Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Testing Archaeological Theory Using Agentic-ABM: A Case from Nuragic Sardinia

Mark Altaweel

m.altaweel@ucl.ac.uk

University College London https://orcid.org/0000-0003-1807-3737

Lucio Pascarelli

Nurnet APS https://orcid.org/0009-0000-8062-5809

Zeynep Zelal Keles

University College London

Adel Khelifi

Abu Dhabi University https://orcid.org/0000-0002-7844-9452

Research Article

Keywords: archaeological theory, Nuragic Sardinia, nuraghi, agent-based modelling, agentic AI, large language models, architectural energetics, construction labour, cooperation, reproducibility

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11851470/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare potential competing interests as follows: MA is a guest editor of this Special Issue and was not involved in handling this manuscript.

PREPRINT

Testing Archaeological Theory Using Agentic-ABM: A Case from Nuragic Sardinia

Mark Altaweel[1] | Lucio Pascarelli[2] | Zeynep Zelal Keles[3] | Adel Khelifi[4]

[1]Institute of Archaeology, University College London, London, UK | [2]Nurnet APS, Sardinia, Italy | [3]Department of Electronic and Electrical Engineering, University College London, London, UK | [4]Department of Computer Science and Information Technology, Abu Dhabi University, Abu Dhabi, United Arab Emirates

Correspondence: Mark Altaweel (m.altaweel@ucl.ac.uk); Lucio Pascarelli (lucio@pascarelli.com); Zeynep Zelal Keles (zeynep.keles.26@ucl.ac.uk); Adel Khelifi (adel.khelifi@adu.ac.ae)

Funding: Abu Dhabi University.

Keywords: archaeological theory | Nuragic Sardinia | nuraghi | agent-based modelling | agentic AI | large language models | architectural energetics | construction labour | cooperation | reproducibility

ABSTRACT

Archaeology has many competing theories but, at the meso-scale where most interpretation occurs, few sound tests. We advance an agentic agent-based modelling (agentic-ABM) framework in which an AI research agent works with an archaeologist throughout testing. The AI agents stay outside the model, whose simulated agents follow fixed rules. The research agent synthesises literature, formalises published positions as rules, labels every parameter as cited or assumed, executes scenarios, runs sensitivity analyses, and docks the model: two independent AI agents on different large language models re-implement it blind. The archaeologist directs research, approves rules, tests code and evaluates outcomes. The case is a long-standing debate: the labour and cooperation behind Bronze Age Sardinia's 7,167 nuraghi (Nurnet data); 26 of 29 scenarios draw on published work, 13 directly. We develop a protocol for testing theory with agentic-ABM, extending docking to AI agents. Both re-implementations reproduce the labour results in sampled runs of both engines; construction-time differences trace mostly to specification ambiguities, and the second arm falls just short of the time criterion. Docking verifies implementation, not historical validity; results are bounds under stated assumptions. Agentic-ABM turns speculative theory into auditable, replicable, comparative tests of which pasts are more likely, not which happened.

1 | Introduction

Archaeology has no shortage of theory but few ways of testing it. Although a distinct discourse of archaeological theory exists, the discipline lacks a unified theoretical consensus, with questions often settled locally rather than through broader frameworks [1]. Plog [2] adds that the discipline has "a multitude of theoretical approaches, but remarkably few theoretical programmes that outline how these approaches can best be implemented and tested", so that case studies too often read as "self-fulfilling prophecies". Much of the theory archaeologists use is of this local kind: case-based or meso-scale propositions that sit between grand social theory and the empirical record [3, 4], and cases can generate and challenge theory rather than merely illustrate it [5]. Such propositions are usually stated in prose, rest on quantities seldom made explicit, and are rarely tested against one another on equal terms [6].

This paper uses the term "agent" in two senses, and this study involves both: the simulated agents that act within its agent-based model (ABM), and the AI agents that built and tested that model. The simulated agents follow fixed rules, with any randomness controlled by a seed, and in this paper involve no large language model (LLM), although some recent studies do use LLMs to drive simulated agents themselves [7, 8]. ABM makes archaeological propositions testable by expressing them as explicit rules and comparing the consequences those rules produce [9]. Building such models is time-consuming [10], and the choice of rules can reflect an individual researcher's interpretation and assumptions. Coupling ABM with agentic artificial intelligence (AI) could support a more systematic exploration of competing interpretations,

or "potential pasts" [4]. We examine one part of that programme. An agentic system can retrieve the relevant literature, trace each claim to its source, translate competing claims into runnable scenarios, and implement, run and check the resulting simulations, turning propositions long debated in the literature into experiments. Its value depends on whether researchers can inspect, audit and replicate each step. Without that transparency, the workflow risks becoming a black box [11].

We demonstrate such a coupling of agentic AI and ABM (agentic-ABM) on a long-running case: the building of the nuraghi, the Bronze Age stone towers of Sardinia. Nurnet's data [12] list 7,167 of them. Interpretations of who built them range from household self-build through cooperation between communities to mobilisation by chiefs, and rest on rarely explicit labour, transport and time estimates (Section 2). A research agent working under our direction turned the published positions on construction, labour and cooperation, and our own propositions, into separate scenario families, then built, ran, documented and verified a spatially explicit ABM of construction for every nuraghe on the island, while we, as the humans in the loop, tested, checked and debugged the code. Two docking agents on different LLMs then re-implemented the model blind from its written specification. Five AI agents, each a software system built around an LLM, took part on one agentic harness (Appendix S1); none acts inside the simulation.

Little work has applied AI to archaeological inference, specifically to formalising and testing theory, or used LLMs in interpretation. This paper makes three contributions. We show how a workflow of AI agents can make meso-scale theory testable while keeping each rule traceable to a source or labelled as an assumption. New for archaeology, we extend classic model docking to persistent, tool-using AI agents on different LLMs that re-implement a whole island-scale ABM from its specification, though one implementation per LLM cannot show vendor independence. We develop a transferable protocol for testing archaeological theory with agentic-ABM, from cited specification through layered verification and blind docking to sensitivity-based pruning, and release the code with Colab notebooks that reproduce the runs.

2 | Background Literature

2.1 | Testing archaeological theory

Archaeological theory is plural and shows little sign of converging. The debate between processual and post-processual archaeology is, in Hodder's account, now largely historical: many archaeologists "pick and choose" between approaches pragmatically, and he describes archaeology as a historical science that works by "fitting" many kinds of data into a coherent account, not by experiment [13].

Two strands of the theoretical literature point to where rival theories can be compared. The first is middle-range theory. It was proposed to link high-order theories with low-order empirical studies, and its later narrowing in archaeology to site-formation processes has been criticised [3]; forms of it "are in fact common to both" processual and post-processual camps [14]. The second is the method of rival hypotheses. Chamberlin [15] urged investigators to replace the single "ruling theory" with a "family of hypotheses", each developed and tested in parallel, and hypothesis testing can serve any theoretical approach, not only the processual one in which it was formalised [16].

A family of hypotheses about a meso-scale question, such as how a class of monuments was built, involves many interacting quantities: labour, materials, distance, population, time and social organisation. Working out the consequences of each hypothesis, under the same constraints and data, is beyond narrative argument. Many of archaeology's grand challenges will require both sophisticated modelling and large-scale synthesis [17].

2.2 | The Nuragic period

The nuraghi are dry-stone towers that give their name to the Nuragic period of Sardinia, the island's Bronze Age and earliest Iron Age. Their origin is placed in the Middle Bronze Age with the classic forms appearing by about 1600–1500 BCE [18, 19]. The earliest form, the protonuraghe or corridor nuraghe, is a squat building crossed by corridors; the classic nuraghe that followed is a truncated-cone tower roofed by a corbelled (*tholos*) chamber, often with an intramural stair. Many single towers were later enlarged into multi-towered complexes with bastions, outer walls and surrounding villages [20]. Nuragic society flourished roughly between 1350 and 1100 BCE, when complex nuraghi, sanctuaries and sacred wells multiplied [21]. Published estimates count about 6,500 known sites, many badly ruined; the 7,167 used here come from Nurnet's database [12], which merges several inventories. What the towers were for is debated: fortified farmsteads [22], the seats of a hierarchy of cantons, or the monuments of cooperating communities [23].

2.3 | Building the nuraghi

Evidence for how the nuraghi were built remains indirect and debated. At Duos Nuraghes (Borore), the builders quarried the fissured basalt knoll beneath the site along its natural joints, and the quarry scars match the stones of the tower walls [24, 19]. Webster [25] describes the same quarrying and reports a survey of Marghine (Gallin, 1989, cited there) in which only 7 of 113 nuraghi used stone from more than 2 km away. Some large complexes in stone-poor zones are exceptions: S'Urachi was built with basalt brought from outcrops more than 2 km away in a straight line [26], and at other complexes stone was deliberately brought from several kilometres [27]. Most blocks were set with little dressing in roughly coursed cyclopean masonry; fine, squared and isodomic work is concentrated in the upper courses and on complex nuraghi [28], and was arguably present from the first building phases [29].

How blocks were raised is the least literature-constrained step. Proposals include external earth ramps, counterweights, wooden ramps set in putlog holes, winches, and Laner's reading of the intramural stair as the builders' own ramp, which makes the tower a "machine of itself"; an external ramp would need about as much material as the tower [30]. Course-by-course construction is the consensus [31, 25, 32].

The social side of construction is open. Webster [25] modelled nuraghe building as household or kin-group work spread over years and concluded that the towers of the Borore group would have taken only about 2.5% of adult male labour time. Lilliu [21] proposed a servile workforce led by specialised builders, though the size and status of the workforce are unknown. Ugas [18] reconstructs a hierarchy of cantons and tribes; Perra counts at least 350 territorial systems island-wide (Perra, 2008, cited in Bernardini [33]) and describes a society in which kinship egalitarianism masked inequality [34]. Araque González [23] reads nuraghi and sanctuaries as monuments of cooperating communities, and survey and spatial studies find little sign of top-down control in settlement location [35, 36, 37].

These positions have proved hard to test. The labour figures in use are few and local (essentially Webster's estimates for Duos and the Borore group [25]). The rates behind them are borrowed from Mesoamerican experiments [38] and have been criticised for Sardinian stone; Turner [39] notes that these are tuff rates applied to basalt, which is about twice as dense. And the step from a per-tower cost to a social claim needs assumptions about population, phasing and cooperation that are usually left implicit.

Architectural energetics converts buildings into labour by combining volumes with task rates [38, 40, 41, 42]. Studies of Mycenaean walls [43, 44], Stonehenge [45] and Neolithic megaliths [46] have used such estimates to ask whether a single community could build a monument; for the megaliths, what one Funnel Beaker community could supply is put at about 30,000–50,000 person-hours. Static energetics, however, gives a total, not a process: it does not say who worked when, how crews formed around heavy blocks, how long a site took with its own people, or how communities interacted.

2.4 | Agent-based models

ABM represents autonomous, heterogeneous agents interacting in an environment, so that system-level patterns emerge from local rules rather than being specified in advance [47, 48]. This is "generative" social science: it shows how individual rules "map up" to macroscopic regularities, and it lets the modeller subject core theories to stress [9]. Archaeologists have used it for over two decades [49, 50, 51, 52]. ABM can test many scenarios at once in a simulated laboratory and generate hypotheses that focus fieldwork, though it is no substitute for fieldwork [53].

Two limits restrict what an ABM can test. The first is truth. A model that reproduces an observed pattern shows that its mechanism is sufficient, not that it operated in the past, and verification (the code does what the model says) is distinct from validation (the model matches its target) [48, 54]. For the deep past, where behaviour cannot be observed and the targets are few and coarse, validation is often thin [4]. The second limit is authorship: archaeological ABMs usually encode behaviour from the modeller's chosen analogy or premise, which can build in bias or miss behaviours, and rival ideas are seldom implemented in the same framework and data, so they cannot be compared on equal terms. Apart from a few classic models, most ABMs have never been replicated by anyone but their developer [55].

ABMs demand more data and more explicit theory than equation-based models, and their use must be justified by the question [56]. Construction labour justifies it: the outcome depends on discrete local events (a block needing more haulers than are free, a stage waiting for stone, a neighbour refusing a request), on heterogeneity between sites and on interaction between communities. ABMs oblige the modeller to state every rule, opening assumptions to test [57, 53], and the Overview, Design concepts and Details (ODD) protocol supports reproduction [58, 59].

2.5 | Agentic tools

Agentic AI addresses the second limit more directly than the first. Autonomous agents built on LLMs are now the subject of a large literature on their construction and applications [60]. In the natural sciences, such systems have designed, planned and carried out experiments [61]. In simulation, much of this work places the LLM inside the simulated agents as the source of their decisions, from generative agents that remember, reflect and plan [7] to whole agent-based models driven this way, whose open problems include alignment with human behaviour and evaluation [8]. For archaeology, such a decision engine has been proposed together with an orchestration layer that generates and runs different ABMs, each testing a different idea about the past, under goals set by the researcher [4]. This study keeps language models outside the simulation and leaves that engine to future work (Section 5.7).

We deploy the research agent as an assistant because it lets us test many rival ideas. Once the literature is retrieved, recording its data, coding, running scenarios, and analysing and documenting results can be delegated, so turning an idea into a tested scenario costs far less. The researcher's effort shifts to which questions matter, which readings of the literature are defensible, which assumptions to accept, and interpreting results.

3 | Methods

3.1 | Agentic architecture

We implement an agentic-ABM framework [4]: a system that builds, runs and checks a conventional ABM under the researchers' direction. The simulated agents follow explicit, seeded rules (Section 3.2), so the model remains a reproducible stochastic ABM.

Research agent. The model was specified, coded, run, documented and checked by a single persistent research agent, built on Vesta, an open-source agentic harness [62], around a general-purpose LLM (Claude Opus 5.5, Anthropic), one isolated container per agent. Every modelling decision was logged verbatim with its timestamp. Its configuration, tools, memory, literature search and standing rules are in Appendix S1.

Human in the loop. The research team took every modelling decision. The research agent proposed a rule or value with its

source or reason, and the team approved, amended or rejected it before it entered the code. Approval was required for every scenario, every rule without a published source, every change of rule, every parameter fixed or dropped after the sensitivity screen, and the docking criteria.

Provenance. Every modelled value or rule is either cited or labelled. A parameter audit gives each parameter of the headline scenario (MR_P10) one of five statuses, from *sourced* (read at a page in the primary text) to *the research agent's design choice, accepted by the authors*; unsupported items read "not supported by the literature: assumption for now" (Appendices S1 and S3).

3.2 | Repast ABM

The model is documented with the updated ODD protocol [59]. An ODD description also makes a model replicable: of 18 models re-implemented from their ODD descriptions, 15 replications succeeded [63]. Our docking re-implementations (Section 3.3) apply the same idea with AI agents as the replicators.

The model has two stages (Fig. 1). A one-off preprocessing step finds, for every nuraghe, the cheapest sledge route from a permitted stone outcrop. The ABM then simulates each nuraghe's construction day by day, 100 times with different random seeds. Two engines share inputs and rules: engine A, which runs the headline, books each community's labour on a chain of tasks within building seasons; engine B, a day-based engine of builders and crews, runs the Webster baseline, the cooperation levels and the literature family (Appendix S1). It is written in Python with Repast4py 1.3.1 [64], numba and numpy. The full ODD is in the code repository and Appendix S1.

