URBANSURFACE-AGENT: PHYSICS-GATED SCIENTIFIC REASONING FOR URBAN SURFACE PARAMETERIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Urban land-surface numerical models are important tools for simulating urban energy exchanges and local climate. However, improvements in model performance do not necessarily imply scientifically valid parameterization, because compensating errors among coupled processes can make inappropriate parameter changes appear beneficial. To address this issue, we introduce URBANSURFACE-AGENT, a physics-gated scientific agent that formalizes parameter discovery as an experience-guided proposal–verification process. Model–observation discrepancies are used to formulate mechanism-guided parameter interventions with expected process responses, which are evaluated by SUEWS and independently adjudicated by a PHYSICS JUDGE using semantic, process-wise, and temporal criteria. Experimental outcomes, including successful, failed, and boundary-case interventions, are retained in scientific memory to guide subsequent proposals. We evaluate URBANSURFACE-AGENT on 20 Urban-PLUMBER flux-tower sites under a sealed temporal protocol. Across 1,410 adaptation candidates, 350 improve the aggregate objective but fail process consistency, and 89 further fail temporal consistency, demonstrating that numerical improvement alone is insufficient for reliable parameter discovery. The framework freezes interventions at 16 sites, with 14 satisfying practical success criteria on the sealed future period. These results demonstrate that reliable scientific-agent parameter discovery requires more than numerical optimization: proposed corrections must be tested against the physical processes they are intended to represent, and transferable scientific knowledge is better represented by context-dependent parameter–process evidence than by universal numerical corrections.

1 INTRODUCTION

Urban land-surface models represent the coupled energy and water exchanges that shape local climate, but improving their fit to observations does not necessarily improve their physical explanation. Different parameter changes can produce similar aggregate scores, while an apparent improvement in one output may compensate for errors in another process or for structural model deficiencies (Huang et al., 2024; Raoult et al., 2024; Bernhard et al., 2025). Surface parameters also continue to rely substantially on lookup tables or values calibrated elsewhere (Cheng et al., 2024; Sun et al., 2025b), although their effective values depend on model structure, environmental context, and observational support (Xia et al., 2025; Ahmad et al., 2025). Model observation discrepancies therefore raise a scientific question before an optimization question: which process should be tested, and what evidence would support or contradict a proposed correction?

Addressing this question requires more than minimizing a scalar objective. A parameter intervention may reduce average error while degrading an individual flux, and favorable performance over a full period may conceal temporally localized failure. Scientific model refinement needs an executable protocol that specifies which changes are permitted and evaluates their consequences independently of the proposer (Romera-Paredes et al., 2024; Novikov et al., 2025). In this work, *scientific admissibility* means satisfying predefined semantic, numerical, process-wise, and temporal checks. These checks can expose conflicting responses and limit unsupported claims, although passing them can-

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.

not establish a unique physical cause or rule out every compensating error. A failed intervention remains informative when its tested direction, output responses, and failure conditions are recorded.

LLM-based scientific agents offer a way to connect discrepancy diagnosis, mechanism hypotheses, numerical experiments, and accumulated experience (Wang et al., 2024; Ding et al., 2025; Wu et al., 2026; Zhao et al., 2026). Their potential lies in formulating targeted experiments and revising subsequent proposals in light of prior outcomes (Ghafarollahi & Buehler, 2025; Kao et al., 2026). For this purpose, memory must preserve the tested hypothesis, site context, process responses, and reasons for acceptance or rejection, rather than only the best-fitting parameter value. Proposal and acceptance must remain separate: an agent-generated explanation cannot validate its own intervention (Almansoori et al., 2025; Wang et al., 2025). We therefore ask: *Can an autonomous agent turn model observations discrepancies into testable physical hypotheses, distinguish admissible corrections from apparent improvements caused by process trade-offs, and accumulate evidence about when those hypotheses hold or fail?*

We introduce URBANSURFACE-AGENT, a simulator-in-the-loop framework that treats parameter interventions as experiments and records *conditional parameter process evidence* (Fig. 1). A typed ontology separates observation-constrained surface facts, searchable parameters, and evolving model states; parameter meanings and equation-level roles remain fixed (Tucker et al., 2022), while revisions to surface facts require documented provenance (Imhoff et al., 2020; Leo et al., 2024). The LLM uses model observation discrepancies and prior trials to propose bounded interventions and expected responses. Deterministic code generates numerical candidates, SUEWS evaluates their effects (Grimmond et al., 2026), and an independent PHYSICS JUDGE applies fixed acceptance criteria. Successful, failed, and boundary-case trials inform subsequent hypotheses without changing those criteria. Crucially, an accepted intervention supports its proposed explanation only when the predicted response is also observed.

We evaluate the framework at 20 Urban-PLUMBER flux-tower sites (Lipson et al., 2022a) under a sealed temporal protocol. Among 1,410 completed adaptation candidates, 350 first fail process consistency despite improving the aggregate objective, and a further 89 first fail temporal consistency. These outcomes separate numerical improvement from admissibility. The framework freezes parameter interventions at 16 sites, of which 14 meet practical success criteria on the held-out future period. Site-level failure signatures and a memory-informed hypothesis revision illustrate what is learned beyond the final parameter values. Opposite correction directions within the same parameter family further motivate context-conditioned evidence records; they do not, by themselves, establish a predictive cross-site rule or identify the cause of each site's discrepancy.

Our contributions are threefold:

- **An executable scientific-agent protocol.** We connect process-level diagnosis to bounded parameter experiments within an existing simulator, separating LLM-generated hypotheses, numerical search, and independent acceptance.
- **Evidence-preserving hypothesis revision.** Scientific memory retains successful, failed, and boundary-case trials together with their context and process responses, supporting subsequent hypothesis revision while preserving abstention when no tested intervention is admissible.
- **Evidence-constrained surface-parameter correction.** Across 20 Urban-PLUMBER sites, we show how process-wise and temporal evidence distinguishes physically admissible parameter corrections from aggregate improvements driven by conflicting responses. Site-level failure analyses characterize the limits of these corrections, while sealed future-period evaluation establishes which site-specific interventions retain their benefits beyond adaptation.

2 METHOD

We formulate model diagnosis and adaptation as *simulator-in-the-loop hypothesis testing*: an LLM proposes sparse parameter interventions, a deterministic process model produces numerical evidence, and an independent PHYSICS JUDGE decides admissibility (Romera-Paredes et al., 2024). The numerical experiment remains constrained parameter search; its scientific record additionally

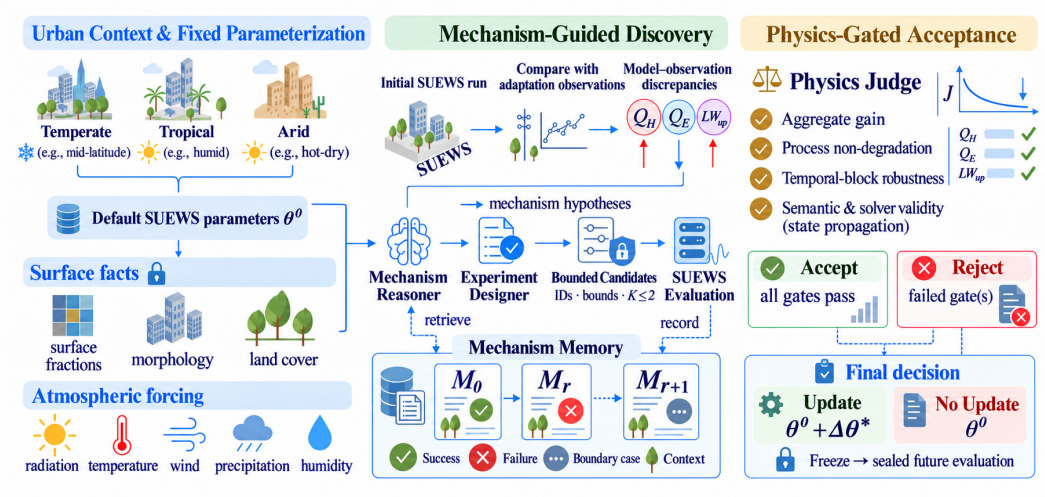

Figure 1: Overview of URBANSURFACE-AGENT. Site evidence and the default parameterization define a reference SUEWS simulation. Model observation discrepancies guide mechanism-based intervention search, SUEWS provides numerical evidence, and an independent Physics Judge determines acceptance. Outcomes are stored as scientific memory, while abstention retains the default configuration.

links the hypothesis and site context to process effects and gate outcomes. Fig. 1 summarizes the loop. We define the numerical task (Sec. 2.1) and intervention semantics (Sec. 2.2), then describe hypotheses and memory (Sec. 2.3), independent verification (Sec. 2.4), and leakage-free refinement and evaluation (Sec. 2.5).

2.1 PROBLEM FORMULATION

Consider an urban site s with atmospheric forcing $\mathbf{x}_{s,1:T}$, site descriptors $\mathbf{c}_s$, default SUEWS parameters $\boldsymbol{\theta}_s^0$, and observations $\mathbf{y}_{s,1:T_a}$ accessible only during an adaptation period of length T_a. SUEWS configuration and model variables are detailed in App. B. We seek a sparse, interpretable intervention $\Delta\boldsymbol{\theta}_s$ that improves the prediction of sensible heat flux Q_H, latent heat flux Q_E, and upward longwave radiation LW_{up}. We write the stateful simulator as

$$(\hat{\mathbf{y}}_{s,t},\, \mathbf{h}_{s,t}) = \mathcal{M}_{\text{SUEWS}}\big(\mathbf{x}_{s,t},\, \mathbf{c}_s,\, \boldsymbol{\theta}_s^0 + \Delta\boldsymbol{\theta}_s,\, \mathbf{h}_{s,t-1}\big)\,, \tag{1}$$

where $\mathcal{M}_{\text{SUEWS}}$ is the SUEWS simulator, $\hat{\mathbf{y}}_{s,t}$ its predicted outputs at time t, and $\mathbf{h}_{s,t}$ its prognostic state (heat storage, soil moisture, surface-water storage), propagated recursively from $\mathbf{h}_{s,t-1}$. Observations are used only to discover and evaluate parameters: they are never fed into Eq. 1 or used to correct dynamic states, so the deployed model remains a forcing-only forecasting system.

The task cannot be posed as unconstrained scalar optimization: an intervention may improve one output while degrading another, and different parameters may achieve similar losses by compensating for structural model error rather than correcting it. We therefore retain constrained optimization as the numerical subproblem used to instantiate and select admissible interventions:

$$\begin{aligned} \min_{\Delta\boldsymbol{\theta}_s} \quad & J_s\big(\boldsymbol{\theta}_s^0 + \Delta\boldsymbol{\theta}_s\big) && && (2)\\ \text{s.t.} \quad & \Delta E_{s,y} \leq 0, && \forall y \in \mathcal{Y}_s, \\ & \boldsymbol{\theta}_s^0 + \Delta\boldsymbol{\theta}_s \in \Theta_s^{\text{phys}}, \\ & \|\Delta\boldsymbol{\theta}_s\|_0 \leq K, \\ & G_s^q\big(\boldsymbol{\theta}_s^0 + \Delta\boldsymbol{\theta}_s\big) = 1, && q \in \{\text{semantic}, \text{solver}, \text{temporal}\}. \end{aligned}$$

Here $J_s(\cdot)$ is the aggregate objective and $\Delta E_{s,y}$ the relative per-output error change, both formally defined in Sec. 2.4, Eq. 5; $\mathcal{Y}_s \subseteq \{Q_H, Q_E, LW_{\text{up}}\}$ denotes the outputs evaluated at site s; Θ_s^{phys} is the semantically and physically admissible parameter space defined by the ontology of Sec. 2.2;

$\|\cdot\|_0$ counts nonzero entries, so $K \leq 2$ bounds the number of simultaneously modified parameters; and $G_s^q(\boldsymbol{\theta}) \in \{0, 1\}$ denotes the semantic, numerical, and temporal validity gates defined in Sec. 2.4. An intervention is admissible only when all applicable gates are satisfied.

2.2 Hierarchical and Semantically Grounded Surface Representation

We partition surface information into three layers by epistemic role, ordered from fixed to searchable. **Layer I** holds observable facts - seven-surface fractions, building and tree height, and LAI - derived from remote sensing, GIS, LiDAR, or local records. The agent may quality-control and translate them but cannot optimize them. **Layer II** holds uncertain physical controls (radiative properties, vegetation conductance, aerodynamic parameters, storage coefficients, hydrological capacities); only this layer is searchable. **Layer III** holds evolving soil-moisture, surface-water, phenological, and thermal states, propagated by SUEWS (i.e., $\mathbf{h}_{s,t}$ in Eq. 1) rather than fitted at each time step.

Every searchable variable is registered in a canonical parameter ontology with a complete provenance and execution record (full schema in App. D.1). The LLM, memory, surrogate, and Judge communicate only through canonical parameter IDs. Before simulation, a semantic firewall rejects unknown IDs, unit/support mismatches, out-of-range values, and incomplete adapters. Multi-parameter interventions must compile in full, and frozen artifacts store absolute physical values rather than ambiguous multipliers.

2.3 Mechanism Hypotheses and Scientific Memory

We first run SUEWS with the default parameters $\boldsymbol{\theta}_s^0$. The resulting hourly residuals, $r_{s,y,t} = \hat{y}_{s,y,t}(\boldsymbol{\theta}_s^0) - y_{s,y,t}$, are summarized overall and by season, day/night, and wetness regime. Given these diagnostics, site context $\mathbf{c}_s$, and retrieved records, the Mechanism Reasoner proposes

$$H = (p, \{\mathrm{id}_j, \mathrm{dir}_j, T_j\}_j, \{\rho_y\}_{y \in \mathcal{Y}_s}, \mathcal{F}), \tag{3}$$

where p denotes the suspected limiting process, and each parameter proposal specifies a canonical ID, direction, and bounded trust region. The pre-experiment expectations ρ_y specify a targeted error reduction ($\Delta E_{s,y} < 0$), non-degradation ($\Delta E_{s,y} \leq 0$), or uncertainty for each evaluated output. They refer to prediction errors rather than necessarily to the physical direction of a flux. The recorded failure conditions $\mathcal{F}$ identify responses that would limit the hypothesis; they do not modify the fixed Physics Judge (Sec. 2.4). SUEWS evaluates each compiled intervention, and the Judge assesses its output-wise error effects and gate outcomes. We retain each trial as

$$m_i = (\mathbf{c}_i, H_i, \Delta\boldsymbol{\theta}_i, \Delta\mathbf{E}_i, \mathbf{g}_i, o_i), \tag{4}$$

where $\Delta\mathbf{E}_i$ contains the relative error changes for Q_H, Q_E, and LW_{up}, $\mathbf{g}_i$ records the gate results, and o_i summarizes whether the observed response supports, fails to support, contradicts, or limits the stated expectation. Uninterpretable tests remain inconclusive. This hypothesis status does not replace Judge acceptance or the sealed-evaluation grade; the full effects and failure rationale remain in the record (App. D.2).

Retrieval selects records from relevant mechanism and parameter families and ranks them by site context. Successful and failed trials inform the next proposal, but their numerical values are not adopted without local SUEWS evaluation. The record m_i, rather than an isolated parameter value, therefore represents evidence conditional on the tested intervention and site context.

2.4 Physics-Gated Candidate Verification

For output y, define

$$E_{s,y}(\boldsymbol{\theta}) = \frac{\sqrt{N_{s,y}^{-1} \sum_t (\widehat{y}_{s,t}(\boldsymbol{\theta}) - y_{s,t})^2}}{\mathrm{SD}(y_{s,t})}, \qquad J_s(\boldsymbol{\theta}) = \frac{1}{|\mathcal{Y}_s|} \sum_{y \in \mathcal{Y}_s} E_{s,y}(\boldsymbol{\theta}), \tag{5}$$

where $N_{s,y}$ is the number of valid observation simulation pairs for output y at site s, $\mathrm{SD}(y_{s,t})$ the standard deviation of the observed series (placing all outputs on a common dimensionless scale), and $|\mathcal{Y}_s|$ the number of evaluated outputs. Let $\Delta E_{s,y} = [E_{s,y}(\boldsymbol{\theta}) - E_{s,y}(\boldsymbol{\theta}_s^0)]/E_{s,y}(\boldsymbol{\theta}_s^0)$ denote the

relative change in per-output error and ΔJ_s the analogous relative change in the aggregate objective J_s, both measured against the default-parameter baseline $\boldsymbol{\theta}_s^0$.

The PHYSICS JUDGE applies acceptance criteria in three layers, each targeting a distinct failure mode: aggregate improvement alone cannot rule out cross-process compensation, full-period averages can mask temporally localized degradation, and neither guards against invalid or unstable candidates. First, it requires material aggregate improvement $\Delta J_s \leq -5 \times 10^{-4}$ and full-period non-degradation $\max_y \Delta E_{s,y} \leq 0$. Second, it evaluates four contiguous adaptation blocks and rejects any candidate whose error on any process increases by more than 10^{-3} within a block. Third, semantic validity, parameter bounds, finite output, solver completion, and state propagation are enforced as hard constraints.

Operational hypothesis outcomes. $\{\rho_y\}$ and $\mathcal{F}$ are interpreted using the same simulator effects and Judge checks, without introducing a new statistical test, threshold, or selection objective. Each tested intervention receives one of five statuses:

- `SUPPORTED`: the stated target response and non-degradation expectations hold, and all applicable Judge gates pass. Support is specific to the tested simulator, intervention, outputs, and context.
- `UNSUPPORTED`: valid evidence shows the targeted response is absent; failing to improve aggregate error alone does not identify which expectation failed.
- `CONTRADICTED`: aggregate gain accompanies degradation of an evaluated process, contradicting the proposed non-degrading correction - an operational trade-off, not proof of its physical cause.
- `CONTEXT_LIMITED`: the full-period response matches the hypothesis, but a temporal-block check fails, limiting support to within the tested period.
- `INCONCLUSIVE`: evidence is insufficient or ambiguous, the pre-experiment expectation is missing, or validity failures prevent an interpretable test. A favorable but sub-material response does not qualify as support.

When a single summary status is needed, invalid or uninterpretable evidence is marked inconclusive first; a process contradiction then takes precedence over an absent target response, followed by temporal limitation where the full-period response is otherwise as expected. Support requires both the stated response and all gates to pass; all other cases are inconclusive. Every applicable failure reason is retained even when only one label is reported. These statuses annotate existing evidence rather than add a new gate - a Judge pass with no recorded, testable expectation is not automatic mechanism support - and leave reported acceptance counts and temporal grades under their original definitions, not reclassified as hypothesis-status counts.

