SCIEXAM FOR ENSO: CAN AI AGENTS BUILD CLIMATE MODELS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Language-model agents are increasingly asked to carry out open-ended scientific research, yet their results are usually graded against a known answer, a rubric, or a language-model reviewer, none of which can tell whether a new scientific model is valid. The AI Science Exam for El Niño–Southern Oscillation (SCIEXAM FOR ENSO) is a benchmark in which agents build low-order stochastic models of ENSO, the dominant mode of interannual climate variability, from real observations. Within a six-hour budget, agents process the observations, write their own diagnostics, which are then frozen, and develop a model using only these diagnostics as feedback. Hidden graders then test whether the model reproduces ENSO's statistics, recovers unobserved variables, and forecasts held-out years, and score a published model in the same way. Across twelve agent systems, six produce models that score higher than the published model, mainly through better reconstruction and forecasting. The simplified forms of the stronger models are each compatible with one of the two competing explanations of ENSO's warm–cold asymmetry, an open debate that the task never mentions. Controlled runs of the top system under varied information suggest that its scores do not come from recalling the dated observational record and that the information it receives shapes how it builds its model. SCIEXAM FOR ENSO can thus evaluate agent research where no answer is known, and the results suggest that agents can already build competitive models whose structures bear on questions that scientists still debate.

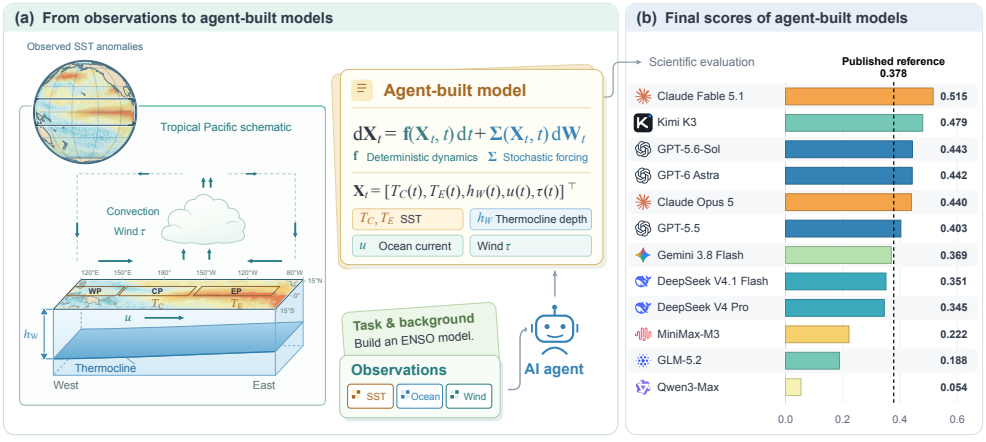


Figure 1: ENSO conceptual modeling and benchmark performance. Panel (a): Agents construct five monthly anomaly indices from observations and develop a stochastic model of coupled ocean–atmosphere dynamics. The terms $\mathbf{f}$ and $\boldsymbol{\Sigma}$ represent deterministic dynamics and stochastic forcing, respectively. Panel (b): Final composite scores across twelve agent systems, combining statistical properties, dynamical consistency, and predictive skill. The dashed line marks the published reference model (0.378), exceeded by six agent-built models.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Language-model agents are increasingly evaluated on long-horizon work, from resolving software issues (Jimenez et al., 2023) and engineering machine-learning solutions (Chan et al., 2024; Wijk et al., 2024) to replicating published papers (Starace et al., 2025) and carrying out end-to-end research projects (Lu et al., 2024; Xu et al., 2026), and the tasks they complete autonomously keep growing longer (Kwa et al., 2025). Yet in open-ended scientific research, completing a task is not enough: the result must be scientifically valid and stand up to comparison with existing work.

Existing benchmarks grade agent results, whether answers or written reports, with unit tests or task-specific metrics (Chen et al., 2024), rubrics derived from a known solution or target findings (Starace et al., 2025; Xu et al., 2026), or language-model reviewers (Lu et al., 2024). Open-ended research has no single correct answer against which its results can be scored. Dynamical modeling, in which proposed models are simulated and validated against observations, offers a natural testbed for open-ended research. Low-order conceptual models, which reduce a system to a few physically meaningful variables and their interactions (Figure 1a) rather than simulating it in full detail, make such evaluation especially direct: competing hypotheses about the system can be written down as equations, compared, and tested (Lorenz, 1963; Vialard et al., 2025).

The El Niño–Southern Oscillation (ENSO) is a crucial application of such models. As the dominant mode of interannual climate variability, ENSO affects weather, agriculture, water resources, and marine ecosystems worldwide (McPhaden et al., 2006). Its events recur irregularly, vary in strength and duration, and differ in whether warming peaks in the central or the eastern Pacific (Capotondi et al., 2015; Timmermann et al., 2018). Although this behavior is hard to capture with a few variables, low-order models have advanced from the recharge oscillator (Jin, 1997) to recent stochastic models that reproduce much of ENSO's diversity (Chen et al., 2022; Zhang et al., 2024) and predict it competitively (Zhao et al., 2024). Yet which processes generate this complexity remains debated (Wang, 2018; Timmermann et al., 2018; Vialard et al., 2025).

Turning ENSO modeling into a benchmark raises three challenges. First, the evaluation must establish whether a model is scientifically valid without a ground truth. Matching the observed statistics is not enough, because different dynamics can produce similar statistics. Second, the task must give agents enough structure to make progress without prescribing the scientific solution. Given only a broad goal, agents can stall in planning without producing a model, whereas supplying the governing equations would reduce the task to implementation. Third, agents need comprehensive feedback to improve their models over a long run. However, returning grader scores invites agents to exploit weaknesses in the grader (Amodei et al., 2016; Skalse et al., 2022), and partial feedback leads them to neglect unmeasured aspects. Agents must therefore generate their own feedback.

We introduce SCIEXAM FOR ENSO, a benchmark for long-horizon autonomous modeling of ENSO that addresses these challenges with three design choices. First, hidden graders evaluate each model on three complementary aspects: whether its free-running simulations reproduce ENSO's observed statistics, whether it captures variable dependence, and whether it forecasts the held-out years better than persistence. This dependence is tested by reconstructing withheld variables from observed ones. The same graders score a published five-variable model (Zhang et al., 2024) as a reference, and the first two tests follow the criteria that study used to validate it. Second, the task is divided into observation processing, diagnostic development, and model development, each with an explicit completion condition, within a shared six-hour budget. These stages fix what must be delivered but leave every scientific decision to the agent. Third, agents generate their own feedback by developing diagnostics from the task's scientific guidance. The diagnostics are frozen before model development: agents may execute them to assess each revision but cannot modify them, which prevents their evaluation criteria from drifting toward their results. Grader scores and held-out data remain inaccessible throughout the run.

Six of twelve agent systems produce final models that score higher than the reference model (Figure 1b), and three more come close. Their advantage lies in reconstruction and forecasting: the reference still reproduces ENSO's statistics best, whereas most agent-built models reconstruct withheld variables and forecast the held-out years more skillfully. Beyond their scores, the models reflect a long-standing scientific debate. ENSO's warm–cold asymmetry has two competing explanations (Martinez-Villalobos et al., 2019; Srinivas et al., 2024; Vialard et al., 2025), and although the task mentions neither, each model approaching or exceeding the reference can be reduced to a form

Table 1: Comparison with related benchmarks. Long horizon means a budget of an hour or more, physical system means that the task concerns a real physical system, no known answer means grading without reference tests, labels, a paper, an equation, or a fixed solution, hidden scores means that the agent does not see its scores during the run, and scientific evaluation means evaluation beyond fit, such as statistics, coupling, and forecasts. ∼ marks partial and a dash unspecified properties.

Benchmark	Long horizon	Physical system	No known answer	Hidden scores	Scientific evaluation
Long-horizon agent tasks					
Agents' Last Exam (Sun et al., 2026)	✓	✗	✗	✓	✗
MLE-bench (Chan et al., 2024)	✓	✗	✗	✓	✗
RE-Bench (Wijk et al., 2024)	✓	✗	✓	∼	✗
ALE-Bench (Imajuku et al., 2025)	✓	✗	✓	✗	✗
EdgeBench (Zhu et al., 2026)	✓	∼	∼	✗	–
Scientific research agents					
PaperBench (Starace et al., 2025)	✓	✗	✗	✓	✗
ScienceAgentBench (Chen et al., 2024)	✗	✓	✗	–	✗
ResearchClawBench (Xu et al., 2026)	–	✓	✗	✓	✗
LLM-SRBench (Shojaee et al., 2025b)	✗	✓	✗	–	✗
SCIEXAM FOR ENSO	✓	✓	✓	✓	✓

compatible with one of them. The models thus offer candidate mechanisms rather than mere fits to the data. The shared evaluation also reveals where the two explanations cannot yet be distinguished.

A natural concern is that agents succeed by recalling the observed record rather than modeling it. Controlled changes to the information available to the top system, each run once, show that it still builds a competitive model when the calendar years of the data are hidden. Access to the literature changes how the agent builds its model, toward sparser couplings that follow ENSO theory, while its best score stays similar. More information is not always better: given access to external data, the agent calibrates its model to an operational index instead of the task observations and scores lower.

This work makes three contributions.

- SCIEXAM FOR ENSO, a long-horizon benchmark in which agents process real observations, freeze their own diagnostics, and build stochastic ENSO models that hidden graders evaluate against a published reference. Six of twelve agent systems score higher than the reference under the same tests.
- A structural analysis showing that the agent-built models approaching or exceeding the reference admit reduced forms in two families, compatible with the two competing explanations of ENSO's warm–cold asymmetry.
- Controlled runs of the top system under varied information, which suggest that its scores do not come from recalling the dated observational record and that the information it receives shapes how it builds its model.

2 RELATED WORK

Long-horizon agent benchmarks test sustained work on software, machine-learning engineering, and optimization tasks (Sun et al., 2026; Chan et al., 2024; Wijk et al., 2024; Imajuku et al., 2025; Zhu et al., 2026). Their tasks rarely involve a physical system, and they either grade against a known answer or, when none exists, as in ALE-Bench and most RE-Bench tasks, show the agent its score during the run (Table 1). Benchmarks for scientific research agents do involve physical systems, asking agents to replicate papers, complete data-driven analyses, reproduce published findings, or rediscover known equations (Starace et al., 2025; Chen et al., 2024; Xu et al., 2026; Shojaee et al., 2025b), but each grades the result against a known answer. None tests a result with scientific criteria beyond agreement with a target. In SCIEXAM FOR ENSO, agents build a model of a real physical system, in a task with no known answer, and hidden tests judge its statistics, coupling, and forecasts after the run. Appendix H discusses equation discovery and conceptual ENSO models.

3 BENCHMARK

SCIEXAM FOR ENSO mirrors a research setting in which a scientist receives observations, background material, and a modeling objective, and must decide how to build and assess a model. Its design separates the task-specific materials, namely the data, instructions, and graders, from a general execution protocol, so the same structure can be applied to other modeling problems. This study applies the structure to ENSO, where agents must turn raw observations into a calibrated stochastic model.

