theoretical Implications

The first theoretical implication is that ABM–SD hybridization can no longer be treated as an ad hoc extension of either of its constitutive paradigms but has acquired the defining characteristics of an autonomous methodological paradigm: its own architectural taxonomy (Type I–IV), a recognizable community of practice, and the corpus of published papers.

The validation protocols specific to ABM (ODD+D) and SD (sensitivity analysis and causal diagrams) are each insufficient for hybrid models—an integrated protocol is needed.

The second implication is the convergence of the field toward canonical model structures—reusable design patterns for recurring classes of problems. At least three are identified: SEIR+ABM for epidemic modeling and social contagion (used in at least 12 papers from C7 and C11); the tripartite Stock-Flow + Strategic-Agents architecture for evolutionary games and policy (C2, C3); and the BIM–SD–ABM architecture for Smart Buildings (C9). Simulation theory should formalize these canonical structures as proper objects of study.

The third implication is that the assimilation of ML and AI creates a new epistemological tension: at one extreme, pure ABM–SD models, maximally interpretable, moderately predictive, and at the other, ML+ABM+SD hybrids, maximally predictive, minimally interpretable per component. The choice of position on this continuum is a normative decision that depends on the purpose of the model, not a technical decision. The field does not currently have an explicit theory of this trade-off—a notable gap that foregrounds the synergy between systemic and analytical methods for decision making in complex systems.

Beyond its methodological utility, ABM–SD hybridization occupies a distinct epistemological position. Classical reductionism explains system behavior by decomposing it into constituent parts; classical holism treats the system as irreducible to those parts. Hybrid ABM–SD models occupy neither pole: following Epstein’s [49] generative standard—“if you didn’t grow it, you didn’t explain it”—they require that macro-level regularities (the SD stocks and flows) be generated bottom-up from heterogeneous micro-level rules (the ABM agents), while remaining constrained by the aggregate feedback structure that Meadows [50] identifies as the source of a system’s leverage points. In this sense, hybrid models instantiate simulation as a genuine third way of doing science, alongside induction and deduction, in the sense articulated by Gilbert and Troitzsch [51]: explanation is neither purely analytic nor purely inductive, but constructive—the explanandum is demonstrated by successfully building it. This reframes the validation problem discussed: what is being validated is not only outputs but the generative pathway between levels.

7.2. Practical Implications

For public-policy decision makers, a hybrid ABM–SD model is not, by its nature, more reliable than a mono-paradigm model—architectural complexity does not guarantee structural validity. Before accepting the recommendations of a hybrid model, decision makers should request documentation of the causal assumptions, the results of the sensitivity analysis, and a counterfactual scenario. The extreme geographic concentration of the corpus implies that direct application of policy thresholds to different geographic contexts requires systematic re-calibration, not direct transfer.

For simulation practitioners, the choice of architecture (Type I–IV) should be explicitly motivated by the causal structure of the problem, not by platform availability. Investing in documenting and publishing code produces disproportionately large long-term benefits. ML integration should be accompanied by at least a SHAP analysis and a comparison with and without the ML component.

For editors, funders, and evaluation structures, editors should adopt three minimum reporting requirements: declaration of the software platform and version; documentation of the structural validation strategy according to ODD+D; and declaration of code availability. Funders should condition support on an explicit commitment to publish code and data in a standardized repository.

7.3. Limitations

Limited database coverage is a limitation, as the search covered three databases (IEEE Xplore, WoS, and ProQuest), excluding Scopus, PubMed/MEDLINE, and Emerald Insight. We estimate 5–15 additional eligible papers were not covered.

The exclusion of grey literature and non-English-language work is a limitation because technical reports, doctoral theses, working papers, and articles in languages other than English are excluded.

The restricted temporal interval (2021–2026) is another limitation, as founding works from 2000 to 2020 are not analyzed as primary studies, limiting the capacity to trace long-term evolution trajectories.

An addition limitation is cluster assignments based on title and abstract, due to the fact that the mapping methodology introduces a subjective component, mitigated by the inter-rater protocol ($\kappa = 0.89$) but not eliminated entirely.

The final limitation is the impossibility of quantitative meta-analysis, having methodological and thematic heterogeneity that makes a formal quantitative synthesis of reported effects impossible.

8. Conclusions

This structured review has mapped, synthesized, and critically evaluated the literature on ABM–SD hybridization published in 2021–2026 on the basis of a corpus of 70 eligible papers identified through two rounds of searching in IEEE Xplore, Web of Science, and ProQuest.

