SNOTRA: CONTINUALLY ADAPTING IRRIGATION WORLD MODELS UNDER PHYSICAL CHANGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Physical agents must revise imperfect world models as the systems they control change. However persistent prediction error alone does not reveal which change is identifiable, which revision deserves belief, or whether that revision matters for action. We study this problem in precision irrigation, where partially observed soil–crop–weather dynamics interact with constrained, long-horizon interventions. We present **Snotra**, an evidence-governed framework that separates world-model revision into three decisions: *KNOW*, whether competing explanations are distinguishishable from available evidence; *BELIEVE*, whether a candidate survives held-out validation, falsification, physical-consistency, safety, and provenance checks; and *ACT*, whether an admitted model changes a legal decision and improves an independently evaluated physical outcome. We further introduce **SnotraBench**, a stage-resolved evaluation protocol for world-model revision under physical change. On 35 matched governance units, Snotra accepts all $10/10$ supported revisions while rejecting all $25/25$ unsupported ones; Heldout-Only accepts the same $10/10$ supported revisions but also $10/25$ unsupported revisions. The paired reduction in unsupported admissions is 0.286 (95% cluster-bootstrap CI $[0.143, 0.429]$). In a one-check-removed replay, identifiability and physical-consistency are the only checks whose removal increases false admissions on this scenario set. Finally, a controlled drift sweep exposes a prediction-to-decision threshold: prediction discrepancy increases as physical drift grows from 0% to 50%, but the irrigation action remains unchanged through 30%; at 50%, revision changes the action by 10 mm and reduces independently evaluated cost by 110 protocol units. These results show that predictive repair, model admission, and decision value are distinct quantities and motivate evaluating adaptive world models by both admission selectivity and prediction-to-decision transmission.

1 INTRODUCTION

World models let an agent predict the consequences of candidate actions before acting and are central to model-based planning and control (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2020; 2021; Schrittwieser et al., 2020; Hansen et al., 2024). Recent work increasingly asks whether learned dynamics are useful in closed-loop decision making rather than merely accurate as predictors (Zhang et al., 2026; Wang et al., 2025; 2026). A different problem appears when an agent remains coupled to a physical system for long periods: the world described by its model can change. Physical parameters drift, operating regimes shift, sensing remains partial, and an abstraction that was previously adequate can become systematically mismatched.

Updating the model whenever prediction error persists is not enough. Fault diagnosis has long distinguished residual generation from fault isolation (Isermann, 2005; Blanke et al., 2016); system identification and adaptive control study estimation under uncertain dynamics (Ljung, 1999; Åström & Wittenmark, 1995); model validation asks whether observed data are consistent with a candidate uncertainty description (Smith & Doyle, 1992); and dual control studies actions that simultaneously regulate and probe an uncertain system (Mesbah, 2018). Likewise, decision-focused learning has shown that predictive error and downstream decision quality need not align (Donti et al., 2017; Wilder et al., 2019; Elmachtoub & Grigas, 2022). These literatures make the individual pieces familiar. The unresolved systems question we study is how to compose them into an auditable *revision lifecycle*:

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.

when does evidence justify changing an executable physical world model, and when should that changed belief receive authority over action?

This lifecycle contains three distinct decisions. First, *what can be known?* A residual may be compatible with several parameter, structural, sensing, or forcing changes. Second, *what should be believed?* A candidate can improve recent fit while overfitting discovery data or violating physical/provenance constraints. Third, *what should be allowed to act?* Even a credible model revision may leave the planner's action unchanged, while a changed action may fail to improve the physical outcome. Treating these steps as one adaptation score hides the location of failure.

Precision irrigation makes the problem concrete. Irrigation decisions depend on root-zone water, crop demand, weather forcing, soil properties, and actuation efficiency. Physically interpretable evapotranspiration and water-balance models are well established (Allen et al., 1998; Pereira et al., 2020), as are model-based irrigation controllers (Lozoya et al., 2014; Delgoda et al., 2016; Bwambale et al., 2023). Yet the same soil-water mismatch can be consistent with changed crop demand, hydraulic response, delivery efficiency, observation error, or forcing error. More importantly, a correction that matters statistically may still be too small to cross an irrigation decision boundary.

We present **Snotra**, an evidence-governed framework for revising versioned executable physical world models. Snotra organizes revision around **KNOW–BELIEVE–ACT**. KNOW tests whether the pre-specified alternatives are distinguishable under the available observations and interventions. BELIEVE separates candidate construction from admission using held-out evidence, falsification, physical consistency, safety, and provenance checks. ACT separates model admission from decision value by measuring whether the admitted revision changes a legal action and whether that action improves an independent physical endpoint. The runtime model and the independent evaluator remain information-isolated.

We also introduce **SnotraBench**, an evaluation protocol that reports the revision chain stage by stage. The irrigation implementation deliberately uses an interpretable gray-box model so that physical changes and candidate mechanisms can be audited; the evaluation contract itself does not require a particular model class and can be applied to analytical, learned, or hybrid world models. This is not another controller leaderboard. Its purpose is to prevent a gain at one stage—for example, prediction—from being silently promoted into a stronger claim about diagnosis, action, or utility. The protocol is instantiated here in irrigation; the empirical claims in this paper are restricted to that domain.

Our experiments use real-world irrigation time series together with explicitly labeled counterfactual stress tests, and answer three questions that directly test this formulation. First, does evidence governance add anything beyond ordinary held-out validation? On 35 matched units, Heldout-Only and Snotra both retain $10/10$ supported revisions, but Heldout-Only also admits $10/25$ unsupported revisions whereas Snotra admits none; the paired reduction is 0.286 with a 95% cluster-bootstrap CI of $[0.143, 0.429]$. Second, what revision capability is actually established, and which admission checks matter in this scenario set? On the scoped capability tasks, the KNOW challenge agrees with the pre-specified status in 8/8 cases, a Bushland-derived crop-coefficient repair reduces held-out MAE from 0.260 to 0.078 mm (a 70% reduction) with 9/9 candidates admitted, and a finite structural task succeeds in 2/4 cases. Removing either identifiability or physical consistency then introduces $5/25$ false admissions; removing validation, falsification, safety, or provenance changes no admission decisions on these cases, so we do *not* claim that every gate is empirically necessary. Third, when does predictive repair become decision-relevant? Across a controlled 0–50% drift sweep, prediction mismatch increases smoothly, yet both evaluated planners keep the same action through 30% drift; at 50%, the revised model changes irrigation by 10 mm and lowers independently evaluated cost by 110 protocol units. This turns the usual "prediction is not decision quality" warning into a measurable transmission threshold on a controlled physical task.