Among candidates that satisfy all requirements, selection is lexicographic: minimize the worst per-output error effect $\max_y \Delta E_{s,y}$, then ΔJ_s, then the number of modified parameters; if none is admissible, the agent abstains and retains the default configuration. Detailed definitions of the temporal robustness checks, uncertainty qualification, and auxiliary diagnostics are provided in App. D.

2.5 ITERATIVE REFINEMENT AND LEAKAGE-FREE EVALUATION

At adaptation round r, the agent maintains a mechanism hypothesis h_r (an instance of H from Eq. 3), a search policy π_r, and an adaptation-safe memory $\mathcal{M}_r$ (accumulating records m_i of Eq. 4). After evaluating a candidate intervention $\Delta\boldsymbol{\theta}_r$ with SUEWS, the PHYSICS JUDGE returns a gate outcome g_r and, when rejected, the violated process constraints. Comparing this evidence with the recorded expectations yields the outcome o_r, which is stored with the trial before the next retrieval.

We enforce a strict temporal separation between spin-up, adaptation, and sealed evaluation:

$$\mathcal{T} = \mathcal{T}_{\text{spin}} \cup \mathcal{T}_{\text{adapt}} \cup \mathcal{T}_{\text{eval}}, \qquad \mathcal{T}_{\text{adapt}} \cap \mathcal{T}_{\text{eval}} = \varnothing, \tag{6}$$

where $\mathcal{T}_{\text{spin}}$, $\mathcal{T}_{\text{adapt}}$, and $\mathcal{T}_{\text{eval}}$ are the spin-up, adaptation, and sealed evaluation intervals, respectively. Spin-up advances the SUEWS state but contributes no evidence to parameter selection. All diagnosis, retrieval, candidate generation, surrogate fitting, simulation, judging, and memory updates are restricted to $\mathcal{T}_{\text{adapt}}$. At the adaptation evaluation boundary, the selected intervention - or abstention decision - is frozen together with the simulator configuration in an immutable manifest. The sealed evaluation then propagates the SUEWS state once through $\mathcal{T}_{\text{eval}}$, with no further search, tuning, or memory updates. This protocol measures future-period generalization at observed sites.

Algorithm 1 (App. F) summarizes the procedure. Implementation details are provided in App. D, with memory and prompt details in App. E.

2.6 EXPERIMENTAL SETUP

Data and temporal protocol. We evaluate URBANSURFACE-AGENT on 20 Urban-PLUMBER urban sites (Lipson et al., 2022b), using the common prediction task of sensible heat flux Q_H, latent heat flux Q_E, and upward longwave radiation $LW_\uparrow$. Quality-controlled meteorological forcing and site surface metadata are supplied to SUEWS; flux observations are used only for diagnosis and scoring. Each site uses a 365-day spin-up (justified in App. C), followed by a chronological 70/30 split of the valid observational record into adaptation and sealed evaluation intervals, with state propagated continuously across all intervals.

Baseline and discovery protocol. The physical baseline is the site-configured SUEWS model with its original lookup-table parameters and observation-constrained surface facts. Search is restricted to registered Layer-II parameters with executable SUEWS adapters (App. D.1); interventions remain within $\pm 5\%$ of the site baseline and modify at most two parameters, unless a narrower physical range applies. The complete method performs at most five diagnosis rounds and 40 candidate SUEWS evaluations per site and seed (full hyperparameters in Tab. 6). The LLM proposes the limiting process, canonical parameter IDs, directions, and trust regions; deterministic code selects numerical values, and the PHYSICS JUDGE alone owns acceptance (Sec. 2.4).

Comparisons and ablations. We compare against fixed SUEWS (Ward et al., 2016), random search (Bergstra & Bengio, 2012), sensitivity-guided search (Morris, 1991), constrained Bayesian optimization (Jones et al., 1998; Gardner et al., 2014), and memory-free LLM search. All methods use the same sites, temporal split, base parameter registry, physical bounds, simulator, and PHYSICS JUDGE, with all screening costs included. Sensitivity is capped at 25 total SUEWS calls and treats each bilateral one-at-a-time (OAT) probe as an admissible candidate. Random search, Bayesian optimization, and the memory-free LLM instead perform a complete OAT screen (mean 24.8 calls) followed by 25 search evaluations. We therefore evaluate all methods using a consistent end-to-end budget accounting protocol in Tab. 2, and report discovery as a function of cumulative SUEWS calls in Fig. 3(b). Complete protocols and component ablations are provided in App. G.5.

Metrics. Candidate selection minimizes the equally weighted mean nRMSE over Q_H, Q_E, and $LW_\uparrow$ (Eq. 5); RMSE, MAE, bias, Pearson correlation are reported as auxiliary diagnostics where defined. Search performance is measured by valid discovery rate, simulation efficiency, and calls to first discovery. Frozen temporal transfer is assessed by aggregate gain, process-wise non-degradation, generalization retention, process consistency, and $D_{\max} = 100 \max_y \max(0, \Delta E_y)$ (%). Sealed-period grades A+–D summarize frozen interventions, with A+/A treated as successful transfers and U denoting abstention; detailed grade definitions are given in App. G.6. Grades are diagnostic only and never feed back into adaptation. Randomized comparisons use five seeds, with uncertainty estimated by resampling sites.

3 RESULTS

We organize the evidence around four questions: whether scalar improvement survives independent adjudication, how failures differ across sites, how stored evidence informs hypothesis revision, and which parameter process relationships remain supported after temporal transfer.

3.1 AGGREGATE IMPROVEMENT AND SCIENTIFIC ADMISSIBILITY DIVERGE

The adaptation archive makes the distinction concrete (Fig. 2). Among 1,410 completed candidates, 862 lack material aggregate gain, 350 improve the aggregate objective but first fail full-period process consistency, 89 first fail temporal consistency, 34 fail a remaining mechanism or validity check, and 75 pass strictly. These are mutually exclusive first-failure categories, not independent counts of every violated constraint; pre-simulation semantic rejections and solver failures are outside this denominator. Thus, a lower scalar error frequently fails to support an admissible intervention: process

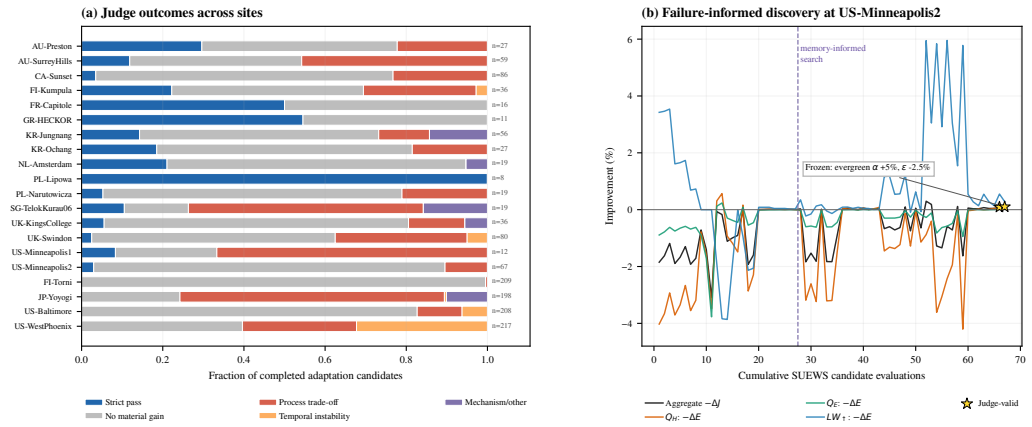

Figure 2: Population- and case-level evidence for physics-gated discovery. **(a)** Primary outcome of all 1,410 completed adaptation candidates, normalized within site. Rejected candidates are assigned to the first violated numerical gate; counts at right report the number of completed candidates. **(b)** Discovery trajectory at US-Minneapolis2. The cold-start search produces no Judge-valid intervention; failure-conditioned memory redirects the subsequent search to the evergreen surface, yielding a frozen 5% albedo increase and 2.5% emissivity decrease. Pre-simulation semantic rejections and solver failures are excluded from panel (a).

Table 1: End-to-end component ablation on the 20-site benchmark. Discovery is measured during adaptation, whereas evaluation gain and deployment outcomes are measured after freezing on the sealed future period.

Variant	Discovery (%)	Eval. gain (%)	Practical success (%)	Sealed failure (%)
Sensitivity search + Judge	25	0.224	80.0	20.0
LLM + Judge, no memory	45	0.204	100.0	0.0
LLM + memory, no Judge	100	0.631	55.0	45.0
Full system	80	0.356	87.5	12.5

Notes. Evaluation gain is averaged over all site runs, with zero gain assigned to abstentions. Practical success requires $-\Delta J_{\text{eval}} \geq 0.05\%$ and $D_{\max,\text{eval}} \leq 0.1\%$, where $D_{\max,\text{eval}} = \max_y \Delta E_{y,\text{eval}}$. Sealed failure denotes an accepted intervention that does not satisfy these future-period criteria.

trade-offs are evidence against accepting the tested correction. A site-level case study in App. H.8 further shows how differences in mechanism prioritization lead to distinct search trajectories.

The component comparison shows the consequence of removing this distinction (Tab. 1). The no-Judge branch freezes updates at all 20 sites, but only 17 retain positive aggregate transfer and only 11 meet the practical sealed-period criteria; six exhibit process trade-offs and three lose positive aggregate transfer (App. H.6). Its higher all-site gain (0.631% versus 0.356%) accompanies a sealed failure rate of 45%, compared with 12.5% for the full system: numerical gain and admissibility answer different questions.

3.2 Distinct Failure Signatures Constrain the Next Hypothesis

Failure evidence distinguishes unresolved sites rather than assigning them a single unsuccessful-search label. FI-Torni is dominated by ineffective interventions (208/209 completed candidates); JP-Yoyogi is dominated by process trade-offs (129/198); and US-WestPhoenix combines ineffective, process-conflicted, and temporally unstable responses (Fig. 2(a); App. H.6). Under Sec. 2.4, an absent target response is unsupported, aggregate gain with process degradation contradicts the non-degradation expectation, and a full-period response that fails temporal checks is context-limited; these first-failure counts identify distinct ways a correction fails the protocol, but do not by themselves distinguish every finer-grained hypothesis outcome.

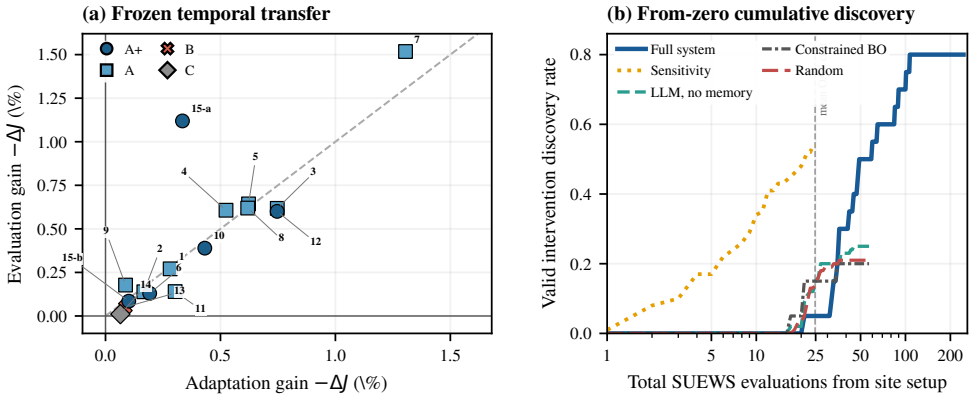

Figure 3: Frozen temporal transfer and from-zero discovery on the 20-site benchmark. **(a):** adaptation versus sealed-evaluation gain for the 16 frozen interventions. Marker shape denotes grade, numeric labels refer to Tab. 8, and the dashed line is $y = x$. **(b):** cumulative Judge-valid discovery rate against total SUEWS calls, including screening.

The refinement protocol uses those distinctions explicitly (App. D.4). Insufficient aggregate gain requires reconsidering the limiting process or direction; process degradation adds the affected output to the retrieval risk set; temporal failure records an applicability boundary. These are operational failure mechanisms within the tested intervention space, not diagnoses of missing physical laws. FI-Torni, JP-Yoyogi, US-Baltimore, and US-WestPhoenix retain their baselines after continued refinement: abstention records the absence of a supported correction within the bounds and budget, not the absence of any useful correction outside that search.

3.3 Scientific Memory Supports Hypothesis Revision

US-Minneapolis2 provides a concrete trace of how retrieved evidence informs the next experiment (Fig. 2(b); App. H.4). The cold-start episode tests 27 candidates without a strict pass. After the generation boundary, local failures and context-conditioned cross-site evidence redirect testing toward the evergreen surface. Two strict candidates appear at calls 39 40 of the memory-informed episode, and the frozen intervention increases evergreen albedo by 5% and decreases emissivity by 2.5%. On the sealed period, the intervention yields 0.085% aggregate gain and grade A+. The trace illustrates revision of the surface-specific hypothesis followed by local testing, rather than direct adoption of a retrieved parameter value. Across sites, frozen coverage rises from 9/20 in the cold start to 14/20 and then 16/20, as the candidate archive grows from 476 to 816 and 972 trials; coverage subsequently remains unchanged through 1,410 trials. Because later generations also receive additional simulator calls, this trajectory reflects the joint contribution of accumulated evidence and continued search rather than an equal-budget isolation of memory's effect.

3.4 Conditional Parameter Process Evidence and Its Temporal Limits

The frozen interventions connect parameter families to different output pathways (App. H.3 and H.7). Radiative interventions produce pronounced changes in $LW_{\uparrow}$: sealed-period error reductions are 16.624% at KR-Jungnang, 4.940% at FR-Capitole, 4.532% at KR-Ochang, and 3.949% at FI-Kumpula. Their building-emissivity directions differ: a decrease at FI-Kumpula contrasts with increases at FR-Capitole, KR-Jungnang, and PL-Lipowa. Conductance interventions primarily affect Q_H Q_E partitioning; conductance decreases at CA-Sunset, PL-Narutowicza, and US-Minneapolis1, but increases at NL-Amsterdam and SG-TelokKurau06. Paired aerodynamic changes occur at GR-HECKOR and UK-Swindon, where morphological support permits those tests. Absolute values and signed changes remain in Tab. 9 and Fig. 5.

These observations support retaining each intervention with its surface class, context, residual signature, and output response, and rule out a single correction direction for the reported sites. They do not identify which contextual variable causes a given direction reversal, nor the separate causal con-

Table 2: Operational search outcomes with all screening and refinement costs included. Budgets differ across policies. Frozen and practical/exact success rates use all sites or site seed episodes, whereas frozen gain conditions on intervention.

Method	Calls/ site	Frozen (%)	Test nRMSE ↓	All-site gain (%)↑	Frozen gain (%)↑	Final A+/A (%)	Exact succ. (%)
Fixed SUEWS baseline	0.0	0	0.6556	0.000	-	-	-
Random + Judge	49.8	21	0.6542	0.186	0.884	18	7
Sensitivity + Judge	25.0	54	0.6538	0.247	0.457	54	25
Constrained BO + Judge	49.8	20	0.6540	0.217	1.084	15	5
LLM, no memory + Judge	49.8	25	0.6540	0.212	0.848	23	10
Full evolving-memory system	**95.3**	**80**	**0.6532**	**0.356**	**0.445**	**70**	**25**

tribution of each parameter in a paired intervention: the conditional knowledge here is a collection of locally tested relationships and observed failures, not a newly inferred parameterization formula.

Temporal evaluation further bounds those relationships (Tab. 8). All 16 frozen interventions retain positive aggregate gain, but only five remain exactly non-degrading (A+); nine meet the practical process margin of 0.1% and material gain threshold of 0.05% (A). These 14 A+/A sites comprise the reported high-confidence set under that operational criterion. UK-KingsCollege has positive aggregate gain but 0.108% longwave degradation (B); UK-Swindon remains within the process margin but has only 0.011% aggregate gain (C). A useful adaptation response thus need not remain exactly process-safe or material in a future period; grades are diagnostic and never trigger retuning or memory updates. This establishes within-site, not cross-city, temporal transfer (App. H.9).

3.5 Search Coverage, Predictive Effects, and Computational Cost

Tab. 2 reports end-to-end operational cost; budgets differ across methods. Sensitivity is the strongest simple baseline, freezing interventions in 54% of site seed episodes with 25 SUEWS calls per site, a test nRMSE of 0.6538, and an all-site gain of 0.247%. Random search, constrained Bayesian optimization, and the memory-free LLM average 49.8 calls per site but freeze updates in only 20 25% of episodes. The full system uses 95.3 calls per site and increases frozen coverage to 80%.

Across its 16 frozen interventions, sealed gains range from 0.011% to 1.518%. Assigning zero gain to abstentions yields an all-site gain of 0.356% and a test nRMSE of 0.6532, compared with 0.6556 for fixed SUEWS. Constrained BO achieves a larger gain conditional on acceptance (1.084% versus 0.445%) but substantially lower coverage. Sensitivity and the full system both attain 25% exact held-out success, while the full system reaches 70% practical success. Fig. 3 shows temporal transfer and cumulative search cost; App. H.1 provides detailed accounting. App. H.5 reports a budget-matched ablation with its site-level heterogeneity and statistical limits.

4 Conclusion

UrbanSurface-Agent turns residual diagnostics into bounded, testable parameter hypotheses and uses a Physics Judge to separate aggregate improvement from process- and time-consistent admissibility. Across 1,410 candidate evaluations at 20 Urban-PLUMBER sites, only 75 satisfy all adaptation-time gates. Sixteen sites receive frozen interventions that retain positive aggregate transfer during the sealed period, with 14 meeting the practical success criteria. The heterogeneous and sometimes opposite correction directions within the same parameter family argue against universal numerical corrections and instead support context-conditioned parameter process evidence. The current results establish within-site temporal transfer but not unique causal identification or generalization to unseen sites. Future work will test cross-site transfer, isolate the contribution of memory under matched budgets, and strengthen causal and conservation-aware verification.

AI USE STATEMENT

Generative AI tools were used for language editing, manuscript framing, and reorganizing the presentation of existing evidence. Within the proposed system, an LLM was used during adaptation as the Mechanism Reasoner to formulate process hypotheses and propose sparse, bounded parameter interventions, as described in Sec. 2 and App. E. The LLM did not generate experimental data, execute SUEWS, determine numerical parameter values, evaluate candidates, or make acceptance decisions. All numerical results were produced by the simulator and verified against the predefined Physics Judge criteria. The authors reviewed all AI-assisted content and take responsibility for the final text, claims, and reported results.