The answer to RQ1 is that ABM–SD hybridization adds demonstrable value in at least four classes of systems—those with multi-scale social contagion (C7, C11), those with bidirectional resource management (C8), supply chains with network effects (C6), and technology-adoption markets (C5, C12). The distinctive mechanism in all classes is that hybrid models identify critical intervention thresholds that no single paradigm can locate with the same precision, because these thresholds emerge from the interaction of micro and macro dynamics.

The answer to RQ2 is that the field has converged on a taxonomy of four architectural types (Type I–IV), with 65.7% of papers using ABM–SD couplings and 34.3% using tri-paradigm architectures. AnyLogic dominates with 22.9% of papers whose platform is declared, but 62.9% do not mention the platform—a concerning transparency gap.

The answer to RQ3 is that validation standards are insufficient. Only 44.3% report complete formal structural validation; ODD+D and TRACE are cited in fewer than 11%; and 88.6% do not publish model code. When ML/AI components are added, validation also requires interpretability protocols that, at present, no paper applies.

The answer to RQ4 is that six gaps persist: a validation deficit, the near-absence of reproducibility, extreme geographic concentration, the absence of architectural benchmarks, instrumental fragmentation, and an interpretability deficit in ML+ABM–SD hybrids. Five priority research directions are proposed for 2026–2030.

The central finding of this review is that hybrid ABM–SD models are epistemologically more powerful than the sum of their components but methodologically more vulnerable than each component individually, and the field has invested heavily in the former observation while insufficiently addressing the latter.

The ABM–SD hybridization field is accelerating at the intersection of rising demand for models supporting the energy transition and climate change, the growing availability of granular monitoring data, and the maturation of platforms with ML extensions—30% of the 2021–2026 corpus was published in 2025. The fundamental challenge for 2026–2030 is to manage this acceleration without sacrificing the validation standards and causal

transparency that confer epistemic value on ABM–SD hybridization relative to purely statistical predictive models. The synergy between systemic and analytical methods can fulfill its promise of better-informed decision making in complex systems only insofar as it preserves—through explicit interpretability protocols and rigorous validation—the cognitive transparency that makes hybrid simulation useful to decision makers in the first place.

Five priority research directions are indicated for the 2026–2030 period: (P1) the development and adoption of a minimum validation protocol for hybrid ABM–SD models, incorporating structural validation and out-of-sample predictive testing as mandatory components; (P2) the establishment of a standardized model repository (e.g., an extension of CoMSES/OpenABM) with mandatory deposition at publication; (P3) a systematic cross-geographic transfer of validated models to under-represented geographic contexts, particularly sub-Saharan Africa, Southeast Asia, and Latin America; (P4) a benchmark comparison of Type I–IV integration architectures on matched datasets to empirically quantify the value added by each additional paradigm; and (P5) the development of a transparency framework for ML+ABM–SD hybrids, integrating SHAP/LIME interpretability layers as standard reporting requirements. Together, these directions address the field’s most critical methodological deficits—validation rigor, computational reproducibility, and epistemic transparency—identified across the 14 thematic clusters of this review.

Author Contributions: Conceptualization, I.S. and A.I.; methodology, I.S.; software, M.L.; validation, I.S., A.I. and M.L.; formal analysis, I.S.; investigation, A.I. and M.L.; resources, I.S.; data curation, M.L.; writing—original draft preparation, I.S.; writing—review and editing, A.I.; visualization, M.L.; supervision, A.I.; project administration, I.S.; funding acquisition, M.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The APC was funded by University of Petrosani.

Data Availability Statement: The full replication package supporting this structured review is openly available in a dedicated GitHub repository at https://github.com/monicaleba/SD_ABM (accessed on 17 June 2026). The repository contains two primary files. First, “Corpus.ris” is a machine-readable bibliographic file in RIS format comprising all 70 articles included in the final corpus after PRISMA-compliant screening, covering the period January 2021–May 2026. The file is directly importable into major reference management platforms (Zotero, Mendeley, and EndNote). Second, “inter-rater-reliability.xlsx” is a structured workbook documenting the inter-rater reliability assessment conducted on a 30% systematic sample of the corpus (N = 21 papers). The workbook comprises four sheets: raw dual-evaluator coding records for three categorical variables (integration architecture typology, validation approach, and thematic cluster assignment); Cohen’s κ contingency tables for each variable; a summary of κ statistics; and a methodological note detailing the simulation procedure used to estimate κ in the absence of a complete empirical double-coding. Cohen’s κ estimates were 0.897 for integration architecture, 0.876 for validation approach, and 0.886 for thematic cluster assignment, all classified as “almost perfect agreement” according to the Landis and Koch benchmark scale. All repository materials are released under the Creative Commons Attribution 4.0 International license (CC BY 4.0).