Contributions.

- **A revision-centric formulation of physical world models.** We separate knowability, belief admission, action sensitivity, and independent utility, connecting ideas from diagnosis, model validation, adaptive control, and decision-focused learning without claiming novelty for those fields individually.

- **Snotra, an evidence-governed revision procedure.** Snotra maintains versioned executable models and makes retain, reject, promote, and rollback explicit outcomes; model admission and actuator authority are separate decisions.
- **SnotraBench and stage-resolved evidence.** We provide matched baselines, a matched Heldout-Only baseline, one-check-removed replay, scoped KNOW/parameter/structure tests, and a drift-to-decision transmission study. The results show both where governance adds value beyond held-out validation and where stronger control/generalization claims remain unsupported.

2 Related Work

We position the work by the decision made about a changing model: diagnosis and identifiability address what can be inferred, model validation and adaptive control address model fit and closed-loop adaptation, decision-focused learning addresses downstream loss, and runtime assurance addresses action safety. Our contribution is the explicit revision lifecycle connecting these questions and evaluating each transition separately.

World models, adaptation, and closed-loop evaluation. Learned world models support latent prediction, planning, and policy learning (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2020; 2021; 2025; Schrittwieser et al., 2020; Hansen et al., 2022; 2024; Micheli et al., 2023; Chua et al., 2018; Janner et al., 2019). Recent work studies adaptation under distribution shift or sequential experience (Wang et al., 2025; 2026; Kessler et al., 2023; Liu et al., 2025), while closed-loop benchmarks show that perceptual or predictive quality need not translate into task success (Zhang et al., 2026). We share the closed-loop concern but focus on a different object: the *revision* of a deployed physical model after mismatch appears.

Diagnosis, identifiability, and active information gathering. Model-based fault detection and isolation distinguishes detection from isolating the underlying cause (Isermann, 2005; Blanke et al., 2016). Structural and practical identifiability similarly ask whether parameters or states can be uniquely recovered from a model and experiment (Walter & Pronzato, 1997; Raue et al., 2009; Villaverde et al., 2016; Lam et al., 2022). When existing observations are insufficient, experiment design and dual control explicitly value actions that reveal uncertain dynamics while controlling the system (Mesbah, 2018). Snotra does not replace these theories; KNOW turns their core question into a precondition for admitting a deployed model revision.

System identification, model validation, and adaptive control. System identification and adaptive control provide mature methods for estimating and controlling uncertain dynamics (Ljung, 1999; Åström & Wittenmark, 1995); MPC makes the quality of a predictive model central to constrained action selection (Morari & Lee, 1999; Rawlings et al., 2017; Hewing et al., 2020). Model invalidation asks whether measured data are compatible with a proposed uncertainty description (Smith & Doyle, 1992), and sparse discovery methods such as SINDy learn parsimonious dynamical models from data (Brunton et al., 2016). Our novelty boundary is therefore not parameter fitting or model validation in isolation, but the explicit separation between inference, revision admission, and action authority in a long-running physical-agent workflow.

Decision-focused learning and distribution shift. Decision-focused and predict-then-optimize methods directly target the misalignment between prediction loss and downstream decision loss (Donti et al., 2017; Wilder et al., 2019; Elmachtoub & Grigas, 2022); concept-drift research studies detection and adaptation when data-generating relations change over time (Gama et al., 2014). These works motivate two parts of our evaluation: prediction improvement is not used as a proxy for utility, and model change is evaluated across a controlled severity sweep. Unlike decision-focused training, however, our primary object is governance of a post-deployment physical model revision rather than learning a predictive model against a downstream loss.

Safety, physics-guided learning, and irrigation. Runtime-assurance and safety-filter architectures constrain actions without asserting that the underlying model is epistemically correct (Seto et al., 1998; Ames et al., 2017; Berkenkamp et al., 2017; Wabersich & Zeilinger, 2021; Brunke et al., 2022). Physics-informed and hybrid modeling inject mechanistic structure into learning (Raissi et al., 2019;

Chen et al., 2018; Karniadakis et al., 2021). Irrigation itself is a mature physical-control domain with established crop-water models, MPC, and crop simulators (Allen et al., 1998; Lozoya et al., 2014; Delgoda et al., 2016; Bwambale et al., 2023; de Wit et al., 2019; Steduto et al., 2009; Jones et al., 2003). We use it as an interpretable setting in which belief revision, action, and realized consequence can be audited separately.

3 Problem Formulation

Let $x_t \in \mathcal{X}$ be latent physical state, $u_t \in \mathcal{U}(x_t)$ a legal action, and e_t exogenous forcing. The physical system evolves as

$$x_{t+1} = F_t^\star(x_t, u_t, e_t) + \xi_t, \qquad o_t = H_t^\star(x_t) + \nu_t, \tag{1}$$

while the agent maintains an executable world model

$$M_t = (m_t, \theta_t), \qquad \hat{x}_{t+1} = F_{m_t}(x_t, u_t, e_t; \theta_t). \tag{2}$$

"Executable" means that M_t can predict physical consequences under candidate actions and declared forcing; it may be analytical, learned, or hybrid.

A persistent residual $r_t = d(o_t, \hat{o}_t(M_t))$ signals mismatch but does not identify its cause (Isermann, 2005; Blanke et al., 2016). We therefore define revision as three decisions rather than one update.

KNOW: is a proposed explanation distinguishable? Let $\mathcal{H}_t = \{M_t^{(1)}, \dots, M_t^{(K)}\}$ be a pre-specified hypothesis family. For a legal observation/intervention sequence $\mathbf{u}$, let

$$\Delta_{ij}(\mathbf{u}) = d\Big(\hat{\mathbf{o}}^{(i)}(\mathbf{u}), \hat{\mathbf{o}}^{(j)}(\mathbf{u})\Big)$$

measure the observable separation of two hypotheses. An operational knowability test

$$K_t = \mathcal{K}(\mathcal{D}_t, \mathcal{H}_t, \mathcal{U}_{\text{legal}}) \in \{\text{IDENTIFIABLE}, \text{AMBIGUOUS}, \text{INSUFFICIENT}\}$$

uses only evidence available at decision time. It is relative to the sensing channels, interventions, evidence window, and hypothesis family; it is not a claim of universal identifiability (Walter & Pronzato, 1997; Raue et al., 2009).