3.2.1 | Data

Sites. The 7,167 nuraghi come from the Nurnet GeoPortal. Its GIS was originally built by Nurnet with the GIS group of CRS4 and went online, in beta, in June 2014, as a participatory GIS in which users add and correct records that an expert validates before publication [65, 66]. The present GeoPortal is its successor. The model uses the number of towers (1 to 7, including the keep); 5,902 sites have one tower. Where the record gives no count, a value was assumed for 3,764 sites (52.5%), all single or two-tower sites (Appendix S1).

Stone. Outcrops come from the regional lithological map at 1:25,000 [67]. A site quarries its recorded building stone where one is known (Nurnet's data), and otherwise the nearest outcrop of one of seven stone classes (Appendix S1).

Terrain. Elevation is the Copernicus GLO-30 surface model at 30 m [68]. A directional cost raster of sledge dragging (friction $\mu = 0.3$ in the baseline, rebuilt for other rules; Table 1) and a multi-source Dijkstra [69] search finds each site's cheapest route from any permitted outcrop.

3.2.2 | Building a nuraghe

The simulated agents are builders grouped in households and work parties at each site and, at the coarser level, the community of each nuraghe (Section 3.2.3). Each nuraghe needs 1,350 t of stone per tower, Webster's [25] estimate for a two-storey single tower, in blocks around a median of 450 kg. A block moves only with its full crew, sized by the load one person can drag, so heavy blocks need many hands at once; crews drag loaded sledges along the least-cost route and walk back empty. Builders who are not hauling quarry, dress, lift and set stone, never running ahead of delivered stone. A working day has 8 hours. Engine B simulates crews day by day and counts labour for the whole work party from start to finish, so crew-level events (blocks waiting for a crew, blocks split for small crews, idle builders) shape its results, including the cooperation levels; engine A books each stage's task labour, divided by the work party for duration. In the Webster baseline no crew is idle, so the accountings agree (Fig. 2).

The baseline stage labour follows Webster's [25] Table 1 (500 man-days quarrying, 1,300 dressing, 700 building per tower, with transport simulated explicitly). A literature family of 22 scenarios replaces one step at a time with rates from the comparative literature (steps S1–S8; Appendix S1). Two combined bounds, low (L) and high (H), stack the cheapest and costliest options at every step.

3.2.3 | Communities and help

Relationships and interaction [47] enter at two levels. Within a site, builders interact through crews and shared stage work, and in the cooperation levels (Fig. 2) a share of builders pool their labour. Between sites, each nuraghe is a community agent that builds its towers as a chain of tasks over building seasons of 90 working days, the communal labour norm Webster borrowed from Renfrew [25]; one season stands for one year. Help results from four rules: a community learns whether to ask for help, asks up to three neighbours within 10 km, is accepted with a probability that rises with trust and falls with the helper's backlog and distance, and is helped from the same daily budget as the helper's own tasks, while help given builds the helper's skill along a Wright [70] learning curve. A further family uses pre-existing networks, such as territorial or tribal groups. The behavioural parameters are free and covered by the sensitivity analysis (Appendix S1).

3.2.4 | The most realistic scenario

We defined a combined scenario, most realistic (MR), that takes, at every step, the option best supported by Sardinian evidence (Table 1). Its technique, realistic (R), is a deterministic per-site rule set for the construction steps (Table 1; Appendix S1). Its social rules are those of the trust-and-learning scenario (A3), with help in "join" mode (below), and coastal sites may move stone by sea [71].

Realism ratings. "Best supported" rests on a five-point rating of every scenario (5 = best supported by Sardinian evidence, 1 =

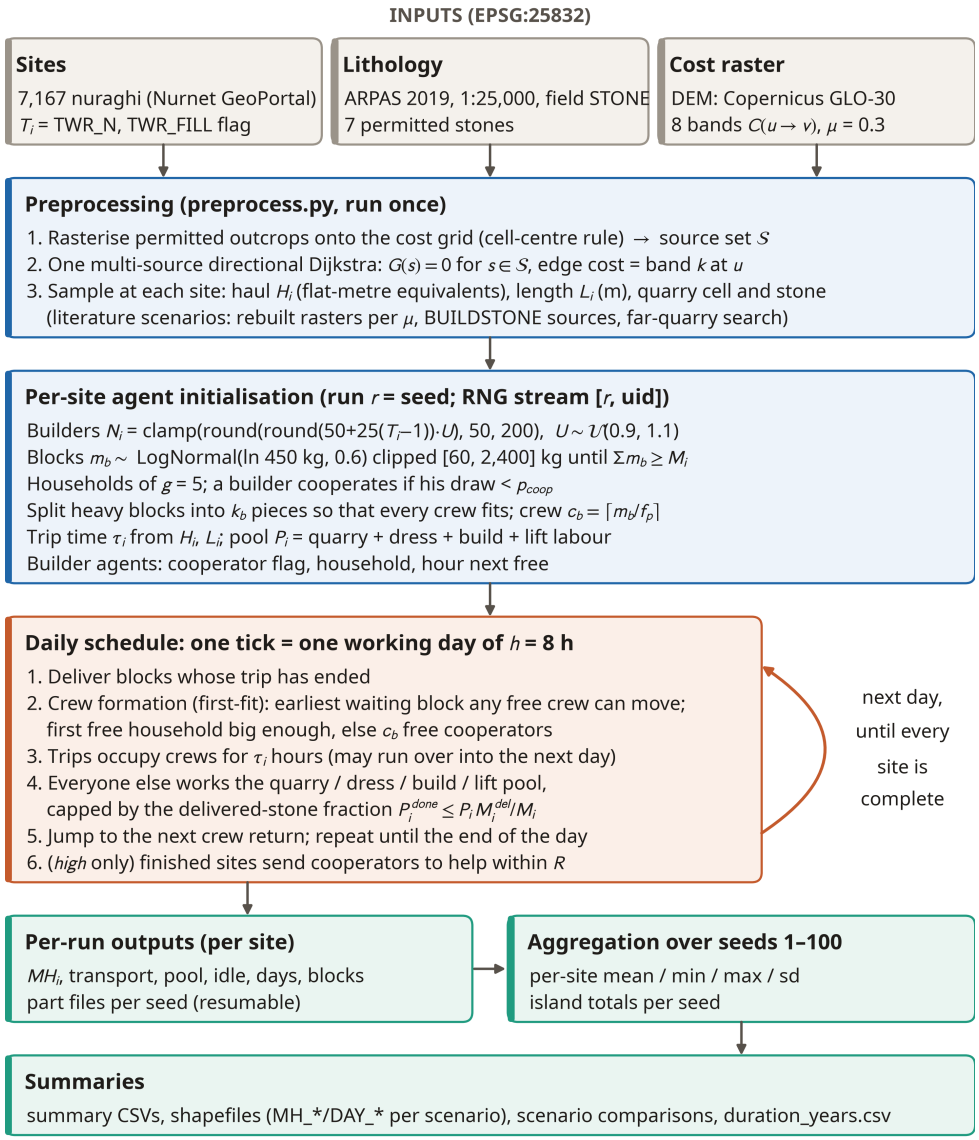


FIGURE 1 | Model workflow. Grey: inputs; blue: preprocessing and per-site initialisation; orange: daily schedule; green: outputs. The schedule shown is engine B (Webster baseline, cooperation levels, literature family) with the default work parties; engine A, which runs the headline, books each community's labour on a chain of tasks per season (Section 3.2). In both engines a block too heavy for the largest crew is split at the quarry into the fewest pieces that fit (one man-hour per cut). Upper-case names (e.g. TWR_N, BUILDSTONE, MH_*) are fields and files of the code repository.

contradicted or implausible). The research agent proposed each rating with a written reason and its sources, and the lead author agreed. The ratings weigh attestation at Sardinian sites, published readings and physical plausibility (Appendix S1). MR takes the highest-rated option at each step, but a combination of individually best-supported options is not thereby shown plausible; otherwise the ratings only label scenarios.

Work parties. Each nuraghe's community is drawn per seed from a resident range for its tower class, anchored on published household, village and cantonal figures [25, 19, 73, 76], and two in five residents work (Webster's one man per five residents, doubled to include women). The mean work party is 9.0 workers for a single tower, 30.0 for two towers, 59.0 for three and up to 146.3 for the largest complexes (Appendix S1). All work parties sum to 115,954 workers, against 418,633 under the default rule used by the other scenario families (50 builders plus 25 per extra tower, randomised; Table 3).

Help: join versus handover. A helper adds its hands to the owner's stage ("join"), up to a physical cap of hands per block. A "handover" variant, in which a helper takes a whole stage over, is run as a comparison, and an "alone" variant with no requests is the baseline for measuring how much a nuraghe needs a network.

Phasing and population. The chronology implies gradual growth, not simultaneous building [74]. Under P10, a random 10% of nuraghi start first, and the next 10% start only when every site in the current cohort is finished; P20 uses 20% cohorts, and P100 starts every site at once. We report P10 as the headline because it is closest to gradual colonisation by fission over centuries (Appendix S1), and give P20 and P100 alongside it. A scenario is a rule set; the variants run under it (P10, P20, P100, factorial points) are its sub-scenarios, and a cell is one sub-scenario run with 100

TABLE 1 | Components of MR (FREE and ASSUMPTION flags as in the configuration; S1–S8, steps of the literature family, see Abbreviations).

Part	Rule	Basis
Technique	R per-site rules: S1 jointed bedrock at every site (in engine-B runs of R, at on-outcrop sites only); S5 recorded stone, else the nearest of seven classes; S2 none (1 tower), Webster (2), fine facing (3 or more); S4 lever for corridor nuraghi, else internal stair; S6 density and hardness; S8 learning	Webster [25]; Lilliu and Zucca [72]; Contu [28]; Laner [30]; tower-count thresholds FREE
Friction	Sledge route > 1 km: tallowed rails, μ 0.1; otherwise dry ground, μ 0.38	Richards and Whitby [45]; Boswinkel [44]; 1 km threshold FREE
Social	Requests, trust, specialisation (help given builds the helper's skill) (A3); help joins the owner's stage	Araque González [23]; behavioural parameters FREE
Sea	Laden craft at 2 knots, 2,000 kg payload, coastal band 2 km	Hazell and Fitzpatrick [71] (Micronesian analogue)
Work party	Residents drawn by tower class; workers = 0.4 × residents (Webster's men = residents/5, doubled for adult women; DERIVED)	Webster [19]; Webster [25]; Webster and Webster [73]; Ugas via a secondary source; interpolation FREE
Phasing	P10: 10% cohorts, each starting when the previous one is finished; P20 and P100 as comparisons	Gradual colonisation by fission [74]
Population ceiling	At most 300,000 residents in all communities, started or waiting (checked once per seed; residents redrawn if exceeded)	Calò et al. [75] (uncited estimate for prehistory to 1700 CE, next to a citation of Day, 1988, on medieval–early modern demography)

seeds. The residents of all communities, started or waiting, may not exceed 300,000, a figure for the island before 1700 CE given without a source [75], used only as an upper reference, not as a Bronze Age population estimate.

3.2.5 | Outputs

Island years (island time) are the finish time of the last nuraghe and of 95% of them, as (finish day + 1)/90, averaged over seeds, and a site's build span is its working days divided by 90. Network need compares each site with itself in the alone run: a site "requires a network" if alone it would take more than 25 years or the network halves its time. Scenarios share seeds 1–100 (common random numbers), so comparisons are paired; seed ranges show stochastic variation, not uncertainty in the assumptions (Appendix S1).

No parameter is assumed unimportant; all are tested in two stages. Morris screening [77] first screened the 32 configurable unsourced factors cheaply, and 8 of them, plus the alliance threshold, were fixed as negligible or within seed noise. The rest, with the help radius (requests within 10 km) and the lift rule (lever for corridor nuraghi, internal stair otherwise, against stair everywhere), pass to Sobol variance decomposition [78, 79], varied over sourced or assumed ranges, not calibrated. The indices depend on these ranges. The analysis of MR_P10 has 26 factors and 14,336 three-seed runs. We ran a feasibility grid over pull, drag speed and tonnage (75 runs), and paired equivalence tests (two one-sided tests) [80] for four parameters that Sobol flagged as weak; all designs are in Appendix S2.

3.3 | Verification

When agentic-ABM generates code faster than a human can read it, verification becomes the main safeguard.

The checks of the research agent's reference code form four complementary layers, none a proof of correctness and each able to find what the others miss, and every result-changing fix was approved by a human first (history in Appendix S3). Throughout, a model rule is correct only if it implements what its cited source says.

- **The research agent's own checks.** Unit tests (fewer test functions than either arm's; Table 2), identity checks after each change, two automatic checks against misquoted numbers and unwanted content in the public code (Appendix S1), and reproducibility notebooks that pass only on a byte-exact match (71 of 71 items). An independent check of 161 claims of the model description found 147 correct, 8 wrong and 6 unverifiable.
- **Manual checks by humans.** The authors' static reading of the code reported 13 findings: 9 confirmed, 3 latent (no current effect) and 1 by design.
- **Another AI agent checking.** Audit agents run by the authors in the OpenCode coding tool (v2.2.4, Big Pickle LLM), outside the research agent's environment, reviewed the code; across the first three reviews, 27 issues were verified and then fixed or documented.
- **LLM docking.** Blind re-implementations on other LLMs, docked against the reference, found bugs on both sides, including reference bugs that its own tests had missed.

Docking. "Docking" [81] establishes whether two models produce the same results, at three levels: numerical identity, not to be expected from stochastic models; distributional equivalence, in which the output distributions cannot be told apart statistically; and relational alignment, in which the models reproduce the same relationships among their results. Docking exposes ambiguity in the target model; the same standards have been adopted for replicating ABMs [55], a task that usually needs a second team of modellers.

LLM docking. With agentic tools this can be delegated: docking agents on the same harness, given the research agent's runtime, skills and data access but not its memory, each implement the model blind from its specification, so that, for arm 2, the LLM was the only intended difference. We ran two arms: arm 1 on Claude Opus 5.5 (Anthropic) and arm 2 on GLM-5.3-Flash (Zhipu AI). Arm 1 shares the research agent's LLM, so its independence lies in the implementation: a separate instance sharing no memory, conversation or files, with code of a different architecture (Table 2). Arm 2, on another developer's LLM, removes the shared LLM and is the stronger check; it also recreated the parts of repast4py it needed instead of using the library. The

arms never received the reference code; they worked from the written specification and the same checksummed input tables (Appendix S1), reached only through numbered, checksummed instructions. Blindness had stated exceptions: two small leaks of reference lines in an early audit, debugging instructions that gave an arm's deviation from the reference results, and two speed-only patches written on the reference side (Appendix S3). Gaps in the specification were closed by addenda with identical rulings to the arms they concerned, two of which went to arm 1 only (Appendix S1). The written specification, provided with addenda a–k and the instruction log (Supplementary Data), is thus the object under test: docking checks that it is implemented faithfully, not that it is conceptually valid, and an error shared by the specification and every code stays invisible.

Gates and criteria. Input gates confirm that each run used the intended code and inputs. We fixed the criteria before any clean-run result was compared, following the docking levels [81]. For man-hours, instead of distributional equivalence, a cell counts as similar enough when its seed mean lies within ±3% of the reference, a margin of our choice, not a statistical test; Kolmogorov–Smirnov tests are reported alongside. Relational alignment requires Spearman's $\rho \geq 0.9$ between the arm's and the reference's scenario effects (each cell relative to that code's own headline), and sign agreement of at least 95% over the cells whose reference effect exceeds twice its standard error, for both island man-hours and island time. The codes *dock* when relational alignment holds and the headline cell is within ±3%, and *dock with named differences* when every cell outside ±3% is also traced to a named reading of the specification.

