Beyond Simulate-Then-Optimize: Geothermal AI for Geothermal Dynamics Prediction, Design, and Discovery

Kunpeng Liu[1], Nori Nakata[2], Jinghan Zhang[1], Guodong Chen[2,3],
Rui Liu[4], Tao Zhe[4], Dongjie Wang[4], Xinyuan Wang[5], Hongyu Cao[5], Yanjie Fu[5,†]

[1]Clemson University, [2]Lawrence Berkeley National Laboratory,
[3]University of California Berkeley, [4]University of Kansas, [5]Arizona State University

ABSTRACT

The central bottleneck in computational geothermal science is not simulator fidelity or data scarcity—it is the abstraction itself. Geothermal energy is increasingly important to the clean energy transition, yet its computational core still follows a legacy simulate-then-optimize paradigm: a deterministic simulator is calibrated to sparse observations and then used to optimize decisions within a fixed model. Hidden inside this pipeline are three commitments—one predicted future, one mostly static operating strategy, and one fitted model per site. We argue that, for next-generation enhanced geothermal systems, the subsurface is partially observed, heterogeneous, and intervention-sensitive, and the information available to characterize it is limited. As a result, forecasting and decision-making must reason over multiple physically plausible futures under uncertainty. Our central claim is that geothermal should be reframed as an adaptive problem of inference, intervention, and discovery. Under this view, simulation becomes conditional generation over plausible reservoir futures rather than point prediction of one trajectory. Operation becomes adaptive decision making over belief states rather than offline scheduling under a presumed known state. Calibration becomes the separation of transferable physical structure from site-specific corrections rather than repeated fitting within a fixed equation class. These are not three independent engineering problems; they are three phases of a single inference cycle. This reframing matters because, in geothermal, uncertainty is not merely something to quantify; it is something operations act upon and reshape. Likewise, persistent model mismatch is not merely an engineering nuisance to suppress; it is the primary scientific signal from which missing or site-modulated physics can be discovered. We therefore organize the paper around three consequences of this reframing: generative world models of reservoir evolution, belief-state policy learning for sustainable operation, and data-to-equation discovery for transferable geophysics. Taken together, these directions define a new agenda for geothermal AI beyond faster surrogate prediction toward adaptive subsurface intelligence where inference, intervention, and discovery are intrinsically coupled.

1. INTRODUCTION

Geothermal energy is emerging as a strategically important pillar of the clean energy transition. Unlike solar and wind, geothermal can provide *firm*, 24/7 carbon-free power with a small land footprint, making it valuable for stabilizing electricity systems dominated by intermittent renewables. Beyond grid decarbonization, geothermal can boost energy security, support industrial growth, enable resilient power for data centers [41], and, in some regions, co-produce critical minerals from subsurface fluids [26].

Recent advances in horizontal drilling, high-rate completions, distributed sensing, and closed-loop field operations are turning geothermal into a more scalable and manufacturable technology stack [21, 39, 54]. Yet the computational logic used to reason about geothermal systems has changed far less. Most development workflows still follow a legacy *simulate-then-optimize* paradigm: a deterministic physics-based simulator is calibrated to sparse observations via history matching and then used to optimize well placement, injection rates, and operating schedules within a fixed model. Hidden inside this pipeline are three commitments: one predicted future, one largely static control strategy, and one fitted model per site. For next-generation enhanced geothermal systems (EGS), these commitments are increasingly untenable. The subsurface does not admit a single inevitable future: sparse observations, heterogeneous rock properties, and tightly coupled thermal–hydraulic–mechanical–chemical processes yield multiple physically plausible reservoir evolutions consistent with the same initial conditions. Operation is not a stationary optimization problem either. Injection and production decisions reshape the reservoir over time, creating path dependence, delayed feedback, and trade-offs among energy yield, reservoir longevity, and safety. Nor is calibration simply parameter fitting inside a universal model class. Geological variability entangles site-specific structure with broader governing physics, limiting transferability across locations [8, 40]. An AI agent has been developed for seamless connection to the knowledge base to digital twins, but subsurface reservoir evolution is currently a missing piece [22, 24].

We argue that geothermal should be reframed not as a forward simulation problem followed by downstream optimization, but as an integrated problem in which subsurface futures are inferred, interventions are adaptively chosen under uncertainty, and persistent mismatch is converted into scientific insight. This shift changes the computational object itself. The goal is no longer to estimate one best future under a fixed model, but to reason over plausible futures, act under partial observability, and update physical understanding as operations unfold. Under this framing, simulation becomes conditional

[†]Corresponding author: yanjie.fu@asu.edu

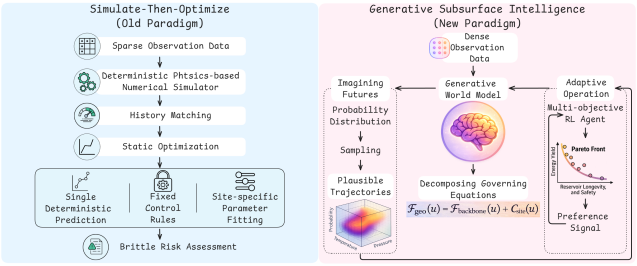

Figure 1: Contrasting geothermal AI paradigms: (A) the traditional simulate-then-optimize pipeline, based on deterministic solvers, history matching, and static control rules; and (B) the generative subsurface intelligence paradigm, which treats the subsurface as a partially observed, uncertain system and integrates generative world models, adaptive decision making, and transferable physics discovery.

generation over reservoir trajectories rather than deterministic point prediction; operation becomes belief-state policy learning rather than static scheduling; and calibration becomes the discovery of transferable structure and site-specific corrections rather than repeated parameter fitting within a fixed equation class.