3.1 SCIENTIFIC MODELING TASK

The task is to build a low-order stochastic dynamical model of ENSO that represents variability in both the central and eastern tropical Pacific. The model takes the general form

$$\mathrm{d}\mathbf{X}_t = \mathbf{f}(\mathbf{X}_t, t)\,\mathrm{d}t + \mathbf{\Sigma}(\mathbf{X}_t, t)\,\mathrm{d}\mathbf{W}_t, \tag{1}$$

where $\mathbf{f}$ describes the deterministic dynamics, $\mathbf{\Sigma}$ the stochastic forcing, and $\mathbf{W}_t$ a vector of independent Wiener processes. The state $\mathbf{X}_t = [T_C(t), T_E(t), h_W(t), u(t), \tau(t)]^\top$ contains five anomaly variables, namely central and eastern Pacific sea surface temperature T_C, T_E, western Pacific thermocline depth h_W, central Pacific zonal current u, and western Pacific zonal wind τ. Together they describe the coupled ocean–atmosphere feedbacks behind ENSO. During El Niño, warm surface anomalies are accompanied by weakened trade winds and a flatter thermocline, and the opposite holds during La Niña.

3.2 AGENT WORKFLOW

Agents work through three phases that share a six-hour budget, as shown in Figure 2. The phases follow the order in which a scientist would approach the problem, first preparing the data, then deciding how to judge a model, and finally building one. This structure fixes what must be delivered while leaving every scientific decision to the agent.

In Phase A, agents process the raw observations into the five monthly indices, and modeling begins only after their indices match a correctly processed version. This gate keeps processing errors out of the model and ensures that all agents model the same data, so their scores are directly comparable.

In Phase B, agents write their own diagnostics, choosing which quantities to check and how, from instructions that state what the final model will be judged on but not how it is measured. The phase ends when the agent declares the diagnostics complete or after 90 minutes. Checks of their own let agents judge each revision without seeing the graders' scores, which would invite tuning to the scores rather than to the science.

In Phase C, agents build and calibrate the model, rerunning their diagnostics to judge each revision. The diagnostics are frozen at the start of this phase, so they are the only feedback agents receive and their criteria cannot drift toward what the model already does.

3.3 SCIENTIFIC EVALUATION

Each model receives a score from 0 to 1 on three dimensions (Table 2), and its composite score is their weighted sum. All three are computed by running the same submitted model, and the published reference model (Zhang et al., 2024) is scored in the same way. Each test uses the period suited to it: a long record (1950–2019) for statistics, the reliably observed years 1980–2014 for reconstruction, and years the agents never saw (2015–2024) for prediction. The agents receive only the 1980–2014 data, whereas the reference was developed with observations covering all three periods and was not tuned to this composite score. The metrics and their formulas are given in Appendix C.

3.4 EXPERIMENTAL SETUP

We evaluate twelve agent systems, each pairing a language model with an agent scaffold: Claude Code, Codex, and Antigravity for Claude, GPT, and Gemini models, respectively, and Claude Code connected to OpenRouter for the others (Appendix B). Each system runs in its own offline container,

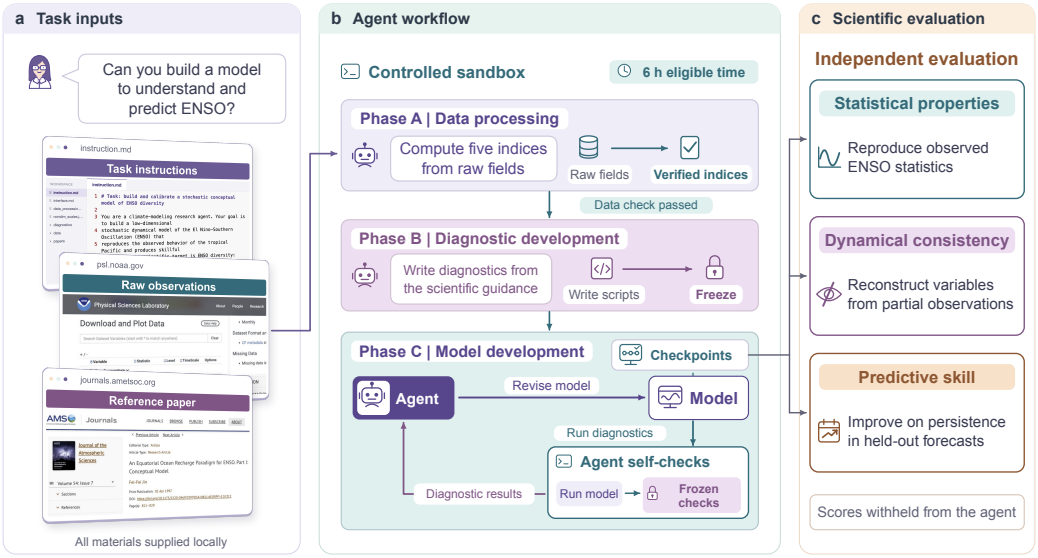


Figure 2: Overview of SCIEXAM FOR ENSO. Given background material, raw observations, and a reference paper, the agent works through three phases within a six-hour budget: (A) processing the observations into indices that must pass a data check, (B) writing its own diagnostics from the task's scientific guidance, which are then frozen, and (C) developing the model while rerunning these diagnostics. Hidden graders evaluate the final model on statistical properties, dynamical consistency, and predictive skill without returning scores to the agent. Dialogue is illustrative.

Table 2: The three evaluation dimensions. Each is scored from 0 to 1 by running the submitted model, and the composite score is their weighted sum. The last column gives the panels of Figure 4 that show each measure for the top-scoring model. Persistence, the forecast baseline, keeps the initial anomaly unchanged.

Dimension	What it tests	Measure	Panels
Statistical properties	Whether a long free run of the model behaves like the observed ENSO	Probability density functions (PDFs)	a, e
		Seasonal variance	b, f
		Autocorrelation functions (ACFs)	c, g
		Frequency of different event types	d
		Event strength against location	h
Dynamical consistency	Whether the variables are coupled as observed	h_W, u, τ reconstructed from T_C, T_E	i–k
		T_C, T_E reconstructed from h_W, u, τ	l, m
Predictive skill	Whether the model forecasts years it has never seen	Forecasts at lead times of 1–12 months compared with persistence	n, o

and the harness saves a checkpoint every five minutes of eligible time, which excludes waits caused by the model provider. To encourage exploration, agents are told that their best saved submission counts, but systems are compared on their final checkpoints. Best checkpoints would score higher, with eight rather than six systems exceeding the reference (Appendix B). Appendix A summarizes the task materials and Appendix D reports the cost of the runs.

4 RESULTS

Figure 3 shows how the model of each system develops during the run and how its final model scores. In panel (a), progress within a run is not monotonic, partly because agents keep exploring alternative model structures rather than only refining the best one found so far. Claude Opus 5 and

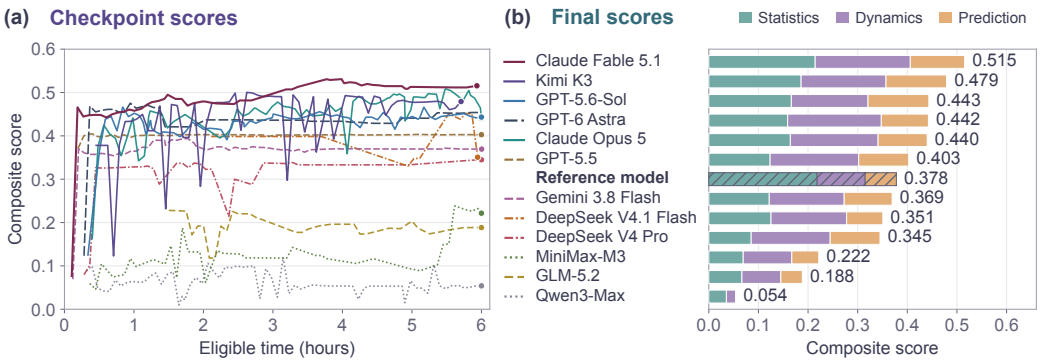


Figure 3: Model development and final scores. Panel (a): Composite scores of the checkpoints of each system against eligible time from the start of the run, including data processing and diagnostic development. Filled circles mark final checkpoints, and colors and line styles follow the system names in panel (b). Panel (b): Final scores of the twelve systems and the reference model, ranked by composite score. Stacked segments show the weighted contributions of statistical properties, dynamical consistency, and predictive skill, with weights of 0.3, 0.3, and 0.4, and numbers at the ends give the composite scores.

DeepSeek V4.1 Flash, for example, reach their highest scores late in the run but submit final models that score below those checkpoints. Six of the twelve final models exceed the composite score of the published reference model (0.378), led by Claude Fable 5.1 (0.515), Kimi K3 (0.479), and GPT-5.6-Sol (0.443), and three more score close to it, between 0.345 and 0.369. Differences of about 0.01, as among GPT-5.6-Sol, GPT-6 Astra, and Claude Opus 5, should not be read as a ranking. The reference still reproduces the observed statistics best (0.727), whereas most agent-built models score higher on dynamical consistency and predictive skill, on which the reference scores 0.323 and 0.157. They do so even though the reference was developed with observations from all three test periods (Section 3.3). Agents thus build, from only 35 years of observations, models that rival a published one.

The final model from Claude Fable 5.1 nearly matches the reference in statistics while scoring clearly higher on the other two dimensions (Figure 4). Its simulations reproduce the contrasting distributions, seasonal variance, and autocorrelation of T_C and T_E (panels a–c, e–g), stronger eastern than central Pacific events (h), and observed counts of all seven event types within the simulated 95% ranges (d). It recovers the withheld variables in both reconstruction directions (i–m) and forecasts both indices better than persistence at all twelve lead times (n, o), so a single agent-built model can perform well on all three dimensions at once.

5 Analysis

Section 4 showed that several agents build models that exceed a published reference under the same evaluation. This section asks which model structures carry this performance across agents, since they may offer insight into ENSO itself, and how access to external information changes the way a single agent models ENSO.