Staged debugging. Each arm, like the reference, went through fix rounds with regression tests until a full round found no result-changing bug; the code was then frozen and the full run list executed (419 rows, one per scenario; 640 cells, 612 for arm 2; 100 seeds each), which covers all scenarios above plus the sensitivity and robustness runs of the appendix. We traced every cell outside the margin in code and classified it as an *arm bug*, *two readings* of the same specification text or a *reference bug*; the reference is not presumed correct. The codes share the specification, the gated inputs and one operator, the research agent (protocol in Appendix S1; results in Appendix S3).

Code variation. Code that runs quickly and returns complete results can still be broken, as found elsewhere for LLM-generated code [82]; arm 2's engine B was such a case before its fixes (Table 5). The 20 traced code differences are classified in Appendix S3.

4 | Results

4.1 | Literature scenarios

4.1.1 | Construction techniques

With Webster's stage rates, the default work parties and every nuraghe built at once, the island needs 303.7 million man-hours (303,705,815, mean of 100 runs; run-to-run standard deviation 6,464). Transport is 43.4% of all labour. The median single tower needs 22,462 man-hours, 6.4% less than Webster's 3,600 man-days (24,000 h at his day lengths, Table 3) [25], all of it in transport, because the sledge rate is faster than the baseline's man-hauling rate [38] and more than two-thirds of sites stand on a permitted outcrop.

The literature scenarios span 72.6 million (L) to 812.4 million man-hours (H), a factor of 11.2 (Fig. 2). Ranked by the spread they produce, transport technology (59.2 percentage points), co-operation, stone source, lifting and dressing (52.9) lead the steps of the chaîne opératoire, and learning (2.5) and Webster's class totals (1.5) trail them (Appendix S2). The R technique needs 283.2 million man-hours, 6.8% less than the Webster baseline: savings on undressed single towers offset the higher costs of fine facing and dense stone on complexes (Table 3).

Against the lower end of the single-community range of 30,000–50,000 person-hours [46], a Neolithic benchmark rather than a Nuragic figure, in the Webster scenario 16.1% of single towers and 100.0% of multi-tower nuraghi exceed it; even under the L bound, 16.0% of multi-tower nuraghi do. Single towers sit near the boundary, depending on how blocks were lifted and dressed; complexes are beyond it in every one-step literature scenario.

4.1.2 | Stone, haul and terrain

Stone sources are local for most nuraghi. 4,941 sites (68.9%) stand on an outcrop of their stone and get the minimum haul; the mean haul is 460 flat-metre equivalents, the median 0 and the 90th percentile 1,076 (Fig. 3). This is consistent with local sourcing in Marghine [25], but follows partly from the rule that a site without a recorded stone quarries the nearest permitted outcrop on a modern map. Where the stone is recorded, the nearest outcrop of the seven classes matches it at 77.2% of single-stone sites (Appendix S2). Long hauls cluster in a few areas, chiefly around the Gulf of Oristano and in the south.

Terrain enters only through haul costs, and its correlations with labour and time are weak (Appendix S2).

4.1.3 | Sea access

Moving stone by boat at 2 knots matters only locally: island labour falls by 2.97% and the finish time, set by inland sites, is unchanged (Appendix S2).

TABLE 2 | The three codes built from one written specification: key structural contrasts (simulation packages; non-blank, non-comment lines; MPI = Message Passing Interface, for parallel copies (ranks) of a program; numba: a compiler that speeds up Python). Engine B and the full comparison: Appendix S3.

	Reference	Arm 1 (Claude Opus 5.5)	Arm 2 (GLM-5.3-Flash)
Frameworks	repast4py, MPI, numba	repast4py, MPI; numba in engine B	repast4py installed but not used; recreated the scheduling and MPI partitioning it needs (mpi4py, numpy)
Engine A	one module; compiled day loop over an island-wide task list	package of 13 modules; continuous-time schedule over each community's pending work	one module; a task list per community, worked against the labour available per season
Parallelism	seeds over MPI ranks	spatial partition designed; seed-parallel single-process workers in practice	seeds over MPI ranks
Tests (files / functions)	6 / 40	115 / 424	4 / 177
Lines of code	4,872	6,772	4,682
Decision log	dated change notes	90 + 31 entries	49 entries

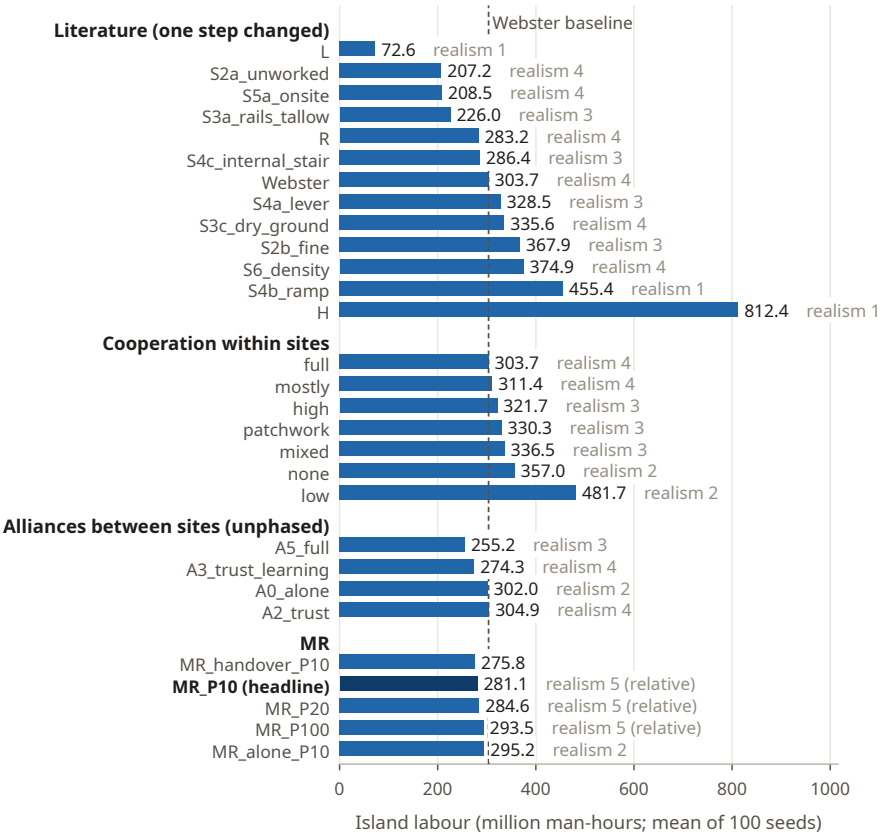


FIGURE 2 | Island labour (million man-hours, mean of 100 seeds) by scenario family, with realism ratings (1 = contradicted or implausible, 5 = best supported; judgements, not results; "5 (relative)": MR combines the highest-rated options; MR_handover is unrated). Labels: Abbreviations. Literature family: 12 of its 22 scenarios (all: Appendix S1); a step's spread is the range of its options' changes against the baseline. Engine B runs literature and cooperation, engine A alliance and MR. Cooperation levels, share of builders pooling labour in shared crews: none 0%, low 25%, mixed 50%, mostly 75%, full 100%; patchwork, 100% or 0% per site at random; high, full plus help from finished sites within 10 km. Alliances use a rolling start (1/450 of sites per season). The 29 scenarios include the MR phasing and help sub-scenarios. Low costs more than none because engine B charges builders left idle while its small pooled crew moves every heavy block (Appendix S2).

4.1.4 | Site-level effects

Fig. 4 aggregates the per-site results of the 84 one-parameter alternatives over 60 territories, clusters of nuraghi that mostly help one another. Averaged over the alternatives, a territory's man-hours change by 4.7 to 25.3% (median 7.8%). The sensitivity follows the stone supply more than the size of the nuraghi: it rises with the mean haul to the outcrop (Spearman $\rho = 0.84$) and is unrelated to the mean tower count ($\rho = 0.10$), and the alternative that moves an area most differs by area (Fig. 4c; Appendix S2).

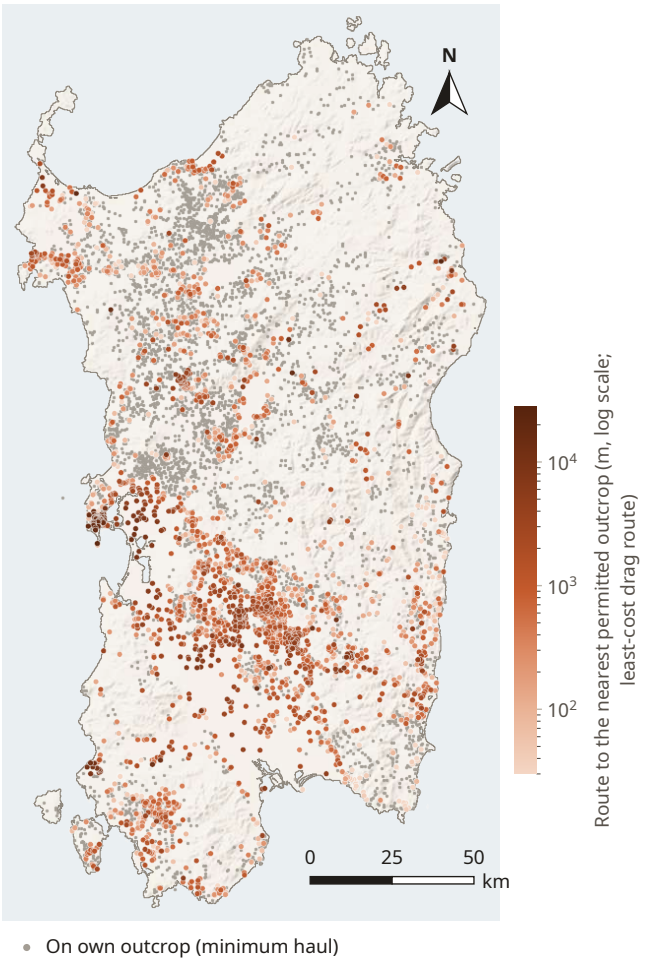


FIGURE 3 | Headline stone rule: horizontal length of each nuraghe's least-cost drag route to the nearest permitted quarry face of its stone (log scale); grey: sites on their own outcrop. Section 4.1.2 reports hauls in flat-metre equivalents, the length of level ground needing the same dragging effort (uphill counts more, downhill less). In the simulation a community may choose among its five cheapest quarry faces (Appendix S1).

4.2 | New scenarios

4.2.1 | The MR scenario

Under the headline MR_P10 (Table 3), the model finishes all 7,167 nuraghi after 482.5 years on average (seed range 356.5–644.3; 95% of them after 436.7 years), with 281.1 million man-hours. Of the 115,954 workers, the largest cohort averages 12,466, excluding helpers from other cohorts. A nuraghe takes 3.59 years on average (median 3.17). The population ceiling never binds: even if every community built at once, residents would number at most 292,627 in any seed (mean 289,886).

Total labour hardly depends on phasing or assistance: it varies between 275.8 and 295.2 million man-hours across the variants, 3–9% below the Webster baseline (303.7). Island time, by contrast, depends strongly on both. Under P100 the island takes 99.16 years; under P20, 303.12; under P10, 482.53. A nuraghe's own work party sets its time: literature-based work parties are much smaller than the default rule (9.0 workers for a single tower, against about 50), so a nuraghe takes about three and a half years rather than under one.

The slowest sites, small single towers whose work parties of a few people face long hauls, set the pace of every cohort (Fig. 5).

4.2.2 | Phasing

Phasing changes the picture of the island more than any technical choice (Fig. 6). With the default work parties, the baseline takes 25.48 years under P100, 84.15 years under P20 and 145.38 years under P10. P10 is not ten times the P100 time (it is 5.7 times in the baseline) because the slowest sites are spread across cohorts. Under phasing, the peak number of builders at once falls from 418,633 to 42,673 (P10, default parties).

The simulated time for MR_P10, 482.5 years on average, is close to the roughly 500-year expansion phase [74]. This shows compatibility, not a reconstruction of the chronology. Cohorts are drawn at random, each starts as soon as the previous one ends and waits for its slowest site, so the time is a lower bound for this cohort rule only. Rules with fewer or no cohorts finish sooner (Section 4.2.1); a rolling start (1/450 of sites per season, as in the alliance rule sets) takes about 455 years, set by its start rate, and starts following population growth take 436 or 583 years (Appendix S2).

On the map (Fig. 7), 16.6% of the nuraghi are finished after 50 years under the headline and 36.4% after 150, scattered over the island rather than advancing as a front. With a published lower pull of 47 kg per hauler (P10, with friction-0.3 movement on all routes), only 13.8% are finished after 150 years and the island takes 1,889.3 years. Under P100, 90.4% are finished after 10 years.

4.2.3 | Cooperation and networks

In the headline run, nuraghi receive 17.9% of their labour from other communities (Table 4). Help shortens the mean build span from 4.85 to 3.59 years overall and from 5.12 to 3.72 years for single towers, which have the smallest work parties and receive the largest share (21.0%). By the criterion of Section 3.2.5, few nuraghi strictly require a network: 1.0% of site-runs would take more than 25 years alone, and help halves the span of 18.0% (21.6% of single towers). Help is useful only where hands are idle, as under P10 (Fig. 8): under P100, nuraghi receive 6.3% of their labour and the mean span barely changes (4.83 against 4.85 years alone). Under handover, complexes receive more labour but build more slowly (8.58 against 3.76 years for five towers), because the reward for asking counts a community's own man-hours, not its time.