Figure 1 contrasts these paradigms. The traditional approach compresses uncertainty into a single forecast and then optimizes against that forecast as if the relevant physics were already known. By contrast, the proposed paradigm models reservoir evolution through probabilistic representations induced by partial observability and unresolved heterogeneity in which inference, intervention, and discovery continually update one another. Uncertainty in reservoir evolution shapes operational decisions; operational decisions alter the latent system from which future knowledge must be inferred; and persistent mismatch between predicted and observed behavior can reveal missing or site-modulated physics.

Geothermal is a particularly revealing testbed for this broader challenge because prediction, control, and scientific discovery are inseparable in practice: uncertainty is not merely something to quantify but something operations act upon and reshape. Geothermal AI will therefore not be transformed by improved deterministic surrogates alone. As long as the field remains organized around prediction within a fixed simulator followed by offline optimization, progress will remain local, brittle, and difficult to transfer. The central limitation is not insufficient data, compute, or simulator fidelity. It is that the dominant abstraction no longer matches the geothermal systems we seek to model and operate.

The remainder of this paper develops this argument in four steps. We first show that the dominant simulate-then-optimize paradigm is insufficient. We then propose a closed-loop reframing centered on inference, intervention, and discovery. From this reframing, three research frontiers follow naturally: generative world models of reservoir evolution, belief-state policy learning for adaptive operation, and equation discovery for transferable geothermal physics. We conclude by discussing how evaluation must change if this new framing is taken seriously, and by outlining the broader scientific and deployment implications of that shift.

2. WHY THE CURRENT FRAMING IS INSUFFICIENT

The *simulate-then-optimize* paradigm has long underpinned computational geothermal science. Its limits are now often described as matters of simulator fidelity, data scarcity, or computational expense. We argue that this diagnosis is too shallow. The deeper problem is that the paradigm encodes the hidden assumptions about what geothermal systems are: it presumes one forecastable future, one mostly fixed control logic, and one locally fitted model per site. No amount of incremental model improvement fully resolves a framing mismatch of this kind. To make that claim concrete, we identify three structural mismatches that are not merely engineering bottlenecks, but conceptual failures in how the problem itself is posed.

2.1 Single-Trajectory Prediction vs. Multiple Plausible Futures

The first hidden assumption concerns prediction. At its core lies a single-trajectory worldview: given initial conditions, boundary conditions, and calibrated parameters, the simulator should return one best estimate of the future. That abstraction is reasonable when the subsurface is sufficiently characterized and uncertainty is limited. In such settings, the central scientific question is: *what will happen next?*

However, this is no longer the right question for many EGS settings. Real geothermal reservoirs are not only complex; under sparse observations and unresolved heterogeneity, they admit multiple physically plausible future evolutions. The same observable starting state can correspond to multiple physically plausible futures, as key subsurface properties are uncertain, sparsely measured, and heterogeneous. Permeability varies across fractured rock volumes, fracture connectivity is only partly known, and thermo–hydro–mechanical–chemical (THMC) processes interact nonlinearly over time [48, 51]. Small differences in local conditions can trigger qualitatively different system responses: a slight thermal perturbation may induce phase transitions, causing abrupt pressure changes; mechanical deformation can open or close flow paths, altering transport in ways that are path-dependent and partly irreversible.

A concrete example illustrates the point. At The Geysers in Northern California, the world's largest geothermal complex, injection-induced seismicity remains notoriously difficult to predict. Identical injection protocols applied to neighboring wells can produce qualitatively different seismic responses, because subsurface fracture connectivity and stress conditions differ in ways that no deterministic model can resolve from sparse surface observations [38]. This is not a failure of calibration; it is a failure of the single-trajectory abstraction. More broadly, ensemble simulations with TOUGH2 under different permeability realizations at EGS sites routinely show trajectory divergence exceeding an order of magnitude in predicted flow rates over decadal horizons [7,44], underscoring that the deterministic framing systematically understates subsurface ambiguity [4, 61].

As a result, no single trajectory is sufficient. Even if a deterministic simulator is accurate on average, it maps a structured set of plausible futures into a single forecast. Deterministic AI surrogates inherit the same limitation: they may emulate simulators efficiently, but still return point predictions, whereas the underlying scientific object is a *distribution* over possible futures [30, 60]. This distinction is critical because geothermal decisions are high-stakes and long-horizon. Operators must reason not only about what

one model predicts, but about what *could* happen: which futures are plausible, which are risky, and how uncertainty propagates over decades of operation.

The core limitation, therefore, is not any specific simulator, but the deterministic prediction abstraction itself. When the system admits many physically plausible evolutions, the computational task should be reframed from predicting a single trajectory to describing a conditional family of futures. The cost of retaining the deterministic abstraction is not merely reduced realism; it is a systematic compression of risk, in which structurally different but plausible reservoir futures are collapsed into a single forecast.

2.2 Static Control vs. Adaptive, Multi-Objective Operation

The second hidden assumption concerns control. The dominant paradigm treats operation as a scheduling problem: devise a plan largely offline, perhaps revise it occasionally, and then execute it through fixed schedules, threshold rules, or limited re-optimization. That abstraction is reasonable when reservoir conditions evolve slowly, observations are sufficiently informative, and objectives are narrow and stable. In such settings, the central control question is: *what is the best schedule under the current model?*

However, EGS operation is not a one-time scheduling problem. It is a long-horizon decision problem in which actions reshape the system being controlled. Injection and production decisions alter pressure fields, temperature gradients, fracture behavior, and long-term reservoir sustainability.

Moreover, the objective is inherently multi-dimensional [12]. Operators must balance energy yield, reservoir longevity, pressure stability, and mechanical integrity, often under evolving economic, regulatory, and safety constraints [42]. These trade-offs evolve over time: strategies that maximize short-term heat extraction may accelerate thermal depletion, destabilize pressure, or increase downstream risk over decades [40].