BELIEVE: should a candidate replace the active model? A candidate M', constructed from discovery evidence $\mathcal{D}^{\text{disc}}$, is admitted only if it passes a pre-specified admission contract:

$$B(M') = G_{\text{id}} \wedge G_{\text{val}} \wedge G_{\text{fals}} \wedge G_{\text{phys}} \wedge G_{\text{safe}} \wedge G_{\text{prov}}. \tag{3}$$

The gates respectively test knowability, held-out support, falsification, physical consistency, safety admissibility, and provenance/leakage. Candidate construction and acceptance are separate operations, analogous in spirit to model validation but with an explicit versioned lifecycle (Smith & Doyle, 1992).

ACT: is the admitted revision decision-relevant? For planner π, let $u(M) = \pi(o_{\le t}, e_{t:t+H}, M)$ and

$$\Delta_{\text{act}} = d_u(u(M_t), u(M')) \,.$$

If $\Delta_{\text{act}} = 0$, a predictive repair is action-insensitive under the current planner. If actions differ, the relevant outcome is not the candidate model's own score but realized cost under an independent plant $P^\star$:

$$J^\star(\mathbf{u}) = J(P^\star, \mathbf{u}).$$

Evaluator-only state is never available to revision or action ranking. The scientific object is thus the sequence

$$\text{physical change} \rightarrow \text{KNOW} \rightarrow \text{BELIEVE} \rightarrow \text{ACT} \rightarrow \text{independent utility},$$

with a separate measurement at each transition rather than an assumed implication.

4 Snotra: Evidence-Governed World-Model Revision

Snotra implements the formulation above with one design rule: *detecting mismatch, constructing a candidate, admitting a model, selecting an action, and evaluating a physical outcome are separate operations*. Figure 1 shows the information and authority boundaries.

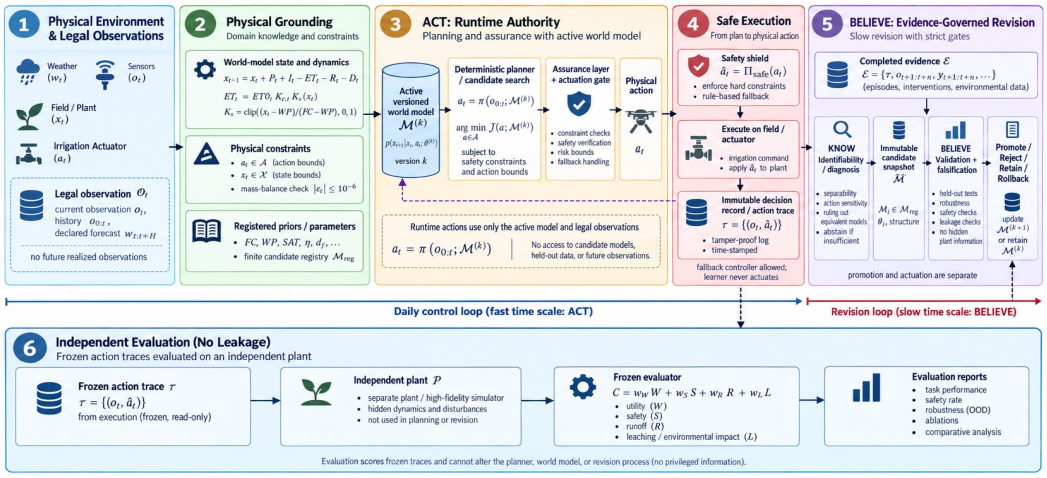


Figure 1: **Snotra separates model belief from physical authority.** Runtime actions are generated only from the active model and legal observations. Completed evidence may trigger a slower revision cycle. Independent evaluation scores frozen action traces without leaking hidden plant information into revision or planning.

4.1 Executable Gray-Box World Model

In the irrigation instantiation, the state is root-zone water storage x_t, and commanded irrigation a_t yields delivered irrigation $I_t = \eta a_t$. With precipitation P_t, reference evapotranspiration $ET_{0,t}$, crop coefficient $K_{c,t}$, field capacity FC, wilting point WP, saturation SAT, and drainage factor d_f,

$$K_s(x_t) = \text{clip}\left(\frac{x_t - WP}{FC - WP}, 0, 1\right), \tag{4}$$

$$R_t = \max(0, x_t + P_t + I_t - SAT), \tag{5}$$

$$D_t = d_f\,[x_t - FC]_+, \qquad [z]_+ \equiv \max(0, z), \tag{6}$$

$$x_{t+1} = \text{clip}(x_t + P_t + I_t - ET_{0,t} K_{c,t} K_s - R_t - D_t, 0, SAT)\,. \tag{7}$$

The conservation relations, state bounds, and action semantics form the fixed physical backbone. Parameters such as crop demand or delivery efficiency and a finite set of structural alternatives form the revisable component. This gray-box representation makes the revision auditable; gray-box modeling itself is not our novelty.

4.2 Revision Lifecycle

At runtime, the fast loop uses the currently active model to predict, plan, and pass an assurance layer before action. Completed observations may trigger a slower revision loop. If the candidate explanations are not distinguishable, the active model is retained. Otherwise, discovery evidence constructs an immutable candidate and held-out evidence is used for admission. A promoted candidate becomes a new model version; a failed candidate is rejected, and a later lifecycle failure can roll back to a parent version.

The algorithm deliberately stops at authority assignment. Independent physical utility is an evaluation quantity measured after a frozen action trace is executed; it is never fed back as privileged truth during the same decision.

4.3 Why Multiple Admission Checks?

The six checks in Eq. 3 are not assumed to be equally necessary in every environment. Identifiability and physical consistency can block candidates for reasons that ordinary held-out fit cannot detect; conversely, safety or provenance may be inactive on a clean controlled scenario set. We therefore evaluate both a strong Heldout-Only baseline and one-check-removed variants rather than treating the

Algorithm 1 Evidence-governed revision in Snotra

Require: active model M, completed evidence $\mathcal{D}$, candidate family $\mathcal{H}$
if mismatch trigger does not fire **then**
  return RETAIN(M)
end if
evaluate $K = \mathcal{K}(\mathcal{D}, \mathcal{H}, \mathcal{U}_{\text{legal}})$
if $K \neq$ IDENTIFIABLE **then**
  return RETAIN/ABSTAIN(M)
end if
construct immutable candidate M' from discovery evidence
evaluate Eq. 3 on held-out / independent checks
if any required admission check fails **then**
  return REJECT(M')
end if
promote M' as the active belief version
compute legal action through planner + assurance using M'
record action sensitivity relative to parent model for evaluation
return PROMOTE(M')

checklist itself as evidence. Section 7 reports exactly which checks change decisions in the current experiments.