REPRODUCIBILITY STATEMENT

The simulator configuration, parameter ontology, candidate-generation protocol, Physics Judge criteria, prompt and output contract, search budgets, data partitions, and site-level frozen interventions are documented in Apps. B H. All reported interventions are selected using adaptation-period data and evaluated without modification on the sealed future period. Code and configuration files will be released in an anonymous supplementary repository.

REFERENCES

Shahryar K. Ahmad, Sujay V. Kumar, Clara Draper, and Rolf H. Reichle. Challenges in unifying physically based and machine learning simulations through differentiable modeling: A land surface case study. *Geophysical Research Letters*, 52(4):e2024GL112893, 2025. doi: https://doi.org/10.1029/2024GL112893. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2024GL112893. e2024GL112893 2024GL112893.

Mohammad Almansoori, Komal Kumar, and Hisham Cholakkal. Medagentsim: Self-evolving multi-agent simulations for realistic clinical interactions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 362–372. Springer, 2025.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(10):281–305, 2012. URL https://www.jmlr.org/papers/v13/bergstra12a.html.

Fabian Bernhard, James Knighton, Stefan Seeger, Peter Waldner, and Katrin Meusburger. Lwf-brook90. jl—including stable water isotopes in a soil vegetation atmosphere transport model to constrain vertical root water uptake dynamics. *Journal of Advances in Modeling Earth Systems*, 17(11):e2024MS004445, 2025.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023. doi: 10.1038/s41586-023-06792-0. URL https://doi.org/10.1038/s41586-023-06792-0.

Yifan Cheng, Lei Zhao, Tirthankar Chakraborty, Keith Oleson, Matthias Demuzere, Xiaoping Liu, Yangzi Che, Weilin Liao, Yuyu Zhou, and Xinchang Li. U-surf: a global 1 km spatially continuous urban surface property dataset for kilometer-scale urban-resolving earth system modeling. *Earth System Science Data Discussions*, 2024:1–38, 2024.

Keyan Ding, Jing Yu, Junjie Huang, Yuchen Yang, Qiang Zhang, and Huajun Chen. Scitoolagent: a knowledge-graph-driven scientific agent for multitool integration. *Nature Computational Science*, 5(10):962–972, 2025.

Peilin Feng, Zhutao Lv, Junyan Ye, Xiaolei Wang, Xinjie Huo, Jinhua Yu, Wanghan Xu, Wenlong Zhang, Lei Bai, Conghui He, et al. Earth-agent: Unlocking the full landscape of earth observation with agents. In *International Conference on Learning Representations*, volume 2026, pp. 55700–55764, 2026.

Jacob R. Gardner, Matt J. Kusner, Zhixiang Eddie Xu, Kilian Q. Weinberger, and John P. Cunningham. Bayesian optimization with inequality constraints. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 937–945, 2014. URL https://proceedings.mlr.press/v32/gardner14.html.

Alireza Ghafarollahi and Markus J Buehler. Sciagents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22):2413523, 2025.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, Anil Palepu, Keran Rong, Ryutaro Tanno, Khaled Saab, Fan Zhang, Jacob Blum, Andrew Carroll, Kavita Kulkarni, Nenad Tomašev, Dina Zverinski, Ivor Rendulic, Elahe Vedadi, Florian Hasler, Luka Rimanic, Marina Boia, Ivan Budiselic, Ben Feinstein, Mathias Bellaiche, Tom Sheffer, Jan Freyberg, Jeremy Ratcliff, Ottavia Bertolli, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R. D. Costa, José R. Penadés, Gary Peltz, Yossi Matias, James Manyika, Demis Hassabis, Yunhan Xu, Pushmeet Kohli, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Accelerating scientific discovery with Co-Scientist. *Nature*, 655(8122):487–496, July 2026. ISSN 1476-4687. doi: 10.1038/s41586-026-10644-y. URL https://doi.org/10.1038/s41586-026-10644-y.

Sue Grimmond, Ting Sun, Lewis Blunn, Vitor Lavor, Fredrik Lindberg, Yiqing Liu, Keisuke Nakao, Matthew Paskin, Silvia Rognone, Xiaoxiong Xie, Leena Järvi, Zhenkun Li, Thomas Loridan, Brian D. Offerle, Hamidreza Omidvar, and Helen Ward. Suews: Surface urban energy and water balance scheme, June 2026. URL https://doi.org/10.5281/zenodo.20555317.

Hoshin V Gupta, Harald Kling, Koray K Yilmaz, and Guillermo F Martinez. Decomposition of the mean squared error and nse performance criteria: Implications for improving hydrological modelling. *Journal of hydrology*, 377(1-2):80–91, 2009.

Hoshin Vijai Gupta, Soroosh Sorooshian, and Patrice Ogou Yapo. Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. *Water Resources Research*, 34(4):751–763, 1998. doi: https://doi.org/10.1029/97WR03495. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/97WR03495.

P. Huang, A. Ducharne, L. Rinchiuso, J. Polcher, L. Baratgin, V. Bastrikov, and E. Sauquet. Multi-objective calibration and evaluation of the orchidee land surface model over france at high resolution. *Hydrology and Earth System Sciences*, 28(19):4455–4476, 2024. doi: 10.5194/hess-28-4455-2024. URL https://hess.copernicus.org/articles/28/4455/2024/.

R. O. Imhoff, W. J. van Verseveld, B. van Osnabrugge, and A. H. Weerts. Scaling point-scale (pedo)transfer functions to seamless large-domain parameter estimates for high-resolution distributed hydrologic modeling: An example for the rhine river. *Water Resources Research*, 56(4):e2019WR026807, 2020. doi: https://doi.org/10.1029/2019WR026807. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019WR026807. e2019WR026807 10.1029/2019WR026807.

Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998. doi: 10.1023/A:1008306431147.

H. J. Jongen, M. Lipson, A. J. Teuling, S. Grimmond, J.-J. Baik, M. Best, M. Demuzere, K. Fortuniak, Y. Huang, M. G. De Kauwe, R. Li, J. McNorton, N. Meili, K. Oleson, S.-B. Park, T. Sun, A. Tsiringakis, M. Varentsov, C. Wang, Z.-H. Wang, and G. J. Steeneveld. The water balance representation in urban-plumber land surface models. *Journal of Advances in Modeling Earth Systems*, 16(10):e2024MS004231, 2024. doi: https://doi.org/10.1029/2024MS004231. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2024MS004231. e2024MS004231 2024MS004231.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory OS of AI agent. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25961–25970, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1318. URL https://aclanthology.org/2025.emnlp-main.1318/.

Chia Hsiang Kao, Wenting Zhao, Cheryl Lam, Aarush Umap, Shreelekha Revankar, Samuel Speas, Snehal Bhagat, Rajeev Datta, Cheng Perng Phoo, Utkarsh Mall, et al. Towards llm agents for earth observation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 2597–2611, 2026.

TV Kokkonen, C Sue B Grimmond, O Räty, HC Ward, A Christen, TR Oke, Simone Kotthaus, and L Järvi. Sensitivity of surface urban energy and water balance scheme (suews) to downscaling of reanalysis forcing data. *Urban Climate*, 23:36–52, 2018.

Simone Leo, Michael R. Crusoe, Laura Rodríguez-Navas, Raül Sirvent, Alexander Kanitz, Paul De Geest, Rudolf Wittner, Luca Pireddu, Daniel Garijo, José M. Fernández, Iacopo Colonnelli, Matej Gallo, Tazro Ohta, Hirotaka Suetake, Salvador Capella-Gutierrez, Renske de Wit, Bruno P. Kinoshita, and Stian Soiland-Reyes. Recording provenance of workflow runs with ro-crate. *PLOS ONE*, 19(9):e0309210, 2024. ISSN 1932-6203. doi: 10.1371/journal.pone.0309210. URL http://dx.doi.org/10.1371/journal.pone.0309210.

M. Lipson, S. Grimmond, M. Best, W. T. L. Chow, A. Christen, N. Chrysoulakis, A. Coutts, B. Crawford, S. Earl, J. Evans, K. Fortuniak, B. G. Heusinkveld, J.-W. Hong, J. Hong, L. Järvi, S. Jo, Y.-H. Kim, S. Kotthaus, K. Lee, V. Masson, J. P. McFadden, O. Michels, W. Pawlak, M. Roth, H. Sugawara, N. Tapper, E. Velasco, and H. C. Ward. Harmonized gap-filled datasets from 20 urban flux tower sites. *Earth System Science Data*, 14(11):5157–5178, 2022a. doi: 10.5194/essd-14-5157-2022. URL https://essd.copernicus.org/articles/14/5157/2022/.

Mathew Lipson, Sue Grimmond, Martin Best, Winston Chow, Andreas Christen, Nektarios Chrysoulakis, Andrew Coutts, Ben Crawford, Stevan Earl, Jonathan Evans, Krzysztof Fortuniak, Bert G. Heusinkveld, Je-Woo Hong, Jinkyu Hong, Leena Järvi, Sungsoo Jo, Yeon-Hee Kim, Simone Kotthaus, Keunmin Lee, Valéry Masson, Joseph P. McFadden, Oliver Michels, Wlodzimierz Pawlak, Matthias Roth, Hirofumi Sugawara, Nigel Tapper, Erik Velasco, and Helen Claire Ward. Data for "harmonized gap-filled dataset from 20 urban flux tower sites" for the urban-plumber project, September 2022b. URL https://doi.org/10.5281/zenodo.7104984.

Mathew J Lipson, Sue Grimmond, Martin Best, Gab Abramowitz, Andrew Coutts, Nigel Tapper, Jong-Jin Baik, Meiring Beyers, Lewis Blunn, Souhail Boussetta, et al. Evaluation of 30 urban land surface models in the urban-plumber project: Phase 1 results. *Quarterly Journal of the Royal Meteorological Society*, 150(758):126–169, 2024.

Tennison Liu, Nicolás Astorga, Nabeel Seedat, and Mihaela van der Schaar. Large language models to enhance bayesian optimization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=OOxotBmGol.

Pingchuan Ma, Tsun-Hsuan Wang, Minghao Guo, Zhiqing Sun, Joshua B. Tenenbaum, Daniela Rus, Chuang Gan, and Wojciech Matusik. LLM and simulation as bilevel optimizers: A new paradigm to advance physical scientific discovery. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 33940–33962. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/ma24m.html.

Max D. Morris. Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2):161–174, 1991. doi: 10.1080/00401706.1991.10484804.

J Eamonn Nash and Jonh V Sutcliffe. River flow forecasting through conceptual models part i—a discussion of principles. *Journal of hydrology*, 10(3):282–290, 1970.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

Siru Ouyang, Jun Yan, I-Hung Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long Le, Samira Daruki, Xiangru Tang, Vishy Tirumalashetty, George Lee, Mahsan Rofouei, Hangfei Lin, Jiawei Han, Chen-Yu Lee, and Tomas Pfister. Reasoningbank: Scaling agent self-evolving with reasoning memory. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 94327–94354, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/980ea04d23d1f6908964eba2a74afe45-Paper-Conference.pdf.

Karl Pearson. Note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 58:240–242, 1895.

N. Raoult, S. Beylat, J. M. Salter, F. Hourdin, V. Bastrikov, C. Ottlé, and P. Peylin. Exploring the potential of history matching for land surface model calibration. *Geoscientific Model Development*, 17(15):5779–5801, 2024. doi: 10.5194/gmd-17-5779-2024. URL https://gmd.copernicus.org/articles/17/5779/2024/.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 625:468–475, 2024. doi: 10.1038/s41586-023-06924-6.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Chuanneng Sun, Songjun Huang, and Dario Pompili. Llm-based multi-agent decision-making: Challenges and future directions. *IEEE Robotics and Automation Letters*, 10(6):5681–5688, 2025a.

Ting Sun and Sue Grimmond. A python-enhanced urban land surface model supy (suews in python, v2019. 2): development, deployment and demonstration. *Geoscientific Model Development*, 12 (7):2781–2795, 2019.

Ting Sun, Hamidreza Omidvar, Zhenkun Li, Ning Zhang, Wenjuan Huang, Simone Kotthaus, Helen C Ward, Zhiwen Luo, and Sue Grimmond. Wrf (v4. 0)–suews (v2018c) coupled system: development, evaluation and application. *Geoscientific Model Development*, 17(1):91–116, 2024.

Yuan Sun, Keith W. Oleson, Lei Zhao, Gerald Mills, Cenlin He, Matthias Demuzere, David O. Topping, Ning Zhang, and Zhonghua Zheng. Enhancing global-scale urban land cover representation using local climate zones in the community earth system model. *Journal of Advances in Modeling Earth Systems*, 17(11):e2025MS004934, 2025b. doi: https://doi.org/10.1029/2025MS004934. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2025MS004934. e2025MS004934 2025MS004934.

G. E. Tucker, E. W. H. Hutton, M. D. Piper, B. Campforts, T. Gan, K. R. Barnhart, A. J. Kettner, I. Overeem, S. D. Peckham, L. McCready, and J. Syvitski. Csdms: a community platform for numerical modeling of earth surface processes. *Geoscientific Model Development*, 15(4):1413–1439, 2022. doi: 10.5194/gmd-15-1413-2022. URL https://gmd.copernicus.org/articles/15/1413/2022/.

Sanidhya Vijayvargiya, Aditya Soni, Xuhui Zhou, Zora Zhiruo Wang, Nouha Dziri, Graham Neubig, and Maarten Sap. Openagentsafety: A comprehensive framework for evaluating real-world ai agent safety. In *International Conference on Learning Representations*, volume 2026, pp. 99089–99114, 2026.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models, 2023. URL https://arxiv.org/abs/2305.16291.

Hanchen Wang, Yichun He, Paula P Coelho, Matthew Bucci, Abbas Nazir, Bob Chen, Linh T Trinh, Serena Zhang, Ziyu Lu, Kexin Huang, et al. Spatialagent: an autonomous ai agent for spatial biology. *bioRxiv*, pp. 2025–04, 2025.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. SciMON: Scientific inspiration machines optimized for novelty. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 279–299, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.18. URL https://aclanthology.org/2024.acl-long.18/.

Helen C. Ward, Simone Kotthaus, Leena Järvi, and C. S. B. Grimmond. Surface urban energy and water balance scheme (suews): Development and evaluation at two uk sites. *Urban Climate*, 18: 1–32, 2016. doi: 10.1016/j.uclim.2016.05.001.

Bingxian Wu, Yu Zhang, Zonghao Guo, Tang Liu, Chen Qian, Yuxiang Lu, Xingbo Du, Yanghao Li, Yidan Zhang, Chi Chen, Ling Yao, Chenghu Zhou, and Maosong Sun. RSMeM: Knowledge-enhanced memory evolution for remote Sensing agents with systematic evaluation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32899–32915, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1519. URL https://aclanthology.org/2026.acl-long.1519/.

Jiyang Xia, Fenghua Ling, Zhenhui Jessie Li, Junjie Yu, Hongliang Zhang, David Topping, LEI BAI, and Zhonghua Zheng. Learning urban climate dynamics via physics-guided urban surface–atmosphere interactions. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 125813–125840. Curran Associates, Inc., 2025. doi: 10.52202/085713-4194. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/b71cfefae46909178603b5bc6c11d3ae-Paper-Conference.pdf.

Zidi Xiong, Yuping Lin, Wenya Xie, Pengfei He, Zirui Liu, Jiliang Tang, Himabindu Lakkaraju, and Zhen Xiang. How memory management impacts LLM agents: An empirical study of experience-following behavior. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 623–645, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.27. URL https://aclanthology.org/2026.acl-long.27/.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 17577–17604. Curran Associates, Inc., 2025. doi: 10.52202/085713-0593. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/19909c36f51abc4856b4560aff3d36d6-Paper-Conference.pdf.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 12028–12068, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/3339f19c5fcee3ad74502947a32be9e6-Paper-Conference.pdf.

Sijie Zhao, Feng Liu, Xueliang Zhang, Hao Chen, Xinyu Gu, Zhe Jiang, Fenghua Ling, Ben Fei, Wenlong Zhang, Junjue Wang, et al. Openearth-agent: From tool calling to tool creation for open-environment earth observation. *arXiv preprint arXiv:2603.22148*, 2026.

A Background and Related Work

A.1 Urban Surface Parameterization and Semantic Interoperability

Urban land-surface models such as SUEWS (Sun et al., 2024; Lipson et al., 2024; Jongen et al., 2024) represent coupled energy and water exchanges across heterogeneous urban surfaces, but model skill varies across outputs and does not consistently increase with structural complexity (Lipson et al., 2024) - motivating process-wise evaluation over a single aggregate score, echoing multi-objective hydrological calibration (Gupta et al., 1998). We repurpose this principle as a gate rather than a loss term: non-degradation is a condition for acceptance, not one objective traded off against others. Semantic standards such as CSDMS Standard Names (Tucker et al., 2022) address variable *interoperability*, but not *admissibility* - whether a proposed change is physically meaningful and allowed. UrbanSurface-Agent targets admissibility directly by separating observation-constrained facts, uncertain physical parameters, and prognostic states, so that interventions are validated deterministically before execution.

A.2 LLM-Assisted Optimization and Scientific Agents

LLMs have been used for numerical search, proposal generation, and iterative scientific discovery: OPRO (Yang et al., 2024), LLAMBO (Liu et al., 2024), and FunSearch (Romera-Paredes et al., 2024) illustrate language-model-guided optimization, while SGA (Ma et al., 2024) and Coscientist (Boiko et al., 2023) extend this paradigm to simulation- and experiment-driven workflows. Our focus differs in explicitly constraining the intervention space and separating proposal from independent acceptance (Feng et al., 2026; Sun et al., 2025a; Vijayvargiya et al., 2026). Agent memory further supports experience reuse through reflection, executable skills, and structured records (Kang et al., 2025; Xiong et al., 2026; Shinn et al., 2023; Wang et al., 2023; Xu et al., 2025), with systems such as Co-Scientist retaining and refining hypotheses across iterations (Gottweis et al., 2026; Ouyang et al., 2026). UrbanSurface-Agent instead ties each memory record to the context and physical admissibility of a tested intervention, so that independent adjudication prevents aggregate gains from overriding physical-validity violations.