Compounding this challenge, the system is only *partially observed*. Unlike many engineered systems, the geothermal reservoir state is not directly observable. Measurements are sparse, noisy, and indirect, and key variables must be inferred rather than directly measured. As a result, geothermal control is fundamentally a problem of acting under uncertainty about the current subsurface state, not merely optimizing over a known state. Existing strategies such as fixed injection rates or threshold-based adjustment rules are poorly suited to this setting [18, 25]: they do not explicitly reason about uncertainty, they do not adapt as the reservoir evolves, and they often encode implicit short-term objectives at the expense of long-term system health.

The limitation, therefore, is not that current controllers are insufficiently tuned, but the current control is the inappropriate abstraction for a partially observed, evolving, multi-objective system. What is needed is a formulation in which policies adapt continuously to uncertain and changing reservoir conditions, while explicitly trading off competing objectives over long operational horizons. The cost of retaining the static-control abstraction is not merely suboptimal scheduling; it is a failure to recognize that operational decisions are epistemic as well as engineering actions, because they reshape the latent system from which future decisions must be made.

2.3 Site-Specific Calibration vs. Transferable Physics

The third hidden assumption concerns knowledge accumulation. The dominant paradigm treats calibration as a local fitting problem: adjust model parameters until simulated outputs match observations at one site. That approach is reasonable if each reservoir is treated as an isolated engineering project and if sufficient local data are available. In that framing, the central question is: *how can we fit this model to this site?*

However, this framing creates a scalability problem. Each new geothermal site requires costly recalibration, often with limited data, because fitted parameters absorb multiple sources of variation at once [16]. They reflect not only universal physical structure, such as conservation laws and Darcy-type flow, but also local geological idiosyncrasies, such as fracture geometry, stress sensitivity, and site-specific permeability corrections. Thus, calibration entangles what should transfer across sites and what should remain site-specific.

This entanglement is costly both scientifically and operationally. A model calibrated at Utah FORGE may not transfer to sites in Nevada or Texas, even when the underlying physics is largely shared. The issue is not that one site obeys different laws of nature than another, but that the current calibration pipeline lacks a mechanism to separate reusable physical structure from local corrections. Consequently, each site is treated almost as a new problem. Progress becomes site-locked: knowledge accumulates locally with weak transfer across sites.

For geothermal to scale as a science and an industry, calibration must become more than parameter estimation. The computational goal is to decompose governing behavior into two components: a transferable physical backbone that captures shared structure, and site-specific components that can be learned or discovered independently. Without such a separation, AI-for-geothermal remains trapped in a cycle of local fitting rather than cumulative scientific learning. The cost of retaining the site-specific calibration abstraction is not merely repeated labor; it is the inability of the field to accumulate knowledge across projects, because reusable physical structure remains entangled with local corrective fitting.

These three hidden assumptions define the ceiling of the current paradigm. The central bottleneck is not the fidelity of any individual simulator, controller, or calibration routine, but the abstraction that organizes them. Next-generation geothermal systems require distributional reasoning, adaptive decision-making under partial observability, and separable, transferable physics. To move beyond that ceiling, the field needs not just better tools, but a different computational formulation. We turn to that formulation next.

3. OUR PERSPECTIVE: GEOTHERMAL AS AN ADAPTIVE, COUPLED PROBLEM OF INFERENCE, INTERVENTION, AND DISCOVERY

3.1 Core Reframing

The structural mismatches identified above point to a deeper conclusion: geothermal is not best understood as a pipeline of simulation, optimization, and calibration, but as a coupled

and iterative system in which these functions continually inform and update one another. Uncertainty in subsurface evolution shapes operational decisions; operational decisions alter the reservoir state and the information subsequently observed; and persistent mismatch between predicted and observed behavior can reveal missing or site-modulated physics. In this sense, geothermal is fundamentally a problem of inference, intervention, and discovery. Simulation, control, and calibration should therefore be understood not as three independent engineering problems, but as three phases of a single inference cycle.

The value of AI is therefore not simply to accelerate existing modules, but to support a different computational object altogether. Rather than predicting one best future and optimizing against it, the field should reason over families of plausible futures, choose interventions under partial observability, and treat residual mismatch as a source of scientific learning. This reframing leads to three coupled consequences: (1) simulation should be treated as conditional generation over physically plausible reservoir trajectories; (2) operation should be treated as adaptive policy learning over belief states under multiple objectives; and (3) calibration should be treated as the discovery of transferable physical structure plus site-specific corrections, rather than repeated fitting within a fixed equation class.

3.2 Conceptual Mapping: Old Objects, New Roles

Under this perspective, every core object in geothermal computational science acquires a new role:

- **Simulation** is no longer a deterministic forward solve of a numerical PDE system, but *conditional distribution learning* over feasible spatiotemporal trajectories, $p_\theta(\mathbf{u}_{0:T}|\mathbf{c})$, where $\mathbf{u}_{0:T}$ denotes the evolution of temperature, pressure, phase saturation, and fluid velocity, and $\mathbf{c}$ encodes geological conditions and boundary controls.
- **Operation/control** is no longer the application of fixed injection schedules or threshold-based rules, but *adaptive multi-objective policy learning* under latent-state uncertainty, $\pi_\phi(\mathbf{a}_t|\mathbf{b}_t)$, where $\mathbf{b}_t = p(\mathbf{z}_t|o_{0:t}, \mathbf{a}_{0:t-1})$ is a belief distribution over latent reservoir states $\mathbf{z}_t$ inferred from sparse, noisy observations.
- **Calibration/history matching** is no longer parameter fitting within a fixed governing equation, but *automated equation discovery* via a backbone–calibration decomposition, $\mathcal{F}_{\text{geo}}(\mathbf{u}) = \mathcal{F}_{\text{backbone}}(\mathbf{u}) + \mathcal{C}_{\text{site}}(\mathbf{u})$, where the backbone captures universal conservation laws and the site-specific term is discovered through symbolic regression.