5 SnotraBench: Evaluating Revision Stage by Stage

A revision benchmark should identify *where* evidence stops propagating. Prediction-only evaluation cannot reveal whether a candidate was identifiable or decision-relevant; end-to-end cost alone conflates model error, planner insensitivity, unsafe actions, and absence of physical headroom. SnotraBench therefore treats the revision lifecycle as five distinct estimands: **CHANGE** measures whether mismatch is present; **KNOW** tests whether candidate explanations are distinguishable; **BELIEVE** measures supported versus unsupported admission; **ACT** measures whether the revision changes a legal decision; and **UTILITY** evaluates the frozen changed action under an independent physical endpoint. A failure at one stage blocks stronger downstream interpretation rather than being absorbed into a single score.

Matched contracts. A comparison is admissible only when methods receive the same observations, candidate family, action space, planner/evaluation budget, plant, and scoring rule. An evaluator may use hidden plant truth to label or score a completed trajectory, but that information is never available to runtime revision or action selection.

Fail-closed qualification. Some tasks are unsuitable for answering a downstream question: the change may be unidentifiable, the planner may have no action headroom, or the available observations may not support the requested distinction. SnotraBench reports such cases as unqualified rather than assigning a method win/loss. In this paper, all headline evidence is anchored to the real-world irrigation data described in Section 6; controlled perturbations are used only to isolate causal stages of revision.

Scope. The benchmark contract is representation-agnostic—the model may be analytical, neural, or hybrid—but the experiments in this paper instantiate it only in irrigation. We therefore claim a reusable *evaluation abstraction*, not cross-domain empirical generalization.

6 Experimental Setup

Real-world irrigation data and controlled stress tests. The empirical substrate consists of seven field-data configurations from USDA–ARS Bushland lysimeter studies, USDA–ARS Colorado/LIRF studies, and Michigan LOCOMOS/Sackett deployments (Evett et al., 2022; Trout & Bausch, 2017;

Trout et al., 2025; Dong et al., 2024). Common variables include timestamp, soil-water estimate, precipitation, irrigation, reference evapotranspiration, temperature, humidity, site, and year. Missingness and temporal coverage vary substantially (Appendix B), and repeated years from one site are never counted as independent sites. Headline prediction and revision experiments use these observed time series; controlled parameter/drift perturbations are explicitly counterfactual stress tests used to isolate admission and decision-transmission mechanisms. They are not presented as additional field observations or as evidence of field-deployed control efficacy.

Information firewall. The runtime agent receives only declared observations and decision-time forcing. Future realized forcing, hidden plant state, terminal crop outcome, and evaluator internals are scoring-only. Candidate construction and held-out evaluation use disjoint evidence when required. For counterfactual action comparisons, the independent physical evaluator executes a frozen action trace only after the decision is fixed; evaluator-only truth never enters revision or action ranking.

Revision tasks. We evaluate four complementary settings. (i) A known crop-coefficient shift tests whether a represented parameter can be repaired from discovery data and validated on held-out observations. (ii) Eight pre-specified KNOW scenarios cover parameter revision, structural revision, no-revision, insufficient-evidence, and no-admissible-candidate outcomes. (iii) A 35-unit governance set contains 10 supported-revision and 25 unsupported-revision cases. (iv) A controlled severity sweep applies 0, 5, 10, 20, 30, and 50% model drift to study when prediction differences become action- and utility-relevant. A four-scenario structural task is reported as supporting evidence because its candidate family is finite.

Baselines. The central governance comparison includes **Static**, **Direct-Update**, **Heldout-Only**, and **Snotra**. Static never revises. Direct-Update accepts a fitted candidate without the full evidence contract. Heldout-Only accepts candidates that improve a disjoint validation set, providing the critical test of whether Snotra adds value beyond ordinary validation. Evaluator-only Oracle information is used for scoring/qualification where appropriate but is not a comparable acting method. For the severity study, we evaluate a threshold planner and a finite candidate-search planner under the same observations, action set, and independent plant.

Frozen operating points. The admission cutoffs, candidate families, and information contracts used by the main comparisons were fixed before the sensitivity replay. The threshold study re-evaluates those frozen outputs over nearby available settings; no cutoff is re-selected after observing the headline outcomes. This supports a scoped anti-tuning check, not a claim that the chosen thresholds are universally optimal.

Metrics and uncertainty. Prediction is reported as MAE. Governance reports supported-revision recall, unsupported-admission rate, harmful-deployment count, and precision. The 35 matched governance units are also analyzed with a 5,000-replicate cluster bootstrap over the independent unit. Action relevance is measured by irrigation-distance in mm. Physical evaluation reports water, stress/deficit, runoff, leaching, safety, and the pre-specified diagnostic cost $J = W + 5S + 10R + 10L$ (lower is better); because these weights are application-specific, Appendix F reports a five-setting weight scan rather than treating J as a universal agronomic objective. All headline numbers use the independent unit stated in the corresponding table or figure, following established concerns about variance and pseudo-replication in empirical learning-system evaluation (Henderson et al., 2018; Agarwal et al., 2021).

7 Evaluation

We organize the evidence around four questions: whether governance adds beyond held-out validation, which admission checks contribute on the current scenarios, what revision capabilities are actually established, and when predictive repair becomes decision-relevant. Table 1 makes the claim–evidence boundary explicit before the detailed analysis.

Table 1: **Main evidence map.** Different tasks have different independent units and are interpreted separately; they are not pooled into a single accuracy.

Question	Independent unit	Main result	Supported interpretation
Admission selectivity	35 matched units	Heldout-Only 10/25 unsupported vs. Snotra 0/25; both 10/10 supported	governance adds beyond validation on this scenario set
Scoped knowability	8 pre-specified scenarios	8/8 status agreement; 8/8 information-firewall pass	consistency on registered cases, not general diagnostic accuracy
Parameter revision	9 Bushland-derived episodes	MAE 0.260 $\rightarrow$ 0.078 mm; 9/9 admitted	represented crop-coefficient repair
Finite structural revision	4 pre-specified mechanisms	2/4 pass	preliminary bounded model selection only
Decision transmission	6 drift levels	action unchanged through 30%; 10-mm change and 110-unit cost reduction at 50%	controlled prediction-to-decision boundary

7.1 Does evidence governance add beyond held-out validation?

Table 2 gives the central matched comparison. Direct-Update, Heldout-Only, and Snotra all recover the $10/10$ supported revisions, so the relevant question is not recall but unsupported admission. Direct-Update accepts $15/25$ unsupported revisions and Heldout-Only accepts $10/25$; Snotra accepts none. Relative to Heldout-Only, the paired unsupported-admission-rate reduction is 0.286, with a 95% cluster-bootstrap CI of $[0.143, 0.429]$ over 35 independent governance units (5,000 bootstrap replicates). Heldout-Only also yields 5 harmful deployments among the 35 units, compared with 0 for Snotra. Thus, on this scenario set, the added evidence checks improve admission selectivity without reducing supported-revision recall.