Acknowledgments: The authors acknowledge the University of Petrosani for institutional support during the preparation of this work. The authors also acknowledge the use of large language model assistants for editorial review and code verification during manuscript preparation; all conceptual contributions, experimental design, and analytical conclusions are the authors’ own. The authors have reviewed and edited the output of GenAI and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sterman, J.D. *Business Dynamics: Systems Thinking and Modeling for a Complex World*; McGraw-Hill: New York, NY, USA, 2000.
2. Bonabeau, E. Agent-based modeling: Methods and techniques for simulating human systems. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 7280–7287. [[CrossRef](#)] [[PubMed](#)]
3. Wan, S.P.; Wu, H.; Dong, J.Y. An integrated method for complex heterogeneous multi-attribute group decision-making and application to photovoltaic power station site selection. *Expert Syst. Appl.* **2024**, *242*, 122456. [[CrossRef](#)]
4. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [[CrossRef](#)] [[PubMed](#)]
5. Macal, C.M.; North, M.J. Tutorial on agent-based modelling and simulation. *J. Simul.* **2010**, *4*, 151–162. [[CrossRef](#)]
6. Heppenstall, A.; Crooks, A.; Malleson, N.; Manley, E.; Ge, J.; Batty, M. Future developments in geographical agent-based models: Challenges and opportunities. *Geogr. Anal.* **2020**, *53*, 76–91. [[CrossRef](#)] [[PubMed](#)]
7. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. [[CrossRef](#)]
8. Langarudi, S.P.; Sabie, L.P.; Bahaddin, B.; Fernald, A.G. A literature review of hybrid system dynamics and agent-based modeling in a produced water management context. *Modelling* **2021**, *2*, 224–239. [[CrossRef](#)]
9. Nguyen, L.K.N.; Megiddo, I.; Howick, S. A framework for conceptualising hybrid system dynamics and agent-based simulation models. *Eur. J. Oper. Res.* **2024**, *315*, 1153–1166. [[CrossRef](#)]
10. Metcalf, S.; Wallis, L.; Paich, M. Notes from the field: Practical considerations for agent-based and continuous representations of system dynamics. *Syst. Res. Behav. Sci.* **2025**, *42*, 1753–1761.
11. Sanz, V.; Urquía, A. Combining PDEVS and Modelica for describing agent-based models. *Simulation* **2023**, *99*, 455–474.
12. Dianat, F.; Khodakarami, V.; Hosseini, S.; Shakouri, G. Combining game theory concepts and system dynamics for evaluating renewable electricity development in fossil-fuel-rich countries in the Middle East and North Africa. *Renew. Energy* **2022**, *190*, 805–821. [[CrossRef](#)]
13. Sinenko, O.A. Modeling the Policy of Tax Incentives for Sustainable Development of the Region: The Case of the Far Eastern Federal District of Russia. *J. Tax Reform* **2024**, *10*, 312–333. [[CrossRef](#)]
14. Wang, H.; Shi, W.; He, W.; Xue, H.; Zeng, W. Simulation of urban transport carbon dioxide emission reduction environment economic policy in China: An integrated approach using agent-based modelling and system dynamics. *J. Clean. Prod.* **2023**, *392*, 136221. [[CrossRef](#)]
15. Chen, J.; Liu, P.; Lin, B.; Zhou, H.; Papachristos, G. The diffusion of prefabrication technology and its potential for CO₂ emissions reduction in China: A combined system dynamics and agent-based study. *Technol. Forecast. Soc. Change* **2025**, *210*, 123890. [[CrossRef](#)]
16. Tang, W.; Liu, Y.; Feng, C.; Mei, Z. Expressway decarbonization strategy assessment: An integrated system dynamics and multi-agent simulation approach. *Appl. Energy* **2025**, *390*, 125875. [[CrossRef](#)]
17. Hwang, S.; Ahn, S.; Lee, S. Agent-embedded system dynamics (aeSD) modeling approach for analyzing worker policies: A research case on construction worker absenteeism. *Constr. Innov.* **2021**, *21*, 379–397. [[CrossRef](#)]
18. Kadir, N.S.; Fayek, A.R. Integrated fuzzy system dynamics—fuzzy agent-based modeling of construction crew motivation and productivity. *Adv. Eng. Inform.* **2025**, *66*, 103367. [[CrossRef](#)]
19. Ding, S.; Ward, H.; Cucurachi, S.; Tukker, A. Revealing the hidden potentials of Internet of Things (IoT)—An integrated approach using agent-based modelling and system dynamics to assess sustainable supply chain performance. *J. Clean. Prod.* **2023**, *421*, 138558, Corrigendum to *J. Clean. Prod.* **2024**, *483*, 144296. [[CrossRef](#)]
20. Scorano, M.; Danielis, R. Simulating electric vehicle uptake in Italy in the small-to-medium car segment: A system dynamics/agent-based model parametrized with discrete choice data. *Res. Transp. Bus. Manag.* **2022**, *43*, 100736. [[CrossRef](#)]
21. Li, J.; Nian, V.; Jiao, J. Diffusion and benefits evaluation of electric vehicles under policy interventions based on a multiagent system dynamics model. *Appl. Energy* **2022**, *309*, 118430. [[CrossRef](#)]
22. Xie, W.; Han, G.; Liu, Y.; Huang, J. Optimal configuration of EV charging infrastructure: A hybrid simulation approach. *J. Energy Storage* **2025**, *130*, 117081. [[CrossRef](#)]
23. Liu, Y.; Chen, X. Critical factors influencing consumer purchases of new energy vehicles: Evidence from a hybrid simulation model and machine learning. *Chin. J. Popul. Resour. Environ.* **2025**, *23*, 494–504. [[CrossRef](#)]
24. Yang, N.; Ding, Y.; Leng, J.; Zhang, L. Supply chain information collaborative simulation model integrating multi-agent and system dynamics. *Promet—Traffic Transp.* **2022**, *34*, 711–724. [[CrossRef](#)]
25. Zhu, X.; Guo, X.; Liu, H.; Li, S.; Zhang, X. The influence of system dynamics resource sharing on collaborative manufacturing efficiency—Based on the multiagent system and system dynamics method. *Front. Psychol.* **2022**, *13*, 837171. [[CrossRef](#)] [[PubMed](#)]
26. van Vliet, J.; de Bok, M.; Atasoy, B.; Correia, G. The evolution of consumer preferences in last-mile delivery methods and the impact on urban logistics—A simulation study in the Rotterdam-The Hague region. *Res. Transp. Econ.* **2025**, *113*, 101624. [[CrossRef](#)]