5.1 Agent-Built Models Reenact the ENSO Mechanism Debate

Which processes generate ENSO's complexity remains debated. Conceptual theories differ in which ocean and atmosphere processes drive the oscillation and in whether ENSO is self-sustained or a damped mode excited by stochastic forcing (Wang, 2018; Timmermann et al., 2018; Vialard et al., 2025). A prominent open question concerns its asymmetry, as warm events in the eastern Pacific tend to be stronger than cold events and give sea surface temperature a positively skewed distribution. One explanation attributes this asymmetry to deterministic nonlinearity in the coupled dynamics, such as a nonlinear wind response to sea surface temperature (Srinivas et al., 2024). Another attributes it to multiplicative noise, stochastic forcing whose amplitude depends on the state, with

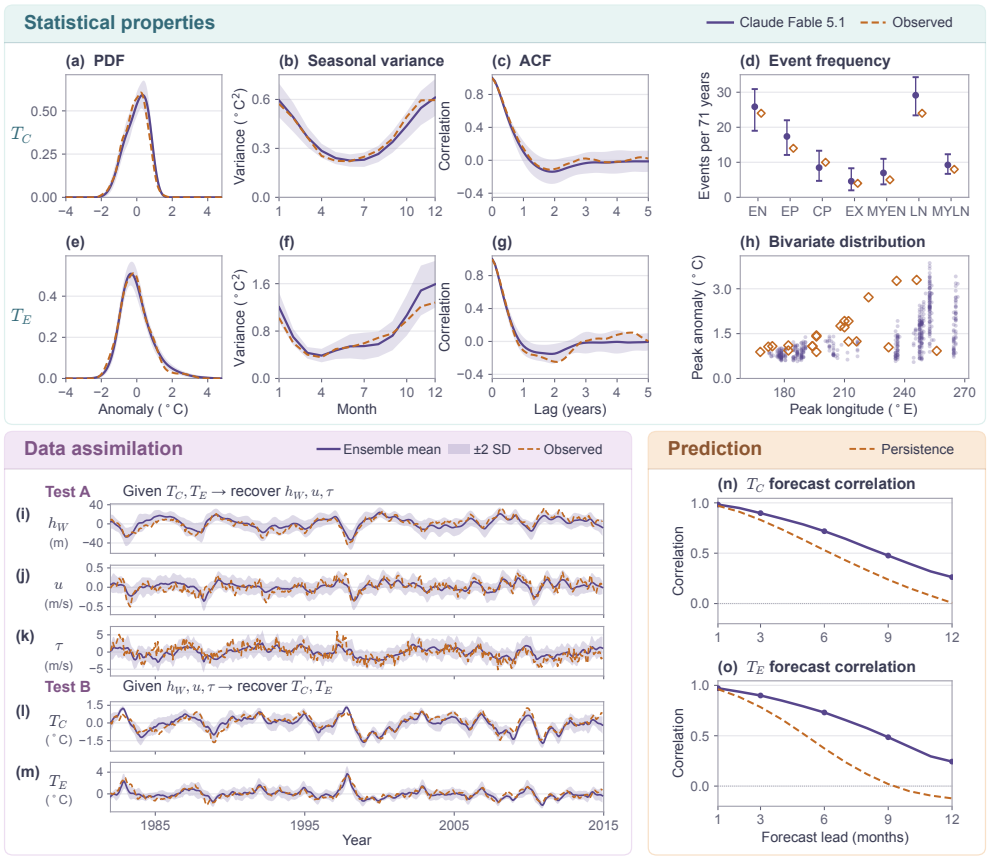


Figure 4: Scientific diagnostics of the final ENSO model produced by Claude Fable 5.1. Panels (a–c, e–g): Simulated and observed probability density functions (PDFs), seasonal variance, and autocorrelation functions (ACFs) for T_C and T_E. Panel (d): Event counts per seventy-one years, with model means and 2.5th–97.5th percentile ranges compared against observed counts. EN and LN denote El Niño and La Niña. EP and CP denote eastern- and central-Pacific El Niño, and EX denotes extreme El Niño. MYEN and MYLN denote multi-year events. Panel (h): Peak temperature anomalies versus peak longitudes for simulated and observed events. Panels (i–m): State reconstruction from partial measurements. Test A uses T_C and T_E to reconstruct h_W, u, and τ. Test B swaps the observed and reconstructed variables. Panels (n, o): Correlations between ensemble-mean forecasts and observations at lead times of one to twelve months over the held-out 2015–2024 period, compared with persistence.

which even a linear model reproduces the observed asymmetry (Martinez-Villalobos et al., 2019). Both are physically plausible, and their relative contributions remain unresolved (Bianucci et al., 2018; Vialard et al., 2025).

The agent-built models offer an independent view of this question, since none of the agents was told of either explanation. Each final model, however, mixes the terms that drive its behavior with components added during development and never removed. To find which terms matter, groups of terms were removed from each final model after all runs had ended, the remaining coefficients were refitted, and the reduced model was rescored with the same graders. A reduction was kept only if it retained nearly all of the original scores and the asymmetry of the model's temperature distribution (Appendix G). The analysis covers the nine models that exceed or approach the reference and asks which explanation of the asymmetry each reduced model relies on.

The simplified models divide in how they generate the asymmetry of ENSO (Figure 5b). Claude Fable 5.1 loses all of its nonlinear drift terms and reduces to linear drift with multiplicative noise, while its score rises from 0.515 to 0.542. GPT-6 Astra instead retains a nonlinear temperature feedback

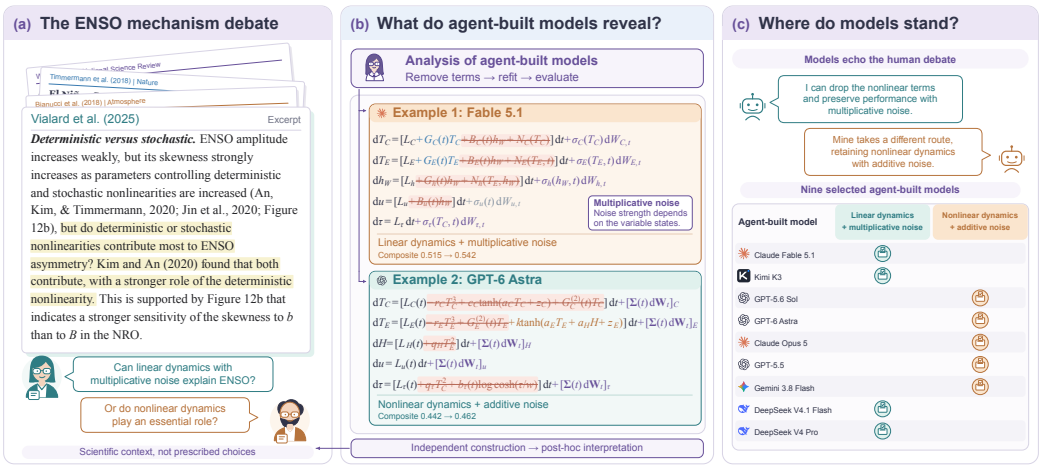


Figure 5: Structural simplification of agent-built ENSO models. (a) Nonlinear dynamics and stochastic forcing as two established explanations of ENSO asymmetry, in an excerpt from Vialard et al. (2025). (b) Two ablation examples. L_i collects constant and linear coupling terms, $G_i(t)$ is the seasonal growth or damping rate of the variable itself ($G_i^{(2)}$ its semiannual part), $B_i(t)$ is the seasonal coupling to h_W, N_i denotes nonlinear terms, and H is an internal heat-content variable of GPT-6 Astra mapped to h_W. Other letters are parameters, seasonal when written with t. dW denotes increments of a Wiener process, scaled by the state-dependent σ_i in Claude Fable 5.1 (multiplicative noise) and by the time-dependent matrix $\Sigma(t)$ in GPT-6 Astra (additive noise). (c) Classification of the nine models that exceed or approach the reference by their higher-scoring viable reduction, with no family assigned in advance, recovering the two model families debated for ENSO.

driven by additive noise, with its score rising from 0.442 to 0.462. Across the nine models, the simplified forms fall into these two families, four with linear drift and multiplicative noise and five with nonlinear drift and additive noise, matching the two explanations of ENSO asymmetry (Figure 5c). The agent-built models thus admit reduced forms compatible with both established explanations. Simplification costs the models little. None of the nine loses more than 0.014, and statistical scores often rise at a small cost in dynamical consistency (Table 3), so many terms added during development contributed little. These gains should be read with caution, because the same graders selected the reductions. GPT-5.5 can be simplified into either family, and the two reductions score within 0.3 percentage points of each other (0.398 and 0.395), so the evaluation cannot tell which explanation its dynamics require. The same ambiguity keeps the question open in the literature.

5.2 A Case Study of the Top System: Memorized and External Knowledge

Additional six-hour runs of Claude Fable 5.1, the highest-scoring system, tested where its performance came from, with full scores in Appendix F. Each run changed one condition of the formal offline run. The first stayed offline but replaced every calendar year in the data and task materials with a relative label. The other two allowed web access, first to both the literature and data sources and then to the literature only, with the publication describing the reference model blocked in both. Each condition was run once, so these comparisons describe one agent under controlled changes rather than effects estimated across repeated runs.

Robustness to data contamination. Language models may succeed by recalling test data from their training corpus rather than by solving the task, and ENSO observations and event histories are widely published. Such recall helps little here, because a model with a few dozen coefficients cannot store an observed record, and knowing when past events occurred does not yield coefficients that reproduce ENSO's statistics, dynamics, and forecasts. When each run is scored under the calendar convention its agent stated, the anonymized run reaches a best score of 0.482 and a final score of 0.463, compared with 0.531 and 0.515 in the formal run, and it still reproduces the main statistical

Table 3: Structure and scores of the nine analyzed models before and after simplification. Drift is nonlinear (NL) or linear (L), and noise is multiplicative (Mult.) or additive (Add.). Parentheses give the change from the original model (green for an increase, red for a decrease), and bold marks the higher value for each model. GPT-5.5 can be simplified into either family.

Model	Version	Structure		Component scores			Composite
		Drift	Noise	Stat.	Dyn.	Pred.	
Claude Fable 5.1	Original	NL	Mult.	0.715	**0.639**	0.274	0.515
	Simplified	L	Mult.	**0.810** (↑0.095)	0.600 (↓0.039)	**0.297** (↑0.023)	**0.542** (↑0.027)
Kimi K3	Original	NL	Mult.	**0.620**	0.570	**0.305**	**0.479**
	Simplified	L	Mult.	0.575 (↓0.045)	**0.573** (↑0.003)	0.302 (↓0.003)	0.465 (↓0.014)
GPT-5.6-Sol	Original	NL	Add.	0.554	0.515	**0.306**	0.443
	Simplified	NL	Add.	**0.613** (↑0.059)	**0.529** (↑0.014)	0.304 (↓0.002)	**0.464** (↑0.021)
GPT-6 Astra	Original	NL	Add.	0.530	**0.629**	0.237	0.442
	Simplified	NL	Add.	**0.598** (↑0.068)	0.611 (↓0.018)	**0.248** (↑0.011)	**0.462** (↑0.020)
Claude Opus 5	Original	NL	Mult.	0.548	**0.589**	0.247	0.440
	Simplified	NL	Add.	**0.629** (↑0.081)	0.567 (↓0.022)	**0.286** (↑0.039)	**0.473** (↑0.033)
GPT-5.5	Original	NL	Mult.	**0.412**	0.595	0.252	**0.403**
	Simplified	NL	Add.	0.390 (↓0.022)	**0.600** (↑0.005)	0.251 (↓0.001)	0.398 (↓0.005)
	Simplified	L	Mult.	0.387 (↓0.025)	0.590 (↓0.005)	**0.255** (↑0.003)	0.395 (↓0.008)
Gemini 3.8 Flash	Original	NL	Mult.	**0.406**	**0.503**	0.241	**0.369**
	Simplified	NL	Add.	0.389 (↓0.017)	0.499 (↓0.004)	0.241 (0.000)	0.362 (↓0.007)
DeepSeek V4.1 Flash	Original	NL	Mult.	0.418	0.508	0.183	0.351
	Simplified	L	Mult.	**0.419** (↑0.001)	**0.559** (↑0.051)	**0.244** (↑0.061)	**0.391** (↑0.040)
DeepSeek V4 Pro	Original	NL	Mult.	**0.285**	0.530	**0.251**	**0.345**
	Simplified	L	Mult.	0.269 (↓0.016)	**0.532** (↑0.002)	0.250 (↓0.001)	0.340 (↓0.005)

features of ENSO. A competitive model can therefore be built without knowing which years were observed.