B SUEWS Simulator

B.1 Simulator Overview

The Surface Urban Energy and Water Balance Scheme (SUEWS) is a process-based urban land-surface model that simulates energy and water exchanges between heterogeneous urban surfaces and the atmosphere (Sun & Grimmond, 2019; Kokkonen et al., 2018; Grimmond et al., 2026). Within UrbanSurface-Agent, SUEWS instantiates the simulator $\mathcal{M}_{\mathrm{SUEWS}}$ of Eq. 1 and serves as the deterministic process model providing numerical evidence for candidate interventions. The language model neither predicts corrected fluxes nor modifies SUEWS outputs directly. Each candidate is instead compiled into valid native SUEWS inputs, simulated over the registered period, and evaluated by the independent Physics Judge (App. D.3).

SUEWS represents the urban surface using seven classes - paved surfaces, buildings, evergreen trees, deciduous trees, grass, bare soil, and water (**?**) - whose area fractions sum to one. Surface-specific radiative, thermal, vegetation, and hydrological properties are combined with site-scale morphological and aerodynamic parameters, all registered in the ontology of App. D.1. Given atmospheric forcing, SUEWS resolves surface-energy and water-balance processes, including heat storage, evapotranspiration, interception, drainage, and soil-moisture evolution. Candidate effects therefore arise through the model equations rather than through direct statistical correction of the evaluated observations.

For readability, this appendix follows the SUEWS literature's convention and suppresses the site subscript s used throughout the main text (e.g., $\mathbf{F}_t$ below corresponds to $\mathbf{x}_{s,t}$ in Eq. 1 for a fixed site s); all quantities are understood to be computed per site unless stated otherwise.

B.2 Forcing, Surface Data, and Prognostic States

We distinguish externally supplied forcing, surface data, and internally propagated state variables.

Atmospheric forcing. At time t, the meteorological forcing is

$$\mathbf{F}_t = (K_\downarrow, L_\downarrow, T_a, q_a, U, P, p_a)_t \,, \tag{7}$$

where $K_\downarrow$ and $L_\downarrow$ are incoming shortwave and longwave radiation, T_a is air temperature, q_a denotes atmospheric humidity, U is wind speed, P is precipitation, and p_a is air pressure; $\mathbf{F}_t$ plays the role of $\mathbf{x}_{s,t}$ in the main text. Variables are converted to the units and timestamp convention required by SUEWS before model execution. Flux observations are excluded from $\mathbf{F}_t$: they are used only for adaptation-period diagnosis and subsequent evaluation, never as model input (Sec. 2.1).

Surface parameters. The surface configuration contains (i) the fractions of the seven surface classes; (ii) morphological descriptors, such as building and vegetation height; (iii) radiative properties, including surface-indexed albedo and emissivity; (iv) site-scale aerodynamic quantities, including momentum roughness length and displacement height; and (v) vegetation, thermal-storage, interception, drainage, and soil-hydraulic parameters. Surface fractions and other externally observed properties are Layer-I quantities (App. D.1) and are fixed when supported by remote sensing, GIS, LiDAR, or site metadata. Only the uncertain Layer-II parameters registered in the canonical ontology may be modified by the discovery system. Every intervention is checked for unit, surface support, physical range, and native SUEWS mapping before execution (the six-check semantic firewall of App. D.1) and Tab. 5.

Prognostic states. SUEWS propagates the prognostic state vector as

$$\mathbf{h}_t = \mathcal{H}\left(\mathbf{h}_{t-1}, \mathbf{F}_t, \mathbf{c}, \boldsymbol{\theta}\right), \tag{8}$$

where $\mathbf{h}_t$ is the model state, $\mathbf{c}$ the fixed site context, $\boldsymbol{\theta}$ the model parameters, and $\mathcal{H}$ the SUEWS state-transition component of the simulator $\mathcal{M}_{\mathrm{SUEWS}}$ introduced in Eq. 1; that is, $\mathcal{M}_{\mathrm{SUEWS}}$ jointly returns the predicted outputs $\hat{\mathbf{y}}_t$ and the updated state $\mathbf{h}_t = \mathcal{H}(\cdot)$ at each step. Prognostic quantities, including soil moisture, intercepted water, phenological states, and heat storage, are evolved dynamically rather than optimized independently (Layer III, App. D.1). The final spin-up state initializes adaptation, and state evolution continues through the adaptation evaluation boundary without reset: the terminal adaptation state becomes the initial state for sealed evaluation, as justified in App. C.

B.3 Evaluated Outputs

SUEWS produces many diagnostic and prognostic variables. The 20-site benchmark restricts evaluation to the three outputs available under a common observation protocol across all sites:

$$\mathcal{Y} = \{Q_H, Q_E, LW_\uparrow\}. \tag{9}$$

Q_H is the turbulent sensible heat flux and responds to surface temperature, heat storage, and aerodynamic exchange. Q_E is the latent heat flux associated with evaporation and transpiration and depends on water availability, vegetation state, surface conductance, and aerodynamic transfer. $LW_\uparrow$ is outgoing longwave radiation and constrains effective surface temperature and emissivity. Their joint evaluation connects radiative, thermal, and hydrological behavior and exposes compensating changes that can be hidden by a single aggregate score - the central motivation for the multi-process Judge (Sec. 2.4).

B.4 Temporal Execution and Data Alignment

Each site has three registered chronological intervals, matching $\mathcal{T}_{\mathrm{spin}}$, $\mathcal{T}_{\mathrm{adapt}}$, and $\mathcal{T}_{\mathrm{eval}}$ in Eq. 6. The *spin-up period* advances thermal and hydrological states and contributes no selection metrics. The *adaptation period* is the only interval available for residual diagnosis, memory retrieval, candidate search, and PHYSICS JUDGE adjudication. The subsequent *evaluation period* remains sealed until a site-level intervention or baseline-retention decision has been frozen. Baseline and candidate simulations use identical forcing, initial-state provenance, timestamps, and output masks.

Model outputs are aligned to observations by site timestamp. A metric is computed only from finite baseline candidate observation triplets, ensuring both simulations are compared on exactly the same records. Missing observations are masked rather than interpolated into evaluation targets. The chronological split, rather than a random hourly split, prevents leakage through weather persistence, seasonality, and hydrological state memory.

Before the sealed run, an immutable manifest records the selected canonical values, SUEWS adapters, temporal boundaries, Judge thresholds, configuration, and code and input hashes (full freezing contract in App. D). Evaluation continues the initialized state exactly once; the resulting observations and metrics cannot trigger a new proposal, alter a parameter range, or update scientific memory.

B.5 Diagnostic Metrics

For each output y with N_y aligned model observation pairs, we compute

$$\mathrm{RMSE}_y = \sqrt{\frac{1}{N_y}\sum_{t=1}^{N_y}(\widehat{y}_t - y_t)^2}, \qquad \mathrm{Bias}_y = \frac{1}{N_y}\sum_{t=1}^{N_y}(\widehat{y}_t - y_t), \tag{10}$$

$$\mathrm{MAE}_y = \frac{1}{N_y}\sum_{t=1}^{N_y}|\widehat{y}_t - y_t|, \qquad \mathrm{nRMSE}_y = \frac{\mathrm{RMSE}_y}{\mathrm{SD}(y)}. \tag{11}$$

$\mathrm{nRMSE}_y(\boldsymbol{\theta})$ is identical to $E_{s,y}(\boldsymbol{\theta})$ defined in Eq. 5 for a fixed site s; we use the explicit RMSE/SD decomposition here to make the diagnostic components (bias, MAE) visible, and reuse the two symbols interchangeably elsewhere. We additionally report Pearson's correlation coefficient r (Pearson, 1895), Nash Sutcliffe efficiency (NSE) (Nash & Sutcliffe, 1970), and Kling Gupta efficiency (KGE) (Gupta et al., 2009) as diagnostic statistics when well defined. Pearson's r measures the linear association between simulated and observed temporal variability, complementing error-based metrics by capturing agreement in temporal co-variation rather than absolute error magnitude. These three diagnostics are reported for interpretation only and play no role in $J_s(\boldsymbol{\theta})$, Eq. 5, or in any PHYSICS JUDGE decision.

Adaptation-time acceptance is governed by the three-layer PHYSICS JUDGE of App. D.3 (Eq. 33), and near-Pareto candidates undergo the moving-block bootstrap qualification of the same appendix (Eq. 34); both apply unchanged to SUEWS simulations and are not restated here.

B.6 Sealed-Period Generalization

For a frozen intervention, we distinguish two held-out criteria that serve different purposes. *Aggregate temporal generalization* is declared when $\Delta J_s^{\mathrm{eval}} < 0$ (any held-out improvement, however small), while *strict temporal generalization* additionally requires

$$\Delta E_{s,y}^{\mathrm{eval}} \le 0, \qquad \forall y \in \mathcal{Y}_s. \tag{12}$$

Both are weaker than the *practical success* criterion used for grading (App. G.6), which requires $-\Delta J_s^{\mathrm{eval}} \ge 0.05\%$; aggregate/strict generalization report whether the sign of the held-out effect is favorable at all, whereas practical success additionally requires the effect to exceed a material threshold. If the agent abstains at a site, the original SUEWS configuration is retained; such cases are reported as baseline-retention decisions (grade U) and are excluded from intervention-level generalization rates.

C Justification of the Spin-Up Protocol

We test whether a fixed 365-day spin-up (used throughout Sec. 2.1 and App. B.4 as $\mathcal{T}_{\mathrm{spin}}$) is sufficient to stabilize the model outputs used in this study. Spin-up is driven only by meteorological forcing and surface data; flux observations are neither accessed nor assimilated. In the production workflow, the spun-up state is serialized as a typed SuPy checkpoint and restored without reinitialization. For the sensitivity analysis below, each configuration is instead integrated continuously from the start of spin-up through a common comparison period, with no intermediate state resets.

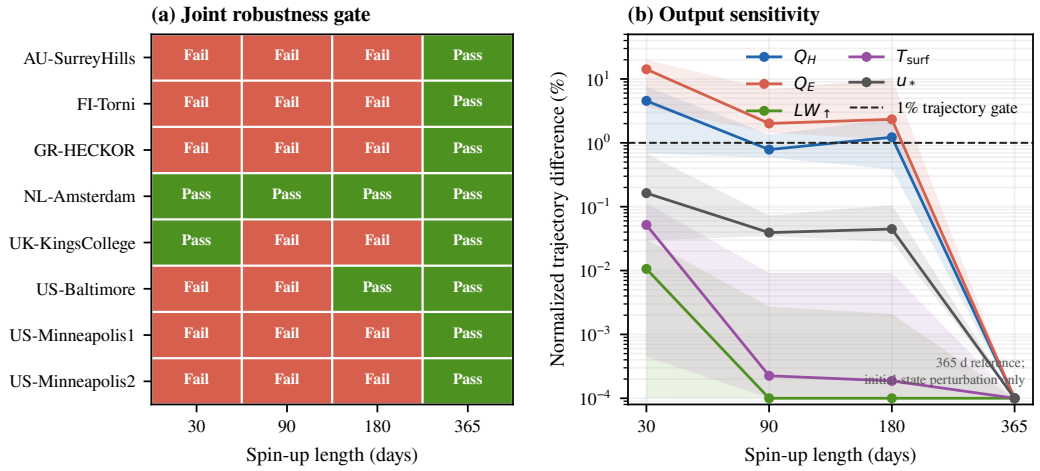

Figure 4: Spin-up robustness across the eight sites with nonzero open-water fractions. **(a)** Joint pass/fail outcomes under the preregistered robustness criteria. **(b)** Normalized output differences relative to the 365-day reference; the dashed line denotes the 1% trajectory threshold. T_{surf} and u_* provide auxiliary thermal and aerodynamic diagnostics, beyond the three-output set $\mathcal{Y}_s$ used in the main benchmark (App. B.3). Only the 365-day spin-up passes at all sites.

This analysis is independent of parameter discovery: its outputs are not used for residual diagnosis, candidate generation, parameter selection, candidate promotion, or memory updates (App. D.2). The purpose is solely to assess the sensitivity of simulated trajectories to spin-up duration and open-water initialization, and thereby justify the fixed choice of $\mathcal{T}_{\mathrm{spin}}$ used elsewhere in the paper.

Experimental setup. We consider the eight sites with nonzero open-water fractions and combine four spin-up durations (30, 90, 180, and 365 days) with three initial open-water storage conditions (low, default, and high), giving

$$8 \times 4 \times 3 = 96 \tag{13}$$

simulations.

Let W_0 denote the configured initial open-water storage and $W_{\max}$ its registered state limit. We define

$$W_0^{\mathrm{low}} = \max\left(0.01W_{\max},\, W_0 - 0.1W_{\max}\right), \tag{14}$$

$$W_0^{\mathrm{default}} = W_0, \tag{15}$$

$$W_0^{\mathrm{high}} = \min\left(0.99W_{\max},\, W_0 + 0.1W_{\max}\right). \tag{16}$$

All configurations are evaluated over the same post-spin-up comparison period. The 365-day simulation initialized from the configured default storage serves as the reference. Because all spin-up periods terminate at the same comparison boundary, changing the spin-up duration also changes the integration start date and seasonal history. The experiment therefore measures practical sensitivity to the combined spin-up history rather than a monotonic convergence timescale.

Output-level sensitivity. For each diagnostic output

$$y \in \{Q_H,\, Q_E,\, LW_{\uparrow},\, T_{\mathrm{surf}},\, u_*\},$$

let $y_t^{[c]}$ denote the simulated trajectory of output y at time t under spin-up configuration $c \in \{1, \ldots, 96\}$, and let $y_t^{[0]}$ denote the 365-day, default-storage reference trajectory. We define the normalized trajectory difference

$$S_y^{[c]} = \frac{\mathrm{RMSE}\left(y^{[c]}, y^{[0]}\right)}{\max\left[\mathrm{RMS}\left(y^{[0]}\right), \epsilon\right]}, \tag{17}$$

Table 3: Spin-up robustness across the eight sites with nonzero open-water fractions. A site passes only if all three open-water initializations satisfy every applicable trajectory and diurnal-cycle criterion (Eq. 17 and the seasonal and diurnal thresholds above).

Spin-up duration	30 days	90 days	180 days	365 days
Sites passing	2/8	1/8	2/8	**8/8**

where $\mathrm{RMSE}(\cdot,\cdot)$ is the same root-mean-square operator used in App. B.5, here applied between two simulated trajectories rather than between simulation and observation; $\mathrm{RMS}(y^{[0]}) = \sqrt{\mathrm{mean}_t(y_t^{[0]})^2}$ is the root-mean-square magnitude of the reference trajectory itself (not an error), used only to normalize $S_y^{[c]}$ onto a relative scale; and ϵ avoids instability when the reference magnitude is close to zero. We write $[c]$ rather than a bare superscript (s) to avoid confusion with the site index s used throughout the main text - every quantity in this appendix is computed at a single, fixed site from the eight-site open-water subset.

This spin-up sensitivity check additionally includes T_{surf} and u_* as diagnostic outputs; the main three-output discovery benchmark $\mathcal{Y}_s = \{Q_H, Q_E, LW_\uparrow\}$ used elsewhere in the paper (App. B.3) is unaffected by this broader check. We require

$$S_y^{[c]} \le 0.01 \tag{18}$$

over the full comparison period and

$$S_{y,k}^{[c]} \le 0.02, \qquad k \in \{\mathrm{DJF}, \mathrm{MAM}, \mathrm{JJA}, \mathrm{SON}\}, \tag{19}$$

for each season, where $S_{y,k}^{[c]}$ is $S_y^{[c]}$ restricted to the timestamps of season k.

We also compare mean diurnal cycles. The maximum absolute differences must not exceed

$$\begin{aligned} 2\,\mathrm{W\,m^{-2}} &\quad \text{for } Q_H \text{ and } Q_E, \\ 1\,\mathrm{W\,m^{-2}} &\quad \text{for } LW_\uparrow, \\ 0.2\,^\circ\mathrm{C} &\quad \text{for } T_{\mathrm{surf}}. \end{aligned} \tag{20}$$

No separate diurnal-cycle threshold is imposed on u_*. A site passes for a given spin-up duration only if all three open-water initializations satisfy every applicable full-period, seasonal, and diurnal-cycle criterion.

The preregistered protocol also targeted downstream decision stability, but the present experiment compares output trajectories only and does not rerun the full discovery pipeline (Algorithm 1) for every spin-up configuration; decision stability is therefore excluded from the pass rates reported here.

Sensitivity across spin-up durations. Tab. 3 shows that only one or two sites satisfy all robustness criteria after 30 180 days of spin-up, with the largest discrepancies in Q_E and Q_H and comparatively stable $LW_\uparrow$ and T_{surf}. In contrast, the 365-day protocol passes at all eight sites under all three open-water initializations, so all

$$8 \times 3 = 24$$

annual-spin-up configurations satisfy the predefined output-level criteria.

For a fixed spin-up duration, the three open-water initializations produce numerically identical diagnostic outputs in the archived simulations, so the failures at shorter durations cannot be attributed to the tested open-water perturbations; they are instead consistent with differences in the broader thermal, soil-moisture, vegetation, and seasonal forcing history. Because spin-up duration and starting season vary jointly, the nonmonotonic pass rates for 30, 90, and 180 days should not be interpreted as a convergence curve.

Implications and scope. Among the tested configurations, only the full-year spin-up eliminates the material output differences seen at shorter durations, and it is also insensitive to the tested open-water initial conditions. We therefore fix $\mathcal{T}_{\mathrm{spin}}$ to 365 days throughout the main experiments (Sec. 2.1, Eq. 6).

D Additional Method Details

This appendix specifies the executable contracts underlying the method in Sections 2.2 2.5. It describes the current 20-site, three-output protocol with $\mathcal{Y} = \{Q_H, Q_E, LW_\uparrow\}$. Historical experiments with a different site set, output set, or candidate archive are not used to define the method below.

D.1 Parameter Ontology and Executable Adapters

Typed records. The ontology separates measured evidence, searchable parameters, and dynamic model states by construction. An observation-constrained Layer-I record is

$$r_i = (v_i, \mu_i, \mathrm{src}_i, \sigma_i, a_i, \tau_i), \tag{21}$$

where v_i is the reported value, μ_i its unit, src_i its provenance, σ_i its uncertainty, a_i its spatial or surface-class applicability, and τ_i its temporal support. Examples include surface fractions, building and tree height, and remotely sensed LAI. Layer-I records may be quality controlled and translated into a SUEWS configuration, but they are not free optimization variables.