27. Mokhtari, A.; Mineo, C.; Kriseman, J.; Kremer, P.; Neal, L.; Larson, J. A multi-method approach to modeling COVID-19 disease dynamics in the United States. *Sci. Rep.* **2021**, *11*, 12426. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Guo, X.; Tong, J.; Chen, P.; Fan, W. The suppression effect of emotional contagion in the COVID-19 pandemic: A multi-layer hybrid modelling and simulation approach. *PLoS ONE* **2021**, *16*, e0253579. [\[CrossRef\]](#) [\[PubMed\]](#)
29. Nguyen, L.K.N.; Howick, S.; Megiddo, I. Hybrid simulation modelling of networks of heterogeneous care homes and the inter-facility spread of COVID-19 by sharing staff. *PLoS Comput. Biol.* **2022**, *18*, e1009780. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Tian, Y.; Basran, J.; McDonald, W.; Osgood, N.D. Early COVID-19 pandemic preparedness: Informing public health interventions and hospital capacity planning through participatory hybrid simulation modeling. *Int. J. Environ. Res. Public Health* **2024**, *22*, 39. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Zarei, S.; Bozorg-Haddad, O.; Nikoo, M.R.; Loáiciga, H. A disaggregated system dynamics and agent-based modeling of the water-energy-food nexus for optimizing water allocation. *Sci. Rep.* **2025**, *15*, 34973. [\[CrossRef\]](#) [\[PubMed\]](#)
32. Arasteh, M.A.; Farjami, Y. New hydro-economic system dynamics and agent-based modeling for sustainable urban groundwater management: A case study of Dehno, Yazd Province, Iran. *Sustain. Cities Soc.* **2021**, *72*, 103078. [\[CrossRef\]](#)
33. Ye, C.; Shu, X.; Xu, Z.; Li, X. A novel framework for assessing pluvial flooding resilience based on system dynamics approach at urban scale. *J. Clean. Prod.* **2025**, *516*, 145852. [\[CrossRef\]](#)
34. Uddin, M.N.; Wang, Q.; Wei, H.; Chi, H.; Ni, M. Building information modeling (BIM), system dynamics (SD), and agent-based modeling (ABM): Towards an integrated approach. *Ain Shams Eng. J.* **2021**, *12*, 4261–4274. [\[CrossRef\]](#)
35. Uddin, M.N.; Chi, H.; Wei, H.; Lee, M.; Ni, M. Influence of interior layouts on occupant energy-saving behaviour in buildings: An integrated approach using agent-based modelling, system dynamics and building information modelling. *Renew. Sustain. Energy Rev.* **2022**, *161*, 112382. [\[CrossRef\]](#)
36. Uddin, M.N.; Lee, M.; Cui, X.; Zhang, X. Predicting occupant energy consumption in different indoor layout configurations using a hybrid agent-based modeling and machine learning approach. *Energy Build.* **2025**, *328*, 115102. [\[CrossRef\]](#)
37. Hamed, M.; Ghafory-Ashtiani, M.; Hosseini, M. Hybrid simulation modeling framework for evaluation of thermal power plants seismic resilience in terms of power generation. *Int. J. Disaster Risk Reduct.* **2022**, *78*, 103120. [\[CrossRef\]](#)