These remappings carry immediate corollaries: uncertainty quantification becomes intrinsic to the generative framework rather than a post hoc add-on; site characterization becomes representation learning over multimodal evidence; and cross-site transfer becomes systematic comparison of discovered $\mathcal{C}_{\text{site}}$ terms rather than ad hoc expert judgment (Table 1). The three core mappings are the most consequential because, together, they instantiate the closed loop of inference, intervention, and discovery.

3.2.1 Simulation as conditional distribution learning

The central insight is that high-fidelity geothermal simulation is computationally prohibitive, real-world data are sparse, and subsurface physics admits multiple physically plausible evolutions under the same conditions, because subsurface structure and state are only partially characterized, the governing system admits multiple physically plausible evolutions consistent with the available evidence [7, 13, 48].

Diffusion-based generative models provide a natural solution: they treat generation as a stochastic search over a probability landscape, learning to reverse a noise-corruption process to sample from a conditional trajectory distribution. Such generative models has demonstrated their capabilities to extract subsurface elastic properties [5, 46]. Just as diffusion models in computer vision generate diverse, high-quality images from noise, a geothermal diffusion model generates diverse, physically plausible reservoir trajectories conditioned on geological parameters and boundary controls. Each trajectory represents a distinct hypothesis of subsurface evolution rather than a noisy variant of a single prediction. This reframes uncertainty quantification from a separate analytical step into the generation process itself.

3.2.2 Operation as belief-state policy learning

Geothermal operation is a long-horizon decision process with delayed and often irreversible consequences. Injection decisions alter temperature gradients, fracture permeability, and mechanical stress, with effects that may only manifest years later [40]. Under this framing, we model the system as a partially observable Markov decision process (POMDP), where policies act on *belief states*—probability distributions over latent reservoir conditions—rather than fully observed states. Multi-objective learning makes trade-offs between energy yield, pressure stability, thermal longevity, and mechanical integrity explicit, yielding families of Pareto-efficient policies rather than a single operating strategy. Operators can then select among these policies based on real-time constraints and risk tolerance, shifting from reactive, site-specific heuristics to proactive, generalizable, and uncertainty-aware control.

3.2.3 Calibration as equation discovery

Traditional calibration adjusts parameters within a fixed set of governing equations, conflating universal physical laws (e.g., conservation of energy, Darcy flow) with site-specific geological effects (e.g., fracture geometry and stress-dependent permeability). Under the new framing, we explicitly decompose the governing equations:

$$\mathcal{F}_{\text{geo}}(\mathbf{u}) = \mathcal{F}_{\text{backbone}}(\mathbf{u}) + \mathcal{C}_{\text{site}}(\mathbf{u}), \tag{1}$$

where $\mathcal{F}_{\text{backbone}}$ encodes universal laws shared across sites, and $\mathcal{C}_{\text{site}}$ is a *learnable* site-specific correction discovered via data-driven equation discovery (e.g., symbolic regression). The resulting $\mathcal{C}_{\text{site}}$ terms are interpretable expressions rather than opaque neural weights, enabling comparison across sites, identification of shared mechanisms, and construction of transferable geothermal knowledge. This shift elevates calibration from numerical fitting to scientific discovery.

3.3 Why This Is a Paradigm Shift, Not a Better Tool

This is not merely a stronger surrogate model applied to the same task. It changes the unit of modeling—from a single trajectory to a distribution; the role of control—from static

Dimension	Simulate-then-Optimize	Generative SubsurfaceIntelligence
Problem formulation	Single deterministic prediction	Conditional distribution over trajectory space
Representation	Fixed PDE discretization	Learned latent dynamics with physics structure
Optimization target	Maximize heat extra under fixed mode	Multi-objective policy over belief states
Uncertainty handling	Post hoc sensitivity analysis	Intrinsic: each sample is a hypothesis
Calibration	Parameter fitting in fixed equations	Equation discovery: backbone + site terms
Cross-site transfer	Re-calibrate from scratch	Compare discovered calibration terms
Evaluation criteria	Prediction error at one site	Adaptation, robustness, physical consistency, transfer

Table 1: Comparison between the legacy simulate-then-optimize pipeline and the proposed closed-loop subsurface intelligence framing for geothermal systems.

optimization to belief-state-conditioned policy learning; and the interface between learning and physics—from parameter fitting to equation discovery.

Under the old paradigm, the computational challenge was "solve this PDE faster." Under our perspective, it becomes "learn the distribution of what the subsurface *could do*, reason about what it *should do*, and discover *why* it behaves differently across sites."

This distinction reshapes the research agenda for geothermal AI: not faster solvers for fixed equations, but new formulations, learning architectures, and evaluation criteria.

A concrete litmus test clarifies the stakes: under the old framing, a perfectly calibrated deterministic simulator at one site would be considered a solved problem. Under ours, it would be considered a *failure mode*—because it has absorbed site-specific effects into opaque parameters, foreclosed distributional reasoning, and produced knowledge that cannot transfer to the next site. If this claim is correct, then the community's default measure of success (single-site prediction error) is not merely incomplete but actively misleading, because minimizing it drives the field deeper into the paradigm we argue should be replaced.

3.4 What This Perspective Is Not Claiming

This perspective does not argue that numerical simulators should be discarded, or that geothermal can be treated as a purely data-driven problem. Physics-based simulators remain essential sources of structure, supervision, and validation. Nor do we claim that geological variation can be removed through AI alone. The challenge is not to eliminate site specificity, but to represent uncertainty, intervention dependence, and transferable physical structure more faithfully than the current pipeline allows.