A searchable Layer-II parameter is represented as

$$p_j = (\mathrm{id}_j, \mu_j, k_j, \tau_j, [\ell_j, u_j], p_j^0, A_j, e_j), \tag{22}$$

where id_j is a canonical physical identifier, k_j is its surface-class or equation support, τ_j specifies whether the parameter is static or seasonally supported, $[\ell_j, u_j]$ is the admissible physical interval, p_j^0 is the site baseline, A_j is the executable adapter to the native SUEWS field, and e_j records the prior and data provenance. We use μ_j for units and $[\ell_j, u_j]$ for bounds throughout. Layer-III soil moisture, surface water, phenology, and thermal stores are advanced by SUEWS; the agent cannot write them as static parameter interventions.

Six semantic-firewall checks. For a proposed intervention $\Delta\boldsymbol{\theta}$, the firewall evaluates the following checks before a SUEWS call is permitted:

1. **Identity:** every parameter has exactly one registered canonical ID; aliases and unknown native field names are rejected.
2. **Unit:** the proposed quantity is convertible to the registered unit μ_j without an implicit scale or offset.
3. **Support:** the surface class, spatial support, and surface index agree with k_j; a building parameter cannot be silently applied to all seven surface classes.
4. **Temporal role:** a static Layer-II parameter cannot overwrite a Layer-III state, and the intervention respects τ_j.
5. **Range and sparsity:** the absolute value lies in $[\ell_j, u_j]$, the proposed trust region is a subset of that interval, and at most $K = 2$ parameters are active in the main experiment.
6. **Executability:** A_j supports both forward and audit mapping, preserves the intended equation-level action, and compiles the complete multi-parameter intervention atomically. Partial application is forbidden.

The compiler stores absolute physical values, the originating canonical IDs, the native SUEWS fields, and the inverse audit mapping. Consequently, a frozen result does not depend on an ambiguous multiplier or on an LLM-specific name. Data preparation, baseline simulation, diagnosis, retrieval, compilation, SUEWS execution, judging, and freezing form an incremental DAG. A cached node is reused only when its input hashes, temporal registry, ontology version, code version, simulator settings, and parameter values all match. Sealed-evaluation artifacts occupy a separate read-only branch and cannot be dependencies of adaptation nodes.

D.2 Memory Retrieval and Scoring

Memory record. Each adaptation-period simulation produces the record introduced in Eq. 4, expanded here as

$$m_i = (\mathbf{c}_i, H_i, \Delta\boldsymbol{\theta}_i, \Delta\mathbf{E}_i, \mathbf{g}_i, o_i, u_i, a_i, f_i). \tag{23}$$

Here, $\mathbf{c}_i$ contains the site, climate regime, morphology, wetness and stability descriptors available for that experiment; H_i is the mechanism-hypothesis instance under test (Eq. 3) that specifies the limiting process, parameter IDs, directions, trust regions, expected process responses, and operational falsification conditions; $\Delta\boldsymbol{\theta}_i$ contains canonical absolute and baseline-relative changes; and

$$\Delta\mathbf{E}_i = (\Delta E_{Q_H}, \Delta E_{Q_E}, \Delta E_{LW_\uparrow}) \tag{24}$$

is retained in full. The gate vector $\mathbf{g}_i$ records semantic, solver, state, aggregate, process, and temporal-block outcomes. The status o_i records `SUPPORTED`, `UNSUPPORTED`, `CONTRADICTED`, `CONTEXT_LIMITED`, or `INCONCLUSIVE` under Sec. 2.4, with its supporting or limiting evidence retained in the record. It neither replaces the gate vector nor certifies a unique cause. For an archived trial whose pre-experiment expectations are unavailable, the status is inconclusive: expectations must not be reconstructed from a favorable result. This explicit schema does not assert that the reported archive has been re-annotated or supply new status-frequency results. The remaining fields store uncertainty u_i, the applicability domain a_i, and the explicit failure cause f_i, including the degraded output and boundary status. The existing success/failure retrieval groups continue to refer to intervention and gate outcomes, rather than being redefined by the hypothesis-status annotation. Context is attached to every record.

Eligible records and retrieval. The initial generation is a cold start: $\mathcal{M}_0 = \varnothing$. Memory is then constructed only from completed adaptation experiments. For a query $q = (s, c, \mathcal{I}, \mathcal{R})$ containing target site s, climate class c, allowed canonical IDs $\mathcal{I}$, and currently risky outputs $\mathcal{R}$, an eligible record must (i) come from SUEWS, (ii) come from an adaptation period, and (iii) involve at least one ID in $\mathcal{I}$. Evaluation-period records are never eligible.

The implemented benchmark score is

$$S(m_i, q) = 4\,\mathbb{I}[s_i = s] + 2\,\mathbb{I}[c_i = c] + |\mathcal{F}_i \cap \mathcal{R}|\,, \tag{25}$$

where $\mathcal{F}_i$ is the set of outputs that failed in record i. The first term prioritizes evidence accumulated at the current site during earlier rounds, the second supports climate-conditioned cross-site reuse, and the last retrieves failures relevant to the current process risk. Morphology, wetness, stability, and residual signatures are retained in $\mathbf{c}_i$ and shown to the reasoner as applicability evidence; missing descriptors never receive an imputed similarity score. Ties are resolved deterministically by memory ID. The prompt receives at most five successful and five failed records, selected separately, rather than the entire archive.

Effect-aware reranking. Memory also supplies a soft ordering prior over deterministic candidates. For a candidate $\boldsymbol{\theta}$ and a record with the same canonical parameter set, let

$$d_i(\boldsymbol{\theta}) = \sum_{j \in \mathcal{I}} |\delta_j(\boldsymbol{\theta}) - \delta_{i,j}|\,, \qquad w_i(\boldsymbol{\theta}) = \exp[-d_i(\boldsymbol{\theta})/0.03] \begin{cases} 1, & c_i = c, \\ 0.5, & c_i \neq c, \end{cases} \tag{26}$$

where $\delta_j(\boldsymbol{\theta})$ is the relative change candidate $\boldsymbol{\theta}$ proposes for parameter j and $\delta_{i,j}$ the corresponding change stored in record i, so d_i measures parameter-space distance and w_i downweights both distant and cross-climate records. The weighted memory prediction is $\widetilde{\Delta E}_y(\boldsymbol{\theta}) = \sum_i w_i \Delta E_{i,y} / \sum_i w_i$, and its risk score is

$$\rho(\boldsymbol{\theta}) = \mathrm{mean}_y\, \widetilde{\Delta E}_y(\boldsymbol{\theta}) + 2 \max\left(0, \max_y \widetilde{\Delta E}_y(\boldsymbol{\theta})\right), \tag{27}$$

where the second term penalizes an apparently good mean response that sacrifices one process. Memory rank is blended with the deterministic proposer rank using

$$\alpha(n_{\mathrm{local}}) = \max(0.10, 0.50 \exp[-n_{\mathrm{local}}/4])\,, \tag{28}$$

where n_{local} is the number of local (same-site) trials completed so far and α is the weight assigned to the cross-site memory prior. Thus memory has its largest effect before local evidence is available and cannot override accumulating target-site simulations. Retrieval changes which experiment is attempted first; it neither supplies an accepted numerical value nor changes the Judge.

D.3 Acceptance Criteria: Full Specification

Let $E_{s,y}$ and J_s be defined by Eq. 5. All relative changes compare a candidate $\boldsymbol{\theta}$ with the fixed site baseline $\boldsymbol{\theta}_s^0$:

$$\Delta E_{s,y}(\boldsymbol{\theta}) = \frac{E_{s,y}(\boldsymbol{\theta}) - E_{s,y}(\boldsymbol{\theta}_s^0)}{E_{s,y}(\boldsymbol{\theta}_s^0)}, \qquad \Delta J_s(\boldsymbol{\theta}) = \frac{J_s(\boldsymbol{\theta}) - J_s(\boldsymbol{\theta}_s^0)}{J_s(\boldsymbol{\theta}_s^0)}. \tag{29}$$

The adaptation Judge is a conjunction of three layers.

Layer 1: semantic admissibility. $G_s^{\mathrm{sem}}(\boldsymbol{\theta}) = 1$ only if all six checks in App. D.1 pass. Otherwise the candidate is rejected before simulation.

Layer 2: executable-model validity. $G_s^{\mathrm{run}}(\boldsymbol{\theta}) = 1$ requires successful solver completion, finite target outputs, a valid typed checkpoint, continuous state propagation from spin-up through adaptation, and an unchanged forcing, surface-fact, and scoring mask relative to the baseline. This is an execution-validity gate; it is not presented as an independent proof of energy or water conservation.

Layer 3: multi-process adaptation evidence. With $\epsilon_J = 5 \times 10^{-4}$ and $\epsilon_{\mathrm{blk}} = 10^{-3}$, define

$$G_s^{\mathrm{agg}}(\boldsymbol{\theta}) = \mathbb{I}[\Delta J_s(\boldsymbol{\theta}) \le -\epsilon_J], \tag{30}$$

$$G_s^{\mathrm{proc}}(\boldsymbol{\theta}) = \mathbb{I}[\max_{y \in \mathcal{Y}_s} \Delta E_{s,y}(\boldsymbol{\theta}) \le 0], \tag{31}$$

$$G_s^{\mathrm{blk}}(\boldsymbol{\theta}) = \mathbb{I}\left[\max_{b \in \{1,\dots,4\}} \max_{y \in \mathcal{Y}_s} \Delta E_{s,y}^{(b)}(\boldsymbol{\theta}) \le \epsilon_{\mathrm{blk}}\right], \tag{32}$$

where b indexes the four chronological adaptation blocks and $\Delta E_{s,y}^{(b)}$ is the relative per-output error change computed within block b alone. The strict adaptation decision is therefore

$$G_s^{\mathrm{adapt}}(\boldsymbol{\theta}) = G_s^{\mathrm{sem}} G_s^{\mathrm{run}} G_s^{\mathrm{agg}} G_s^{\mathrm{proc}} G_s^{\mathrm{blk}}. \tag{33}$$

No weighted average can compensate for a failed factor. Among strict candidates, selection is lexicographic: minimize the worst process effect, then ΔJ_s, and finally the number of modified parameters.

Near-Pareto uncertainty qualification. A separate, explicitly labeled fallback may be enabled when a candidate passes semantic, run, and aggregate gates but a very small process effect is close to zero. Baseline and candidate time series are paired by timestamp. For each y, moving blocks of length $L \in \{24, 168\}$ hours are resampled with replacement, and $\Delta E_{s,y}$ is recomputed for $10{,}000$ replicates. Let $U_{s,y}^{(L)}$ denote the resulting 95% upper confidence bound. The candidate is uncertainty-qualified only if

$$G_s^{\mathrm{NP}}(\boldsymbol{\theta}) = \mathbb{I}\left[\max_{L \in \{24,168\}} \max_{y \in \mathcal{Y}_s} U_{s,y}^{(L)} \le 10^{-3}\right] = 1. \tag{34}$$

Strict candidates always have priority over uncertainty-qualified candidates, and the two routes are reported separately. The bootstrap does not inspect or guarantee future-period performance. Sealed-evaluation grading and the resulting deployment metrics are reported in App. G.6 (Sec. 3).

D.4 Iterative Refinement Procedure $\mathcal{R}$

Let h_r be the structured hypothesis at round r, π_r the deterministic candidate policy, $\mathcal{M}_r$ the adaptation-safe memory, and g_r the Judge evidence produced by tested interventions $\Delta\boldsymbol{\theta}_r$. The refinement operator is

$$(h_{r+1}, \pi_{r+1}, \mathcal{M}_{r+1}) = \mathcal{R}(h_r, \pi_r, \mathcal{M}_r, \Delta\boldsymbol{\theta}_r, g_r). \tag{35}$$

The hypothesis outcome o_r summarizes the recorded expectations, effect vector, and Judge results; it does not change the following existing deterministic update rules before another LLM query:

1. A semantic failure repairs or removes the invalid canonical ID; it cannot enlarge the registry or consume a SUEWS call.

2. A solver or checkpoint failure blacklists the exact compiled intervention and contracts the affected trust region. It is not treated as evidence for or against the proposed physical mechanism.
3. If $G^{\text{agg}} = 0$, the next query must reconsider the limiting process or direction rather than refine the same ineffective value.
4. If aggregate error improves but output y degrades, y is added to the query's risk set. Retrieval prioritizes failures involving y, and the next proposal must include a falsifiable explanation for avoiding that trade-off.
5. A temporal-block failure adds the responsible season or regime to the applicability boundary and either contracts the trust region or abstains.
6. A strict pass terminates discovery. Repeated boundary optima are stored as ambiguity between a narrow prior and structural compensation; they never authorize automatic bound expansion.

For a valid LLM plan, the numerical search remains deterministic. In the first round, each active parameter is centered within the proposed trust region. In round r, the interval width is contracted by 2^{r-1} around the best previous candidate, ranked by worst process effect, aggregate effect, and sparsity. A one-parameter plan evaluates up to five evenly spaced values. A two-parameter plan evaluates four axial counterfactuals and one center point, thereby testing each direction without incurring a full Cartesian grid. All absolute values are clipped to registered physical bounds, duplicates are removed, and the global simulator budget is enforced.

Every completed adaptation simulation is represented in $\mathcal{M}_{r+1}$ by its hypothesis (including expected responses and falsification conditions), intervention, full effect vector, gate outcome, and contextual hypothesis status before the next retrieval. Both accepted and failed records are retained. The loop stops after a strict discovery, five diagnosis rounds, budget exhaustion, or an explicit ABSTAIN. Neither the frozen evaluation result nor any evaluation-derived grade is written back to memory.

E PROMPT AND LLM CONTRACT

The Mechanism Reasoner is invoked only during adaptation to formulate process hypotheses and propose physically motivated parameter adjustments. The following listings document the fixed system prompt, the state-conditioned user-message template, and the response schema. Numerical candidate generation, SUEWS execution, memory retrieval, and intervention acceptance are handled externally.

E.1 VERBATIM SYSTEM MESSAGE

Listing 1 reproduces the fixed `SEARCH_PLAN_PROMPT` used in the reported process-adaptive experiments.

```
You are a scientific diagnosis agent, not a numerical optimiser.
Using adaptation-only residual metrics and the small retrieved memory,
    output only: the
most likely limiting process and process family, one or two canonical
    candidate parameters,
the physically justified direction of change, a conservative baseline-
    relative trust region,
the expected sign of the effect on QH, QE, and LWup, and failure patterns
     that must not be
repeated. Select parameters outside the sensitivity shortlist when the
    residual mechanism
supports doing so. Mark the process unrepairable when forcing, footprint,
     or structural error
is more plausible than any allowed surface parameter. Do not propose
    exact parameter values
or accept any intervention. A deterministic local search will choose
    numerical values, SUEWS
```

```
will produce evidence, and a separately specified Physics Judge will
    decide acceptance.
```

Listing 1: Verbatim system message used by the Mechanism Reasoner.

E.2 SCHEMATIC USER-MESSAGE STRUCTURE

The user message is serialized as a JSON object containing `task_input` and `required_output_schema`. The task input provides site context, adaptation residuals, retrieved experience, completed trial outcomes, and the allowed parameter registry. Listing 2 preserves the implemented field names and nesting; angle-bracketed entries represent runtime values.

```
{
  "task_input": {
    "site_id": "<SITE_ID>",
    "climate_regime": "<CLIMATE_REGIME>",

    "baseline_metrics": {
      "QH": "<ADAPTATION_METRICS>",
      "QE": "<ADAPTATION_METRICS>",
      "LWup": "<ADAPTATION_METRICS>"
    },

    "site_context": {
      "residual_signature": {
        "QH": {
          "nrmse": "<VALUE>",
          "bias": "<VALUE>"
        },
        "QE": {
          "nrmse": "<VALUE>",
          "bias": "<VALUE>"
        },
        "LWup": {
          "nrmse": "<VALUE>",
          "bias": "<VALUE>"
        }
      }
    },

    "retrieved_memory": {
      "mechanism_summary": "<RETRIEVED_SUMMARY>",
      "successful_episodes": [
        "<RETRIEVED_SUCCESS_RECORD>"
      ],
      "failed_episodes": [
        "<RETRIEVED_FAILURE_RECORD>"
      ]
    },

    "previous_effect_vectors": [
      {
        "call": "<CALL_INDEX>",
        "relative_changes": {
          "<CANONICAL_PARAMETER_ID>": "<RELATIVE_CHANGE>"
        },
        "effect_vector": {
          "QH": "<RELATIVE_NRMSE_CHANGE>",
          "QE": "<RELATIVE_NRMSE_CHANGE>",
          "LWup": "<RELATIVE_NRMSE_CHANGE>"
        },
        "aggregate_change": "<OBJECTIVE_CHANGE>",
        "physics_judge_pass": "<BOOLEAN>",
        "mechanism_consistent": "<BOOLEAN>"
      }
    ],

    "forced_process_family":
      "<NULL_OR_DETERMINISTICALLY_ROUTED_FAMILY>",

    "instruction":
      "Select the limiting process and expand beyond sensitivity "
      "Top-3 when justified. If forced_process_family is set, use "
      "that family because the previous family repeatedly degraded "
      "a required output.",
```

```
    "allowed_parameters": {
      "<CANONICAL_PARAMETER_ID>": {
        "baseline": "<VALUE>",
        "maximum_relative_change": "<VALUE>",
        "process_family": "<PROCESS_FAMILY>",
        "sensitivity_per_fraction": "<VALUE_OR_NULL>",
        "preferred_direction_from_oat": "<VALUE_OR_NULL>",
        "adapter": "<EXECUTABLE_ADAPTER>"
      }
    }
  },

  "required_output_schema":
    "<SCIENTIFIC_SEARCH_PLAN_JSON_SCHEMA>"
}
```

Listing 2: Schematic user-message structure. Angle-bracketed entries are runtime values; this listing is not a verbatim per-call transcript.

The same message structure is used for both initial diagnosis and subsequent refinement. In the first round, `previous_effect_vectors` is empty and `forced_process_family` is null. Later rounds incorporate only completed adaptation trials. If material candidates repeatedly degrade a required output, an external deterministic routing rule may restrict the next round to a compatible process family. This restriction does not alter the acceptance criteria.