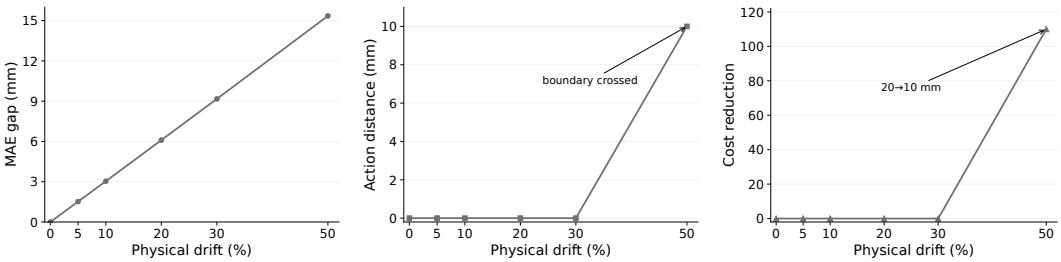
Table 2: **Matched governance comparison.** Denominators distinguish 25 unsupported-revision cases from 10 supported-revision cases. Precision is defined over admitted revisions. Mean physical cost is shown to make the limitation explicit: stronger admission does *not* imply lower control cost.

Method	Unsupported accepted $\downarrow$	Supported accepted $\uparrow$	Precision $\uparrow$	Harmful / 35 $\downarrow$	Mean cost $\downarrow$
Static	0/25	0/10	–	0	3218.6
Direct-Update	15/25	10/10	0.40	10	2548.0
Heldout-Only	10/25	10/10	0.50	5	**2453.7**
Snotra	**0/25**	**10/10**	**1.00**	**0**	2820.6

The cost column is deliberately not used as a win for Snotra: Heldout-Only has lower mean cost on these scenarios. Our supported conclusion is therefore narrower—Snotra improves *revision admission* beyond ordinary held-out validation under the matched contract, not overall irrigation-control performance.

7.2 Which admission checks matter on these scenarios?

The one-check-removed replay removes one admission check at a time while replaying the same 35 matched decision records. Removing the identifiability check produces $5/25$ false admissions; removing physical consistency produces another $5/25$. Removing held-out validation, falsification, safety, or provenance changes no admission decisions on this particular scenario set. The correct interpretation is therefore *not* that six gates are all empirically necessary: only two have measured marginal effect here. The remaining checks are part of the general contract but require targeted failure scenarios before claiming empirical necessity.



(a) Prediction repair. (b) Action transfer. (c) Independent utility.

Figure 2: **Prediction-to-decision transmission under controlled drift.** Prediction discrepancy grows continuously, but action and physical cost change only at 50% drift. Both evaluated planners select the same actions on this fixture.

Full one-check-removed counts are reported in Appendix C; no aggregate "gate importance" score is formed because the removals are evaluated on the same fixed replay.

7.3 What revision capabilities are established?

The capability tests deliberately separate knowability from repair. On eight pre-specified KNOW cases spanning supported revision, no-revision, insufficient-evidence, and no-admissible-candidate outcomes, the gate agrees with the registered status in 8/8 cases and the information firewall passes in 8/8. These eight cases are too few, and not constructed as a balanced binary diagnostic set, to support TPR/TNR or AUROC claims. On nine Bushland-derived crop-coefficient episodes, held-out prediction MAE falls from 0.260 to 0.078 mm (approximately 70%), and all 9/9 candidates satisfy the registered admission contract. This establishes represented-parameter repair, not general online adaptation. Finally, only 2/4 pre-specified structural alternatives pass; we therefore describe structural revision as preliminary finite model selection rather than mechanism discovery. These scoped results are summarized in Table 1, with full denominators in Appendix D.

7.4 When does predictive repair become decision-relevant?

A prediction repair may be physically real yet too small to cross the planner's decision boundary. We test this by sweeping controlled model drift from 0 to 50% with plant, forcing, action set, and evaluator fixed. The prediction gap grows from 1.52 to 15.35 mm over 5–50% drift, but the selected action remains 20 mm through 30%. At 50%, the repaired model changes irrigation from 20 to 10 mm and lowers independent cost from 3999.9 to 3889.9 (110 protocol units) with no recorded safety violation. Figure 2 therefore identifies a controlled prediction-to-decision threshold rather than a universal law.

A threshold policy and finite candidate-search planner select identical actions at all six drift levels, and five pre-specified objective-weight settings do not change the action on the robustness scenario. These checks narrow obvious alternative explanations on this fixture but do not establish planner- or objective-independent behavior (Appendix F).

8 Conclusion

Evidence for revising a physical world model and evidence for using that revision to act are not the same object. On the matched irrigation scenarios, ordinary held-out validation retains all supported revisions but still admits 10/25 unsupported ones, whereas the full evidence contract admits none; a separate severity study shows that increasingly better model repair can remain action-insensitive until the planner's decision boundary is crossed. These results motivate two reusable evaluation targets for adaptive world models—*admission selectivity* and *prediction-to-decision transmission*—and a fail-closed rule when the evidence required for a stronger claim is absent. Snotra and SnotraBench instantiate this revision-centric view on real-world irrigation time series with controlled counterfactual stress tests; general diagnosis, field-control superiority, open-ended discovery, and cross-domain generalization remain outside the demonstrated scope.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Richard G. Allen, Luis S. Pereira, Dirk Raes, and Martin Smith. Crop evapotranspiration: Guidelines for computing crop water requirements. Technical Report 56, Food and Agriculture Organization of the United Nations, Rome, 1998.

Aaron D. Ames, Xiangru Xu, Jessy W. Grizzle, and Paulo Tabuada. Control barrier function based quadratic programs for safety critical systems. *IEEE Transactions on Automatic Control*, 62(8): 3861–3876, 2017. doi: 10.1109/TAC.2016.2638961.

Karl J. Åström and Björn Wittenmark. *Adaptive Control*. Addison-Wesley, 2 edition, 1995.

Felix Berkenkamp, Matteo Turchetta, Angela P. Schoellig, and Andreas Krause. Safe model-based reinforcement learning with stability guarantees. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Mogens Blanke, Michel Kinnaert, Jan Lunze, and Marcel Staroswiecki. *Diagnosis and Fault-Tolerant Control*. Springer Berlin Heidelberg, 3 edition, 2016. doi: 10.1007/978-3-662-47943-8.

Lukas Brunke, Melissa Greeff, Adam W. Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P. Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5:411–444, 2022. doi: 10.1146/annurev-control-042920-020211.