Labour sharing. Help is not scripted; it emerges from the rules of Section 3.2.3. The links present in at least 5% of seeds carry 78.7% of the help, 39.5 of 50.2 million man-hours per run, and join 7,105 of the 7,167 nuraghi into one network (Fig. 9a; Appendix S2). The median link is 4.63 km long. Complexes of three or more towers give 64.5% of all help and receive 30.5%, and 83.9% of them are net lenders, whereas single towers give 16.3% and receive 52.5% (Fig. 10). Aggregated into the 60 territories and

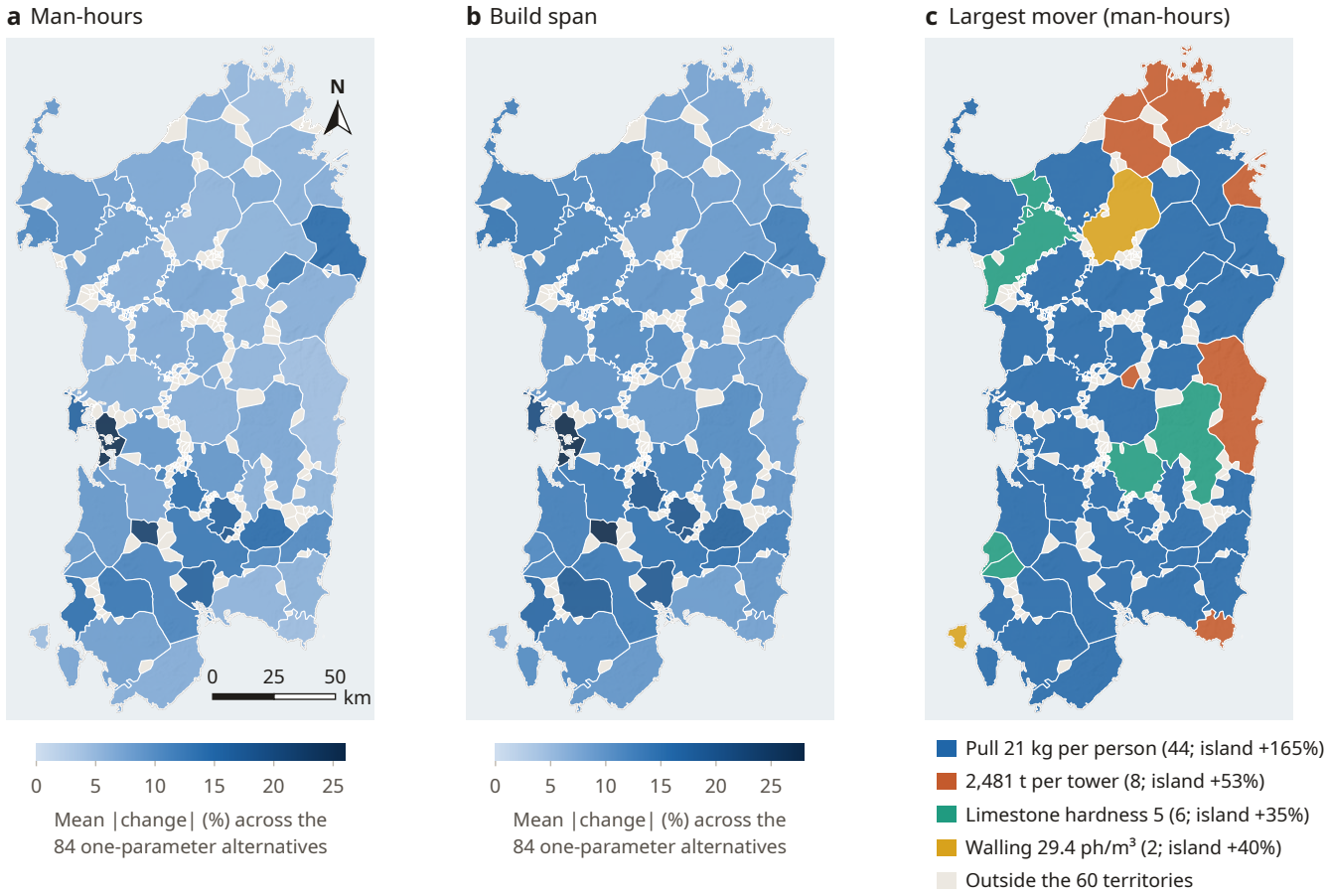


FIGURE 4 | One-parameter alternatives by territory: 84 alternatives, each replacing one headline (MR_P10) parameter (74 from published figures, 10 our assumptions or range ends; 100 seeds). They differ from the literature family (Fig. 2), which changes one Webster step at a time. (a) Mean absolute change in a territory's man-hours across the alternatives; (b) the same for build span (working days / 90); (c) the alternative that changes each territory most (brackets: territories; island change): pull 21 kg per hauler; 2,481 t of stone per tower; limestone and marble at five times the base quarrying effort; walling 29.4 person-hours per m^3. Territories: the 60 groups of nuraghi that help one another most, stable across the 100 seeds (6,718 nuraghi; Fig. 9); grey: outside them. Per-site changes: Appendix S2. Site data: Nurnet.

TABLE 3 | MR and its variants (full island, 100 seeds; years = seasons of 90 communal working days, mean over seeds; P10 = 10% cohorts, P20 = 20%, P100 = all at once; MH = man-hours; M = million; p95 = 95th percentile). Largest cohort: its workers, mean over seeds, excluding helpers from other cohorts. R and Webster: engine B, default work parties (50 builders plus 25 per extra tower, times a per-seed factor of 0.9–1.1, capped at 50–200). Webster's 3,600 man-days per single tower equal 24,000 h at his day lengths (5 h quarrying and transport, 8 h dressing and building). Verdict: grade of Section 5.3 (seeds of 100 finishing within 500 years, or builders at once against the 300,000-resident reference; cutting at a quarter of seeds instead of half changes no verdict).

Scenario	Island MH (M)	Largest cohort (workers)	Island years, max / p95	Per-nuraghe years, mean / median	Verdict
MR_P10 (headline)	281.1	12,466	482.53 / 436.74	3.59 / 3.17	open (63)
MR_P20	284.6	24,080	303.12 / 248.89	3.88 / 3.47	consistent with the window (100)
MR_P100	293.5	115,954	99.16 / 14.29	4.83 / 4.47	consistent with the window (100)
MR_alone_P10 (no help)	295.2	12,466	612.27 / 551.30	4.85 / 4.46	less likely (6)
MR_handover_P10	275.8	12,466	625.37 / 565.12	3.79 / 3.21	less likely (3)
R (technique only, default parties, all at once)	283.2	418,633	12.75 / 2.00	0.78 / 0.60	less likely (builders)
Webster (baseline, all at once)	303.7	418,633	25.48 / 1.99	0.92 / 0.62	less likely (builders)

7 super-regions of the help network (Leiden clusters [83]), help rarely crosses a super-region boundary: 381 thousand man-hours per run, 0.76% of all help (Fig. 9b). Both the single network and the rarity of help between super-regions, which are detected from the help network itself, follow largely from the 10 km help radius and the density of nuraghi; they are properties of the model, not independent findings.

TABLE 4 | Network need by tower class, headline rule set (join help), in 10% cohorts (P10) and all at once (P100): mean build span with help and alone, and share of a site's labour received from other communities. Last two columns, each site against itself in the alone run (same seed): share of site-runs whose span with help is less than half the span alone, and share that would take more than 25 years alone. Full island, 100 seeds per run.

Towers	Sites	Mean work party	P10 years, help / alone	P10 received	P100 years, help / alone	P100 received	Help halves span (P10)	>25 years alone
1	5,902	9.0	3.72 / 5.12	21.0%	5.07 / 5.11	3.6%	21.6%	1.2%
2	788	30.0	2.67 / 3.39	17.0%	3.40 / 3.39	7.9%	2.4%	0.04%
3	230	59.0	3.64 / 4.47	15.7%	4.37 / 4.47	9.1%	0.5%	0%
4	128	89.0	3.06 / 3.60	14.5%	3.88 / 3.60	8.4%	0.03%	0%
5	111	117.9	3.30 / 3.76	13.3%	4.38 / 3.76	9.0%	0%	0%
6	7	145.9	2.72 / 2.96	12.2%	3.96 / 2.96	8.5%	0%	0%
7	1	146.3	2.08 / 2.20	10.1%	2.87 / 2.20	7.1%	0%	0%
All	7,167	16.2	3.59 / 4.85	17.9%	4.83 / 4.85	6.3%	18.0%	1.0%

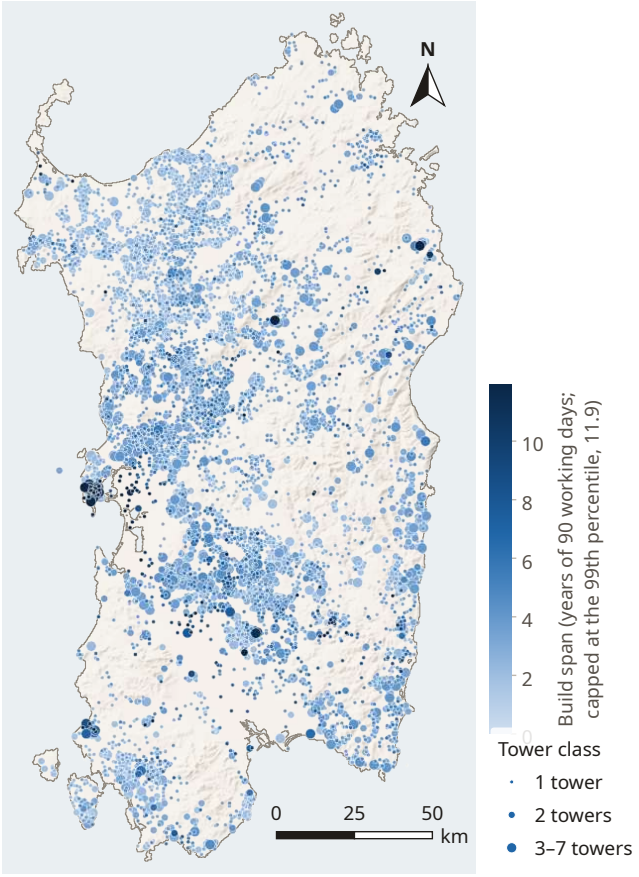


FIGURE 5 | MR_P10: build span of each nuraghe, from start to completion (working days / 90, mean over 100 seeds; colour capped at the 99th percentile), on a digital elevation model hillshade; marker size by tower class.

How the network forms. Under gradual building the network grows cohort by cohort (Fig. 11): under P10 the largest connected help network of a run contains a median of 1.7% of the nuraghi after 50 years, 23.6% after 200 and 67.8% at the end. Under P100, the largest network holds 8.0% of the nuraghi and help totals 18.5 million man-hours per run. Gradual colonisation thus produces an island-wide web of short help links, and simultaneous building fragmented local clusters.

Pre-existing networks, tested against a run without networks, change where and when labour flows more than the island total (Appendix S2).

4.2.4 | Feasibility

A grid of 75 runs over the sourced ranges of pull (20–60 kg per person), drag speed (0.19–1.0 m/s) and tonnage (1,350, 1,909 and 2,481 t per tower) finishes at every point (Appendix S2), but its slowest corner (20 kg, 0.19 m/s) takes 2,839–5,145 years (seed 1), far beyond the 500-year window. Morris screening ranks pull first for island labour, then tonnage and drag speed; loading time has no effect (it is charged per tonne).

4.2.5 | Drivers

In the Sobol analysis, island labour depends on the stone: the hard-stone hardness, the dressing rule and the stone-source rule have the largest total-order indices for island man-hours (0.49, 0.24 and 0.12). Island time, counted in working days, depends on the number of cohorts (0.40), the help mode (0.31), the stone rule (0.19) and the dressing rule (0.16); season length's index on island years (0.41) only rescales days into years; it has no effect on working days. The help radius, the lift rule and the acceptance parameters stay below 0.01. In a separate Sobol analysis of engine B's cooperation levels, pull leads for labour and drag speed ranks fourth or fifth (Appendix S2).

Time depends on different terms (Fig. 12). The number of cohorts barely affects island labour (total-order index below 0.01) but most affects island time, and time scales with the hours worked per day. Of four parameters with weak Sobol indices, none was equivalent to the baseline within ±1% on both island labour and island time in all six cooperation scenarios; household size and block mass shift island time by up to about 46% in the low-cooperation scenario.

4.3 | Docking results

After the fixes and the ruling that learning does not discount lifting, we docked both arms against the reference on a common sample of 60 cells of the run list (48 on engine A, 12 on engine B, 100 seeds each), with the gates and criteria of Section 3.3 (Table 5).

On engine A, 47 of 48 cells lie within ±3% of reference man-hours in each arm, both headline cells match, and Spearman's ρ and sign agreement for man-hours meet the criteria in both arms.

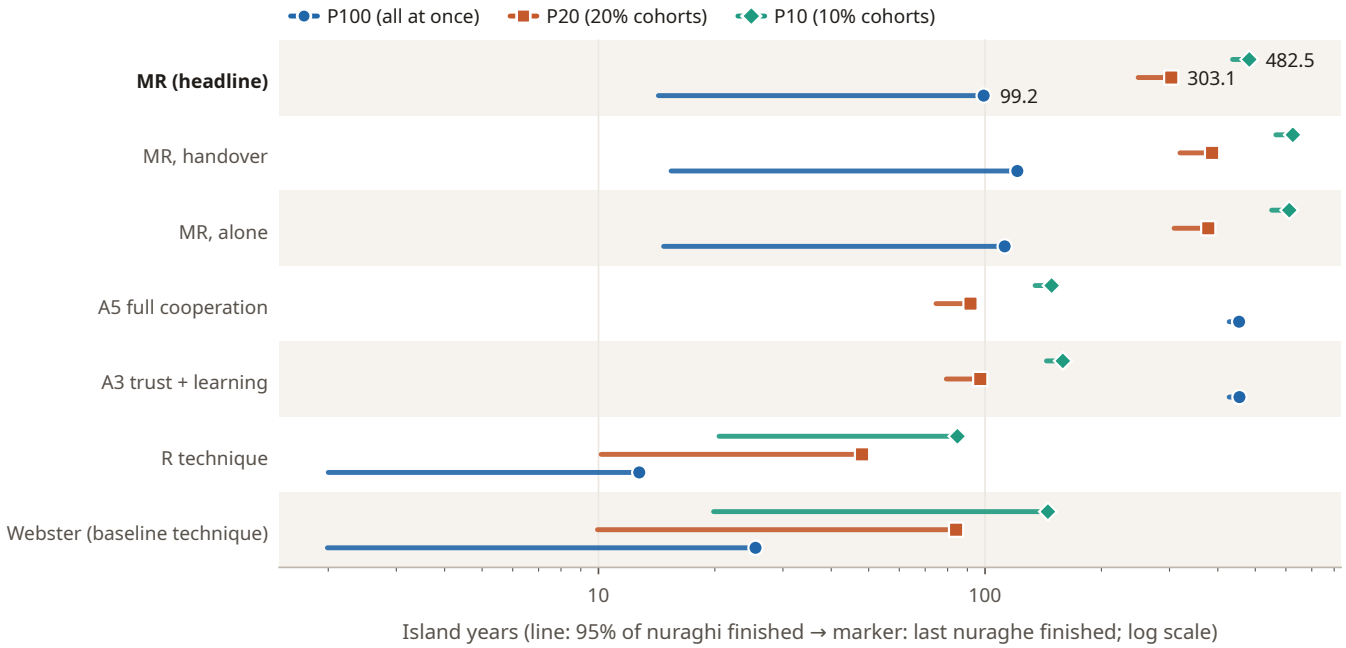


FIGURE 6 | Time to build all 7,167 nuraghi, all at once (P100) and in 20% and 10% cohorts (P20, P10); line from 95% finished to the last nuraghe (marker); years of 90 communal working days, log scale. Webster and R (no help between sites) are run once and their site times grouped into cohorts afterwards; scenarios with help are simulated cohort by cohort. A3 and A5 use a rolling start (1/450 of sites per season, shown as "P100 (all at once)"), so the start rate sets their time. "A5 full cooperation" is A5 (Abbreviations), not the cooperation level "full". Webster, R, A3 and A5 use the default work parties, MR the literature-based ones (Section 3.2.4).

On engine B, arm 1 matches all 12 cells and arm 2 its eight full-cooperation cells exactly; its four mixed-cooperation cells differ by up to 5.5% in man-hours, a minor difference traced to one arm error (Appendix S3). The one engine-A cell outside the margin is the same in both arms (−4.84% and −3.40%): it turns on whether a community that has not yet started its own nuraghe may help, a difference kept by ruling; in arm 2 it carries a small untraced transport offset (−2.91% without it). Arm 1 therefore docks with named differences. Arm 2 does not dock by these criteria: its man-hours meet them, but its island-time ρ is 0.898, just below 0.9. The share of labour received as help does not dock: it is within ±3% of the reference in 1 and 2 of 46 cells (Appendix S3). The gap traces at least partly to two readings of the help rules in the specification, whether a community not yet building may help and when help is requested, which a controlled vocabulary would close (Section 5.7).