3.5 The Structural Opportunity: Why Now

This reframing is now possible due to the convergence of three developments. *First*, advances in generative AI—particularly diffusion models, score-based generative modeling, and physics-informed neural operators—have demonstrated the ability to learn complex, high-dimensional distributions over spatiotemporal fields, from weather prediction to molecular dynamics [23, 29, 43, 50, 66]. These tools can be adapted to geothermal systems, provided appropriate inductive biases (e.g., physical constraints and coupling-aware architectures) are incorporated. *Second*, the geothermal sensor ecosystem is undergoing rapid transformation. Dense fiber-optic sensing (distributed temperature, acoustic, and strain sensing), microseismic monitoring, and downhole geochemical sampling now generate high-resolution spatiotemporal data streams at sites such as Utah FORGE [54]. Data density is approaching the threshold at which learning-driven approaches become practical rather than aspirational. *Third*, the urgency of clean energy deployment creates both economic pull and policy support. The DOE Enhanced Geothermal Shot, targeting $45/MWh by 2035, demands not incremental improvements in drilling or simulation but a step change in how geothermal systems are computationally understood, operated, and scaled [52]. The transition from oil and gas workforces to geothermal industries further amplifies the need for AI-enabled, transferable operational knowledge. These developments make it both possible and necessary to move from deterministic, site-locked, and static computation to generative, transferable, and adaptive subsurface intelligence.

4. RESEARCH FRONTIERS OPENED BY THIS REFRAMING

The three directions below are not parallel wish lists or loosely related technical opportunities. They are the three necessary research consequences of the coupled reframing above. Once geothermal is treated as a problem of inference, intervention, and discovery, the field must learn to: (1) generate plausible reservoir futures rather than point forecasts; (2) choose actions over belief states rather than fixed schedules; and (3) convert persistent residual mismatch into interpretable physical insight rather than absorb it into opaque calibration. Taken together, they form a closed-loop intelligence architecture in which each component changes the object that the next component must reason about, aligning naturally with the recent shift toward agentic AI systems that couple reasoning, planning, and iterative feedback [17, 32, 35, 36, 56, 57, 71, 75, 76]. Read this section not as a menu of tools, but as a decomposition of that architecture.

4.1 World Models That Generate Reservoir Futures

Once geothermal simulation is reframed as reasoning over plausible futures rather than forecasting a single trajectory, the first technical frontier is the construction of *generative world models* for subsurface systems. Their role is not merely to emulate a simulator faster, but to represent the conditional distribution of what the reservoir *could do* under uncertain geology and chosen controls. Figure 2 illustrates this generative world-modeling perspective, in which sparse site information and control conditions define a conditional distribution over physically plausible reservoir trajectories. This matters because failing to reason over the *distribution* of reservoir trajectories, rather than a single prediction, produces brittle risk assessments, overconfident operational

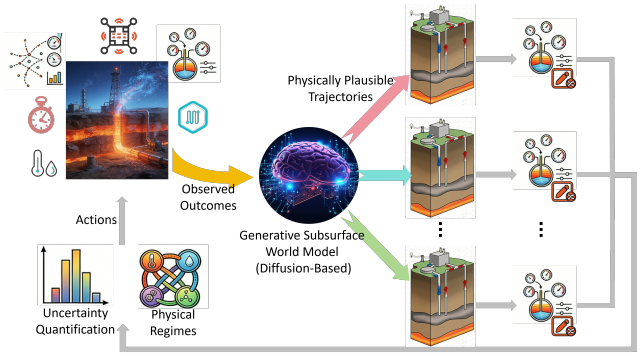

Figure 2: A generative world model for subsurface geothermal systems. Given initial conditions, control parameters, and sparse site data, a diffusion-based model generates 100+ physically plausible trajectories within a fractured subsurface volume, constrained by conservation laws and thermodynamic consistency to enable reliable uncertainty quantification.

planning, and a false sense of certainty about long-horizon reservoir behavior.

The core question is therefore not "how can we speed up simulation?" but: *how can we learn conditional generative models* $p_\theta(\mathbf{u}_{0:T}|\mathbf{c})$ *over high-dimensional, multi-physics geothermal trajectories that are diverse, physically consistent, and computationally useful for downstream decisions?*

Why this frontier is now technically plausible. Diffusion models have shown a remarkable ability to learn complex, high-dimensional distributions [23,50]. Recent work on latent diffusion [3,47] and physics-informed score matching suggests that operating in a learned latent space can significantly reduce computational cost while preserving physical fidelity. Closely related work demonstrates that diffusion-based generation can be coupled with causal stability constraints to yield robust selections under distribution shift [55]. Preliminary work on Brownian Bridge-augmented frameworks for CO_2 storage simulations demonstrates that generative models can produce physically consistent trajectories with higher fidelity than deterministic surrogates [1]. Work on the supply chain also demonstrates the importance of the combination of simulation and generative models [2, 9, 10]. Neural operators (Fourier Neural Operators, DeepONet) provide resolution-independent function mappings and can serve as efficient backbone architectures [29, 37].