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

Erion Bwambale, Felix K. Abagale, and Geophrey K. Anornu. Data-driven model predictive control for precision irrigation management. *Smart Agricultural Technology*, 3:100074, 2023. doi: 10.1016/j.atech.2022.100074.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Allard de Wit, Hendrik Boogaard, Davide Fumagalli, Sander Janssen, Rob Knapen, Daniel van Kraalingen, Iwan Supit, Raymond van der Wijngaart, and Kees van Diepen. 25 years of the WOFOST cropping systems model. *Agricultural Systems*, 168:154–167, 2019. doi: 10.1016/j.agsy.2018.06.018.

Dilini Delgoda, Hector Malano, Syed K Saleem, and Malka N Halgamuge. Irrigation control based on model predictive control (mpc): Formulation of theory and validation using weather forecast data and aquacrop model. *Environmental Modelling & Software*, 78:40–53, 2016.

Younsuk Dong, Benjamin Werling, Zhichao Cao, and Gen Li. Implementation of an in-field iot system for precision irrigation management. *Frontiers in Water*, 6:1353597, 2024. doi: 10.3389/frwa.2024.1353597.

Priya L. Donti, Brandon Amos, and J. Zico Kolter. Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5484–5494, 2017.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1): 9–26, 2022. doi: 10.1287/mnsc.2020.3922.

Steven R. Evett, Karen S. Copeland, Brice B. Ruthardt, Gary W. Marek, Paul D. Colaizzi, Terry A. Howell, and David K. Brauer. The bushland, texas maize for grain datasets. USDA National Agricultural Library, Ag Data Commons, 2022. Field lysimeter, water-balance, soil-water, weather, and agronomic data.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):44:1–44:37, 2014. doi: 10.1145/2523813.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565. PMLR, 2019.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *International Conference on Learning Representations*, 2021.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025. doi: 10.1038/s41586-025-08744-2.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024.

Nicklas A. Hansen, Hao Su, and Xiaolong Wang. Temporal difference learning for model predictive control. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8387–8406. PMLR, 2022.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11694.

Lukas Hewing, Kim P. Wabersich, Marcel Menner, and Melanie N. Zeilinger. Learning-based model predictive control: Toward safe learning in control. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:269–296, 2020. doi: 10.1146/annurev-control-090419-075625.

Rolf Isermann. Model-based fault-detection and diagnosis–status and applications. *Annual Reviews in Control*, 29(1):71–85, 2005. doi: 10.1016/j.arcontrol.2004.12.002.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

James W. Jones, Gerrit Hoogenboom, Cheryl H. Porter, Kenneth J. Boote, William D. Batchelor, L. A. Hunt, Paul W. Wilkens, Upendra Singh, Arjan J. Gijsman, and Joe T. Ritchie. The DSSAT cropping system model. *European Journal of Agronomy*, 18(3–4):235–265, 2003. doi: 10.1016/S1161-0301(02)00107-7.

George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3:422–440, 2021. doi: 10.1038/s42254-021-00314-5.

Samuel Kessler, Mateusz Ostaszewski, Michał Paweł Bortkiewicz, Mateusz Żarski, Maciej Wołczyk, Jack Parker-Holder, Stephen J. Roberts, and Piotr Miłoś. The effectiveness of world models for continual reinforcement learning. In *Proceedings of the 2nd Conference on Lifelong Learning Agents*, volume 232 of *Proceedings of Machine Learning Research*, pp. 184–204. PMLR, 2023.

Nicholas N. Lam, Paul D. Docherty, and Rua Murray. Practical identifiability of parametrised models: A review of benefits and limitations of various approaches. *Mathematics and Computers in Simulation*, 199:202–216, 2022. doi: 10.1016/j.matcom.2022.03.020.

Zichen Liu, Guoji Fu, Chao Du, Wee Sun Lee, and Min Lin. Continual reinforcement learning by planning with online world models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 38397–38423. PMLR, 2025.

Lennart Ljung. *System Identification: Theory for the User*. Prentice Hall, 2 edition, 1999.

Camilo Lozoya, Carlos Mendoza, Leonardo Mejia, Jesus Quintana, Gilberto Mendoza, Manuel Bustillos, Octavio Arras, and Luis Solis. Model predictive control for closed-loop irrigation. *IFAC Proceedings Volumes*, 47:4429–4434, 2014. doi: 10.3182/20140824-6-ZA-1003.02067.

Ali Mesbah. Stochastic model predictive control with active uncertainty learning: A survey on dual control. *Annual Reviews in Control*, 45:107–117, 2018. doi: 10.1016/j.arcontrol.2017.11.001.

Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. In *International Conference on Learning Representations*, 2023.

Manfred Morari and Jay H. Lee. Model predictive control: Past, present and future. *Computers & Chemical Engineering*, 23(4–5):667–682, 1999. doi: 10.1016/S0098-1354(98)00301-9.

Luis S. Pereira, Paula Paredes, and Nebojsa Jovanovic. Soil water balance models for determining crop water and irrigation requirements and irrigation scheduling focusing on the FAO56 method and the dual Kc approach. *Agricultural Water Management*, 241:106357, 2020. doi: 10.1016/j.agwat.2020.106357.

Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. doi: 10.1016/j.jcp.2018.10.045.

Andreas Raue, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula Klingmüller, and Jens Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, 2009. doi: 10.1093/bioinformatics/btp358.

James B. Rawlings, David Q. Mayne, and Moritz M. Diehl. *Model Predictive Control: Theory, Computation, and Design*. Nob Hill Publishing, 2 edition, 2017. ISBN 9780975937730.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588:604–609, 2020. doi: 10.1038/s41586-020-03051-4.

Danbing Seto, Bruce H. Krogh, Lui Sha, and Alongkrit Chutinan. Dynamic control system upgrade using the simplex architecture. *IEEE Control Systems*, 18(4):72–80, 1998. doi: 10.1109/37.710880.

Roy S. Smith and John C. Doyle. Model validation: A connection between robust control and identification. *IEEE Transactions on Automatic Control*, 37(7):942–952, 1992. doi: 10.1109/9.148346.

Pasquale Steduto, Theodore C. Hsiao, Dirk Raes, and Elias Fereres. AquaCrop—the FAO crop model to simulate yield response to water: I. concepts and underlying principles. *Agronomy Journal*, 101(3):426–437, 2009. doi: 10.2134/agronj2008.0139s.

Thomas J. Trout and Walter C. Bausch. Usda-ars colorado maize water productivity data set. *Irrigation Science*, 35:241–249, 2017. doi: 10.1007/s00271-017-0537-9.