E.3 Machine-Validated Response Contract

Listing 3 presents the JSON Schema exported by `ScientificSearchPlan.model_json_schema()` and supplied to the structured-output caller. The corresponding Pydantic model is used to parse each response.

```
{
  "$defs": {
    "PlannedParameter": {
      "properties": {
        "parameter_id": {
          "title": "Parameter Id",
          "type": "string"
        },
        "direction": {
          "enum": [
            "increase",
            "decrease"
          ],
          "title": "Direction",
          "type": "string"
        },
        "trust_region_relative": {
          "maxItems": 2,
          "minItems": 2,
          "prefixItems": [
            {
              "type": "number"
            },
            {
              "type": "number"
            }
          ],
          "title": "Trust Region Relative",
          "type": "array"
        },
        "mechanism": {
          "title": "Mechanism",
          "type": "string"
        },
        "risk_outputs": {
          "items": {
            "type": "string"
          },
          "title": "Risk Outputs",
          "type": "array"
        },
        "expected_effects": {
          "additionalProperties": {
            "enum": [
              "improve",
              "neutral",
              "degrade",
              "uncertain"
            ],
            "type": "string"
```

```
          },
          "title": "Expected Effects",
          "type": "object"
        }
      },
      "required": [
        "parameter_id",
        "direction",
        "trust_region_relative",
        "mechanism",
        "risk_outputs"
      ],
      "title": "PlannedParameter",
      "type": "object"
    }
  },

  "properties": {
    "limiting_process": {
      "title": "Limiting Process",
      "type": "string"
    },
    "process_family": {
      "default": "radiative",
      "enum": [
        "radiative",
        "aerodynamic",
        "vegetation_hydrology",
        "heat_storage",
        "unrepairable"
      ],
      "title": "Process Family",
      "type": "string"
    },
    "evidence": {
      "items": {
        "type": "string"
      },
      "minItems": 1,
      "title": "Evidence",
      "type": "array"
    },
    "parameters": {
      "items": {
        "$ref": "#/$defs/PlannedParameter"
      },
      "maxItems": 2,
      "title": "Parameters",
      "type": "array"
    },
    "prohibited_failure_patterns": {
      "items": {
        "type": "string"
      },
      "title": "Prohibited Failure Patterns",
      "type": "array"
    },
    "uncertainty": {
      "maximum": 1.0,
      "minimum": 0.0,
      "title": "Uncertainty",
      "type": "number"
    },
    "abstain_if_unrepairable": {
      "default": true,
      "title": "Abstain If Unrepairable",
      "type": "boolean"
    }
  },

  "required": [
    "limiting_process",
    "evidence",
    "prohibited_failure_patterns",
    "uncertainty"
  ],

  "title": "ScientificSearchPlan",
  "type": "object"
}
```


Listing 3: Machine-generated JSON Schema for the LLM response.

Additional semantic validation. After parsing, the caller checks that each `parameter_id` belongs to the supplied allow-list, that no parameter is repeated, and that no more than two active parameters are returned. A repairable process must include at least one parameter. Adjustment directions must agree with the signs of the corresponding trust regions. Each trust region is intersected

Algorithm 1 Physics-Gated Surface Parameter Discovery

Require: Forcing $\mathbf{x}$, surface facts $\mathbf{c}$, baseline $\boldsymbol{\theta}^0$, ontology Ω, adaptation observations D^a, sealed observations D^e, memory $\mathcal{M}$, budget B, max rounds $R \leq 5$
Ensure: Frozen decision d^* and adaptation-safe memory $\mathcal{M}$

$z^{\text{spin}} \leftarrow \textsc{SpinUp}(\mathbf{x}, \mathbf{c}, \boldsymbol{\theta}^0)$
$(\widehat{\mathbf{y}}^0, z^{0,a}) \leftarrow \textsc{RunAdapt}(\boldsymbol{\theta}^0, z^{\text{spin}})$
$D_0 \leftarrow \textsc{ResidualDiagnostics}(\widehat{\mathbf{y}}^0, D^a)$
$\mathcal{A} \leftarrow \varnothing, \quad b \leftarrow 0$
for $r = 1, \ldots, R$ **do**
  $(M^+, M^-) \leftarrow \textsc{Retrieve}(\mathcal{M}, D_{r-1}, \mathbf{c}, \Omega)$
  $h_r \leftarrow \textsc{MechanismReasoner}(D_{r-1}, \mathbf{c}, M^+, M^-, \Omega)$
  if $h_r = \textsc{Abstain}$ **then**
    break
  end if
  $C_r \leftarrow \textsc{DesignAndFilter}(h_r, \Omega, \mathcal{A}, B - b)$
  $C_r \leftarrow \textsc{RankCandidates}(C_r, \mathcal{M})$
  for all $\boldsymbol{\theta} \in C_r$ **while** $b < B$ **do**
    $(\widehat{\mathbf{y}}_{\boldsymbol{\theta}}, z^{\theta,a}) \leftarrow \textsc{RunAdapt}(\boldsymbol{\theta}, z^{\text{spin}})$
    $b \leftarrow b + 1$
    $g_{\boldsymbol{\theta}} \leftarrow \textsc{JudgeAdaptation}(\boldsymbol{\theta}, \widehat{\mathbf{y}}_{\boldsymbol{\theta}}, D^a, \Omega)$
    $o_{\boldsymbol{\theta}} \leftarrow \textsc{AnnotateOutcome}(h_r, \Delta\mathbf{E}_{\boldsymbol{\theta}}, g_{\boldsymbol{\theta}})$
    $\mathcal{A} \leftarrow \mathcal{A} \cup \{(h_r, \boldsymbol{\theta}, z^{\theta,a}, \Delta\mathbf{E}_{\boldsymbol{\theta}}, g_{\boldsymbol{\theta}}, o_{\boldsymbol{\theta}})\}$
    $\textsc{WriteMemory}(\mathcal{M}, \mathbf{c}, h_r, \boldsymbol{\theta}, \Delta\mathbf{E}_{\boldsymbol{\theta}}, g_{\boldsymbol{\theta}}, o_{\boldsymbol{\theta}})$
  end for
  if $\textsc{StrictCandidateExists}(\mathcal{A})$ **then**
    break
  end if
  $D_r \leftarrow \textsc{Refine}(h_r, D_{r-1}, \mathcal{A}, \mathcal{M})$
end for
$(d^*, z^{*,a}) \leftarrow \textsc{SelectOrAbstain}(\mathcal{A}, z^{0,a})$
$\textsc{Freeze}(d^*, z^{*,a}, \Omega, \text{periods, thresholds, hashes})$
$\textsc{EvaluateOnce}(d^*, z^{*,a}, D^e)$ ▷ No tuning, retrieval, or memory write
return $(d^*, \mathcal{M})$

with the registered relative and physical bounds; an infeasible intersection is rejected before SUEWS execution.

If validation fails, the caller appends the validation error and the following repair instruction:

```
Return a corrected plan. Copy parameter_id exactly from
    allowed_parameters;
do not invent aliases. Direction and trust-region signs must agree.
```

Listing 4: Verbatim repair instruction appended after response validation fails.

At most three schema or semantic repair attempts are permitted. The validated response specifies a process hypothesis, canonical parameter identifiers, adjustment directions, expected output-effect signs, and bounded relative search regions. The `expected_effects` field encodes the hypothesized response signs ρ_y. The numerical counterfactual set $\mathcal{F}$ is generated deterministically from the validated directions, trust regions, and parameter registry. SUEWS provides the process-effect evidence, and the independently specified Physics Judge determines acceptance.

F Algorithm

Algorithm 1 instantiates the temporal partition in Eq. 6. Spin-up advances SUEWS state, adaptation is the only searchable interval, and sealed evaluation is opened exactly once after the decision has been frozen. AnnotateOutcome denotes the evidence interpretation in Sec. 2.4, not a new statistical test or acceptance step; the stopping, selection, and numerical experiment rules remain unchanged.

Computational cost. Let $C_{\mathcal{M}}$ denote one SUEWS execution. Because the executable budget is explicit, simulator cost is bounded by $\mathcal{O}(BC_{\mathcal{M}})$ per site, irrespective of the number of LLM hypotheses. Ontology checks are linear in the number of proposed parameters, and memory retrieval is linear in the eligible archive before sorting. Deterministic caching removes repeated parameter–forcing–configuration tuples. Any surrogate or memory ranker only changes the evaluation order; scientific evidence and the dominant computational cost remain the SUEWS executions actually performed.

G Detailed Experimental Protocol

This appendix expands Sec. 2.6. It defines the data, simulator baseline, online comparison, ablations, and metrics used by the current 20-site experiment. Historical 21-site experiments, archive replay, and the separate u_* aerodynamic extension are not included in the main benchmark denominators.

G.1 Sites, Forcing, and Observations

Site cohort. The main cohort contains the 20 Urban-PLUMBER sites for which Q_H, Q_E, and $LW_\uparrow$ can be evaluated under a common protocol, so $\mathcal{Y}_s = \{Q_H, Q_E, LW_\uparrow\}$ for every site s in this cohort. The sites span multiple climates and urban morphologies; their coordinates, climate classes, registered dates, and valid sample counts are reported in the site-level data table. Friction velocity u_* is omitted from the main task because it is unavailable at many sites and is analyzed separately.

Forcing. The forcing vector $\mathbf{F}_t$ follows the definition of Eq. 7 (App. B.2). We use the quality-controlled local forcing distributed with the benchmark and its registered gap-filling products. Flux observations are never inserted into $\mathbf{F}_t$, used to correct dynamic states, or reconstructed as residual energy terms.

Temporal aggregation and masks. Half-hourly variables are converted to hourly resolution using means for meteorological and flux quantities and sums for precipitation. Timestamps and units are converted before simulation. Evaluation targets are not gap-filled: missing or quality-rejected flux observations are masked. For every output, metrics use the intersection of finite baseline, candidate, and observation records, so paired effects cannot arise from different masks.

Surface information. Layer-I surface facts (App. D.1) include the fractions of paved, building, evergreen-tree, deciduous-tree, grass, bare-soil, and water surfaces, together with available building height, vegetation height, LAI, and site morphology, validated to sum to one and held fixed during discovery. Dynamic soil water, interception storage, phenology, and thermal states are advanced by SUEWS rather than tuned hourly (Layer III, App. D.1).

G.2 Spin-up, Adaptation, and Sealed Evaluation

Each site is executed over the three ordered periods of Eq. 6. The fixed 365-day spin-up (App. C) advances thermal and hydrological states without contributing to candidate scores. The valid observation period is then divided chronologically: the first 70% is adaptation and the final 30% is sealed evaluation, avoiding a random hourly split that would leak weather persistence, seasonality, and surface-water memory across the boundary.

All baseline and candidate adaptation simulations start from the same typed spin-up checkpoint. For a frozen intervention, the corresponding end-of-adaptation checkpoint is propagated continuously into evaluation. Before the sealed run, the system freezes the decision, absolute canonical values, native adapters, period registry, Judge thresholds, software versions, and SHA-256 hashes of all relevant inputs (App. D). The sealed runner performs no diagnosis, retrieval, surrogate fitting, parameter selection, threshold adjustment, or memory write.

G.3 Physical Baseline and Searchable Parameter Space

The baseline uses the frozen SUEWS environment, original lookup-table values, and site-specific Layer-I facts. Forcing, surface fractions, anthropogenic-heat priors, checkpoint provenance, and

Table 4: Search policies evaluated from the original site configuration. OAT denotes bilateral one-at-a-time sensitivity screening. Memory contains adaptation-period records only.

Method	Candidate-selection policy
Random + Judge	Constructs an OAT screen, retains the three most responsive parameters, and randomly samples valid one- and two-parameter interventions.
Sensitivity + Judge	Treats each OAT probe as an admissible candidate and uses any remaining budget for deterministic refinement along observed response directions.
Constrained BO + Judge	Constructs an OAT screen and sequentially models aggregate improvement and process-wise feasibility over the retained local space.
LLM, no memory + Judge	Uses site residuals, context, and the OAT screen to propose mechanisms and bounded search directions, with an empty retrieval set in every round.
Full evolving-memory system	Starts with empty memory and iteratively refines unresolved sites using context-conditioned records of completed successes, failures, and boundary cases. Registered parameter-family expansion is permitted, but the Judge remains fixed.

scoring masks are identical between the baseline and every intervention; no tower flux is used to calibrate the baseline.

All methods receive the same canonical Layer-II : seven-surface albedo and emissivity parameters, site-scale momentum roughness length and displacement height, and a vegetation-conductance scale for a diagnosed latent-heat limitation. Unknown IDs, incompatible units, invalid surface support, out-of-range values, and incomplete adapters are rejected before simulation (the six-check semantic firewall, App. D.1). A candidate changes at most two parameters; the default relative trust region is $[-0.05, 0.05]$, and repeated boundary optima are flagged for local verification rather than automatically expanding the range.

All methods share a bilateral one-at-a-time sensitivity probe over eligible parameters. Because this preprocessing is identical, its SUEWS cost is reported once and separately from the online candidate budget. Sensitivity does not restrict the complete method to a fixed top-k: the MECHANISM REASONER may activate another registered family when residual evidence and identifiability checks support it.

G.4 COMPARED SEARCH POLICIES

Tab. 4 summarizes the end-to-end search policies. All methods start from the original site configuration, use adaptation data only, and share the base parameter registry, physical bounds, SUEWS simulator, and PHYSICS JUDGE. Parameter-family selection, registered search-space expansion, candidate ordering, and stopping are method-specific. No method can access sealed observations or candidate outcomes before evaluation.

All preprocessing simulations are included in the reported cost. Random search, constrained BO, and the memory-free LLM use a complete bilateral OAT screen (16 32 calls per site; mean 24.8) followed by 25 search evaluations. Sensitivity includes its probes within a 25-call total budget. The complete system uses 1,410 search simulations across its cold-start and refinement generations; including the 496 OAT simulations gives 1,906 calls, or 95.3 per site on average. Tab. 2 therefore reports operational end-to-end outcomes rather than an equal-budget ranking, while Fig. 3(b) enables comparison at any cumulative SUEWS budget.

Randomized comparators use seeds $\{11, 23, 47, 89, 131\}$, producing 100 site seed episodes per method. The complete system is evaluated as one frozen 20-site trajectory rather than a post-hoc best-of-five result. End-to-end SUEWS-call accounting is reported in Tab. 2.

The LLM variants use the same frozen Kimi-K2.6 route and prompt contract (App. E), with at most five LLM calls per adaptation episode. The LLM proposes mechanisms, parameter IDs, directions, and trust regions but never supplies an accepted numerical value. Deterministic code constructs the counterfactuals, and all numerical evidence and acceptance decisions come from SUEWS and the PHYSICS JUDGE.

G.5 Component Ablations

We evaluate evolving memory and the Physics Judge.

Evolving memory. We compare the complete system with its cold-start, memory-free counterpart. Both use the same LLM contract, parameter ontology, candidate compiler, SUEWS simulator, and Physics Judge. The complete system additionally retrieves completed successes, failures, and boundary cases by site context. Memory may change mechanism priority and search regions, but never the acceptance criteria.

Physics Judge. The no-Judge variant retains the same proposer, candidate compiler, and SUEWS simulator, but selects interventions using aggregate adaptation improvement alone. We apply the complete adaptation-time Judge offline to these decisions to measure unsafe acceptance, and evaluate all frozen decisions on the same held-out period.

G.6 Evaluation Grades and Deployment Metrics

Let $\Delta E_{s,y}^{\mathrm{eval}}$ denote $\Delta E_{s,y}$ (Eq. equation 29) computed on the sealed evaluation period, and define the worst-case held-out degradation

$$D_{\max,s} = 100 \max\left(0,\ \max_{y \in \mathcal{Y}_s} \Delta E_{s,y}^{\mathrm{eval}}\right). \tag{36}$$

We report both exact zero-degradation ($D_{\max,s} = 0$) and a practical future-period non-inferiority criterion $\max_y \Delta E_{s,y}^{\mathrm{eval}} \leq 10^{-3}$ (0.1%). The latter accounts for negligible numerical or sampling variation in the held-out interval but does not retroactively relax the adaptation-time acceptance gate G_s^{adapt} (App. D.3, Eq. equation 33).

Sites are assigned one of six grades:

- **A+**: material aggregate improvement with exact process-wise non-degradation ($D_{\max,s} = 0$).
- **A**: material improvement with worst-process degradation at most 0.1% ($0 < D_{\max,s} \leq 0.1\%$).
- **B**: aggregate improvement with a larger process trade-off ($D_{\max,s} > 0.1\%$).
- **C**: an improvement below the 0.05% practical-success threshold.
- **D**: no aggregate improvement, or an invalid evaluation.
- **U**: abstention; the default baseline configuration is retained.

Grades A+ and A jointly constitute practical temporal success. Abstention sites (U) contribute zero gain when computing all-site deployment utility, but are excluded from denominators that evaluate the quality of tested interventions specifically (i.e., they are counted in deployment-level statistics but not in per-intervention success rates).

H Extended Experimental Results

This appendix expands Tab. 8, 2, and 1 and Fig. 3 using the same 20-site, three-output benchmark. No historical cohort or archive-replay result enters the reported denominators. Positive $-\Delta J$ denotes aggregate improvement, whereas positive ΔE_y denotes degradation of output y.

H.1 From-Zero Search Outcomes

The policy comparison starts from the original site configuration. Random search, constrained Bayesian optimization, and the memory-free LLM each construct a bilateral OAT screen before 25 additional search evaluations, giving 49.8 SUEWS calls per site on average. Sensitivity-guided search treats the probes themselves as candidates and is capped at 25 total calls. The complete system performs 1,410 search simulations after screening; including the 496 screening simulations gives 1,906 calls, or 95.3 per site.

Table 5: Typed parameter registry and admissible intervention space. Layer I contains observation-constrained facts, Layer II searchable physical parameters, and Layer III prognostic states. Physical domains define broad validity bounds, whereas search domains specify the local intervention ranges used in the experiments.