Open Questions. Despite recent progress, three core challenges remain unresolved. The first concerns how to construct latent representations that respect the heterogeneous coupling structure of THMC processes, where slow thermal diffusion, fast pressure propagation, and discrete phase transitions interact across scales [11, 65]. Recent advances in RL-guided Transformer feature construction [20] and graph-walk-based feature–variable alignment [19] illustrate complementary strategies for representing such heterogeneous coupling within learned latent spaces. Closely related is the question of how physical constraints—including conservation laws, thermodynamic consistency, and stress limits—can be built directly into generative dynamics, rather than imposed as external corrections. More broadly, it remains unclear how generative models can represent regime-dependent behavior (e.g., liquid, two-phase, and steam systems) without blurring physically distinct modes or collapsing diversity across operating conditions. Analogous challenges in language-model-driven generation have been addressed through diversity-controlled augmentation [59] and structured exploration of under-covered hypothesis regions [68], suggesting transferable strategies for preventing mode collapse in physical trajectory generation. One promising formulation is a regime-conditioned mixture:

$$p_\theta(\mathbf{u}_{0:T}|\mathbf{c}) = \sum_r p_\theta(\mathbf{u}_{0:T}|r,\mathbf{c})\, p_\theta(r|\mathbf{c}), \tag{2}$$

where $r \in \{\text{liquid, two-phase, steam}\}$ indexes thermodynamic regimes and $p_\theta(r|\mathbf{c})$ is a learned regime classifier conditioned on site context. This decomposes the generation problem into regime identification and within-regime trajectory sampling, preventing cross-regime mode collapse while preserving the ability to reason about regime transitions under changing operational conditions.

What would count as a real shift. Success would not simply be faster simulators, but a qualitative shift in how geothermal systems are modeled and used in practice. Instead of producing a single trajectory over days of computation, models would generate diverse, physically consistent futures on demand, each reflecting different plausible evolutions of the reservoir under uncertainty. Engineers could interrogate these trajectory ensembles to assess risk, compare intervention strategies, and reason about system behavior across a range of operating conditions. In this setting, simulation becomes a tool for exploring distributions of outcomes, rather than committing to a single deterministic forecast.

4.2 Belief-State Policy Learning for Sustainable Reservoir Management

Once geothermal operation is reframed as acting under partial observability rather than executing a fixed schedule, the second frontier is the development of *belief-state policies* that adapt to evolving reservoir conditions over decades-long horizons. Their role is not merely to optimize extraction, but to decide what the system *should do* when the state is uncertain, objectives conflict, and today's action changes tomorrow's reservoir. Figure 3 illustrates this shift from static scheduling to belief-state policy learning, emphasizing adaptive action under partial observability, multi-objective trade-offs, and long-horizon reservoir outcomes. This matters because current practice—static injection rules, threshold heuristics, and single-objective optimization—systematically sacrifices long-term reservoir sustainability for short-term energy yield, a trade-off that becomes more costly as EGS deployments scale.

The core question is: *how can we learn families of Pareto-efficient policies* $\pi_\phi(\mathbf{a}_t|\mathbf{b}_t)$ *that reason over belief states, balance energy yield, reservoir longevity, pressure stability, and mechanical safety, and generalize across geological settings?*

Why this frontier is now technically plausible. Multi-objective reinforcement learning (MORL) provides frameworks for learning Pareto-optimal policy families parameterized by preference vectors [15, 31, 33, 34, 58, 62, 63]. Complementary techniques control exploration depth based on belief uncertainty itself, deciding when to deepen, expand, or terminate trajectory exploration [70]. Generative world models from Direction 1 can serve as interactive environments for policy training, enabling efficient evaluation of long-horizon outcomes without costly numerical simulations. Belief-state

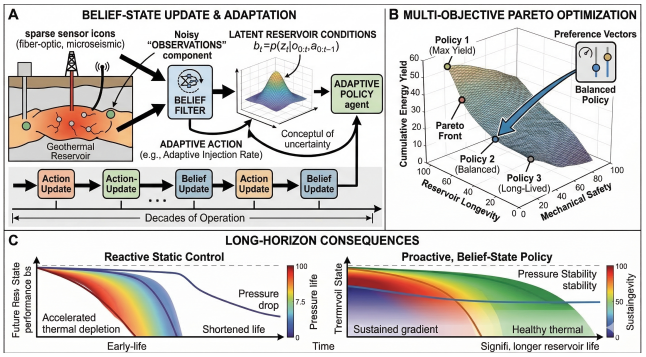

Figure 3: Belief-state policy learning for geothermal operation. (A) A belief filter updates latent reservoir states (b_t) from sparse, noisy observations to produce adaptive actions. (B) Multi-objective optimization identifies Pareto-efficient policies balancing energy yield, reservoir longevity, and mechanical safety. (C) Over long horizons, proactive belief-state policies maintain thermal and pressure stability, achieving longer reservoir life than reactive static control.

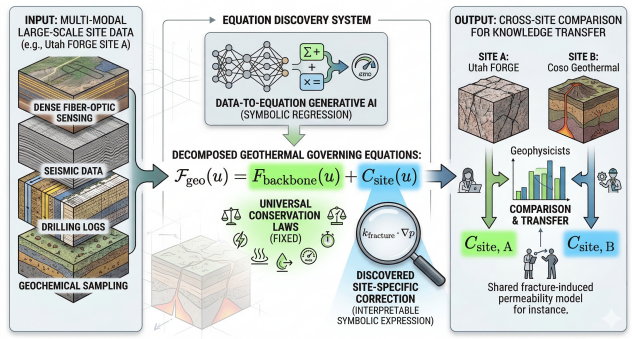

Figure 4: Data-to-equation scientific discovery workflow. Multi-modal site data (e.g., Utah FORGE) are processed by symbolic regression-based generative AI to decompose geothermal governing equations into fixed universal backbone laws ($F_{backbone}$) and site-specific corrections (C_{site}). This interpretability enables cross-site knowledge transfer by allowing geophysicists to compare discovered terms and identify shared geological mechanisms across distinct reservoirs.

methods from the POMDP literature provide principled mechanisms for decision-making under partial observability [28]. In data-sparse regimes, prototype-based reward modeling reduces sample complexity of policy alignment while preserving fidelity to preference signals [69], a property essential when geothermal operational data are too scarce to support fully online reward estimation. Distributionally robust optimization provides tools to ensure policy transfer across sites by optimizing worst-case performance over subsurface distributions [45]. Adaptive weighting strategies allow value-based policies to track non-stationary environments online, a property essential under decadal reservoir drift [75].