Most island-time gaps are a finding about the specification, not a failure of the codes. One addendum lets a helper work "whenever the stage's upstream allows": the reference reads this as overlap (a downstream stage may advance as far as its upstream has progressed), arm 1 as completion (a stage opens only when the previous one is finished), and arm 2, which never received it (Section 3.3), hands a stage to its helper from the next season. Man-hours agree, because help changes who works and when, not how much work there is; island time moves, by up to about 20% in arm 2. The research team ruled that the reference's reading stands. Docking found a coding error: arm 2's relay helpers ignore the worker cap of the food-limited variant (island time −18.5% in that cell). Most other island-time gaps beyond ±3% reflect a second named difference, request timing (the reference asks for help with every stage at tower start, the arms as each stage is reached), which matters only where the slowest sites are not transport-bound (Appendix S3).

Apart from arm 2's coding errors, wherever the text is explicit both arms implemented the theory as written.

5 | Discussion

5.1 | How nuraghi were built

The model does not reconstruct the building of any particular nuraghe; it shows what the island's geology, terrain, site sizes and the best-supported construction evidence imply when every nuraghe is built explicitly.

First, building the nuraghi was a large but not extraordinary undertaking, below the Webster baseline in every MR variant and consistent with Webster's [25] inference that the Borore towers needed only a small fraction of adult male time. Second, given the literature-based work parties, the share of residents who build and the 90-day season, time, not labour, is the binding dimension. Among the 26 factors analysed jointly, time depends most on phasing and help, and labour on the stone; pull, tonnage and work-party size were varied separately, so whether social or physical factors weigh more awaits a joint analysis. Third, cooperation is useful but, for most nuraghi, not necessary, and its form matters. Help that joins the owner's work speeds construction under gradual building (Table 4), and help that takes over whole stages can slow complexes. Fourth, construction is tied to landscape: most nuraghi stand on or near their building stone and long hauls are

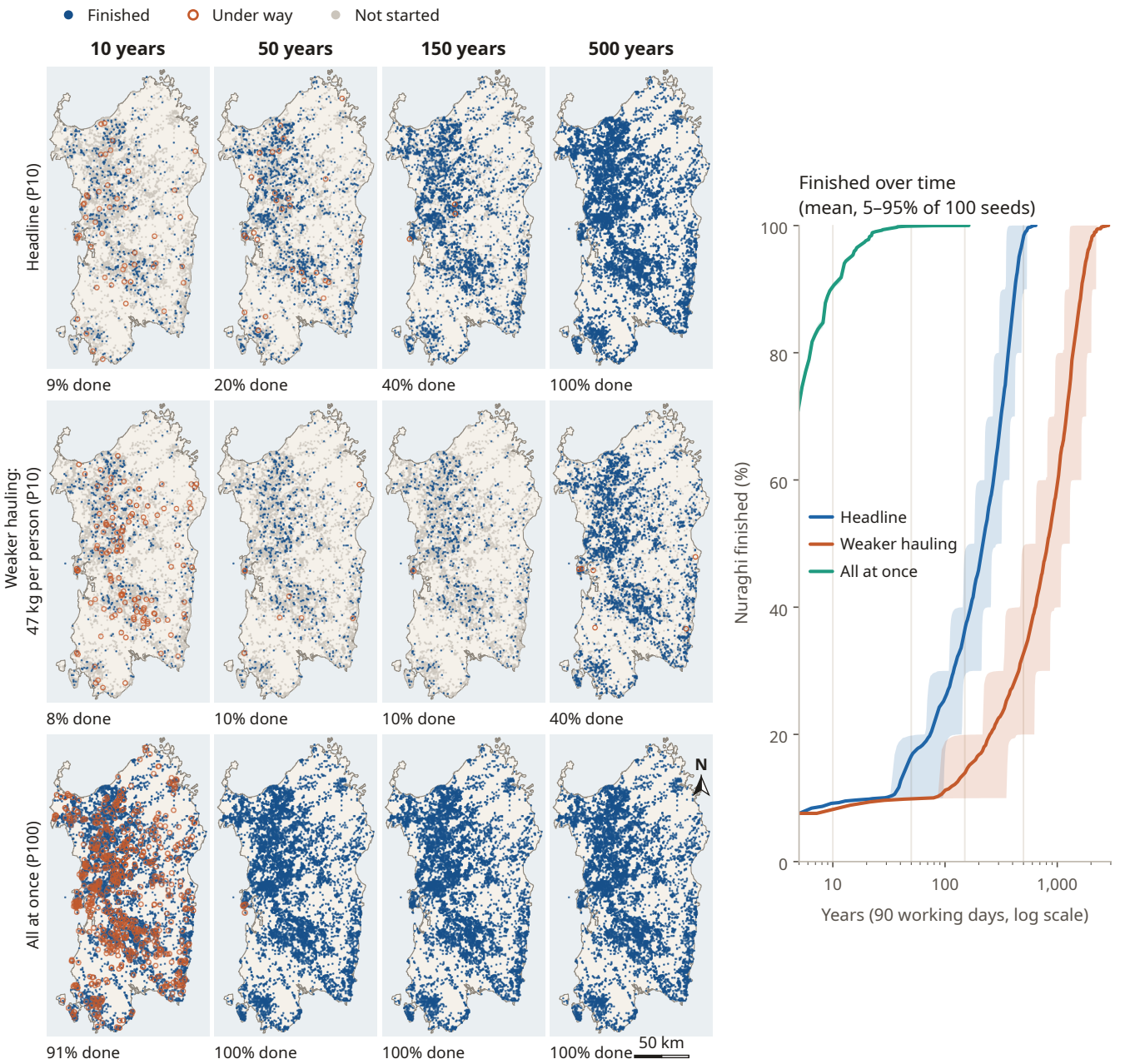


FIGURE 7 | Building the island over time: headline (P10), weaker hauling (47 kg per person, P10; the headline sizes crews for about 79 kg of load per person on dry ground, 300 kg on rails) and all at once (P100); phased engine-A runs. Maps: per scenario, the seed with island time closest to the 100-seed median (473.7, 1,876.7 and 100.2 years), after 10, 50, 150 and 500 years; label: share finished. Right: share finished, mean and 5–95% range over 100 seeds (log scale). Years of 90 working days, (finish day + 1)/90 (Section 3.2.5). Site data: Nurnet.

TABLE 5 | Docking against the reference (the research agent's code): common sample after the fixes and the lift ruling (lifting priced without the learning and skill discount); full round of the whole run list before them. 100 seeds per cell (one scenario setting). Common sample: fixed-seed draw from 24 of the run list's 25 model families (all but the engine-B cooperation levels), always with the headline and the cells with open questions. MH = man-hours; ρ = Spearman's rank correlation of scenario effects; sign agreement: man-hours / island time; headline: arm minus reference, island man-hours. Arm 2's engine B, common sample: 10 seeds per cell, the reference's seeds 1–10; island time identical in its eight full-cooperation cells (Appendix S3). Full round: arm 2's 28 engine-A cells refused at gate 0a excluded (525 of 553). Island time: island seasons (text: (finish day + 1)/90).

		Common sample, after the fixes				Full round, before the fixes
Arm	Engine	Within ±3%	ρ MH / time	Sign MH / time (%)	Headline	Within ±3%; headline
1	A	47/48	0.968 / 0.976	95.1 / 100.0	−0.35%	505/553; −0.41%
1	B	12/12	1.000 / 1.000	100.0 / 100.0	<0.001%	85/87; <0.001%
2	A	47/48	0.967 / 0.898	95.1 / 96.7	+0.001%	495/525; +0.75%
2	B	10/12	0.955 / 0.973	100.0 / 100.0	<0.001%	32/87; +3.45%



FIGURE 8 | Network need by tower class, headline rule set: (a) mean build span alone, with join help (headline) and with handover help, P10; (b) share of a site's labour received from other communities, P10 and all at once (P100). Full island, 100 seeds per run.

regional and make the slowest sites. The model turns the relation between settlement and stone, often discussed qualitatively [24, 23], into site-level quantities comparable with survey data.

5.2 | Social organisation

The literature offers at least seven readings of who built the nuraghi (Section 2.3), from household self-build through heterarchical cooperation and a two-tier mode of local craft builds and mobilised workforces for complexes (Melis, in Moravetti et al. [84]) to servile labour under master builders (each read against the model in Appendix S1). Labour totals alone cannot distinguish them, because the model fixes the workforce instead of deriving it from authority, surplus or obligation; it counts hours, not whether they were free or coerced.

What the model shows is narrower. Single towers are within the reach of a household or extended family over years and of a village within a season. In the Webster and one-step literature scenarios complexes exceed one community's range (Section 4.1.1), which locates the question of mobilisation where Melis placed it: at the complexes, not the towers; with MR's work parties, however, no complex would take more than 25 years alone (Table 4). Cooperative mechanisms of the kind Araque González [23] proposes, based on trust and reciprocity without command, are sufficient in the model to supply that labour and to produce hubs at large complexes. This shows that cooperation could have built the nuraghi, not that it did: a hierarchical system could produce a similar pattern of short help links and hubs (Section 4.2.3) for different reasons, and distinguishing them requires evidence the model does not use, such as differential provisioning, storage or prestige goods at the hubs. The scenarios tested do not cover every possible past.

If complexes were built as unified projects by mobilised workforces, their build spans should be short and their learning effects strong; if they were enlarged in phases over generations [72] by local crews (Melis, in Moravetti et al. [84]), the relevant unit is the phase, not the complex, and the model's per-complex learning is optimistic. Dated building phases at more sites would therefore test the model more directly than any further estimate of rates.

5.3 | More and less likely pasts

The model cannot be validated against observed construction (Section 5.6), but it can grade pasts against stated constraints: the roughly 500-year expansion phase [74] and completion, with 300,000 residents [75] as an upper reference. Every scenario in Table 3 finishes all nuraghi in every seed run, so time decides. One rule assigns four grades by the share of seeds finishing within 500 years: *ruled out*, none, and the fastest seed beyond 750 years; *less likely*, under half; *consistent with the window*, at least 95%, with builders below the reference; *open*, the rest. Grades depend on phasing: under P10 every cohort waits for its slowest site, and P20 and P100 finish sooner (Section 4.2.2). Ruled out are the slowest corner of the feasibility grid (Section 4.2.4; 2,839–5,145 years, seed 1) and the published pull of 47 kg per hauler with friction-0.3 movement on all routes (1,318.8–2,866.7 years), the latter only jointly with P10, the headline's work parties and the 90-day season. Building without help or with handover help is less likely (6 and 3 of 100 seeds within the window), as is simultaneous building with the default work parties, which needs 418,633 builders, above the reference before any non-builder is counted. The headline is open (63 seeds; 356.5–644.3 years); MR_P20, MR_P100 and phased Webster runs are consistent with the window. Grades rest on results only; the realism ratings (Fig. 2) are judgements.

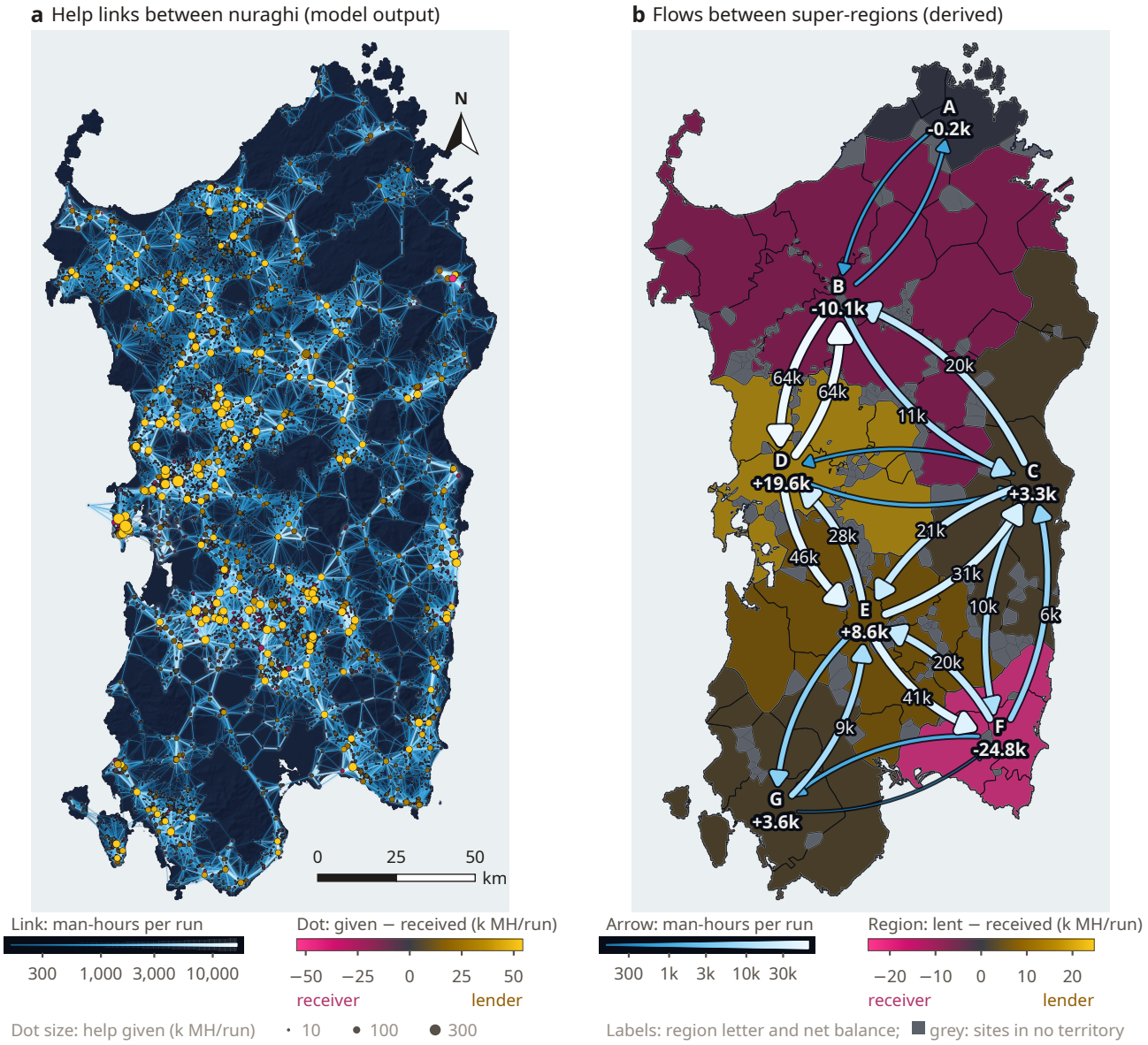
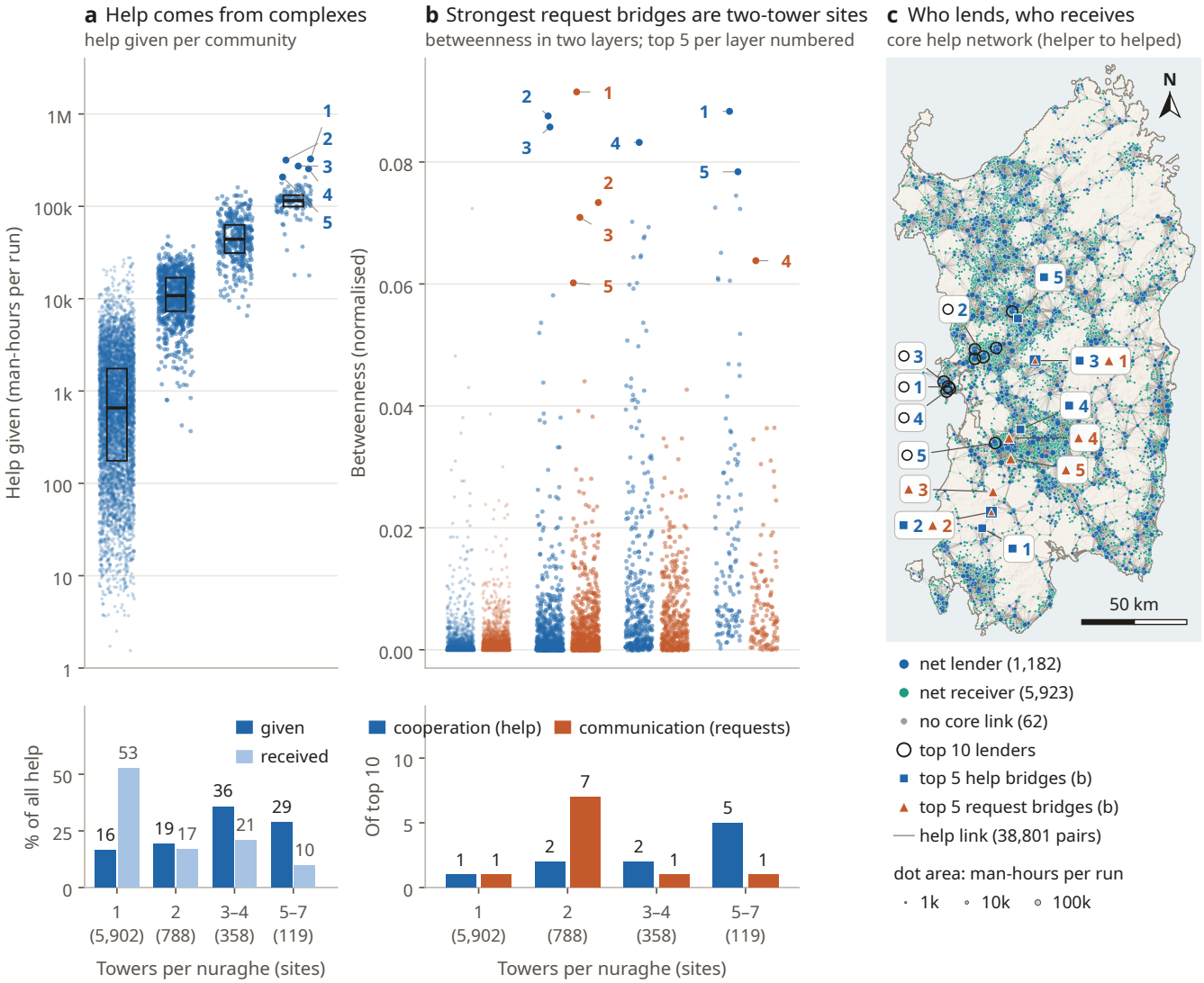