Thomas J. Trout, Louise H. Comas, Kendall C. DeJonge, and Huihui Zhang. Colorado sunflower water use, physiology and productivity dataset. *Data in Brief*, 62:111959, 2025. doi: 10.1016/j.dib.2025.111959.

Alejandro F. Villaverde, Antonio Barreiro, and Antonis Papachristodoulou. Structural identifiability of dynamic systems biology models. *PLOS Computational Biology*, 12(10):e1005153, 2016. doi: 10.1371/journal.pcbi.1005153.

Kim Peter Wabersich and Melanie N. Zeilinger. A predictive safety filter for learning-based control of constrained nonlinear dynamical systems. *Automatica*, 129:109597, 2021. doi: 10.1016/j.automatica.2021.109597.

Eric Walter and Luc Pronzato. *Identification of Parametric Models from Experimental Data*. Springer-Verlag, Berlin, Heidelberg, 1997. ISBN 9783540761198.

Fan Wang, Zhiyuan Chen, Yuxuan Zhong, Sunjian Zheng, Pengtao Shao, Bo Yu, Shaoshan Liu, Jianan Wang, Ning Ding, Yang Cao, and Yu Kang. Context and diversity matter: The emergence of in-context learning in world models. In *International Conference on Learning Representations*, 2026.

Hang Wang, Xin Ye, Feng Tao, Chenbin Pan, Abhirup Mallik, Burhaneddin Yaman, Liu Ren, and Junshan Zhang. AdaWM: Adaptive world model based planning for autonomous driving. In *International Conference on Learning Representations*, 2025.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1658–1665, 2019. doi: 10.1609/aaai.v33i01.33011658.

Jiahan Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal M. Patel, Paul Pu Liang, Daniel Khashabi, Cheng Peng, Rama Chellappa, Tianmin Shu, Alan Yuille, Yilun Du, and Jieneng Chen. World-in-world: World models in a closed-loop world. In *International Conference on Learning Representations*, 2026.

A Scope and Limitations

The experiments support a narrow revision-governance claim, not a general adaptive-control claim. The 35-unit governance study is a matched, pre-specified scenario set rather than a sample of independent agricultural sites; the one-check-removed study is a counterfactual replay over those fixed records rather than six fresh physical rollouts. The eight KNOW scenarios are a scoped challenge set, not a balanced diagnostic benchmark with sufficient support for TPR/TNR/AUROC claims. The parameter task assumes that the changed quantity is represented in the candidate family. The structural task searches only a finite pre-specified family. The controlled drift transmission study uses one independent-plant scenario and two planners that happen to choose identical actions on all evaluated drift levels. Consequently, the paper does not claim general diagnosis, universal gate necessity, planner-independent transmission, cross-site/season control superiority, positive continual-learning transfer, open-ended mechanism discovery, or real-field efficacy.

These boundaries are scientifically informative because they localize what additional evidence would be required. General diagnosis requires a larger balanced set of identifiable and confounded mechanisms; gate necessity requires targeted fresh rollouts that activate each failure mode; planner-general transmission requires planners with meaningfully different decision surfaces; and control superiority requires independently qualified tasks with non-vacuous headroom across multiple environments.

B Dataset Coverage

Table 3: Real-world irrigation configurations used as the empirical substrate for prediction and controlled revision studies. Availability in the study does not imply redistribution permission.

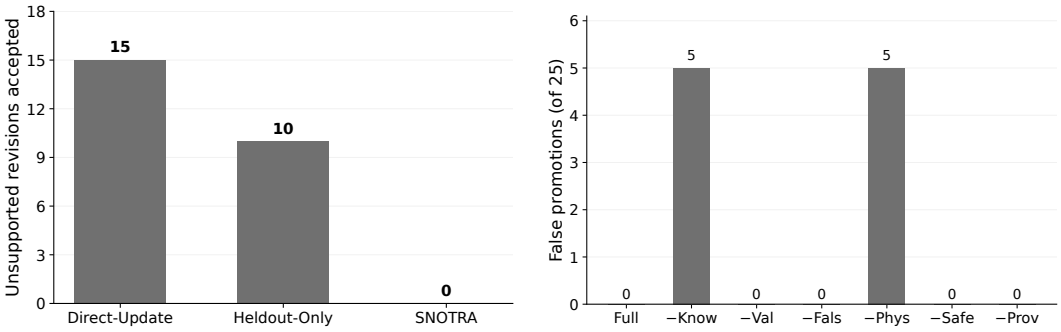
Configuration	Rows	Temporal coverage	Primary data note
Bushland 2016	8,785	2016–2017	soil/temp/humidity missingness present
Bushland multi-year	12,312	2016–2018	overlaps temporally with 2016 subset
Colorado sunflower	30,744	2008–2016	field-average series; no major missing fields
LoCoMoS 2022	456	2022 season	short-season series
Sackett 2022	480	2022 season	short-season series
LIRF 5-year	43,849	2012–2017	soil-water missingness 62.1%
LIRF 9-year	78,912	2008–2017	soil-water missingness 61.1%

Common variables are timestamp, soil-water estimate, precipitation, irrigation, reference evapotranspiration, temperature, humidity, site, and year. The study does not reinterpret repeated years from a site as independent sites, and these series do not by themselves provide parcel-level yield ground truth.

C Detailed Gate Ablation and Capability Tables

Table 4: One-check-removed replay over the 35 matched governance units. This is a counterfactual replay over fixed records, not a fresh plant-level causal rollout.

Admission rule	Unsupported accepted	Supported accepted	Scope
Full Snotra	0/25	10/10	reference
Without identifiability	5/25	10/10	active on these cases
Without held-out validation	0/25	10/10	no marginal effect in replay
Without falsification	0/25	10/10	no marginal effect in replay
Without physical consistency	5/25	10/10	active on these cases
Without safety	0/25	10/10	no marginal effect in replay
Without provenance	0/25	10/10	no marginal effect in replay



(a) Admission beyond held-out validation.

(b) One-check-removed replay.

Figure 3: **Governance diagnostics.** Left: unsupported admissions under Direct-Update, Heldout-Only, and Snotra. Right: only removal of identifiability or physical consistency changes admissions in the fixed replay.

D Scoped Knowability and Revision Results

Table 5: Knowability and revision evidence at distinct scopes. Different denominators and task definitions should not be collapsed into one diagnostic accuracy.