Open Questions. Despite recent progress, key challenges emerge when moving from modeling to decision-making. A central difficulty is maintaining coherent belief representations under partial observability, where geothermal measurements are sparse, indirect, and temporally irregular, making state estimation inherently uncertain and history-dependent. At the same time, policies must remain reliable under substantial variation across reservoirs, raising the question of how control strategies can generalize despite differences in permeability structure, fracture geometry, and stress conditions. Finally, geothermal operation unfolds over decades, forcing a tight coupling between learning and control: actions not only extract energy but also shape future system knowledge, making it unclear how to balance information acquisition with long-term productivity. For cross-site transfer specifically, a minimax robust formulation provides a concrete starting point:

$$\max_{\pi} \min_{k \in \mathcal{K}} J^{(k)}(\pi), \tag{3}$$

where $\mathcal{K}$ indexes a family of site-specific world models and $J^{(k)}(\pi)$ is the multi-objective return under model k. This transforms the vague desideratum of "robustness" into a well-defined optimization problem whose solution is a policy that performs acceptably across geological settings, even if it sacrifices peak performance at any single site.

What would count as a real shift. In practice, success would be reflected in a shift from a set of fixed operating rules to adaptive, state-aware decision-making. Operators would no longer rely on predetermined injection schedules, but instead adjust actions in response to evolving reservoir beliefs, with an explicit understanding of the trade-offs between energy production, reservoir longevity, and safety. Rather than committing to a single operating strategy, they could navigate a spectrum of policies, selecting or adapting strategies as new information becomes available. Over time, such policies would not only improve immediate performance but also steer the reservoir toward more stable and sustainable operating regimes.

4.3 From Calibration to Discovery: Data-to-Equation Generative Physics

Once calibration is reframed as the interpretation of structured residuals rather than the repeated fitting of local parameters, the central scientific challenge becomes extracting reusable physical insight from the prediction–reality gap. The third frontier is therefore developing AI systems that can *automatically discover interpretable expressions* for site-specific geothermal physics while preserving a shared physical backbone. Their role is not merely to fit one site better, but to explain *why* a reservoir behaves differently and to turn residual mismatch into transferable scientific knowledge. Figure 4 summarizes this data-to-equation workflow, where structured residuals are used to discover interpretable site-specific corrections on top of a shared physical backbone. This matters because current practice—site-by-site parameter fitting—absorbs missing structure into opaque local parameters and thereby blocks cumulative learning across geothermal projects.

The core question is: *how can we decompose geothermal governing equations into backbone physics $\mathcal{F}_{backbone}$ and site-specific terms $\mathcal{C}_{site}$, discover $\mathcal{C}_{site}$ as interpretable symbolic expressions from data, and use cross-site comparison to extract generalizable geophysical insights?*

Why this frontier is now technically plausible. Symbolic regression has advanced rapidly, with methods from ge-

netic programming to neural-guided search and transformer-based equation generation [6,14,27]. Recent data-to-equation approaches suggest that foundation models can be adapted to low-data symbolic regression settings, while reinforcement feedback can further align equation generation with downstream numerical fitness and domain-specific structure [64,67]. Beyond raw equation search, retrieval- and LLM-augmented feature generation pipelines provide structured priors that constrain the discovery space to interpretable, domain-meaningful terms [73, 74]. The backbone–calibration decomposition $\mathcal{F}_{\text{geo}}(\mathbf{u}) = \mathcal{F}_{\text{backbone}}(\mathbf{u}) + \mathcal{C}_{\text{site}}(\mathbf{u})$ provides a structural prior that constrains search space: the backbone is fixed by conservation laws, and only the residual site-specific term must be discovered. Physics-informed residuals from the generative simulator (Direction 1) and policy-highlighted anomalies from the adaptive controller (Direction 2) guide discovery to physically meaningful regions of equation space.

Open Questions. What distinguishes equation discovery from conventional calibration is not only expressiveness, but the need for interpretability and scientific validity. In practice, current symbolic methods struggle to move beyond simple or weakly coupled forms, raising the question of how complex, multi-term interactions across the THMC state space can be discovered without losing tractability. Even when candidate expressions are found, their status remains ambiguous: fitting observational data is insufficient, yet there is no clear criterion for when a discovered equation should be regarded as physically meaningful rather than incidental. A further complication is that these expressions are inherently site-specific; without a systematic way to relate them across reservoirs, it is unclear how individual discoveries accumulate into broader geological understanding.

A key insight, largely absent from current symbolic regression practice, is that calibration terms are not unique: multiple functional forms may explain the same observations equally well. This means equation discovery should itself be *generative*—learning a distribution over candidate equations $p(\mathcal{C}_{\text{site}}|\text{data})$ rather than returning a single best-fit expression. A distributional treatment would quantify epistemic uncertainty over governing physics, enable model averaging for more robust prediction, and expose structural degeneracies that point to which additional measurements would most effectively disambiguate competing hypotheses.

What would count as a real shift. Success would be evident not in improved predictive accuracy alone, but in how results are used and interpreted. Instead of treating each site as an isolated calibration problem, practitioners would obtain explicit mathematical descriptions of site-specific behavior that can be interrogated, compared, and debated. These expressions would serve as hypotheses about underlying mechanisms, guiding further analysis rather than acting as fixed outputs. Over time, patterns across sites will reveal recurring structures—shared functional forms for stress-dependent permeability, common fracture-flow corrections—enabling domain experts to move from empirical fitting toward a unified, transferable understanding of subsurface physics.