FIGURE 9 | Labour sharing, headline run (100 seeds). (a) Model output: help links (helper → helped) in at least 5% of seeds, weighted by mean man-hours per run; nuraghi sized by help given, coloured by net balance. (b) Derived view: flows between the seven super-regions (labelled from 5,000 man-hours per run) and their net balances. Territories: groups of nuraghi helping one another more than others, from Leiden clustering [83] of each seed's help network, merged into one consensus grouping; super-regions ("Region" in b): Leiden clusters of the territory network; grey: sites in no territory. The grouping is an analytical aggregation assuming that labour exchange tracks political ties, not a model input. Site data: Nurnet.

These verdicts hold only under the model's assumptions and constraints, such as the share of residents who build (limitation (1), Section 6). Ruling out still makes the approach worthwhile without validation, because each failed scenario narrows what the field must test. The remaining scenarios make site-level predictions that fieldwork can falsify: build sequences for radiocarbon dating, and stone sources and haul distances for provenance studies, such as the long hauls around the Gulf of Oristano or help hubs at Cabras.

5.4 | Why model construction

Modelling construction explicitly has three benefits for mesoscale theory. It makes assumptions explicit, so a disagreement can be located and measured, showing where field or experimental work would be most informative [57, 59]. It bounds what was feasible (Section 4.2.4). Per-site outputs can be mapped and compared with landscape and settlement studies, such as those of Sinis [37] and Sulcis [36]. To the energetics studies of Section 2.3, the Nuragic case adds the role of phasing in turning a large total into a manageable burden and of inter-community help where spare labour exists. To our knowledge this is the first ABM of monument construction for an entire island corpus; the closest construction ABMs address a single infrastructure project [85] or



a Top lenders (comune), towers
1 S'Argara (Cabras), 5
2 Cobulas (Milis), 6
3 Monti Chibuddas (Cabras), 5
4 Leporada (Cabras), 5
5 sa Domu Beccia (Uras), 6

b Top help bridges (comune), towers
1 Sa Domu 'e s'Orcu (Domusnovas), 5
2 Genna Uraxi (Villacidro), 2
3 Filighe (Ula Tirso), 2
4 Soru (Curcuris), 4
5 Porcarzos (Borore), 5

b Top request bridges (comune), towers
1 Filighe (Ula Tirso), 2
2 Genna Uraxi (Villacidro), 2
3 Cuccuru 'e Crabas (Villacidro), 2
4 Onigu (Masullas), 5
5 Perra (Sardara), 2

FIGURE 10 | Who lends and who bridges, headline (MR_P10, 100 seeds); dots: communities (nuraghi) by tower count. (a) Help given (log scale; boxes: median and quartiles; the 267 communities giving none, 266 of them single towers, cannot be drawn); bars: each class's share of help given and received. (b) Normalised betweenness, how often a community lies on shortest paths between others, in the cooperation layer (helper to helped) and communication layer (asker to asked), each reduced to its core network (links in at least 5% of seeds, undirected, unweighted); bars: each class's count among a layer's top ten. (c) Core cooperation network: lines, help links (helper to helped) in at least 5% of seeds (38,801 pairs), wider for more man-hours; dots, net lenders (blue; help given exceeds received; 1,182), net receivers (green; 5,923) and no core link (grey; 62), area by help given (lenders) or received (others); rings, the ten largest lenders. Numbered: top five lenders (circle), help bridges (square) and request bridges (triangle), listed below (nuraghe, comune, towers). All ten largest lenders have 5–6 towers, four in Cabras, on the Sinis; eight of the ten strongest communication bridges have one or two towers (three of ten in the cooperation layer), though typical betweenness rises with tower class. Full lists: Appendix S2. Site data: Nurnet.

treat monuments as an aggregate drain on a regional economy [86].

5.5 | What agentic-ABM offers

The case shows that agentic-ABM can provide the kind of programme for implementing and testing theory that Plog [2] found lacking, at least for a meso-scale question. Five features of the workflow carry over to other monument traditions and to logistics and governance questions [4].

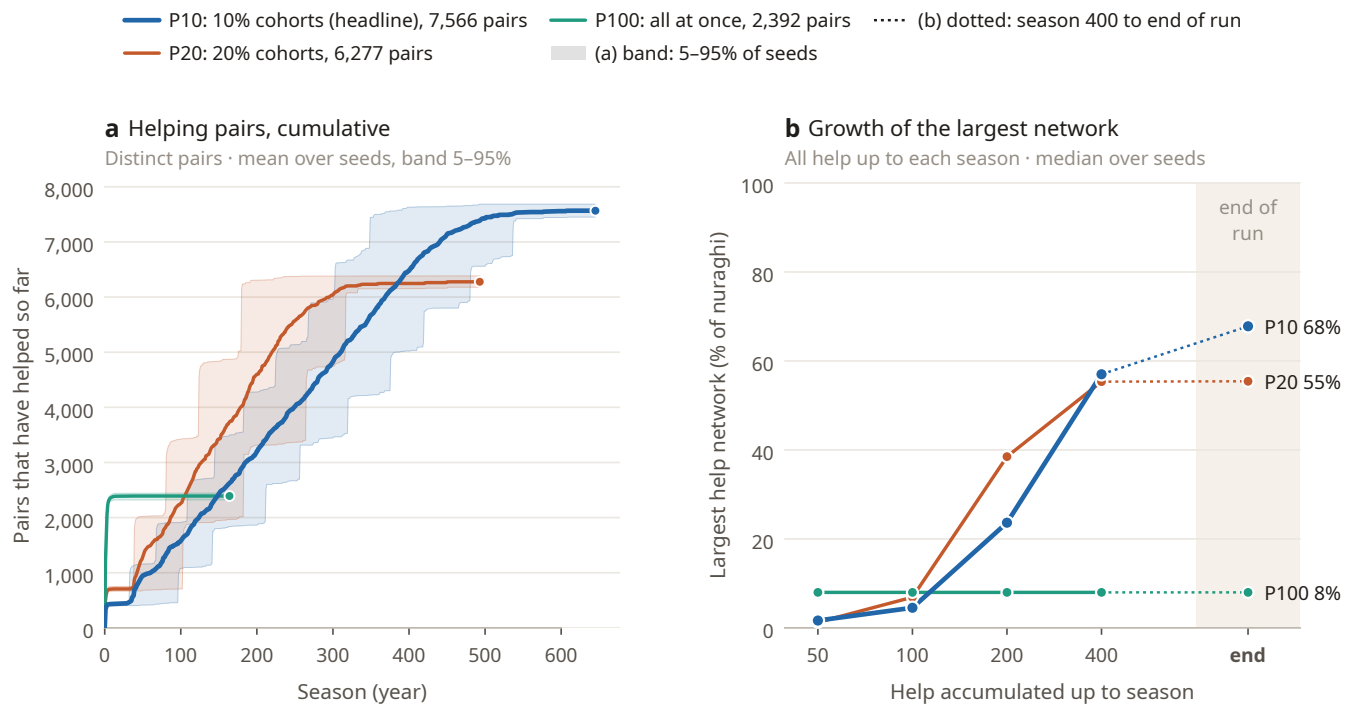


FIGURE 11 | How the help network forms under the headline rule set (join help), P10 (headline), P20 and P100: (a) cumulative number of community pairs that have helped one another, by season (= model year; mean and 5–95% over 100 seeds); (b) largest connected help network as a share of all nuraghi, for help accumulated up to seasons 50, 100, 200, 400 and the end (median over seeds).



FIGURE 12 | Drivers of labour and time: Sobol indices (95% bootstrap intervals) for island man-hours and island years, MR_P10 (26 factors; Saltelli design, N = 512 base rows, 14,336 three-seed runs); dashed line: 0.05. S_T: a factor's effect including its interactions; S_1: its effect alone. Hard-stone hardness: quarrying effort of basalt and granite (headline 2). Help mode: join or handover. Skill specialisation: whether help given builds the helper's skill. Season length: working days per season (40–120; headline 90); its index on years is a unit conversion; it has no effect on working days (Appendix S2). Number of cohorts: 10 in P10, varied 5–15. Stone-source rule: recorded stone where known, else the nearest of seven classes, against nearest class only. Dressing rule: by tower count, against Webster's everywhere.

Traceable rules. Of the 74 parameter rows audited (Section 3.1; audited before the lift ruling), 25 are sourced in the primary text, 2 known only through a second author and 12 derived; 35 are unsupported, of which 5 were specified by the authors and 30 were the research agent's design choices accepted by the authors. The unsupported rows form an agenda for fieldwork and experiment, and distinguishing our assumptions from the research agent's makes the human–machine division of interpretive labour visible, which algorithmic tools in archaeology often hide [11].

Literature and new ideas. The 13 scenarios of Fig. 2 that implement published positions directly (Section 4.1) are reported apart from the 13 indirect ones and three of our own (Section 4.2), following Chamberlin's [15] demand that rival hypotheses be developed in parallel and heeding the warning against a rhetorical middle ground [1]. Because the research agent does the implementation, a field's positions can be tested together, which eases the second limit of ABM (Section 2.4) without removing it: the realism ratings and 30 parameter values are the research agent's proposals accepted by the authors, so authorship bias moves to the review of those proposals.

Parameters. The model has more than 70 parameters, but they do not let it produce any outcome, because none is tuned to fit. Sourced parameters are fixed, the rest varied over stated ranges rather than calibrated, and screening and variance decomposition (Section 3.2.5) let the analysis, not the modeller, decide which mechanisms matter [87, 88]. Every switchable unsourced parameter was screened, and those left free change the results.

Cross-vendor docking. Although the research agent ran on a proprietary LLM (Claude Opus 5.5), arm 2 shows that an open-weight LLM can do the same work: GLM-5.3-Flash re-implemented the ABM from its specification and, once its bugs were fixed, matched the reference's man-hours within ±3% in 47 of 48 engine-A cells, so those labour results can be replicated without a proprietary-LLM subscription. Its weights can be archived and pinned [89], so arm 2's docking agent can be rerun years later on institutional hardware, without vendor deprecation, at lower cost and open to inspection [90]. With one run per LLM, this suggests but does not establish vendor independence.

Reproducibility. Archaeological computation is rarely exposed for others to inspect or reproduce [91]. Both re-implementations matched the headline man-hours within ±3% (Section 4.3). Docking exposed weaknesses and ambiguities in our instructions (start order, request timing); the remaining gaps are mostly readings of unclear instructions, apart from coding errors in arm 2. Docking thus verifies the code, catches bugs, tests whether results replicate, checks the logic of the ideas and helps rule out pasts. Independent docking could become routine for archaeological ABMs.

A transferable protocol. Our steps form a protocol that other archaeological modellers can adopt:

1. The research agent searches the literature, and humans curate the theories to be tested.
2. Every rule and parameter of the specification is anchored in a citation or labelled as an assumption.
3. The research agent codes the model with its own tests and audits.
4. Humans review the code, and an independent AI agent on a different tool or LLM audits it.
5. AI agents on different LLMs re-implement the model blind from the specification, through numbered, checksummed instructions. Every gap is traced, the bug fixed or the reading ruled in an addendum, and the runs repeated until criteria fixed in advance (Section 3.3) are met or each remaining gap is named. Every command is logged.
6. Sensitivity analysis, Morris screening and then Sobol indices over stated ranges, fixes or drops factors whose effect is negligible and keeps those that matter.

5.6 | Challenges

Validation. Docking shows that independent codes reproduce the specified model, not that the model describes Bronze Age Sardinia. The only direct comparison point is Webster's single-tower estimate (Section 4.1.1); build spans, work parties and help have no excavated counterpart. Scenarios can now be generated faster than the field supplies evidence to judge them, and a scenario that fits is sufficient, not proven. The results are therefore bounds and expectations, not full reconstructions.

LLM error modes. LLMs produce fluent text that can be unfaithful to its sources [92]. In our audit, errors were mostly misreadings of a specification, a wrong input table, a mechanism left unimplemented, and a comparison script that took the wrong measure of island time. They are caught by the checks of Section 3.3, and the research agent's documentation needs them as much as its code: about one claim in eleven of the model description was wrong or unverifiable (Section 3.3). Confirming a citation is not confirming a number.