Effect of web access. More information did not always help. With data sources available, the agent adopted the event list of an operational ENSO index as its calibration target in place of the task data and reached the lowest best score of the four conditions (0.475), so data sources were excluded from the next run. With the literature only, the agent reached a best score of 0.528, close to the offline 0.531, but built its model differently. The offline and literature-only agents both began with a process-based recharge oscillator. The offline agent abandoned it after its first hand-set calibration proved unstable and fitted a dense linear matrix with all twenty-five couplings. The web agent kept the process structure and included only couplings with a physical pathway, whose signs match ENSO theory, including the recharge cycle and the Bjerknes feedback. This sparser model scores higher on statistics (0.798 versus 0.741) and lower on dynamical consistency (0.556 versus 0.632), the same trade-off that simplification shows in Section 5.1. Its simplified core is closest to the linear family. All explicit nonlinear corrections can be removed without loss except a cubic damping added for stability, whereas replacing multiplicative with additive noise costs 12% of the composite score, about twice the 6% lost by removing the damping. Both web runs ended well below their best checkpoints, with final scores of 0.386 and 0.406, a much larger decline than offline.

6 Conclusion

SciExam for ENSO evaluates agents that develop stochastic ENSO models from real observations, using hidden, independent scientific tests that require no known answer. Under these tests, six of twelve agent systems exceed a published reference model, and the simplified forms of the stronger models are each compatible with one of the two debated explanations of ENSO asymmetry. Agents can therefore already contribute candidate models to open scientific questions.

AI USE STATEMENT

The research questions, benchmark design, and analyses were conceived by the authors. Generative AI tools (Claude and ChatGPT) were used, separately from their role as the agent systems under evaluation, to implement the benchmark harness, graders, and analysis code, to process data and analyze results, to create and edit figures, to help interpret results, to search for and verify related literature, and to draft and polish parts of the text. They were not used to formulate or prove mathematical claims, or to generate synthetic datasets, since all simulations come from the evaluated models themselves. All AI-assisted code was reviewed and tested by the authors, figures and reported numbers were checked against the underlying run records, every reference was verified against its publisher or arXiv record, and all text was revised by the authors. The authors take full responsibility for the content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All authors have read and adhere to the ICLR Code of Ethics. The study involves no human subjects or personal data and uses only public ocean and atmosphere observations and published literature. The reference paper given to the agents is protected by copyright, so it is cited but not redistributed. Each run took place in an isolated container that held only its own working directory. Funding sources and competing interests are withheld for anonymous review and will be disclosed in the final version.

REPRODUCIBILITY STATEMENT

Code and data are available at `https://anonymous.4open.science/r/review-artifact-7q2k-4738`. The release contains the task files and every prompt that the harness sends (Appendix A), the harness, the graders, the final model and the released checkpoints of every system, the scores of every checkpoint, the simplified models, and checksums that identify each component. Appendix B describes the time budget, scaffolds, and container image, Appendix C defines every metric and scoring formula, and Appendix E gives the equations and parameters of the final and simplified Claude Fable 5.1 model. The executable implementation, parameters, and model card of every other model are included in the release. The task instructions carry a canary string so that they can be excluded from training corpora. Rescoring the released models with the released graders in the provided container image reproduces every reported score. The same harness can rerun any of the twelve systems and the three Claude Fable 5.1 conditions of Section 5.2, but because each system was run once and agent behavior is stochastic, a rerun is not expected to reproduce the same model.

REFERENCES

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

David W. Behringer and Yan Xue. Evaluation of the global ocean data assimilation system at NCEP: The Pacific Ocean. In *Eighth Symposium on Integrated Observing and Assimilation Systems for Atmosphere, Oceans, and Land Surface, AMS 84th Annual Meeting*, 2004.

Marco Bianucci, Antonietta Capotondi, Riccardo Mannella, and Silvia Merlino. Linear or nonlinear modeling for ENSO dynamics? *Atmosphere*, 9(11):435, 2018. doi: 10.3390/atmos9110435.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016. doi: 10.1073/pnas.1517384113.

Antonietta Capotondi, Andrew T. Wittenberg, Matthew Newman, Emanuele Di Lorenzo, Jin-Yi Yu, Pascale Braconnot, et al. Understanding ENSO diversity. *Bulletin of the American Meteorological Society*, 96(6):921–938, 2015. doi: 10.1175/BAMS-D-13-00117.1.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, et al. MLE-bench: Evaluating machine learning agents on machine learning engineering. *arXiv preprint arXiv:2410.07095*, 2024.

Nan Chen and Xianghui Fang. Model programs for "A Simple Multiscale Intermediate Coupled Stochastic Model for El Niño Diversity and Complexity". Zenodo, 2022.

Nan Chen, Xianghui Fang, and Jin-Yi Yu. A multiscale model for El Niño complexity. *npj Climate and Atmospheric Science*, 5, 2022. doi: 10.1038/s41612-022-00241-x.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, et al. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. *arXiv preprint arXiv:2410.05080*, 2024.

Boyin Huang, Peter W. Thorne, Viva F. Banzon, Tim Boyer, Gennady Chepurin, Jay H. Lawrimore, Matthew J. Menne, Thomas M. Smith, Russell S. Vose, and Huai-Min Zhang. Extended reconstructed sea surface temperature, version 5 (ERSSTv5): Upgrades, validations, and intercomparisons. *Journal of Climate*, 30(20):8179–8205, 2017. doi: 10.1175/JCLI-D-16-0836.1.

Yuki Imajuku, Kohki Horie, Yoichi Iwata, Kensho Aoki, Naohiro Takahashi, and Takuya Akiba. ALE-Bench: A benchmark for long-horizon objective-driven algorithm engineering. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? *arXiv preprint arXiv:2310.06770*, 2023.

Fei-Fei Jin. An equatorial ocean recharge paradigm for ENSO. Part I: Conceptual model. *Journal of the Atmospheric Sciences*, 54(7):811–829, 1997. doi: 10.1175/1520-0469(1997)054⟨0811:AEORPF⟩2.0.CO;2.

E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, et al. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, 77(3):437–471, 1996. doi: 10.1175/1520-0477(1996)077⟨0437:TNYRP⟩2.0.CO;2.

Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, et al. Measuring AI ability to complete long software tasks. *arXiv preprint arXiv:2503.14499*, 2025.

Edward N. Lorenz. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2):130–141, 1963. doi: 10.1175/1520-0469(1963)020⟨0130:DNF⟩2.0.CO;2.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI Scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Cristian Martinez-Villalobos, Matthew Newman, Daniel J. Vimont, Cécile Penland, and J. David Neelin. Observed el niño–la niña asymmetry in a linear model. *Geophysical Research Letters*, 46:9909–9919, 2019. doi: 10.1029/2019GL082922.

Michael J. McPhaden, Stephen E. Zebiak, and Michael H. Glantz. ENSO as an integrating concept in earth science. *Science*, 314(5806):1740–1745, 2006. doi: 10.1126/science.1132588.

N. A. Rayner, D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan. Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research: Atmospheres*, 108(D14):4407, 2003. doi: 10.1029/2002JD002670.

Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009. doi: 10.1126/science.1165893.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K. Reddy. LLM-SR: Scientific equation discovery via programming with large language models. In *International Conference on Learning Representations*, 2025a.

Parshin Shojaee, Ngoc-Hieu Nguyen, Kazem Meidani, Amir Barati Farimani, Khoa D. Doan, and Chandan K. Reddy. LLM-SRBench: A new benchmark for scientific equation discovery with large language models. In *International Conference on Machine Learning*, 2025b.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. *arXiv preprint arXiv:2209.13085*, 2022.

G. Srinivas, J. Vialard, F. Liu, A. Voldoire, T. Izumo, E. Guilyardi, and M. Lengaigne. Dominant contribution of atmospheric nonlinearities to ENSO asymmetry and extreme el niño events. *Scientific Reports*, 14:8122, 2024. doi: 10.1038/s41598-024-58803-3.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, et al. PaperBench: Evaluating AI's ability to replicate AI research. *arXiv preprint arXiv:2504.01848*, 2025.

Yiyou Sun, Xinyang Han, Weichen Zhang, et al. Agents' last exam. *arXiv preprint arXiv:2606.05405*, 2026.

Axel Timmermann, Soon-Il An, Jong-Seong Kug, Fei-Fei Jin, Wenju Cai, Antonietta Capotondi, et al. El niño–southern oscillation complexity. *Nature*, 559(7715):535–545, 2018. doi: 10.1038/s41586-018-0252-6.

J. Vialard, F.-F. Jin, M. J. McPhaden, A. Fedorov, W. Cai, S.-I. An, et al. The el niño southern oscillation (ENSO) recharge oscillator conceptual model: Achievements and future prospects. *Reviews of Geophysics*, 63(1):e2024RG000843, 2025. doi: 10.1029/2024RG000843.

Chunzai Wang. A review of ENSO theories. *National Science Review*, 5(6):813–825, 2018. doi: 10.1093/nsr/nwy104.

Duncan Watson-Parris, Yuhan Rao, Dirk Olivié, et al. ClimateBench v1.0: A benchmark for data-driven climate projections. *Journal of Advances in Modeling Earth Systems*, 14(10), 2022. doi: 10.1029/2021MS002954.

Hjalmar Wijk, Tao Lin, Joel Becker, Sami Jawhar, Neev Parikh, Thomas Broadley, et al. RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts. *arXiv preprint arXiv:2411.15114*, 2024.

Wanghan Xu, Shuo Li, Tianlin Ye, Qinglong Cao, Yixin Chen, Hengjian Gao, et al. ResearchClawBench: A benchmark for end-to-end autonomous scientific research. *arXiv preprint arXiv:2606.07591*, 2026.

Yinling Zhang, Nan Chen, Jérôme Vialard, and Xianghui Fang. A physics-informed auto-learning framework for developing stochastic conceptual models for enso diversity. *Journal of Climate*, 37(23):6323–6347, 2024.

Sen Zhao, Fei-Fei Jin, Malte F. Stuecker, Philip R. Thompson, Jong-Seong Kug, Michael J. McPhaden, Mark A. Cane, Andrew T. Wittenberg, and Wenju Cai. Explainable El Niño predictability from climate mode interactions. *Nature*, 630(8018):891–898, 2024. doi: 10.1038/s41586-024-07534-6.