Quantity	Layer	Unit	Physical domain	Search domain	Parameter source	SUEWS mapping
Surface fraction f_s	I	1	$[0,1]$, $\sum_s f_s = 1$	Fixed	Urban-PLUMBER metadata, remote sensing, and GIS	`sfr_surf[s]`
Albedo α_s	II	1	$[0.05, 0.60]$	$\theta_s^0(1 \pm 0.05)$	Site-specific SUEWS lookup table	`alb[s]`; vegetation albedo fields updated consistently
Emissivity ϵ_s	II	1	$[0.80, 1.00]$	$\theta_s^0(1 \pm 0.05)$	Site-specific SUEWS lookup table	`emis[s]`
Roughness length z_0	II	m	$[0.05, 5.0]$	Conditional $\pm 5\%$	Site morphology and baseline configuration	`z0m_in`
Displacement height d	II	m	$[0, z_{\mathrm{meas}})$	Conditional $\pm 5\%$	Building height, morphology, and baseline configuration	`zdm_in`
Vegetation conductance scale	II	1	$[0.25, 4.0]$	Triggered $[0.95, 1.05]$	Unit multiplier on site-specific vegetation parameters	Scales `max_conductance` and `g_max`
Leaf area index (LAI)	I/III	$\mathrm{m}^2\,\mathrm{m}^{-2}$	≥ 0	Observation constrained	Site observations or remote sensing	Surface-specific LAI and phenology fields
Building / tree height	I	m	≥ 0	Fixed	Site metadata, GIS, LiDAR, or building products	Used to derive or validate `z0m_in` and `zdm_in`
Root-zone water	III	mm	Model-specific capacity	Prognostic	Spin-up checkpoint	Surface soil-water state
Surface water	III	mm	Surface-storage bounds	Prognostic	Spin-up checkpoint	Surface interception/storage state

Notes. $s \in \{$`paved`, `buildings`, `evergreen_trees`, `deciduous_trees`, `grass`, `bare_soil`, `water`$\}$. Candidate values must satisfy both physical and search domains. Aerodynamic search requires geometric-measurement height consistency. Vegetation conductance scaling is activated post latent-heat trigger. See machine-readable registry for baselines and executable mappings.

Fig. 3(b) reports discovery against cumulative from-zero cost. Unsuccessful site episodes remain in the denominator. Sensitivity is particularly effective at low budgets because each probe both measures a local effect and tests an admissible intervention. The complete system requires more computation but ultimately reaches 16/20 sites, compared with 54/100 site seed episodes for sensitivity and at most 25/100 for the other randomized policies.

Held-out results separate coverage from conditional effect size. Constrained Bayesian optimization obtains a 1.084% mean gain among its accepted interventions but freezes updates in only 20% of episodes. The complete system obtains a smaller conditional gain (0.445%) but the largest all-site gain (0.356%) by recovering transferable interventions at more sites. Thus, the main benefit is coverage rather than larger corrections at sites already solved by simpler policies.

H.2 Frozen Site-Specific Parameter Interventions

Tab. 9 reports the complete adaptation-time decision and the physical values passed to sealed evaluation. "Intervene" denotes a configuration frozen before the held-out period; it does not imply an A+/A evaluation grade. Abstention sites retain the original SUEWS configuration. Reporting abso-

Table 6: Principal experimental settings.

Setting	Value
Main sites / outputs	20 / $Q_H, Q_E, LW_\uparrow$
Temporal resolution	1 hour
Spin-up / split	365 days / chronological 70:30
Maximum active parameters	2
Default trust region	$\pm 5\%$ of site baseline
Maximum diagnosis rounds / LLM calls	5 / 5 per site
Primary / equal-budget SUEWS calls	40 / 25 per site seed
Random seeds	11, 23, 47, 89, 131
Material aggregate gain	$-\Delta J \geq 5 \times 10^{-4}$
Full-period process tolerance	0
Temporal blocks / tolerance	4 / 10^{-3}
Bootstrap blocks / replicates	24 h, 168 h / 10,000
Retrieved memory per query	at most 5 successes + 5 failures
Statistical uncertainty	95% site-cluster bootstrap

Table 7: Urban-PLUMBER sites and chronological data partitions used in the 20-site benchmark.

Site	Climate regime	Partition boundaries (YYYY-MM-DD)			
		t_0	t_1	t_2	t_3
AU-Preston	Temperate oceanic	2002-08-11	2003-08-12	2004-07-09	2004-11-28
AU-SurreyHills	Temperate oceanic	2003-02-22	2004-02-23	2004-06-06	2004-07-20
CA-Sunset	Temperate oceanic	2010-12-29	2011-12-31	2015-07-02	2016-12-31
FI-Kumpula	Boreal	2009-12-30	2011-01-01	2013-02-06	2014-01-01
FI-Torni	Boreal	2009-12-30	2011-01-01	2013-02-06	2014-01-01
FR-Capitole	Temperate	2003-02-18	2004-02-20	2004-11-08	2005-03-01
GR-HECKOR	Mediterranean	2018-06-29	2019-07-01	2020-03-13	2020-07-01
JP-Yoyogi	Humid subtropical	2015-03-31	2016-04-01	2019-01-18	2020-03-31
KR-Jungnang	Humid continental	2016-01-24	2017-01-25	2018-08-25	2019-04-29
KR-Ochang	Humid continental	2014-06-06	2015-06-08	2016-12-04	2017-07-26
NL-Amsterdam	Temperate oceanic	2017-12-30	2019-01-01	2020-04-01	2020-10-13
PL-Lipowa	Humid continental	2006-12-30	2008-01-01	2011-07-02	2013-01-01
PL-Narutowicza	Humid continental	2006-12-30	2008-01-01	2011-07-02	2013-01-01
SG-TelokKurau06	Tropical	2005-04-29	2006-05-01	2006-12-14	2007-03-22
UK-KingsCollege	Temperate oceanic	2011-04-03	2012-04-04	2013-06-24	2014-01-01
UK-Swindon	Temperate oceanic	2010-05-09	2011-05-11	2012-09-23	2013-04-25
US-Baltimore	Humid subtropical	2000-12-29	2002-01-01	2005-07-02	2006-12-31
US-Minneapolis1	Humid continental	2005-05-30	2006-06-01	2008-07-05	2009-05-29
US-Minneapolis2	Humid continental	2005-05-30	2006-06-01	2008-07-05	2009-05-29
US-WestPhoenix	Hot desert	2010-12-14	2011-12-16	2012-09-08	2012-12-31

Notes. Spin-up: t_0 t_1; adaptation: t_1 t_2; sealed evaluation: t_2 t_3. Spin-up uses atmospheric forcing only. All interventions are selected during adaptation and frozen before evaluation. Flux observations are used only for scoring, never as SUEWS forcing.

lute values, rather than only relative changes, makes every intervention reproducible and auditable against its site configuration.

The frozen values are heterogeneous in both parameter family and direction. Building emissivity decreases at FI-Kumpula but increases at FR-Capitole, KR-Jungnang, and PL-Lipowa. Vegetation conductance decreases at CA-Sunset, PL-Narutowicza, and US-Minneapolis1 but increases at NL-Amsterdam and SG-TelokKurau06. The experiment therefore does not support a universal cross-city multiplier. Instead, it yields bounded, site-conditioned corrections whose applicability remains tied to local surface composition and residual structure.

Fig. 5 further shows that accepted steps span both signs and several magnitudes, including 1.25%, 1.5%, 2.5%, 3%, and 5%. Radiative and conductance corrections are bidirectional, whereas aero-

Table 8: Site-level frozen interventions and temporal evaluation on the complete 20-site benchmark, selected using only the adaptation period and frozen before evaluation. Grade definitions and the $D_{\max}$ formula are given in App. G.6; U sites retain the original SUEWS baseline.

No.	Site	Adapt. gain (%)	Eval. gain (%)	ΔE_{Q_H} (%)	ΔE_{Q_E} (%)	$\Delta E_{LW_\uparrow}$ (%)	$D_{\max}$ (%)	Worst process	Grade
1	AU-Preston	0.281	0.271	-0.674	+0.046	-0.399	0.046	Q_E	A
2	AU-SurreyHills	0.167	0.139	-0.368	+0.013	-0.105	0.013	Q_E	A
3	CA-Sunset	0.747	0.618	-0.679	-0.730	$+2 \times 10^{-5}$	2×10^{-5}	$LW_\uparrow$	A
4	FI-Kumpula	0.524	0.607	-0.198	+0.033	-3.949	0.033	Q_E	A
-	FI-Torni	-	-	-	-	-	-	-	U
5	FR-Capitole	0.622	0.643	+0.095	+0.005	-4.940	0.095	Q_H	A
6	GR-HECKOR	0.192	0.130	-0.155	-0.146	-0.001	0.000	-	A+
-	JP-Yoyogi	-	-	-	-	-	-	-	U
7	KR-Jungnang	1.305	1.518	-0.926	+0.029	-16.624	0.029	Q_E	A
8	KR-Ochang	0.619	0.619	-0.345	+0.001	-4.532	0.001	Q_E	A
9	NL-Amsterdam	0.087	0.177	-0.618	+0.094	-0.090	0.094	Q_E	A
10	PL-Lipowa	0.432	0.389	-0.120	-0.004	-1.547	0.000	-	A+
11	PL-Narutowicza	0.303	0.140	-0.213	-0.151	$+4 \times 10^{-6}$	4×10^{-6}	$LW_\uparrow$	A
12	SG-TelokKurau06	0.747	0.600	-0.687	-0.734	-0.035	0.000	-	A+
13	UK-KingsCollege	0.086	0.051	-0.163	+0.006	+0.108	0.108	$LW_\uparrow$	B
14	UK-Swindon	0.065	0.011	-0.029	+0.0003	+0.0001	0.0003	Q_E	C
-	US-Baltimore	-	-	-	-	-	-	-	U
15-a	US-Minneapolis1	0.335	1.119	-1.005	-1.675	-0.000	0.000	-	A+
15-b	US-Minneapolis2	0.101	0.085	-0.089	-0.017	-0.262	0.000	-	A+
-	US-WestPhoenix	-	-	-	-	-	-	-	U

$16/20$ frozen interventions; $4/20$ abstentions; A+/A: 14, B: 1, C: 1, D: 0, U: 4.

Table 9: Frozen site-level interventions and worst-process degradation on the 20-site benchmark. All decisions were made using adaptation data and frozen before temporal evaluation. $D_{\max}$ is the largest positive output-wise relative nRMSE change on the sealed period and is zero when no output degrades. Bldg. denotes buildings, α albedo, ϵ emissivity, z_0 momentum roughness length, and d displacement height. A dash denotes baseline retention or the absence of process degradation.

Site	Decision	Frozen intervention	$D_{\max}$ (%)	Worst process
AU-Preston	Intervene	Bldg. $\alpha = 0.1529$, Paved $\alpha = 0.1244$	0.046	Q_E
AU-SurreyHills	Intervene	Bldg. $\alpha = 0.1698$, Paved $\alpha = 0.1381$	0.013	Q_E
CA-Sunset	Intervene	Vegetation conductance scale $= 0.950$	0.00002	$LW_\uparrow$
FI-Kumpula	Intervene	Bldg. $\epsilon = 0.8827$, Paved $\epsilon = 0.9215$	0.033	Q_E
FI-Torni	Abstain	Keep baseline		
FR-Capitole	Intervene	Bldg. $\epsilon = 0.9373$, Bldg. $\alpha = 0.1344$	0.095	Q_H
GR-HECKOR	Intervene	$d = 16.520\,\mathrm{m}$, $z_0 = 0.7189\,\mathrm{m}$	0.000	
JP-Yoyogi	Abstain	Keep baseline		
KR-Jungnang	Intervene	Bldg. $\epsilon = 0.9555$, Paved $\epsilon = 0.9975$	0.029	Q_E
KR-Ochang	Intervene	Paved $\epsilon = 0.9025$, Paved $\alpha = 0.1311$	0.001	Q_E
NL-Amsterdam	Intervene	Conductance scale $= 1.050$, Water $\alpha = 0.0822$	0.094	Q_E
PL-Lipowa	Intervene	Bldg. $\epsilon = 0.9237$, Paved $\epsilon = 0.9785$	0.000	
PL-Narutowicza	Intervene	Vegetation conductance scale $= 0.950$	0.000004	$LW_\uparrow$
SG-TelokKurau06	Intervene	Conductance scale $= 1.025$, Grass $\alpha = 0.2721$	0.000	
UK-KingsCollege	Intervene	Water $\alpha = 0.0989$, Bldg. $\alpha = 0.1230$	0.108	$LW_\uparrow$
UK-Swindon	Intervene	$z_0 = 0.5063\,\mathrm{m}$, $d = 3.325\,\mathrm{m}$	0.0003	Q_E
US-Baltimore	Abstain	Keep baseline		
US-Minneapolis1	Intervene	Vegetation conductance scale $= 0.9875$	0.000	
US-Minneapolis2	Intervene	Evergreen $\alpha = 0.1386$, Evergreen $\epsilon = 0.9555$	0.000	
US-WestPhoenix	Abstain	Keep baseline		

dynamic changes occur only at the two sites with sufficient morphological support. This pattern is consistent with local refinement of heterogeneous surface representations rather than selection of a globally preferred correction.

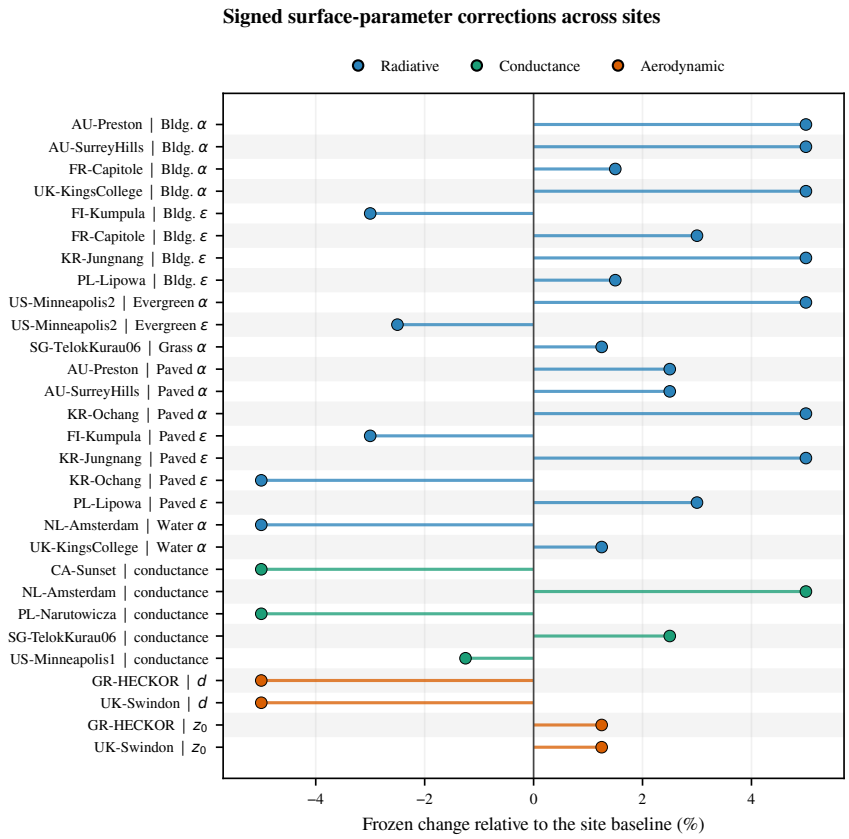

Figure 5: Signed changes in the 29 frozen parameter components. Each segment connects the site baseline to the frozen value, and colour denotes radiative, vegetation-conductance, or aerodynamic control. Opposite directions within the same parameter family demonstrate that the experiment does not support a single cross-city multiplier.

H.3 Output-Wise Temporal Transfer

Fig. 6 separates aggregate transfer from process-wise reliability. All 16 frozen interventions retain positive aggregate gain, but only five are exactly non-degrading across all three outputs. Fifteen remain within the 0.1% practical process margin, and 14 also retain at least 0.05% aggregate gain. These 14 cases receive grades A+ or A.

Most process trade-offs are small changes in Q_E. AU-Preston, FI-Kumpula, KR-Jungnang, KR-Ochang, and NL-Amsterdam each exchange a larger improvement in another output for less than 0.1% degradation in Q_E. UK-KingsCollege is the only intervention with $D_{\max} > 0.1\%$, caused by $LW_{\uparrow}$, whereas UK-Swindon remains process-safe but has only a 0.011% aggregate gain.

The output decomposition also distinguishes parameter mechanisms. Radiative updates produce the largest changes in $LW_{\uparrow}$, including improvements of 16.624% at KR-Jungnang, 4.940% at FR-Capitole, 4.532% at KR-Ochang, and 3.949% at FI-Kumpula. Conductance updates primarily affect the coupled Q_H Q_E partition, most clearly at CA-Sunset, SG-TelokKurau06, and US-Minneapolis1. Different parameter families therefore act through distinct output pathways rather than serving as interchangeable ways to reduce the same scalar objective.

Fig. 7 separates aggregate materiality from process reliability. UK-KingsCollege lies just beyond the practical margin because $LW_{\uparrow}$ degrades by 0.108%, despite a positive aggregate gain. UK-Swindon remains within the process margin but is graded C because its aggregate gain is only 0.011%, below the 0.05% materiality threshold. The practical margin is used only to describe sealed transfer and does not relax the exact adaptation-time acceptance gate.

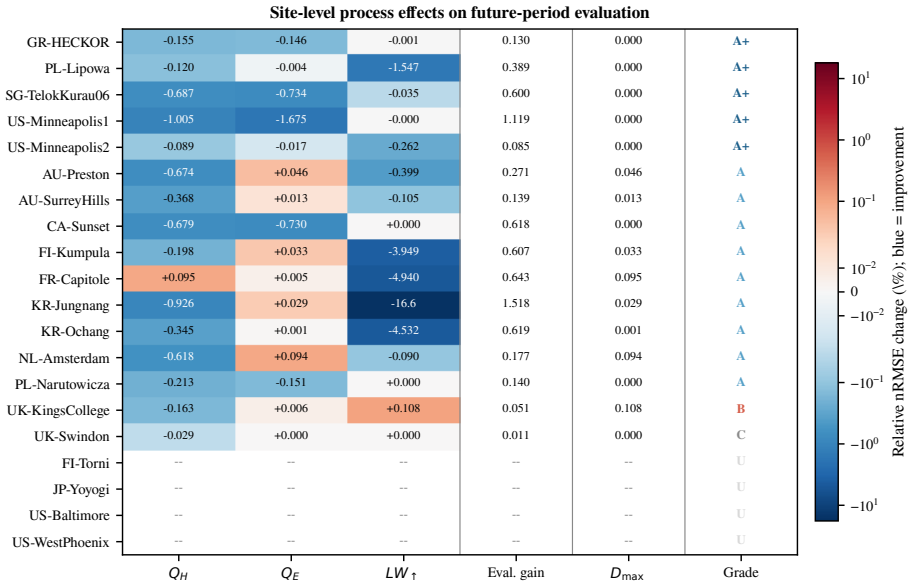

Figure 6: Output-wise effects of the frozen decisions during sealed temporal evaluation. Columns show relative nRMSE changes for Q_H, Q_E, and $LW_\uparrow$, aggregate evaluation gain, worst-process degradation, and grade. Blue denotes improvement, red denotes degradation, and dashes mark baseline retention. The symmetric-log colour scale preserves visibility of near-zero trade-offs without hiding the larger longwave gains.