Cost and compute. The variance-based analysis alone took about 350 hours of summed run time, and LLM inference adds a cost already flagged as a constraint [4] (hardware in Appendix S1). Human hours were not recorded and inference use only roughly (Appendix S1), so the saving is not quantified.

Over-trust. Decision aids that are usually right are the hardest to second-guess [93]. The safeguards are structural: the automatic checks on numbers and public code, independent re-implementation, and reports that keep untraced differences visible.

Human oversight. The research agent shifts disciplinary practice from building models to judging them, and its speed can make the human the bottleneck. Human checks were by the authors and an independent AI agent: independent of the research agent, not of the research team.

5.7 | Future research

The first priority is to link the model to fieldwork. A radiocarbon dating programme on selected Sardinian sites, targeting construction sequences, would test the build spans and phasing directly and give the model its first time-based target. Experimental haulage and lifting would constrain the physical rates to which island labour is most sensitive, and quarry studies would test the stone-sourcing rules. Once such targets exist, calibration methods that compare simulated with observed patterns [88] can replace the present bounds.

The second priority is repeated testing. Because the workflow is cheap to rerun, the cycle of scenario, test, evidence and revision can be repeated as dates and experiments arrive. Each cycle can narrow which ideas remain consistent with the evidence, as Section 5.3 does for the chronology. Coupling an agentic decision engine to the simulated agents would test the remaining part of the agentic-ABM programme [4]. Its call for an ODD [59] extension for agentic-ABM could be met by recording, as here, the provenance of each rule and which party, human or AI agent, proposed it. Docking came close, with 47 of 48 engine-A cells within ±3% of reference man-hours in both arms, and its small discrepancies suggest that the instructions need a more rigid vocabulary for the behaviour of the simulated agents, with controlled terms for timing and order (when help is requested, which site starts first), so that models dock more closely; we are developing one within a formal docking protocol.

6 | Conclusion

Archaeology has many theories and few means of testing them, especially for the case-based and meso-scale propositions most archaeologists use. We show that agentic-ABM can close part of that gap. A research agent, under the researchers' direction, turned the literature on how the nuraghi were built into runnable scenarios kept apart from our own, and independent AI agents on different LLMs re-implemented the resulting model blind and reproduced its labour results in sampled runs of both engines. The model docked on most measures; the few remaining differences, chiefly arm 2's construction time, trace to unclear instructions and coding errors that a clearer specification protocol would resolve. Ideas argued in prose for decades could thus be compared on equal terms, against the same data and constraints.

Such tests deserve trust only within a rigorous framework, whose elements the protocol of Section 5.5 sets out. Here, independent AI agents exposed reference errors that its own tests had missed, and human rulings settled questions of interpretation the AI agents could not. The decisive tests remain in the field, in dated construction sequences and in stone experiments. Even without full validation, the model shows which pasts are more likely under its assumptions and constraints: it rules out the weakest hauling, finds building without help in ten cohorts and simultaneous building less likely, and finds faster phasing consistent with the expansion phase. Agentic-ABM can state what these pasts predict and repeat the exercise as new evidence arrives.

These conclusions have limitations. (1) Many rules and parameters are flagged FREE or ASSUMPTION (Appendix S3, "Parameter provenance"), among them the share of residents who work and most behavioural parameters, so results are conditional on these choices. (2) The realism ratings behind MR are judgements, not results. (3) The modern lithology map and surface model are proxies for Bronze Age sources and terrain. Complexes are modelled by tower count only, imputed for 52.5% of sites; bastions, curtain walls, villages, maintenance, later enlargement and other monuments such as tombs and wells are not modelled, so total labour is a lower bound. (4) Phasing is a stylised stand-in for the chronology, compatible with the expansion phase [74] but not a date range. (5) The building stone is recorded for 31% of sites; elsewhere a seven-class rule applies. (6) Code errors were found by independent checks, not by the agentic workflow alone (Appendix S3), and further errors cannot be excluded.

Supplementary Data

The supplementary material and the project links are:

- Appendix: accompanies this article, in four sections: S1, methods (the AI agents, software, hardware, the full ABM description and the scenario catalogue); S2, sensitivity analyses and supporting results tables; S3, verification and audits; S4, the command log.
- Nurnet: the archaeological and environmental data used are Nurnet's data [12]; Nurnet website, https://www.nurnet.net/; Nurnet GeoPortal, https://geo.nurnet.net/.
- Model code: the Repast4py implementation, configuration, tests, Google Colab notebooks that reproduce the runs in a browser, the full ODD, the parameter audit, the input-data sources and the sensitivity results are in the repository, https://github.com/maltaweel/Repast_Nuraghe; Appendix S1 names the folder for each.
- Docking package: the written specification given to the arms, its addenda a–k and the numbered instruction log.
- Agent layer: Vesta, the open-source agentic harness on which the five AI agents of the study ran (MIT licence; the audit agents ran in OpenCode), https://github.com/elyxlz/vesta [62].

AI-Use Disclosure

The research agent ran on Claude Opus 5.5 (Anthropic). The two independent re-implementations used for docking ran on Claude Opus 5.5 (Anthropic) and GLM-5.3-Flash (Zhipu AI). The independent code audit used OpenCode v2.2.4, an AI coding tool, with its Big Pickle coding LLM. The authors wrote, reviewed and edited all the text; the research agent also edited text, and the authors reviewed every such edit.

Author Contributions

MA designed the paper, checked and tested code, and directed the overall research. LP managed the Vesta deployment used for AI-agent management, helped develop it, and managed the deployment of the AI agents on hardware. MA, LP, ZZK and AK each contributed to planning the study, evaluating code, writing the manuscript and checking the results.

Acknowledgments

We thank Abu Dhabi University for funding support. Special thanks go to Nurnet APS for its data (https://www.nurnet.net/), to Antonello Gregorini, who first created Nurnet and laid the basis for keeping its data clean, and to Giorgio Valdès, who over the years has been the lead reviewer of the GIS data that underpin this study. We also thank Laura Muscas and Roberto Demontis of CRS4 (Center for Advanced Studies, Research and Development in Sardinia), who helped Nurnet build the new GeoPortal on which this study is based.

Conflicts of Interest

MA is a guest editor of this Special Issue and was not involved in handling this manuscript.

Data Availability Statement

The archaeological and environmental data are Nurnet's data, available through the Nurnet GeoPortal (https://geo.nurnet.net/). The model code is available at https://github.com/maltaweel/Repast_Nuraghe. The supplementary appendices (S1–S4) accompany the journal submission and are available from the authors on request.

Abbreviations

A0–A5: alliance rule sets between communities (A0, no help; A1, requests accepted at a fixed rate; A2, reciprocity and trust; A3, trust and learning; A4, trust with trackways; A5, trust, learning, trackways and shared quarries; A0, A2, A3 and A5 are labelled A0_alone, A2_trust, A3_trust_learning and A5_full in Fig. 2; pairs of communities whose mutual trust reaches the alliance threshold, 0.7 both ways, count as allied); ABM: agent-based modelling (or model); agentic-ABM: agentic agent-based modelling; AI: artificial intelligence; ASSUMPTION, DERIVED, FREE: configuration flags for a parameter that is assumed, derived from sourced values, or not constrained by the literature; BCE: before the Common Era; CE: Common Era; cell: one sub-scenario run with 100 seeds (a scenario, one row of the run list, the list of all runs, may have several sub-scenarios, such as P10, P20 and P100 or the points of a factorial); CPU: central processing unit; H: high (bound scenario); L: low (bound scenario); LLM: large language model; MH: man-hours (= person-hours, ph; all workers, women included); MPI: Message Passing Interface; MR: most realistic (scenario), with the suffixes _alone (no help between communities) and _handover (help takes over whole stages), combined with the phasing codes (e.g. MR_P10, MR with P10 phasing); ODD: Overview, Design concepts and Details (protocol); ph: person-hours (= man-hours); P10, P20, P100: phasing in which nuraghi start in successive random cohorts of 10% or

20% of sites, or all at once; R: realistic (technique scenario); S1–S8: steps of the literature family, not Appendix S1–S4 (S1, getting stone; S2, dressing; S3, hauling; S4, lifting; S5, stone source; S6, density and hardness; S7, Webster's class totals, his stage totals per tower-size class; S8, learning), with a letter for the alternative tested, named in the figure label (e.g. S3a_rails_tallow, sledges on tallowed rails); S_T, S_1: total-order and first-order Sobol indices.

References

1. M. H. Johnson. "On the nature of theoretical archaeology and archaeological theory." *Archaeological Dialogues* 13, no. 2 (2006): 117–132. doi: 10.1017/S138020380621208X.
2. S. Plog. "Questions, theory and concepts." *Archaeological Dialogues* 13, no. 2 (2006): 156–158. doi: 10.1017/S1380203806262081.
3. L. M. Raab and A. C. Goodyear. "Middle-range theory in archaeology: a critical review of origins and applications." *American Antiquity* 49, no. 2 (1984): 255–268. doi: 10.2307/280018.
4. M. Altaweel and M. Forte. "Archeological theory and AI: the potential past through agentic systems." *AI & Society* (2026). Online first, published 29 August 2026. doi: 10.1007/s00146-026-03287-0.
5. G. Lucas and B. Olsen. "The case study in archaeological theory." *American Antiquity* 87, no. 2 (2022): 352–367. doi: 10.1017/aaq.2021.151.
6. A. J. Williams and A. Mesoudi. "A formal test using agent-based models of the circumscription theory for the evolution of social complexity." *Journal of Archaeological Science* 172 (2024): 106090. doi: 10.1016/j.jas.2024.106090.
7. J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein 2023. "Generative agents: interactive simulacra of human behavior." In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, 1–22. New York: ACM. doi: 10.1145/3586183.3606763.
8. C. Gao, X. Lan, N. Li, Y. Yuan, J. Ding, Z. Zhou, F. Xu, and Y. Li. "Large language models empowered agent-based modeling and simulation: a survey and perspectives." *Humanities and Social Sciences Communications* 11, no. 1 (2024): 1259. doi: 10.1057/s41599-024-03611-3.
9. J. M. Epstein. "Agent-based computational models and generative social science." *Complexity* 4, no. 5 (1999): 41–60. doi: 10.1002/(SICI)1099-0526(199905/06)4:5<41::AID-CPLX9>3.0.CO;2-F.
10. A. T. Crooks and A. J. Heppenstall 2012. "Introduction to agent-based modelling." In *Agent-Based Models of Geographical Systems*, edited by A. J. Heppenstall, A. T. Crooks, L. M. See, and M. Batty, 85–105. Dordrecht: Springer. doi: 10.1007/978-90-481-8927-4_5.
11. J. Huggett. "Algorithmic agency and autonomy in archaeological practice." *Open Archaeology* 7, no. 1 (2021): 417–434. doi: 10.1515/opar-2020-0136.
12. Nurnet APS 2026. "Nurnet – La rete dei Nuraghi." https://www.nurnet.net/, Website, database and GeoPortal (https://geo.nurnet.net/) of the monuments of Sardinia; accessed 10 November 2026.
13. I. Hodder 2014. "Post-processual archaeology." In *Encyclopedia of Global Archaeology*, edited by C. Smith, 6074–6083. New York: Springer. doi: 10.1007/978-1-4419-0465-2_269.
14. S. Crooks 2014. "Middle-range theory in archaeology." In *Encyclopedia of Global Archaeology*, edited by C. Smith, 4929–4931. New York: Springer. doi: 10.1007/978-1-4419-0465-2_266.
15. T. C. Chamberlin. "The method of multiple working hypotheses." *Science* ns-15, no. 366 (1890): 92–96. doi: 10.1126/science.ns-15.366.92.
16. L. J. Lucero 2014. "Hypothesis testing in archaeological theory." In *Encyclopedia of Global Archaeology*, edited by C. Smith, 3623–3629. New York: Springer. doi: 10.1007/978-1-4419-0465-2_288.
17. K. W. Kintigh, J. H. Altschul, M. C. Beaudry, R. D. Drennan, A. P. Kinzig, T. A. Kohler, W. F. Limp, H. D. G. Maschner, W. K. Michener, T. R. Pauketat, P. Peregrine, J. A. Sabloff, T. J. Wilkinson, H. T. Wright, and M. A. Zeder. "Grand challenges for archaeology." *American Antiquity* 79, no. 1 (2014): 5–24. doi: 10.7183/0002-7316.79.1.5.
18. G. Ugas 2005. *L'alba dei nuraghi*, Cagliari: Fabula.
19. G. S. Webster 2015. *The Archaeology of Nuragic Sardinia*, Sheffield: Equinox. doi: 10.1558/isbn.9781781795309.
20. P. Melis 2003. *Civiltà nuragica*, Sassari: Carlo Delfino.
21. G. Lilliu 2009. *Sardegna nuragica* (Second ed.), Nuoro: Il Maestrale.
22. G. S. Webster 1996. *A Prehistory of Sardinia 2300–500 BC*, Sheffield: Sheffield Academic Press.
23. R. Araque González 2021. "Monuments of cooperating communities: Sardinian nuraghi and sanctuaries." In *Islands in Dialogue (ISLANDIA)*, edited by G. Albertazzi, G. Muti, and A. Saggio, 193–200. Rome: Editoriale Artemide.
24. G. S. Webster 2001. *Duos Nuraghes: A Bronze Age Settlement in Sardinia. Vol. 1: The Interpretive Archaeology*, Number 949 in BAR International Series. Oxford: BAR Publishing. doi: 10.30861/9781841712413.
25. G. S. Webster. "Monuments, mobilization and Nuragic organization." *Antiquity* 65, no. 249 (1991): 840–856. doi: 10.1017/S0003598X0008056X.
26. A. Stiglitz, B. Puliga, A. Usai, S. Carboni, and L. Lecca 2012. "Il complesso di S'Urachi e l'insediamento di Su Padrigheddu (San Vero Milis, OR)." In *Atti della XLIV Riunione Scientifica dell'IIPP: La preistoria e la protostoria della Sardegna (Cagliari, Barumini, Sassari, 23–28 novembre 2009)*, Vol III, 921–926. Florence: Istituto Italiano di Preistoria e Protostoria.
27. F. Lo Schiavo, M. Perra, A. Usai, F. Campus, V. Leonelli, and P. Bernardini. "Sardegna: le ragioni dei cambiamenti nella civiltà nuragica." *Scienze dell'Antichità* 15 (2009): 265–289.
28. E. Contu 2002. *Il Nuraghe Santu Antine*, Sassari: Carlo Delfino.
29. G. Paglietti and A. Mulas. "Statica ed estetica. Architetture isodome e bicrome in età nuragica." *Layers* 8 (2023): 225–254.
30. F. Laner 1999. *Accabadora. Tecnologia delle costruzioni nuragiche*, Milan: FrancoAngeli.
31. W. G. Cavanagh and R. R. Laxton. "An investigation into the construction of Sardinian nuraghi." *Papers of the British School at Rome* 55 (1987): 1–74. doi: 10.1017/S0068246200008941.
32. A. Bortolussi, V. Dentoni, C. Levanti, S. Cara, F. Pinna, and B. Grosso. "A study on the stability and construction techniques of nuragic towers." *International Journal of Architectural Heritage* (2025): 1–16. Online first. doi: 10.1080/15583058.2025.2595146.
33. P. Bernardini. "Le torri di Monte Prama: i nuraghi nel paesaggio culturale dell'età del Ferro." *Layers* 1 (2016): 66–85.
34. M. Perra. "Osservazioni sull'evoluzione sociale e politica in età nuragica." *Rivista di Scienze Preistoriche* 59 (2009): 355–368.
35. D. Schirru and A. Vanzetti. "Climbing the Giara: a quantitative reassessment of movement and visibility in the Nuragic landscape of the Gesturi plateau (South-Central Sardinia, Italy)." *PLOS ONE* 18, no. 8 (2023): e0289023. doi: 10.1371/journal.pone.0289023.
36. M. Rothenberg, T. P. Leppard, E. Weitzel, and E. A. Murphy. "Settlement organization and distribution in Bronze Age Sardinia: utilizing cumulative viewshed analysis and spatial statistics in the Sulcis Plain." *Journal of Field Archaeology* 49, no. 5 (2024): 356–370. doi: 10.1080/00934690.2024.2318682.
37. S. Stoddart, E. D. Aines, and C. Malone (Eds.) 2021. *Gardening Time: Monuments and Landscape from Sardinia, Scotland and Central Europe in the Very Long Iron Age*, Cambridge: McDonald Institute for Archaeological Research.
38. C. J. Erasmus. "Monument building: some field experiments." *Southwestern Journal of Anthropology* 21, no. 4 (1965): 277–301. doi: 10.1086/soutjanth.21.4.3629433.