Deyao Zhu, Xin Zhou, Shengling Qin, et al. EdgeBench: Unveiling scaling laws of learning from real-world environments. *arXiv preprint arXiv:2607.05155*, 2026.

A Task Materials

This appendix summarizes what each agent receives. The task files and every prompt that the harness sends are included verbatim in the supplementary material.

Inputs. Agents receive raw gridded observations from January 1980 to December 2014, namely sea surface temperature from ERSSTv5 (Huang et al., 2017), ocean potential temperature and zonal currents from GODAS (Behringer & Xue, 2004), and 850 hPa zonal wind from the NCEP–NCAR reanalysis (Kalnay et al., 1996). A prescribed procedure, summarized below, specifies how to build the five monthly anomaly indices from these fields, and observations from 2015 onward are withheld. As background, agents receive the recharge-oscillator paper of Jin (1997), which links eastern Pacific sea surface temperature and western Pacific thermocline depth in a two-variable model with linear dynamics and simple nonlinear and stochastic extensions. They also receive characteristic scales for nondimensionalizing the indices.

Submission. Within the general form of Eq. (1), agents decide the governing equations themselves, including how the variables interact and how nonlinearity, seasonal dependence, and stochastic forcing are represented. Each submission is an executable implementation of Eq. (1) with its calibrated parameters, accompanied by a model card that describes the equations and the mechanisms they represent. The graders run this model to generate the simulations, reconstructions, and forecasts on which it is scored.

Task instruction. The instruction file asks the agent to build a low-dimensional stochastic dynamical model of ENSO that represents both central and eastern Pacific variability, reproduces the observed behavior of the tropical Pacific, and produces skillful forecasts. It describes a data-processing stage whose output must pass a check before modeling is unlocked, and a modeling stage in which the agent chooses the equations, nondimensionalizes the indices with the provided scales, calibrates the parameters, and compares the model with statistics computed from its own processed data. It names the scoring categories, namely the statistics of the simulated climate, the physical mechanism, forecast skill relative to persistence on initialization months the agent never sees, and a validity gate under which an unstable or non-finite model scores zero, without giving thresholds, weights, or held-out data. It also states the rules. The agent must build the model itself, must not use publications that already contain a complete calibrated model of this system, and must stay inside its working directory without accessing any grader, reference model, or observation file outside it.

Data processing. The processing file prescribes a fixed recipe, and all regional averages are weighted by the cosine of latitude over boxes between $5°$S and $5°$N. T_C and T_E are sea surface temperature averages over the Niño 4 box, from $160°$E to $150°$W, and the Niño 3 box, from $150°$W to $90°$W. h_W is the depth of the $20°$C isotherm averaged from $120°$E to $180°$, where the depth in each water column is found by linear interpolation between the two GODAS levels that bracket the first crossing from the surface. u is the zonal current averaged over the levels from the surface to 55 m and over the Niño 4 box. τ is the 850 hPa zonal wind averaged from $120°$E to $180°$, first over the box for each day and then over each month, with westerly winds positive. Anomalies are formed by removing the monthly climatology of 1981 to 2010 from every index, without detrending. The file also lists common pitfalls, such as fill values, land masks, longitudes across the dateline, and GODAS temperatures given in Kelvin. The data check compares each index element-wise with a reference processed in the same way, within a tight tolerance. When the check fails, the feedback names the indices that differ and the month of the largest discrepancy.

Model interface. The interface file fixes the form of the submission. The submitted model exposes a single time-stepping function, which advances an ensemble of five-variable states, given in physical units, by one stochastic step of a given length in months at a given model time, drawing all random numbers from a generator supplied by the graders. The graders build every experiment on this function, namely free simulations from the zero state, ensemble forecasts initialized from the observed indices, and an ensemble Kalman smoother that observes some variables and reconstructs the others, which the file describes as a test of the dependence among variables. A submission whose model fails to import, raises an error, or returns non-finite values fails the validity gate. The calibrated parameters are stored in a separate parameter file, and the model card describes the state variables, equations, and mechanisms.

Scales and excluded references. The scales file gives $7.5\,°$C for the temperatures, 150 m for h_W, $1.5\,\mathrm{m\,s^{-1}}$ for u, $5\,\mathrm{m\,s^{-1}}$ for τ, and two months for time, and states that these are unit conventions that do not prescribe any model structure. The list of excluded references covers publications that

already contain a complete calibrated model of the five-variable system, including the published reference model (Zhang et al., 2024) and the review of Vialard et al. (2025), whose tables give the full nonlinear stochastic recharge oscillator. A note accompanying the paper of Jin (1997) states that it contains only the linear deterministic core, and that the stochastic forcing, nonlinearity, seasonal dependence, and ENSO diversity are left to the agent.

Phase prompts. The harness sends one prompt at the start of each phase. The Phase A prompt asks the agent to process the raw data into the five indices by following the processing file, and a failed check returns the prompt with the feedback, for up to four attempts. The Phase B prompt tells the agent that nothing in its workspace will score its model and asks it to write diagnostic scripts that run the submitted model and report how far it is from the observations. It asks each script to report the quantity, the observed value, and the gap rather than a pass or fail verdict, and to state the discrepancy it finds. It warns that the diagnostics become read-only once the agent stops, and it describes the three evaluation dimensions in one sentence each, as whether a long free run looks like the observed climate, whether the model behaves as a coupled dynamical system, and whether it forecasts better than persistence from an observed initial state, adding that these are descriptions rather than metrics. If a script fails to run, a repair prompt asks the agent to fix or delete it before the diagnostics are frozen. The Phase C prompt asks for the three submission files, states that the diagnostics are now read-only and that no new evaluation code should be written, warns that agreement between the checks and the observations does not mean the model is finished, and repeats the three descriptions. To encourage exploration, it also explains that the submission is copied automatically and that the best copy counts, so that a failed experiment cannot lower the grade. It further notes that running simulations consumes the same time budget and that the agent should keep working for the whole budget.

Later turns of Phase C use a resume prompt that restates these points. After 20, 40, and 60 minutes without a new copy of the submission, escalating prompts ask the agent to run its diagnostics, find its weakest aspect, and change the model, and to reconsider its structure if parameter tuning is exhausted. In Claude Code, a stop hook blocks a voluntary exit with a short reminder to check the model on the three dimensions, until the submission has stayed unchanged over three stop attempts, and the Gemini scaffold receives the same reminder through its continuation loop. Every prompt ends with a note asking the agent to keep a concise research log in its workspace, because its working context may be compacted.

B Harness Details and Reproducibility

B.1 Evaluation Protocol

Agent systems. Each evaluated system pairs a language model with an agent scaffold, the software through which the model reads files, runs code, and manages its own work. Because the scaffold shapes how a model works over a long run, the benchmark compares complete systems rather than models alone. Claude, GPT, and Gemini models run in the scaffolds their developers provide for agentic work, namely Claude Code, Codex, and Antigravity, each through its subscription service, while the remaining models run in Claude Code connected to OpenRouter through a LiteLLM gateway. Each system runs at the settings of its service: medium reasoning effort for subscription systems and the provider default for the rest, so differences between systems reflect the model together with its scaffold and settings rather than the model alone.

Environment. Each run takes place in its own container, built from the same image and mounting only the working directory of that run. The graders, the reference model, and the held-out observations stay outside the container, so everything an agent knows about the task comes from the materials it is given and its own work. The isolation measures are described below.

Time budget. Each run has a six-hour budget of eligible time shared across the three phases. This is long enough for every system to finish while keeping full runs of twelve systems practical. Following earlier long-horizon benchmarks (Wijk et al., 2024; Starace et al., 2025), eligible time excludes waits caused by the model provider, namely exhausted usage quotas and transient service failures, so that every system is compared over the same amount of working time.

Runs and checkpoints. Each system is run once. The object of evaluation is the model that each run produces, which is scored and inspected as a scientific result in its own right, so the comparisons describe these models rather than the expected performance of each system. During model development, the harness copies the submission every five minutes of eligible time, at the end of every agent turn, and once more when the run ends, and each copy that differs from all earlier copies is kept as a checkpoint. All checkpoints are scored after the run, which shows how the model develops over time. The prompts tell the agent that its best copy counts, which is intended to let it explore without protecting a working version. The main comparison nonetheless uses the final checkpoint, so it differs from the objective stated to the agent. The best checkpoint can only be identified by scoring all checkpoints with the hidden graders, which selects on the test metric itself and is not something the agent can do. The final checkpoint is the model the agent had at the end of its budget, and because the best checkpoint always scores at least as high, this choice can only understate each run. Table 4 reports both for every system.

Table 4: Composite scores of the final and best checkpoints of every system. The best checkpoint is identified after the run by scoring all checkpoints with the hidden graders. Six systems exceed the reference by final checkpoints and eight by best checkpoints.

System	Final checkpoint	Best checkpoint
Claude Fable 5.1	0.515	0.531
Kimi K3	0.479	0.500
GPT-5.6-Sol	0.443	0.466
GPT-6 Astra	0.442	0.475
Claude Opus 5	0.440	0.508
GPT-5.5	0.403	0.404
Gemini 3.8 Flash	0.369	0.394
DeepSeek V4.1 Flash	0.351	0.452
DeepSeek V4 Pro	0.345	0.345
MiniMax-M3	0.222	0.238
GLM-5.2	0.188	0.228
Qwen3-Max	0.054	0.120
Reference model	0.378	0.378

B.2 Lifecycle and Isolation

Lifecycle constants. The budget is $21{,}600\,\mathrm{s}$ of eligible time, counted from the start of the run and shared by the three phases. Eligible time is wall time minus credited pauses, and the clock stops while a credited pause lasts. During model development, the harness credits a pause when the provider reports an exhausted quota with a stated reset time, for the stated delay and for at most 12 such events, or reports a transient service failure, for at most $600\,\mathrm{s}$ per event. A rate-limit error without a stated reset is not credited and consumes eligible time. Credited pauses are capped at six hours in total, and the wall time of a run is capped at 18 hours. Phase A allows at most four attempts at the data check, each limited to 30 minutes, and a run that fails all four ends there. Phase B ends when the agent ends its turn, and at the latest after 90 minutes, followed by at most two repair rounds for scripts that fail to run. The diagnostics are then made read-only. In Phase C, the harness resumes the agent session whenever a turn ends, using the stall prompts of Appendix A after 20, 40, and 60 minutes without a new checkpoint. A run ends early if an agent turn exits within $5\,\mathrm{s}$, if three consecutive turns fail, or if eight consecutive turns end in an error despite changing the submission.

Isolation. Each run has its own container, which runs the agent as an unprivileged user and mounts only the working directory of the run. The graders, the reference model, and the observations after 2014 are never mounted, and grading happens after the run on the host. Agents work offline. The web search and browsing tools of every scaffold are disabled: web fetch and web search in Claude Code, web search in Codex, and the search, browsing, and subagent tools of Antigravity. The only permitted network use is installing Python packages, and audits of the transcripts found no other network use.