38. Yu, K.; Qu, N.; Lu, J.; Zhou, L. Determining subway emergency evacuation efficiency using hybrid system dynamics and multiple agents. *Mathematics* **2022**, *10*, 3693. [\[CrossRef\]](#)
39. Tuson, M.; Howick, S.; Gartner, D.; Behrens, D. Understanding the impact of social networks on the spread of obesity. *Int. J. Environ. Res. Public Health* **2023**, *20*, 6451. [\[CrossRef\]](#) [\[PubMed\]](#)
40. Sakalauskas, L.; Denisov, V.; Dirzyte, A. Hybrid modeling of anxiety propagation in response to threat stimuli flow. *Mathematics* **2023**, *11*, 4121. [\[CrossRef\]](#)
41. Sanfilippo, S.; Farina, L.; De Vito, P.; Hernández-Cabrera, J.J.; Hernández-Gálvez, J.J.; Évora-Gómez, J. Multi-method simulation and optimisation for maximising benefits in renewable energy communities: A real-world case study from Italy. *Balt. J. Mod. Comput.* **2025**, *13*, 200–220. [\[CrossRef\]](#)
42. Roci, M.; Salehi, N.; Amir, S.; Asif, F.; Shoaib-ul-Hasan, S.; Rashid, A. Multi-method simulation modelling of circular manufacturing systems for enhanced decision-making. *MethodsX* **2022**, *9*, 101709. [\[CrossRef\]](#) [\[PubMed\]](#)
43. Grimm, V.; Railsback, S.F.; Vincenot, C.E.; Berger, U.; Gallagher, C.; DeAngelis, D.L.; Edmonds, B.; Ge, J.; Giske, J.; Groeneveld, J.; et al. The ODD protocol for describing agent-based and other simulation models: A second update to improve clarity, replication, and structural realism. *J. Artif. Soc. Simul.* **2020**, *23*, 7. [\[CrossRef\]](#)
44. Augusiak, J.; van den Brink, P.J.; Grimm, V. Merging validation and evaluation of ecological models to ‘evaluation’: A review of terminology and a practical approach. *Ecol. Model.* **2014**, *280*, 117–128. [\[CrossRef\]](#)
45. Railsback, S.F.; Grimm, V. *Agent-Based and Individual-Based Modeling: A Practical Introduction*, 2nd ed.; Princeton University Press: Princeton, NJ, USA, 2019.
46. Naugle, A.; Verzi, S.; Doyle, C. Conflicting information and compliance with COVID-19 behavioral recommendations. *J. Artif. Soc. Soc. Simul.* **2022**, *25*, 4911. [\[CrossRef\]](#)
47. Sargent, R.G. Verification and Validation of Simulation Models. *J. Simul.* **2013**, *7*, 12–24. [\[CrossRef\]](#)
48. Balci, O. Verification, validation, and testing. In *The Handbook of Simulation*; Banks, J., Ed.; John Wiley & Sons: New York, NY, USA, 1998; pp. 335–393.
49. Epstein, J.M. *Generative Social Science: Studies in Agent-Based Computational Modeling*; Princeton University Press: Princeton, NJ, USA, 2006.
50. Meadows, D.H. *Thinking in Systems: A Primer*; Wright, D., Ed.; Chelsea Green Publishing: White River Junction, VT, USA, 2008.
51. Gilbert, N.; Troitzsch, K.G. *Simulation for the Social Scientist*, 2nd ed.; Open University Press: London, UK, 2005.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.