39. D. R. Turner 2020. *Grave Reminders: Comparing Mycenaean Tomb Building with Labour and Memory*, Leiden: Sidestone Press.
40. E. M. Abrams 1994. *How the Maya Built Their World: Energetics and Ancient Architecture*, Austin: University of Texas Press. doi: 10.7560/704619.
41. J. DeLaine 1997. *The Baths of Caracalla: A Study in the Design, Construction, and Economics of Large-Scale Building Projects in Imperial Rome*, Number 25 in Journal of Roman Archaeology Supplementary Series. Portsmouth, RI: Journal of Roman Archaeology.
42. L. McCurdy and E. M. Abrams (Eds.) 2019. *Architectural Energetics in Archaeology: Analytical Expansions and Global Explorations*, London: Routledge. doi: 10.4324/9781315109794
43. A. Brysbaert. "Set in stone? Socio-economic reflections on human and animal resources in monumental architecture of Late Bronze Age Tiryns in the Argos Plain, Greece." *Arctos* 47 (2013): 49–96.
44. Y. Boswinkel 2021. *Labouring with Large Stones: A Study into the Investment and Impact of Construction Projects on Mycenaean Communities in Late Bronze Age Greece*, Leiden: Sidestone Press.
45. J. Richards and M. Whitby. "The engineering of Stonehenge." *Proceedings of the British Academy* 92 (1997): 231–256.
46. M. Wunderlich, J. Müller, and A. Behrens. "Creating world views: work-expenditure calculations for Funnel Beaker megalithic graves and flint axe head depositions in Northern Germany." *Open Archaeology* 10, no. 1 (2024): 20220357. doi: 10.1515/opar-2022-0357.
47. C. M. Macal and M. J. North 2014. "Introductory tutorial: agent-based modeling and simulation." In *Proceedings of the 2014 Winter Simulation Conference*, 6–20. IEEE. doi: 10.1109/WSC.2014.7019874.
48. N. Gilbert 2020. *Agent-Based Models* (Second ed.), Quantitative Applications in the Social Sciences. Thousand Oaks, CA: SAGE. doi: 10.4135/9781506355580.
49. T. A. Kohler and G. J. Gumerman (Eds.) 2000. *Dynamics in Human and Primate Societies: Agent-Based Modeling of Social and Spatial Processes*, New York: Oxford University Press. doi: 10.1093/oso/9780195131673.001.0001.
50. M. W. Lake. "Trends in archaeological simulation." *Journal of Archaeological Method and Theory* 21, no. 2 (2014): 258–287. doi: 10.1007/s10816-013-9188-1.
51. I. Romanowska, C. D. Wren, and S. A. Crabtree 2021. *Agent-Based Modeling for Archaeology: Simulating the Complexity of Societies*, Santa Fe, NM: SFI Press.
52. T. J. Wilkinson, J. H. Christiansen, J. Ur, M. Widell, and M. Altaweel. "Urbanization within a dynamic environment: modeling Bronze Age communities in Upper Mesopotamia." *American Anthropologist* 109, no. 1 (2007): 52–68. doi: 10.1525/aa.2007.109.1.52.
53. J. D. Rogers and W. H. Cegielski. "Opinion: Building a better past with the help of agent-based modeling." *Proceedings of the National Academy of Sciences* 114, no. 49 (2017): 12841–12844. doi: 10.1073/pnas.1718277114.
54. T. A. Ngo and L. See 2012. "Calibration and validation of agent-based models of land cover change." In *Agent-Based Models of Geographical Systems*, edited by A. J. Heppenstall, A. T. Crooks, L. M. See, and M. Batty, 181–197. Dordrecht: Springer. doi: 10.1007/978-90-481-8927-4_10.
55. U. Wilensky and W. Rand. "Making models match: replicating an agent-based model." *Journal of Artificial Societies and Social Simulation* 10, no. 4 (2007): 2. https://www.jasss.org/10/4/2.html.
56. D. O'Sullivan, J. Millington, G. Perry, and J. Wainwright 2012. "Agent-based models – because they're worth it?" In *Agent-Based Models of Geographical Systems*, edited by A. J. Heppenstall, A. T. Crooks, L. M. See, and M. Batty, 109–123. Dordrecht: Springer. doi: 10.1007/978-90-481-8927-4_6.
57. W. H. Cegielski and J. D. Rogers. "Rethinking the role of agent-based modeling in archaeology." *Journal of Anthropological Archaeology* 41 (2016): 283–298. doi: 10.1016/j.jaa.2016.01.009.
58. V. Grimm and S. F. Railsback 2012. "Designing, formulating, and communicating agent-based models." In *Agent-Based Models of Geographical Systems*, edited by A. J. Heppenstall, A. T. Crooks, L. M. See, and M. Batty, 361–377. Dordrecht: Springer. doi: 10.1007/978-90-481-8927-4_17.
59. V. Grimm, S. F. Railsback, C. E. Vincenot, U. Berger, C. Gallagher, D. L. DeAngelis, B. Edmonds, J. Ge, J. Giske, J. Groeneveld, A. S. A. Johnston, A. Milles, J. Nabe-Nielsen, J. G. Polhill, V. Radchuk, M.-S. Rohwäder, R. A. Stillman, J. C. Thiele, and D. Ayllón. "The ODD protocol for describing agent-based and other simulation models: a second update to improve clarity, replication, and structural realism." *Journal of Artificial Societies and Social Simulation* 23, no. 2 (2020): 7. doi: 10.18564/jasss.4259.
60. L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, and J. Wen. "A survey on large language model based autonomous agents." *Frontiers of Computer Science* 18, no. 6 (2024): 186345. doi: 10.1007/s11704-024-40231-1.
61. D. A. Boiko, R. MacKnight, B. Kline, and G. Gomes. "Autonomous chemical research with large language models." *Nature* 624, no. 7992 (2023): 570–578. doi: 10.1038/s41586-023-06792-0.
62. Vesta contributors 2026. "Vesta: an open-source agent runtime." https://github.com/elyxlz/vesta, MIT licence; accessed 10 November 2026.
63. V. Grimm, U. Berger, J. M. Calabrese, A. Cortés-Avizanda, J. Ferrer, M. Franz, J. Groeneveld, F. Hartig, O. Jakoby, R. Jovani, S. Kramer-Schadt, T. Münkemüller, C. Piou, L. S. Premo, S. Pütz, T. Quintaine, C. Rademacher, N. Rüger, A. Schmolke, J. C. Thiele, J. Touza, and S. F. Railsback. "Using the ODD protocol and NetLogo to replicate agent-based models." *Ecological Modelling* 501 (2025): 110967. doi: 10.1016/j.ecolmodel.2024.110967.
64. N. Collier and J. Ozik 2022. "Distributed agent-based simulation with Repast4Py." In *Proceedings of the 2022 Winter Simulation Conference*, edited by B. Feng et al., 192–206. IEEE. doi: 10.1109/WSC57314.2022.10015389.
65. V. Spanu, E. B. Lorrai, L. Muscas, and R. Demontis. "NURNET – GEOPORTAL." *Archeomatica* 8, no. 5 (2017). http://mediageo.it/ojs/index.php/archeomatica/article/view/1669.
66. R. Demontis. "The Nurnet Geoportal an example of participatory GIS: a review after six years." *Open Access Journal of Archaeology & Anthropology* 2, no. 5 (2020). doi: 10.33552/oajaa.2020.02.000549.
67. ARPAS 2019. "Carta litologica della Sardegna, 1:25.000." https://webgis2.regione.sardegna.it/geonetwork/srv/api/records/R_SARDEG:7a2ce3e7-2686-4c10-8df9-3a3c0ad013f2, Agenzia Regionale per la Protezione dell'Ambiente della Sardegna. CC BY 4.0. Sardegna Geoportale catalogue record; direct download: https://webgis.regione.sardegna.it/scaricocartografiaETL/arpas/carta_litologica_2019.zip; accessed 11 November 2026.
68. ESA and Airbus 2022. "Copernicus DEM." European Space Agency and Airbus. Dataset. doi: 10.5270/esa-c5d3d65, GLO-30 tiles from the open AWS bucket, https://copernicus-dem-30m.s3.amazonaws.com/; accessed 11 November 2026.
69. E. W. Dijkstra. "A note on two problems in connexion with graphs." *Numerische Mathematik* 1, no. 1 (1959): 269–271. doi: 10.1007/BF01386390.
70. T. P. Wright. "Factors affecting the cost of airplanes." *Journal of the Aeronautical Sciences* 3, no. 4 (1936): 122–128. doi: 10.2514/8.155.
71. L. C. Hazell and S. M. Fitzpatrick. "The maritime transport of prehistoric megaliths in Micronesia." *Archaeology in Oceania* 41 (2006): 12–24.

72. G. Lilliu and R. Zucca 2005. *Su Nuraxi di Barumini*, Sassari: Carlo Delfino.
73. G. S. Webster and M. R. Webster. "The Duos Nuraghes Project in Sardinia: 1985–1996 interim report." *Journal of Field Archaeology* 25, no. 2 (1998): 183–201. doi: 10.2307/530578.
74. N. Ialongo. "Crisis and recovery: the cost of sustainable development in Nuragic Sardinia." *European Journal of Archaeology* 21, no. 1 (2018): 18–38. doi: 10.1017/eaa.2017.20.
75. C. M. Calò, A. Melis, G. Vona, and I. S. Piras. "Sardinian population (Italy): a genetic review." *International Journal of Modern Anthropology* 1, no. 1 (2008): 39–64. doi: 10.4314/ijma.v1i1.60356.
76. A. Moravetti, E. Alba, and L. Foddai (Eds.) 2014. *La Sardegna nuragica. Storia e materiali*, Sassari: Regione Autonoma della Sardegna / Carlo Delfino.
77. M. D. Morris. "Factorial sampling plans for preliminary computational experiments." *Technometrics* 33, no. 2 (1991): 161–174. doi: 10.1080/00401706.1991.10484804.
78. I. M. Sobol'. "Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates." *Mathematics and Computers in Simulation* 55, no. 1–3 (2001): 271–280. doi: 10.1016/S0378-4754(00)00270-6.
79. A. Saltelli, P. Annoni, I. Azzini, F. Campolongo, M. Ratto, and S. Tarantola. "Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index." *Computer Physics Communications* 181, no. 2 (2010): 259–270. doi: 10.1016/j.cpc.2009.09.018.
80. D. J. Schuirmann. "A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability." *Journal of Pharmacokinetics and Biopharmaceutics* 15, no. 6 (1987): 657–680. doi: 10.1007/BF01068419.
81. R. Axtell, R. Axelrod, J. M. Epstein, and M. D. Cohen. "Aligning simulation models: a case study and results." *Computational and Mathematical Organization Theory* 1, no. 2 (1996): 123–141. doi: 10.1007/BF01299065.
82. N. Fachada, D. Fernandes, C. M. Fernandes, and J. P. Matos-Carvalho. "Can large language models implement agent-based models? An ODD-based replication study." *Ecological Modelling* 517 (2026): 111624. doi: 10.1016/j.ecolmodel.2026.111624.
83. V. A. Traag, L. Waltman, and N. J. van Eck. "From Louvain to Leiden: guaranteeing well-connected communities." *Scientific Reports* 9, no. 1 (2019): 5233. doi: 10.1038/s41598-019-41695-z.
84. A. Moravetti, P. Melis, L. Foddai, and E. Alba (Eds.) 2017. *La Sardegna nuragica. Storia e monumenti*, Corpora delle antichità della Sardegna. Sassari: Regione Autonoma della Sardegna / Carlo Delfino. Chapter cited: P. Melis, "I nuraghi," pp. 40–42.
85. J. R. Snyder, O. Dilaver, L. C. Stephenson, J. E. Mackie, and S. D. Smith. "Agent-based modelling and construction: reconstructing antiquity's largest infrastructure project." *Construction Management and Economics* 36, no. 6 (2018): 313–327. doi: 10.1080/01446193.2017.1403639.
86. S. Heckbert. "MayaSim: an agent-based model of the ancient Maya social-ecological system." *Journal of Artificial Societies and Social Simulation* 16, no. 4 (2013): 11. doi: 10.18564/jasss.2305.
87. B. Edmonds and S. Moss 2005. "From KISS to KIDS: an 'anti-simplistic' modelling approach." In *Multi-Agent and Multi-Agent-Based Simulation (MABS 2004)*, edited by P. Davidsson, B. Logan, and K. Takadama, Vol 3415 of *Lecture Notes in Computer Science*, 130–144. Berlin, Heidelberg: Springer. doi: 10.1007/978-3-540-32243-6_11.
88. V. Grimm, E. Revilla, U. Berger, F. Jeltsch, W. M. Mooij, S. F. Railsback, H.-H. Thulke, J. Weiner, T. Wiegand, and D. L. DeAngelis. "Pattern-oriented modeling of agent-based complex systems: lessons from ecology." *Science* 310, no. 5750 (2005): 987–991. doi: 10.1126/science.1116681.
89. Zhipu AI 2026. "GLM-5.3-Flash (model weights, MIT licence)." https://huggingface.co/zai-org/GLM-5.3-Flash, Released 25 August 2026; accessed 15 November 2026.
90. A. Spirling. "Why open-source generative AI models are an ethical way forward for science." *Nature* 616, no. 7957 (2023): 413. doi: 10.1038/d41586-023-01295-4.
91. B. Marwick. "Computational reproducibility in archaeological research: basic principles and a case study of their implementation." *Journal of Archaeological Method and Theory* 24, no. 2 (2017): 424–450. doi: 10.1007/s10816-015-9272-9.
92. Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. "Survey of hallucination in natural language generation." *ACM Computing Surveys* 55, no. 12 (2023): 1–38. doi: 10.1145/3571730.
93. R. Parasuraman and D. H. Manzey. "Complacency and bias in human use of automation: an attentional integration." *Human Factors* 52, no. 3 (2010): 381–410. doi: 10.1177/0018720810376055.