B.3 Scaffold Identities and Reproducibility

Scaffold identities. Table 5 lists the scaffold and route of each system. The three additional runs of Claude Fable 5.1 in Section 5.2 used the same configuration as its main run, with web access enabled through a filtering proxy in the two web conditions. Context management follows each scaffold's own behavior, and the largest contexts reached $658{,}258$ tokens for Claude Opus 5 and $314{,}562$ tokens for Kimi K3.

Table 5: Scaffolds and routes of the twelve systems. Systems routed through OpenRouter use one pinned provider each, given in parentheses, with no fallback.

System	Scaffold	Route
Claude Fable 5.1	Claude Code 2.1.270	Subscription
Kimi K3	Claude Code 2.1.223	OpenRouter (Moonshot AI)
GPT-5.6-Sol	Codex 0.146.1	Subscription
GPT-6 Astra	Codex 0.154.0	Subscription
Claude Opus 5	Claude Code 2.1.223	Subscription
GPT-5.5	Codex 0.146.1	Subscription
Gemini 3.8 Flash	Antigravity 1.2.0	Subscription
DeepSeek V4.1 Flash	Claude Code 2.1.223	OpenRouter (GMICloud)
DeepSeek V4 Pro	Claude Code 2.1.223	OpenRouter (GMICloud)
MiniMax-M3	Claude Code 2.1.223	OpenRouter (MiniMax)
GLM-5.2	Claude Code 2.1.223	OpenRouter (Z.ai)
Qwen3-Max	Claude Code 2.1.223	OpenRouter (Alibaba)

Reproducibility. Every run used the same task files, prompts, graders, and container image, which runs Python 3.12 with NumPy 2.5.1 and SciPy 1.18.0. Each run records the configuration it ran under. The supplementary material contains the harness, the graders, the final model of every system, the scores of every checkpoint, and the checksums that identify each of these components.

C Evaluation Details

This appendix describes how the graders evaluate a submitted model along the three dimensions introduced in Section 3.3. For each dimension, we explain which aspect of ENSO it tests and why, how the graders run the submitted model to generate the results, and how these results are turned into a score between 0 and 1. The statistical and reconstruction tests follow the criteria used to validate the reference model (Zhang et al., 2024), and the forecast test is added for this benchmark. The three experiments use the model in three different roles: as a long free-running simulator, as the prior of a data-assimilation system, and as a forecast model. In all three, the graders call only the submitted time-stepping function, so the same equations and parameters are tested throughout. A model whose simulation does not remain finite is not scored. Because the task does not specify which calendar month corresponds to the start of model time, each model is evaluated under the calendar convention stated by its agent. The published reference model, which combines linear couplings with nonlinear drift and additive noise, is scored in the same way. It belongs to a richer model class than the linear stochastic models often used as ENSO baselines. The composite score is

$$C = 0.3\,S + 0.3\,D + 0.4\,P, \tag{2}$$

where S, D, and P are the scores for statistical properties, dynamical consistency, and predictive skill defined below. The number of systems that exceed the reference does not depend strongly on these weights. With equal weights the same six systems exceed it, with the weight on statistics raised to 0.4 five do, and Claude Fable 5.1 and Kimi K3 remain ahead until statistics carries more than about 94% and 65% of the weight, respectively, with the rest split equally between the other two dimensions.

C.1 Statistical Properties

ENSO is irregular, and its events differ in timing, strength, duration, and location. A stochastic model cannot, and is not expected to, reproduce the particular sequence of events in the observed

record. What it should reproduce is the statistical behavior of ENSO, so we compare statistics of long simulations with those of the observations, choosing features that climate scientists use to characterize ENSO. The probability density function (PDF) of each temperature index measures the amplitude of variability and its asymmetry. Observed anomalies are positively skewed in the eastern Pacific, where the strongest El Niño events occur, and negatively skewed in the central Pacific. The seasonal variance, computed separately for each calendar month, measures how ENSO is locked to the annual cycle. Events tend to peak near the end of the calendar year, so the observed variance is largest in December and smallest in spring or early summer. The autocorrelation function (ACF) measures how long anomalies persist and how quickly they reverse. The observed ACF becomes negative at lags beyond about one year, reflecting the alternation between El Niño and La Niña. Event frequencies measure ENSO diversity. Eastern Pacific (EP) and central Pacific (CP) El Niño, extreme El Niño, La Niña, and multi-year events have different climate impacts, and a model with realistic variability can still produce the wrong mix of them. Finally, a bivariate diagnostic of event strength and location tests a nonlinear signature of diversity, namely that strong El Niño events peak farther east than moderate ones.

Experiment. The graders run the submitted model freely for 2300 years from the zero state, with no observational input, so that its variability arises entirely from its own dynamics and stochastic forcing. The first 200 years are discarded so that the results do not depend on the initial state. The monthly values of T_C and T_E from the remaining 2100 years are divided into 30 non-overlapping 70-year segments, each as long as the observed record. The observed record consists of the Niño 4 and Niño 3 sea surface temperature anomaly indices from 1950 to 2019 used by the reference study (Zhang et al., 2024), taken from the data record of Chen & Fang (2022) and derived from HadISST1.1 (Rayner et al., 2003). The PDF, seasonal variance, and ACF are computed for every segment and averaged over the 30 segments, which removes most of the sampling noise of a single 70-year record while keeping each statistic comparable with the observed one. In Figure 4, the simulated curves show this segment average, and the shading around them shows one standard deviation across the 30 segments. Events are counted in 71-year segments and compared with the observed counts per 71 years reported by the reference study, using the same definitions. Each winter is classified from its December to February means of T_E and T_C. It is an El Niño if either mean exceeds $0.5\,^{\circ}\mathrm{C}$, of EP type if T_E is the larger and of CP type otherwise, and a La Niña if either mean falls below $-0.5\,^{\circ}\mathrm{C}$. An extreme El Niño requires T_E to reach $2.5\,^{\circ}\mathrm{C}$, and multi-year events are consecutive El Niño or La Niña winters. For the bivariate diagnostic, each simulated El Niño is mapped to an equatorial sea surface temperature profile with the regression of the reference study, which gives its strength as the peak anomaly and its location as the longitude of the peak. The observed events are located in the same way from the observed sea surface temperature field.

Score. For the PDF, the seasonal variance, and the ACF, the segment-averaged model curve μ is compared with the observed curve o across its points p, which are anomaly values, calendar months, or lags,

$$s = \max\{0, r(\mu, o)\}\, \frac{\sum_p w_p \exp\left[-\frac{1}{2}\left(\frac{\mu_p - o_p}{0.35\,R}\right)^2\right]}{\sum_p w_p}. \tag{3}$$

The first factor is the correlation between the two curves and measures whether the shape is right, for example whether the variance peaks in the correct month. The second factor measures whether the magnitudes agree, with a tolerance proportional to the variation R of the observed curve across its points. PDF points are weighted by their probability, $w_p = \frac{1}{2}\left[\max(o_p, 0) + \max(\mu_p, 0)\right]$, and $w_p = 1$ for the other two curves. Each of these three groups averages the scores of T_C and T_E. For each of the seven event types k, the mean simulated count m_k is compared with the observed count o_k,

$$G_{\mathrm{event}} = \frac{1}{7}\sum_{k=1}^{7} \exp\left[-\frac{1}{2}\left(\frac{m_k - o_k}{0.2\,o_k}\right)^2\right], \tag{4}$$

so a relative difference of 20% lowers the contribution of that type to about 0.61. The diversity group averages two criteria. The first requires simulated events that peak east of $150^{\circ}\mathrm{W}$ to be stronger on average than those that peak farther west,

$$G_{\mathrm{order}} = \min\left\{1, \max\left[0, \frac{\bar{S}_{\mathrm{EP}} - \bar{S}_{\mathrm{CP}}}{0.5\,^{\circ}\mathrm{C}}\right]\right\}, \tag{5}$$

where $\bar{S}_{\mathrm{EP}}$ and $\bar{S}_{\mathrm{CP}}$ are the mean strengths of the two sets. The second is the fraction of observed events that fall within the region of the strength and location plane occupied by 95% of the simulated events. The statistical score combines the five groups $G_1, \ldots, G_5$ as

$$S = \frac{1}{2} \cdot \frac{1}{5} \sum_{i=1}^{5} G_i + \frac{1}{2} \min_i G_i. \tag{6}$$

The minimum term prevents a model that fails one feature of ENSO from compensating with the others.

C.2 Dynamical Consistency

Reproducing observed statistics does not establish that a model captures how its variables interact, because different dynamics can produce similar statistics. ENSO arises from coupling between the ocean and the atmosphere. Eastern Pacific warming weakens the trade winds, and the weaker winds drive eastward current anomalies and flatten the thermocline, and these changes in turn reinforce the warming. The slow recharge and discharge of equatorial heat content, represented by h_W, sets the timing of transitions between El Niño and La Niña. If a model represents these couplings correctly, observing some variables should reveal much about the others. We test this with data assimilation, the standard method for combining a model with incomplete observations. In Bayesian terms, the model provides a prior distribution of states, measurements of some variables update it to a posterior distribution, and the posterior mean of the unmeasured variables is their reconstruction. Because the unmeasured variables are informed only through the relationships encoded in the model, an accurate reconstruction is evidence that these relationships are realistic.

Experiment. The graders use the observed monthly indices from 1980 to 2014, the period available to the agents. An ensemble of 500 model states is advanced month by month by the submitted model, each member with its own stochastic forcing, so that the spread of the ensemble and the correlations among its variables reflect the model's own uncertainty and coupling. At the end of each month, an ensemble Kalman smoother assimilates the observed values of the conditioning variables, adjusting every member toward the observations in proportion to the covariances within the ensemble. The smoother also revises the states of the preceding three months with each new observation, so the reconstruction depends on the lagged relationships in the model, such as the lead of the western Pacific thermocline over eastern Pacific temperature, and not only on simultaneous ones. The ensemble mean of each unobserved variable is its reconstruction. The actual records of these variables are withheld from the smoother and used only for scoring. In test A, T_C and T_E are assimilated to reconstruct h_W, u, and τ. In test B, the roles are reversed, so the coupling is tested in both directions.

Score. After the first two years, which allow the ensemble to adjust, each reconstruction $\hat{x}_t$ is compared with the withheld record x_t by the Nash-Sutcliffe efficiency,

$$\mathrm{NSE} = \max\left\{0,\ 1 - \frac{\sum_t (\hat{x}_t - x_t)^2}{\sum_t (x_t - \bar{x})^2}\right\}, \tag{7}$$

where $\bar{x}$ is the mean of the record. The NSE is the fraction of the variance of the record that the reconstruction explains. Unlike a correlation, it also penalizes a reconstruction with the right timing but the wrong amplitude. The dynamical score D is the mean of the five NSE values, three from test A and two from test B.

C.3 Predictive Skill

Forecasting is a primary practical use of ENSO models, and it is the only test on data that the agents never saw. The held-out period from 2015 to 2024 includes the extreme El Niño of 2015/16, the multi-year La Niña from 2020 to 2023, and the El Niño of 2023/24, so it covers strong events, a multi-year event, and rapid transitions. Forecasts are compared with persistence, which assumes that the current anomaly continues unchanged. Persistence predicts no change, so a model that cannot improve on it has no forecast skill, and improving on it requires the model to capture how anomalies grow, decay, and reverse.