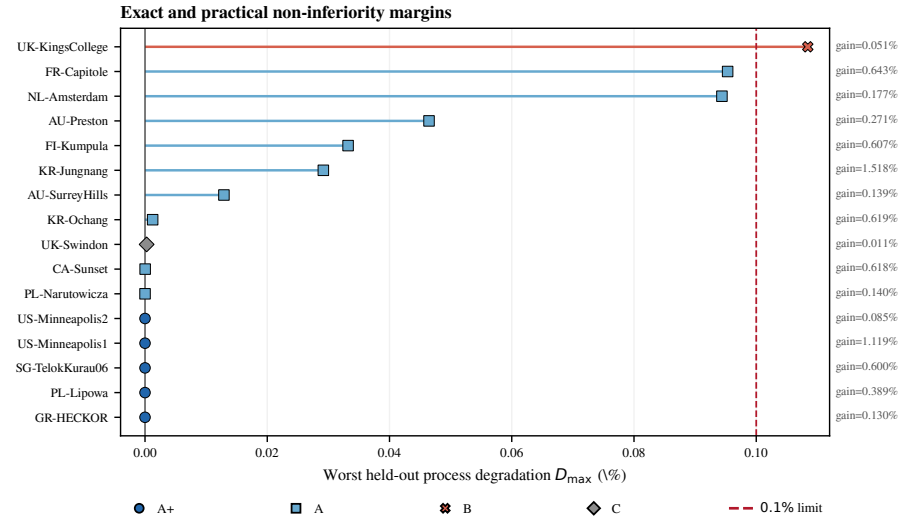

Figure 7: Worst-process degradation of each frozen intervention on the sealed period. Vertical lines mark exact non-degradation and the 0.1% practical margin; text reports aggregate evaluation gain. Marker shape denotes the A+/A/B/C grade.

H.4 Iterative Experience and Discovery Coverage

The cold-start generation evaluates 476 candidates and resolves 9/20 sites. Coverage increases to 14/20 after the first memory-informed generation (816 cumulative candidates) and to 16/20 after the second (972 candidates). It then remains unchanged as the archive grows to 1,118, 1,259, and 1,410 candidates.

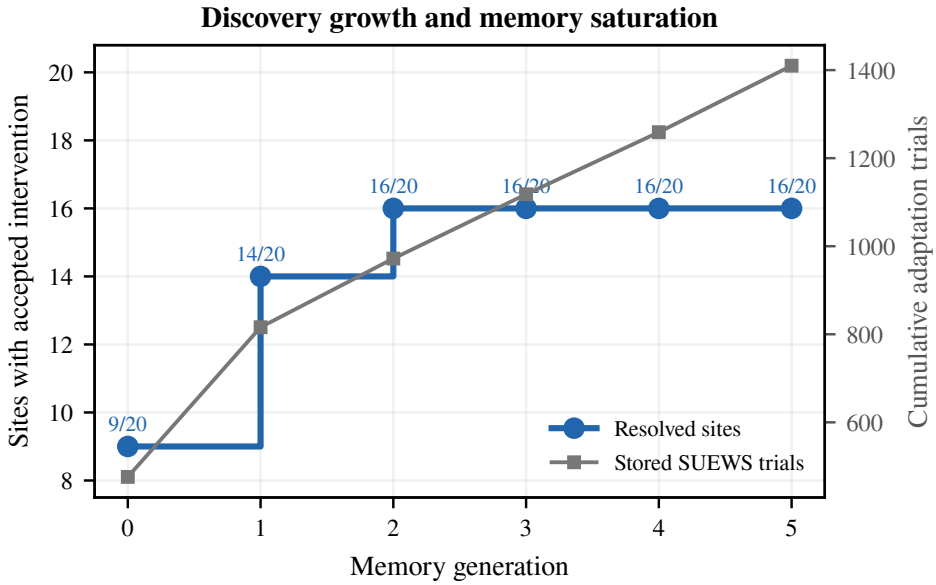

Figure 8: Discovery coverage and accumulated adaptation evidence across iterative generations. The cold-start generation resolves 9/20 sites, and coverage rises to 14/20 and 16/20 in subsequent memory-informed generations. Because cumulative simulator calls also increase, the figure describes the operational trajectory rather than a matched-budget memory effect.

This trajectory is operational evidence for the combined effect of accumulated experience and continued refinement; because later generations also receive additional simulator calls, it is not a matched-budget estimate of memory alone. Nevertheless, the saturation pattern is informative. Additional search does not force updates at FI-Torni, JP-Yoyogi, US-Baltimore, or US-WestPhoenix, indicating that abstention persists even as experimental evidence accumulates.

Fig. 2(a) illustrates the process at US-Minneapolis2. The cold-start episode evaluates 27 candidates without a strict pass. After the generation boundary, the local failure archive and context-conditioned cross-site evidence redirect the search toward the evergreen surface. The subsequent episode finds two strict candidates at calls 39 40 and freezes a 5% albedo increase with a 2.5% emissivity decrease. This case illustrates mechanism redirection rather than direct reuse of a previously accepted numerical value.

H.5 Budget-Matched Memory Ablation: Discovery-Rate Trajectories and Site-Level Heterogeneity

The comparison in Tab. 2 evaluates search policies under different *operational* costs (49.8 vs. 95.3 calls/site), because the compared methods are allowed to stop early once their own protocol terminates. To isolate the effect of cross-site memory access from this budget asymmetry, we run a dedicated matched-budget ablation comparing `LLM + Judge (no cross-site memory)` against `LLM + frozen cross-site memory + Judge`. Both variants share the same Physics Judge, the same candidate compiler, and the same *maximum online simulation budget* of 100 candidate SUEWS evaluations per site–seed episode, evaluated across all 20 sites and the same 5 seeds used elsewhere (App. G.4). This yields $20 \times 5 \times 2 = 200$ site–seed–method trajectories; all 200 complete without a solver failure or sealed-period leakage.

Protocol: frozen, leave-target-site-out cross-site memory. This ablation isolates access to cross-site experience from online memory writes. Memory here is a precomputed snapshot: the frozen generation-5 adaptation archive (1,410 records, App. H.4), with all records from the target site removed before each episode and retrieved but never updated during search. Local trial outcomes feed only within-episode iterative refinement (present in both conditions) and are never written back to

Table 10: Judge-valid discovery rate at matched cumulative candidate budgets, 20 sites × 5 seeds (100 site–seed episodes per condition). Both conditions share the same 100-call maximum budget; realized calls differ because of reasoner-driven early abstention (see protocol above). "Memory" here refers to the frozen, leave-target-site-out cross-site snapshot, not the evolving memory of the complete system.

Budget (cumulative calls)	No cross-site memory (%)	Frozen cross-site memory (%)	Gap (pp)
25	17	29	12
50	28	41	13
100	40	50	10

Table 11: Cross-tabulation of the 100 site–seed episodes by 100-call Judge-valid discovery outcome, no cross-site memory vs. frozen cross-site memory.

	Frozen memory: success	Frozen memory: no success	Total
No memory: success	25	15	40
No memory: no success	25	35	60
Total	50	50	100

the pool or shared across sites or seeds. The complete discovery pipeline reported elsewhere updates memory between adaptation generations; freezing generation 5 here removes confounds from episode ordering and online writes. Both conditions share a 100-call maximum budget per episode but differ slightly in realized calls (9,192 vs. 9,944) because the reasoner may abstain early; we therefore describe this ablation as matched in *maximum* budget, not in realized call count.

Discovery-rate trajectories. Tab. 10 reports the Judge-valid discovery rate — the fraction of the 100 site–seed episodes with at least one strict adaptation-time pass (Eq. 33) — at three cumulative-budget checkpoints. Frozen cross-site memory raises the discovery rate at every checkpoint, with the largest gain at 50 calls (13pp) and a smaller gain at 100 calls (10pp).

Calls to first discovery. Among episodes that eventually succeed within the 100-call budget, the median call index of the first Judge-valid candidate is 28 without cross-site memory and 21 with frozen cross-site memory. Restricting to the 25 site–seed episodes in which *both* conditions succeed (the "both succeed" cell of Tab. 11) makes the comparison paired: the marginal median call index moves from 27 (no cross-site memory) to 17 (frozen cross-site memory), a difference of 10 calls between the two marginals, while the median of the 25 per-episode differences is 13 calls. Both summaries point the same way — memory consistently shortens the path to a valid candidate — and we report both because a difference of medians and a median of differences capture slightly different aspects of that shift.

Outcome heterogeneity at the 100-call horizon. Tab. 11 cross-tabulates the 100 site–seed episodes by whether each condition reaches a strict pass within the shared budget. A quarter of episodes succeed under both conditions and over a third fail under both; the remaining 40 episodes are asymmetric, and frozen cross-site memory resolves substantially more of them (25 memory-only vs. 15 no-memory-only).

Site-level heterogeneity and aggregate effect. The benefit is concentrated, not uniform. Frozen cross-site memory gives a consistent, repeatable advantage at US-Minneapolis1, US-Minneapolis2, CA-Sunset, and PL-Narutowicza, but produces genuine negative transfer at NL-Amsterdam and SG-TelokKurau06, where seeds that succeed without cross-site retrieval more often fail or succeed later once it is consulted — consistent with the effect-aware reranking (Eq. 26–28) occasionally favoring a plausible but locally wrong direction before local evidence outweighs it. A site-clustered bootstrap over the discovery-rate gains in Tab. 10 confirms a positive average effect while showing this heterogeneity is real: the gain is driven by a subset of sites rather than distributed evenly. We therefore read this ablation as showing that cross-site experience is useful on average but must be context-matched, with explicit control for the sites where it transfers negatively.

Table 12: Sealed-evaluation grades for the 20-site component ablation. Counts refer to sites. A+ and A satisfy the reported material-gain and practical process criteria; B and C denote a process trade-off and sub-material gain, respectively; D denotes non-positive aggregate transfer; and U denotes adaptation-time abstention.

Variant	A+	A	B	C	D	U	median $D_{\max}$ (%)
No evolving memory + Judge	4	5	0	0	0	11	0.000004
Evolving memory, no Judge	4	7	6	0	3	0	0.070
Full system	**5**	**9**	1	1	0	4	0.00076

H.6 Component and Failure Diagnostics

Fig. 2(a) shows that failures are not homogeneous. Of 1,410 completed candidates, 862 lack material aggregate gain, 350 improve the aggregate objective but exhibit a full-period process trade-off, 89 first fail temporal consistency, 34 fail a remaining mechanism or validity check, and 75 pass strictly. FI-Torni is dominated by ineffective updates (208/209), JP-Yoyogi by process trade-offs (129/198), and US-WestPhoenix by a mixture of ineffective, process-conflicted, and temporally unstable responses. These distinct signatures justify abstention for different scientific reasons rather than treating every unresolved site as an insufficient-search failure.

Tab. 12 separates search coverage from held-out reliability. The cold-start system without evolving memory freezes nine interventions, all of which satisfy the practical sealed criterion. The complete iterative trajectory increases frozen coverage to 16 sites, of which 14 receive grades A+ or A. The no-Judge branch freezes interventions at all 20 sites, but six exhibit process trade-offs and three fail to retain positive aggregate transfer. Its larger scalar gain therefore comes at the cost of substantially weaker process and temporal reliability.

H.7 Mechanism-Stratified Interpretation

Radiative parameters primarily affect $LW_{\uparrow}$ and available energy, but their useful direction is site-dependent. Vegetation conductance more directly alters Q_H Q_E partitioning, with direction varying across wetness and residual regimes. Paired changes in displacement height and roughness length occur only at GR-HECKOR and UK-Swindon and should be interpreted as locally useful aerodynamic corrections, not newly measured morphological constants.

These results support representing transferable knowledge as a conditional parameter process relationship. They do not support a universal cross-city parameter value or imply that the accepted parameterization is unique.

H.8 Case Study: Failure-Informed Discovery at US-Minneapolis2

Relation to the main results. This case study expands the US-Minneapolis2 result reported as site 15-b in Table 8. It examines the same frozen intervention included in the main-paper site counts, aggregate statistics, and temporal-transfer analysis. Its candidate trajectory is shown in Fig. 2(b).

Cold-start failure. During adaptation, the baseline had mean biases of $+14.11\ \mathrm{W\,m^{-2}}$ for sensible heat flux Q_H, $-7.05\ \mathrm{W\,m^{-2}}$ for latent heat flux Q_E, and $+12.56\ \mathrm{W\,m^{-2}}$ for upward longwave radiation $LW_{\uparrow}$. An initial five-round search without retrieved memory evaluated 27 candidates spanning grass, radiative, and vegetation-conductance parameters. None passed the complete adaptation Judge, so the baseline was retained at the end of the cold-start stage.

Each completed trial was recorded with its parameter changes, output-wise error effects, gate outcomes, and rejection reasons. Among the tested directions, grass and conductance interventions either failed to produce a material aggregate gain or introduced an output-wise trade-off. These failures subsequently provided site-specific evidence about ineffective and process-conflicted directions.

Failure-informed redirection. The subsequent memory-informed generation retrieved both cross-site records and the completed US-Minneapolis2 trials. Its first four rounds tested grass,

paved, building, and deciduous-tree radiative parameters without finding a Judge-valid intervention. In the fifth round, the MECHANISM REASONER proposed a previously untested surface class. The following excerpt shows the decision-relevant fields of its bounded plan:


```
{
  "process_family": "radiative",
  "parameters": [
    {
      "parameter_id": "surface.evergreen_trees.albedo",
      "direction": "increase",
      "trust_region_relative": [0.0, 0.05],
      "expected_effects": {
        "QH": "improve",
        "QE": "improve",
        "LWup": "improve"
      },
      "risk_outputs": ["QH", "LWup"]
    },
    {
      "parameter_id": "surface.evergreen_trees.emissivity",
      "direction": "decrease",
      "trust_region_relative": [-0.05, 0.0],
      "expected_effects": {
        "QH": "improve",
        "QE": "neutral",
        "LWup": "improve"
      },
      "risk_outputs": ["LWup", "QH"]
    }
  ],
  "abstain_if_unrepairable": true
}
```


Listing 5: Decision-relevant excerpt from the fifth-round US-Minneapolis2 parameter proposal.

The LLM specified the parameter identities, directions, bounded trust regions, and expected output-error responses, but did not provide an accepted numerical value. Deterministic code generated counterfactuals within these regions. Candidates were placed in a mechanism-aware deterministic order and softly reranked using retrieved effect memory. This reranking changed their execution priority but did not replace SUEWS evaluation or enter the Judge evidence. As target-site trials accumulated, the contribution of cross-site memory decreased relative to the local response evidence.

Two candidates in the final round passed the complete adaptation Judge. The fixed post-evaluation selection rule froze

$$\alpha_{\text{evergreen}} : 0.1320 \to 0.1386, \qquad \varepsilon_{\text{evergreen}} : 0.9800 \to 0.9555, \tag{37}$$

corresponding to a 5% albedo increase and a 2.5% emissivity decrease.

The intervention was discovered after 67 cumulative candidate evaluations: 27 in the cold-start stage and 40 in the memory-informed generation. The two strict candidates were the 39th and 40th candidates of the latter generation, corresponding to the 66th and 67th cumulative candidate evaluations. This candidate-only count excludes baseline simulations and the 24 bilateral OAT perturbations used to construct the shared sensitivity evidence. Under normalized from-zero accounting with a reusable baseline, the total cost is

$$N_{\text{from-zero}} = 1_{\text{baseline}} + 24_{\text{OAT}} + 67_{\text{candidate}} = 92. \tag{38}$$

Because the archived OAT, cold-start, and memory-informed programs each executed their own baseline, the literal physical execution count is 94.

Adaptation verification. The selected intervention improved the adaptation objective by 0.101%. Its full-period relative nRMSE changes were

$$\Delta E_{Q_H,\text{adapt}} = -0.0886\%, \qquad \Delta E_{Q_E,\text{adapt}} = -0.0483\%, \qquad \Delta E_{LW_\uparrow,\text{adapt}} = -0.2818\%. \tag{39}$$

All three output errors decreased over the full adaptation period, and the candidate passed all four temporal-block checks. The observed error changes were also consistent with the recorded qualitative response expectations. The intervention therefore passed the complete adaptation Judge and was frozen before sealed evaluation.

Sealed temporal transfer. In the one-shot sealed evaluation, the frozen intervention retained an aggregate gain of 0.0849%, rounded to 0.085% in Table 8. Its output-wise relative nRMSE changes were

$$\Delta E_{Q_H,\mathrm{eval}} = -0.0890\%, \qquad \Delta E_{Q_E,\mathrm{eval}} = -0.0169\%, \qquad \Delta E_{LW_\uparrow,\mathrm{eval}} = -0.2615\%. \quad (40)$$

No evaluated output degraded, so $D_{\max} = 0$ and the intervention received grade A+. The absolute frozen values are also reported in Table 9.

Interpretation. This trajectory illustrates how recorded failures and context-matched cross-site effects can inform a subsequent hypothesis and reprioritize candidate execution. The LLM revised the parameter-level hypothesis; the deterministic generator constructed bounded numerical counterfactuals; effect-memory reranking ordered their execution; SUEWS produced the numerical evidence; and the fixed Judge determined admissibility. Retrieved parameter values were never transferred directly.

The case demonstrates failure-informed hypothesis revision and memory-guided search ordering, but does not by itself identify the independent causal contribution of memory or disentangle the physical effects of the two parameter changes.

H.9 Scope and Limitations

These results establish within-site transfer to a held-out future interval; extending the framework to zero-shot generalization across unseen cities, and disentangling the individual causal contribution of each parameter in a paired intervention, are left to future work with an expanded site set. The sparse, bounded corrections improve interpretability over a full re-tuning but do not by themselves guarantee parameter identifiability or uniqueness. The Physics Judge verifies executable semantics, numerical integrity, output-wise error behaviour, and temporal consistency; extending it into an independent test of energy- or water-conservation laws is a natural next step. Finally, the 0.1% practical margin is used only to characterize sealed-period transfer: adaptation-time acceptance remains exactly non-degrading, and sealed outcomes never trigger additional parameter search or memory updates.