Test	Independent unit	Result	Supported interpretation
KNOW challenge	8 scenarios	8/8 status agreement; firewall 8/8	scoped consistency on pre-specified cases
Parameter repair	9 episodes	MAE 0.260 → 0.078 mm; 9/9 admitted	represented parameter revision
Finite structural selection	4 mechanisms	2/4 pass	preliminary bounded selection only
Governance specificity	25 unsupported cases	0/25 admitted by Snotra	scoped admission selectivity

E Admission Thresholds and Sensitivity

The implementation uses fixed operating thresholds for mass balance, mismatch triggering, rank/conditioning, pairwise separation, and downstream task qualification. They are experimental operating points rather than universal constants. We therefore separate two questions: whether a threshold is part of the pre-specified experimental contract and whether the scientific conclusion is stable around that threshold.

The available sensitivity study evaluates 80 governance-threshold combinations and 36 KNOW combinations over the fields present in the fixed experimental outputs. Within those available axes, the evaluation remains fail-closed rather than rewriting thresholds after observing outcomes. We therefore make no broad threshold-robustness claim beyond the scanned scenario fields.

F Objective-Weight and Planner Diagnostics

The diagnostic physical cost in the main study is

$$J = W + 5S + 10R + 10L,$$

where W is irrigation water, S stress/deficit, R runoff, and L leaching/drainage. Because these coefficients encode a task-specific tradeoff rather than a universal agronomic objective, we evaluate five weight settings on the controlled scenario:

$$(1, 1, 5, 5),\ (1, 5, 10, 10),\ (0.5, 5, 10, 10),\ (2, 10, 20, 20),\ (1, 10, 5, 5).$$

All five settings select the same 20 mm action in that scenario while changing the numerical utility value. This establishes local action invariance to the tested weights, not objective-independent control superiority.

Similarly, the threshold and finite candidate-search planners choose the same action at all six drift levels in the transmission experiment. This removes planner choice as an explanation *between these two planners on this scenario*, but it does not establish a general planner-invariant threshold.

G Supporting Revision and Negative Evidence

Finite structural selection. Four pre-specified structural scenarios are evaluated with held-out prediction and falsification checks; two pass. This is evidence for bounded selection within a declared family only. No experiment in this paper demonstrates generation of a previously unspecified physical mechanism.

Finite candidate search. A separate matched-budget search study produces an aggregate relative change of -1.043%, with 11/26 positive and 13/26 non-regressing records. Because action identifiability is incomplete in that study, it is retained as negative supporting evidence and is not used to claim that search is generally ineffective.

History and continual adaptation. Multiple longitudinal/history studies fail to establish stable positive headroom for retained experience on the evaluated tasks. We therefore do not claim a positive continual-learning result. This does not imply that historical context is generally useless; it means that the prerequisite signal was not demonstrated in the tested settings.

Compute accounting. The scoped analyses record terminal execution status together with wall-time and memory wrappers where available. Complete simulator-call counts and per-arm deployment-cost measurements are unavailable, so we make no deployment-efficiency or latency claim.

H Controlled Transmission Example

The 50% drift condition provides a concrete example of the model-to-decision boundary. Before repair, the prediction MAE gap is 15.349 mm and the planner selects 20 mm of irrigation. Under the repaired model, the prediction gap is removed, both evaluated planners select 10 mm, and the independent physical cost changes from 3999.9 to 3889.9 with no recorded safety violation. At 30% drift, by contrast, the prediction gap is already 9.166 mm but the action remains 20 mm and physical cost is unchanged by the repaired model. The comparison illustrates why predictive relevance and decision relevance must be measured separately.

This is a controlled transmission example, not a complete six-check causal walkthrough: the available result does not record a fresh plant rollout for each individual admission check. That stronger worked example remains outside the claims of this paper.

I Statistical Interpretation

The primary governance uncertainty result clusters by the 35 independent governance units and uses 5,000 bootstrap replicates. The Heldout-Only minus Snotra unsupported-admission-rate difference has mean 0.286 and 95% CI $[0.143, 0.429]$; the supported-revision recall difference is zero. These intervals describe the pre-specified governance scenario set and should not be interpreted as cross-site or cross-season population inference.

Exact paired binary tests are available for selected method contrasts, but the paper emphasizes effect sizes, denominators, and confidence intervals because a small paired-test p-value does not establish that the constructed scenario set is representative of a broader deployment population. Numerical results are therefore rounded to a precision commensurate with their experimental scope.

REPRODUCIBILITY STATEMENT

The main paper specifies the independent unit, denominator, information firewall, matched baseline contract, and uncertainty procedure for every headline result. Appendix B documents the real-world field-data coverage and missingness; Appendices C–I provide exact replay counts, scoped capability denominators, threshold/objective diagnostics, supporting negative evidence, and statistical interpretation. Controlled perturbations are labeled separately from observed field data, and no failed or unqualified experiment is converted into a positive claim. The accompanying anonymous supplementary source is sufficient to rebuild the manuscript figures and tables from the reported experimental outputs.

AI USE STATEMENT

Generative AI tools, including OpenAI ChatGPT, OpenAI Codex, and Google Gemini, were used during both the research process and manuscript preparation.

For **research ideation and execution**, these tools were used to support methodological critique, organize and refine research logic, discuss alternative system and experimental designs, assist with selected software implementation and debugging, and help inspect and interpret experimental outputs. The research questions, methodological choices, experimental protocols, evaluation criteria, and final technical decisions were determined by the authors.

For **retrieval and discovery**, generative AI tools were used to assist in identifying potentially relevant prior work, exploring related research directions, and organizing literature for the related-work discussion. AI-suggested references were not treated as authoritative sources. All references included in the manuscript were independently checked by the authors against their original publications for relevance, accuracy, bibliographic information, and support for the associated claims.

For **drafting and manuscript preparation**, generative AI tools were used to help produce and revise portions of the manuscript from author-defined outlines, technical descriptions, experimental results, and scientific arguments. They were also used to aid and polish writing, including improvements to clarity, readability, organization, grammar, scientific presentation, reference formatting, and figure-related text and presentation. All AI-assisted text was reviewed, revised where necessary, and approved by the authors before inclusion in the manuscript.

Generative AI was not used to fabricate field observations, dataset labels, experimental measurements, numerical results, statistical results, or evaluation outcomes. All empirical results reported in this work were obtained from the source data and research datasets described in the paper through the implemented experimental and analysis pipelines.

The authors independently reviewed and verified the study design, source data, experimental configurations, executable analyses, AI-assisted code, numerical results, statistical summaries, mathematical formulations, technical statements, citations, and final tables and figures against the underlying research materials and original sources. AI-generated suggestions, interpretations, or retrieved information were not treated as scientific evidence without independent verification.

The authors retain responsibility for the research questions, methodology, implementation, experimental design, analyses, interpretation of results, scientific claims, conclusions, citations, and all other content presented in this submission.