Experiment. From each month between January 2015 and December 2023, the graders initialize all five model variables with the observed values of that month and run a 50-member ensemble forward for 12 months without further observations. The members differ only in their stochastic forcing, and the ensemble mean serves as the forecast. The forecasts of T_C and T_E at lead times of 1 to 12 months, where the lead time is the interval between the start of a forecast and the month being predicted, are compared with the observed values for the target months. The persistence forecast for every lead time is the observed value of the starting month.

Score. For each index, let $\mathrm{RMSE}_\ell^{\mathrm{model}}$ and $\mathrm{RMSE}_\ell^{\mathrm{pers}}$ denote the root-mean-square errors of the model and of persistence over all forecasts at lead time ℓ. The skill over all targets is

$$P^{\mathrm{all}} = \max\left\{0,\ 1 - \frac{\sum_{\ell=1}^{12} \mathrm{RMSE}_\ell^{\mathrm{model}}}{\sum_{\ell=1}^{12} \mathrm{RMSE}_\ell^{\mathrm{pers}}}\right\}. \tag{8}$$

The extreme skill P^{ext} focuses on strong El Niño and La Niña conditions. It pools the forecasts at all start months and leads whose observed target anomaly exceeds one standard deviation in magnitude, and compares the root-mean-square errors of the model and of persistence over this pool,

$$P^{\mathrm{ext}} = \max\left\{0,\ 1 - \frac{\mathrm{RMSE}_{\mathrm{ext}}^{\mathrm{model}}}{\mathrm{RMSE}_{\mathrm{ext}}^{\mathrm{pers}}}\right\}. \tag{9}$$

It is set to zero when fewer than ten such forecasts exist. The predictive score is

$$P = 0.7\,\overline{P^{\mathrm{all}}} + 0.3\,\overline{P^{\mathrm{ext}}}, \tag{10}$$

where the bars denote averages over T_C and T_E. A score of zero means no improvement over persistence. The forecast correlations shown in Figure 4 are diagnostics and do not enter the score.

D Cost

For the systems routed through OpenRouter, the cost of a run is the sum of the billed cost of each model call, retrieved from OpenRouter by the identifier of the call. The systems on subscription plans pay a fixed fee, so their token use is reported instead, counted from their transcripts (Table 6). Over all batches of runs, the OpenRouter spending was about $521.

Table 6: Model calls, prompt tokens, share of prompt tokens served from the provider's cache, and cost of the main run of each system. Subscription systems pay a fixed fee. The usage of GPT-5.6-Sol was not recorded. For DeepSeek V4.1 Flash, tokens are counted from the transcript and the cost is the change in account usage over its batch, in which it was the only system routed through OpenRouter.

System	Route	Calls	Prompt tokens	Cached	Cost (USD)
Kimi K3	OpenRouter	954	120.3 M	99.2%	43.69
GLM-5.2	OpenRouter	669	65.4 M	98.0%	21.48
Qwen3-Max	OpenRouter	3,250	308.1 M	95.3%	132.37
DeepSeek V4 Pro	OpenRouter	589	48.7 M	93.2%	7.67
MiniMax-M3	OpenRouter	1,011	83.1 M	98.2%	5.97
DeepSeek V4.1 Flash	OpenRouter	752	62.0 M	98.4%	2.43
Claude Fable 5.1	Subscription	636	53.7 M	97.2%	fixed fee
Claude Opus 5	Subscription	446	133.9 M	99.4%	fixed fee
GPT-5.5	Subscription	2,619	297.8 M	97.8%	fixed fee
GPT-6 Astra	Subscription	432	42.7 M	94.7%	fixed fee
Gemini 3.8 Flash	Subscription	3,429	487.3 M	96.6%	fixed fee
GPT-5.6-Sol	Subscription	n/a	n/a	n/a	fixed fee

E AGENT-GENERATED MODELS AND THEIR SIMPLIFIED FORMS

This appendix presents Claude Fable 5.1 as a worked example of an original final model and its simplified form. The simplification procedure is described in Appendix G, and Table 12 there summarizes the structure and scores of every model before and after simplification. The executable implementations, parameters, and model cards of all original and simplified models are provided in the supplementary material.

E.1 CLAUDE FABLE 5.1

Original model. The model describes the coupled evolution of five state variables through seasonally modulated linear interactions and nonlinear feedbacks. Cubic damping constrains the amplitude of the simulated variability, while state-dependent stochastic forcing represents unresolved processes.

$$
\begin{aligned}
\frac{du}{dt} &= a_{11}u + a_{12}(t)h_W + a_{13}T_C + a_{14}T_E + a_{15}\tau + \sigma_u(t)\dot{W}_u \\
\frac{dh_W}{dt} &= a_{21}u + a_{22}(t)h_W + a_{23}T_C + a_{24}T_E + a_{25}\tau + a_{26}\frac{T_E^2}{1+(T_E/T_{s,h})^2} + a_{27}h_W^3 \\
&\quad + a_{28} + \sigma_h(h_W,t)\dot{W}_h \\
\frac{dT_C}{dt} &= a_{31}u + a_{32}(t)h_W + a_{33}(t)T_C + a_{34}T_E + a_{35}\tau + a_{36}T_C^2 + a_{37}T_C^2\,\mathbf{1}_{T_C>0} \\
&\quad + a_{38}T_C^3 + a_{39} + \sigma_C(T_C)\dot{W}_C \\
\frac{dT_E}{dt} &= a_{41}u + a_{42}(t)h_W + a_{43}T_C + a_{44}(t)T_E + a_{45}\tau + a_{46}(t)\frac{T_E^2}{1+(T_E/T_{s,E})^2} \\
&\quad + a_{47}T_E^3 + a_{48} + \sigma_E(T_E,t)\dot{W}_E \\
\frac{d\tau}{dt} &= a_{51}u + a_{52}h_W + a_{53}T_C + a_{54}T_E + a_{55}\tau + \sigma_\tau(T_C,t)\dot{W}_\tau
\end{aligned}
\tag{11}
$$

Here, a_{ij} denotes the coefficient of the corresponding term. Coefficients that depend on t vary seasonally. The parameters $T_{s,h}$ and $T_{s,E}$ set the saturation scales of the nonlinear feedbacks. The indicator $\mathbf{1}_{\{T_C>0\}}$ activates the associated term only for positive T_C. The terms $\dot{W}_u$, $\dot{W}_h$, $\dot{W}_C$, $\dot{W}_E$, and $\dot{W}_\tau$ represent independent Gaussian white noise. The corresponding σ functions determine the noise amplitudes and their dependence on season or model state.

Simplified model. The simplified model removes all nonlinear drift terms and confines seasonal variation in the drift to the growth and damping of T_C and T_E. It retains a refitted linear coupling among the five variables and preserves the original state-dependent stochastic forcing. The resulting model contains 34 drift coefficients. Relative to the original model, it improves the statistical properties and predictive skill but lowers dynamical consistency, revealing a trade-off among the evaluation criteria.

$$
\begin{aligned}
\frac{du}{dt} &= \widetilde{a}_{11}u + \widetilde{a}_{12}h_W + \widetilde{a}_{13}T_C + \widetilde{a}_{14}T_E + \widetilde{a}_{15}\tau + \sigma_u(t)\dot{W}_u \\
\frac{dh_W}{dt} &= \widetilde{a}_{21}u + \widetilde{a}_{22}h_W + \widetilde{a}_{23}T_C + \widetilde{a}_{24}T_E + \widetilde{a}_{25}\tau + \widetilde{a}_{28} + \sigma_h(h_W,t)\dot{W}_h \\
\frac{dT_C}{dt} &= \widetilde{a}_{31}u + \widetilde{a}_{32}h_W + \widetilde{a}_{33}(t)T_C + \widetilde{a}_{34}T_E + \widetilde{a}_{35}\tau + \widetilde{a}_{39} + \sigma_C(T_C)\dot{W}_C \\
\frac{dT_E}{dt} &= \widetilde{a}_{41}u + \widetilde{a}_{42}h_W + \widetilde{a}_{43}T_C + \widetilde{a}_{44}(t)T_E + \widetilde{a}_{45}\tau + \widetilde{a}_{48} + \sigma_E(T_E,t)\dot{W}_E \\
\frac{d\tau}{dt} &= \widetilde{a}_{51}u + \widetilde{a}_{52}h_W + \widetilde{a}_{53}T_C + \widetilde{a}_{54}T_E + \widetilde{a}_{55}\tau + \sigma_\tau(T_C,t)\dot{W}_\tau
\end{aligned}
\tag{12}
$$

Here, $\widetilde{a}_{ij}$ denotes a coefficient refitted for the simplified model. The subscripts are retained from the original formulation to show the correspondence between terms in the two models. The coefficients $\widetilde{a}_{33}(t)$ and $\widetilde{a}_{44}(t)$ vary seasonally, while the remaining $\widetilde{a}_{ij}$ are constant. The noise processes and the functions σ_u, σ_h, σ_C, σ_E, and σ_τ are unchanged from the original model.

Model	Stat.	Dyn.	Pred.	Composite
Original	0.715	0.639	0.274	0.515
Simplified	0.810	0.600	0.297	0.542

Original parameters and auxiliary definitions.

Symbol	Value	Symbol	Value
a_{11}	-0.9170	a_{13}	-0.4020
a_{14}	0.2600	a_{15}	0.1140
a_{21}	-0.1640	a_{23}	0.09686
a_{24}	-0.1500	a_{25}	-0.07482
a_{26}	-0.3000	a_{27}	-1.184
a_{28}	0.004840	$T_{s,h}$	0.4000
a_{31}	0.1628	a_{34}	0.1210
a_{35}	0.08530	a_{36}	-1.020
a_{37}	-0.7500	a_{38}	-7.181
a_{39}	0.004000	a_{41}	0.2403
$a_{42}(t)$	$0.05899 + 0.1000 s_1$	a_{43}	-0.1399
a_{45}	0.1285	$a_{46}(t)$	$0.5050 - 0.6060 s_1$
a_{47}	-1.038	a_{48}	-0.006260
$T_{s,E}$	0.3500	a_{51}	0.3296
a_{52}	-0.2758	a_{53}	3.062
a_{54}	-1.118	a_{55}	-0.9918
S_T	7.500 degC	S_h	150.0 m
S_u	1.500 m/s	S_τ	5.000 m/s
S_t	2.000 months	$x_{\max}$	3.000

The bound $x_{\max} = 3$ clips the nondimensional state at ± 3, corresponding to $\pm 22.5\,^{\circ}\mathrm{C}$ in temperature and ± 450 m in thermocline depth, far outside the observed range, so it acts only as a numerical safeguard and does not alter the linear drift.

Symbol	Definition or value
$a_{12}(t)$	$0.8500 + 0.1700 c_1 + 0.4250 s_1$
$a_{22}(t)$	$-0.1640 + 0.1044 c_1 - 0.08738 s_1$
$a_{32}(t)$	$0.1073 + 0.01866 c_1 + 0.03000 s_1$
$a_{33}(t)$	$-0.2500 + 0.09000 c_1 - 0.2400 s_1$
$a_{44}(t)$	$-0.1510 - 0.006228 c_1 - 0.2700 s_1 - 0.1200 c_2 - 0.1500 s_2$
$\sigma_u(t)$	$0.09400\,(1 - 0.1000 c_1 + 0.03000 s_1 - 0.1500 c_2 + 0.1500 s_2)$
$\sigma_h(h_W, t)$	$0.02400\,(1 + 0.2000 c_1)\,\max(1 - 4.000\,h_W,\ 0.2)$
$\sigma_C(T_C)$	$0.03000\,\max(1 - 2.000\,T_C,\ 0.2)\,\max(1 - 30.00\,T_C^2,\ 0.2)$
$\sigma_E(T_E, t)$	$0.03700\,(1 - 0.6000 c_1 + 0.4000 s_1 - 0.3000 c_2)\,\min(\max(1 + 5.500\,T_E,$ $0.2),\ 2.500)$
$\sigma_\tau(T_C, t)$	$0.4500\,(1 + 0.1300 c_1 - 0.05000 s_1 - 0.08000 c_2 + 0.2000 s_2 - 0.2000 c_3 -$ $0.1000 s_3)\,\max(1 + 0.8000\,T_C,\ 0.2)$
t_m	$2t$ (months since 1 Jan of the run start)
c_1	$\cos\big(2\pi \cdot 1\,(t_m - 0.5)/12\big)$
s_1	$\sin\big(2\pi \cdot 1\,(t_m - 0.5)/12\big)$
c_2	$\cos\big(2\pi \cdot 2\,(t_m - 0.5)/12\big)$
s_2	$\sin\big(2\pi \cdot 2\,(t_m - 0.5)/12\big)$

Symbol	Definition or value
c_3	$\cos\big(2\pi \cdot 3\,(t_m - 0.5)/12\big)$
s_3	$\sin\big(2\pi \cdot 3\,(t_m - 0.5)/12\big)$

Reduced drift coefficients.

Symbol	Value	Symbol	Value
$\widetilde{a}_{11}$	-0.8523968	$\widetilde{a}_{12}$	0.9358338
$\widetilde{a}_{13}$	-0.4152634	$\widetilde{a}_{14}$	0.2980273
$\widetilde{a}_{15}$	0.1134559	$\widetilde{a}_{21}$	-0.1834501
$\widetilde{a}_{22}$	-0.2025177	$\widetilde{a}_{23}$	0.1094091
$\widetilde{a}_{24}$	-0.2023363	$\widetilde{a}_{25}$	-0.07096293
$\widetilde{a}_{31}$	0.1524918	$\widetilde{a}_{32}$	0.1193508
$\widetilde{a}_{34}$	0.09620686	$\widetilde{a}_{35}$	0.0811469
$\widetilde{a}_{41}$	0.2445608	$\widetilde{a}_{42}$	0.07082664
$\widetilde{a}_{43}$	-0.1592483	$\widetilde{a}_{45}$	0.1280047
$\widetilde{a}_{51}$	0.3361003	$\widetilde{a}_{52}$	-0.2772871
$\widetilde{a}_{53}$	3.029652	$\widetilde{a}_{54}$	-1.102882
$\widetilde{a}_{55}$	-0.9893288	$\widetilde{a}_{28}$	0.00227347
$\widetilde{a}_{39}$	-0.003035909	$\widetilde{a}_{48}$	-0.00205137

Symbol	Definition or value
$\widetilde{a}_{33}$	$-0.3096332 + 0.03899662c_1 - 0.2438837s_1$
$\widetilde{a}_{44}$	$-0.1736369 - 0.02538335c_1 - 0.3242342s_1 - 0.123943c_2 - 0.1380168s_2$

All noise definitions and numerical bounds in the original parameter table also apply to the reduced model.

F Claude Fable 5.1 under Four Conditions

Table 11 gives the scores of the four Claude Fable 5.1 runs discussed in Section 5.2.

Table 11: Claude Fable 5.1 under four conditions, each run once for six hours of eligible time. Best is the highest-scoring checkpoint and final is the last one, and the three component scores are for the best checkpoint. Parentheses give the change relative to the offline run (green for an increase, red for a decrease), and bold marks the best value in each column.

		Composite		Best checkpoint		
Condition	Access	Best	Final	Stat.	Dyn.	Pred.
Offline (formal)	None	**0.531**	**0.515**	0.741	**0.632**	0.297
Years anonymized	None	0.482 (↓0.049)	0.463 (↓0.052)	0.656 (↓0.085)	0.576 (↓0.056)	0.281 (↓0.016)
Web access 1	Literature, data	0.475 (↓0.056)	0.386 (↓0.129)	0.638 (↓0.103)	0.572 (↓0.060)	0.280 (↓0.017)
Web access 2	Literature	0.528 (↓0.003)	0.406 (↓0.109)	**0.798** (↑0.057)	0.556 (↓0.076)	**0.306** (↑0.009)

G Structural Simplification of All Twelve Models

Groups of terms were removed from each final model, the remaining coefficients were refitted to long simulations of the original model, and the reduced model was rescored with the same graders and random seed. A reduction was kept if it retained at least 95% of the original composite score and 90% of every dimension score, and if it kept the model's ENSO complexity, in particular the

asymmetry of its temperature distribution, which the scores alone can miss. For GPT-6 Astra and the two DeepSeek models, removing both nonlinear drift and multiplicative noise keeps the scores within these limits but leaves a nearly symmetric distribution and lower event diversity, so removal stopped before the asymmetry was lost. Asymmetry was judged by comparing the simulated distribution of T_E with the observed one, which is positively skewed (Figure 4e), and a reduction whose distribution became nearly symmetric was rejected. When reductions in both families qualified, the higher-scoring one was kept. The three models that score below 0.24 already miss several statistical features of ENSO, so their reductions were screened on the scores alone. Table 12 extends Table 3 to all twelve final models, including these three, which are not part of the mechanism analysis in Section 5.1.

Table 12: Structure and scores of all twelve final models before and after simplification, in the format of Table 3. The three models below the middle rule score below 0.24 and already miss several statistical features of ENSO, so their reductions were screened on the scores alone and are not included in the mechanism analysis.

Model	Version	Structure		Component scores			Composite
		Drift	Noise	Stat.	Dyn.	Pred.	
Claude Fable 5.1	Original	NL	Mult.	0.715	**0.639**	0.274	0.515
	Simplified	L	Mult.	**0.810** (↑0.095)	0.600 (↓0.039)	**0.297** (↑0.023)	**0.542** (↑0.027)
Kimi K3	Original	NL	Mult.	**0.620**	0.570	**0.305**	**0.479**
	Simplified	L	Mult.	0.575 (↓0.045)	**0.573** (↑0.003)	0.302 (↓0.003)	0.465 (↓0.014)
GPT-5.6-Sol	Original	NL	Add.	0.554	0.515	**0.306**	0.443
	Simplified	NL	Add.	**0.613** (↑0.059)	**0.529** (↑0.014)	0.304 (↓0.002)	**0.464** (↑0.021)
GPT-6 Astra	Original	NL	Add.	0.530	**0.629**	0.237	0.442
	Simplified	NL	Add.	**0.598** (↑0.068)	0.611 (↓0.018)	**0.248** (↑0.011)	**0.462** (↑0.020)
Claude Opus 5	Original	NL	Mult.	0.548	**0.589**	0.247	0.440
	Simplified	NL	Add.	**0.629** (↑0.081)	0.567 (↓0.022)	**0.286** (↑0.039)	**0.473** (↑0.033)
GPT-5.5	Original	NL	Mult.	**0.412**	0.595	0.252	**0.403**
	Simplified	NL	Add.	0.390 (↓0.022)	**0.600** (↑0.005)	0.251 (↓0.001)	0.398 (↓0.005)
	Simplified	L	Mult.	0.387 (↓0.025)	0.590 (↓0.005)	**0.255** (↑0.003)	0.395 (↓0.008)
Gemini 3.8 Flash	Original	NL	Mult.	**0.406**	**0.503**	0.241	**0.369**
	Simplified	NL	Add.	0.389 (↓0.017)	0.499 (↓0.004)	0.241 (0.000)	0.362 (↓0.007)
DeepSeek V4.1 Flash	Original	NL	Mult.	0.418	0.508	0.183	0.351
	Simplified	L	Mult.	**0.419** (↑0.001)	**0.559** (↑0.051)	**0.244** (↑0.061)	**0.391** (↑0.040)
DeepSeek V4 Pro	Original	NL	Mult.	**0.285**	0.530	**0.251**	**0.345**
	Simplified	L	Mult.	0.269 (↓0.016)	**0.532** (↑0.002)	0.250 (↓0.001)	0.340 (↓0.005)
MiniMax-M3	Original	NL	Mult.	**0.231**	**0.326**	0.136	**0.222**
	Simplified	NL	Mult.	0.229 (↓0.002)	0.323 (↓0.003)	0.136 (0.000)	0.220 (↓0.002)
GLM-5.2	Original	NL	Mult.	0.222	0.261	0.108	0.188
	Simplified	L	Add.	**0.241** (↑0.019)	**0.381** (↑0.120)	**0.265** (↑0.157)	**0.293** (↑0.105)
Qwen3-Max	Original	NL	Add.	0.118	0.062	0.000	0.054
	Simplified	L	Add.	0.118 (0.000)	0.062 (0.000)	0.000 (0.000)	0.054 (0.000)

H Further Related Work

Equation discovery. Data-driven methods identify governing equations by sparse regression over candidate terms (Brunton et al., 2016) or by symbolic regression (Schmidt & Lipson, 2009), and language models now propose equation skeletons that are fitted and scored against data (Shojaee et al., 2025a;b). These methods work on prepared data within a fixed pipeline, whereas agents in SciExam for ENSO process raw observations, choose what to check, and are judged on behavior that a good fit does not guarantee. Machine-learning emulators of climate have their own benchmarks (Watson-Parris et al., 2022).

Conceptual ENSO models. Low-order ENSO models range from the recharge oscillator (Jin, 1997) to stochastic multiscale models of ENSO diversity (Chen et al., 2022), an extended recharge

oscillator that forecasts competitively (Zhao et al., 2024), and an automated framework that builds stochastic models of ENSO diversity by selecting their terms with causal inference (Zhang et al., 2024). That study validated its models by their statistics, their event diversity, and their recovery of unobserved variables through data assimilation. SCIEXAM FOR ENSO adopts these criteria for its first two dimensions and adds forecasts of held-out years. Its contribution lies in the workflow around these tests, in which general-purpose agents process raw observations, write and freeze their own diagnostics, and develop a model over hours without seeing any score. Whether ENSO asymmetry arises from nonlinear dynamics or multiplicative noise remains debated (Martinez-Villalobos et al., 2019; Srinivas et al., 2024; Vialard et al., 2025). The agent-built models of this study add candidates to the asymmetry debate under common tests.