4.4 Why These Three Cannot Be Separated: The Residual-as-Signal Principle

The three directions above are not independent research programs that happen to share a domain. They are successive phases of one feedback-coupled intelligence cycle, linked by a principle we call *residual-as-signal*:

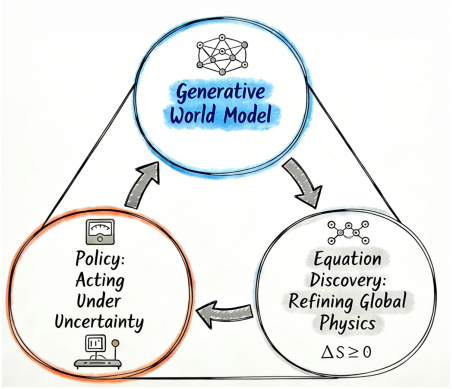

Figure 5: A Coupled and Adaptive Paradigm: from Model to Policy to Equation Discovery.

> *The residual between a world model's predicted distribution and operationally observed outcomes is not merely an error to be minimized—it is the primary scientific signal from which missing site-specific physics can be discovered.*

This is the paper's deepest claim. Residuals are not merely evidence of imperfect fitting; they are structured, decision-dependent traces of what the current representation fails to capture—and they are the only place where scientific discovery enters the computational loop.

This principle creates an irreducible information flow among the three components (Figure 5):

1. **World model → Policy.** The generative world model produces a trajectory distribution $p_\theta(\mathbf{u}_{0:T}|\mathbf{c})$ that serves as the policy's training environment. Without distributional simulation, the policy has no uncertainty-aware sandbox in which to learn.

2. **Policy → Equation discovery.** When the learned policy is deployed, the actions it takes expose a *prediction–reality gap*: systematic discrepancies between the world model's anticipated trajectories and the reservoir's actual response. These residuals are not noise, but structured signatures of physics missing from the world model.

3. **Equation discovery → World model.** Discovered site-specific terms $\mathcal{C}_{\text{site}}$ feed back into the world model, refining its generative distribution. A more accurate world model, in turn, produces better-calibrated training environments for the policy and exposes *subtler* residuals for the next round of discovery.

This closed loop has a concrete consequence: *each component improves the others.* A world model trained in isolation will plateau because it cannot access the operationally induced residuals that reveal missing physics. A policy trained on a static simulator will degrade under distribution shift because it has no mechanism for model refinement. Equation discovery without policy-driven exploration will find only the most obvious corrections, missing the subtle site-specific effects that only emerge under active reservoir management.

The residual-as-signal principle is what distinguishes this agenda from three parallel mini-surveys. It is also the paper's

core differentiator relative to existing geothermal AI reviews: we do not merely propose that AI can help with simulation, control, and calibration separately, but that these three tasks are *informationally coupled* in a way that demands joint treatment.

5. RETHINKING EVALUATION: TOWARD GEOGENBENCH

If the closed-loop perspective is correct, then conventional evaluation protocols for computational geothermal science are not just incomplete; they optimize for the success criterion. Current benchmarks primarily measure *prediction error at a single site*—how closely a simulator or surrogate matches observed data under fixed conditions. But under the new paradigm, prediction accuracy at one site under one condition is necessary and still insufficient. The real question is not whether a model reconstructs one historical trajectory, but whether a coupled system of world modeling, decision making, and discovery produces better uncertainty, better interventions, and more reusable knowledge. We therefore propose **GeoGenBench**, a community benchmark designed to evaluate the full integrated intelligence stack. GeoGenBench is organized around five evaluation dimensions, each with concrete metrics:

- **Distributional fidelity.** Does the generative model produce a diverse, calibrated distribution of trajectories? *Metric:* Continuous Ranked Probability Score (CRPS), which jointly penalizes miscalibration and lack of sharpness, evaluated over held-out THMC trajectories.
- **Physical consistency under sparse data.** Do generated trajectories satisfy conservation laws and coupling constraints even under data scarcity? *Metric:* conservation violation rate (fraction of samples violating energy balance, mass conservation, or thermodynamic monotonicity beyond a tolerance ϵ).
- **Policy robustness and transfer.** Can policies generalize across geological settings? *Metric:* Transferability Score, defined as the ratio of multi-objective return when a policy trained at Site A is deployed at Site B to the return of a policy trained directly at Site B: $\mathrm{TS}(A \to B) = J^{(B)}(\pi_A)/J^{(B)}(\pi_B)$.
- **Multi-objective trade-off quality.** Does the policy produce a well-distributed Pareto front across competing objectives (energy yield, longevity, safety)? *Metric:* Hypervolume indicator, measuring the volume of objective space dominated by the learned Pareto front.
- **Refinement gain from discovery.** Does incorporating discovered $\mathcal{C}_{\text{site}}$ terms improve the world model? *Metric:* Refinement Gain, defined as the reduction in CRPS after augmenting the world model's backbone with discovered site-specific terms: $\mathrm{RG} = 1 - \mathrm{CRPS}_{\text{refined}}/\mathrm{CRPS}_{\text{baseline}}$.

Data and protocol. We envision GeoGenBench built on data from Utah FORGE (the DOE's flagship EGS research site) and The Geysers (the world's largest operating geothermal complex), providing complementary geological settings for transfer evaluation. The evaluation protocol follows the iterative inference–decision–discovery cycle: train a generative world model at Site A → learn a belief-state policy → transfer to Site B → measure transferability score → discover $\mathcal{C}_{\text{site}}$ from the prediction–reality residual → refine the world model → measure refinement gain. This protocol evaluates not only individual components, but their *joint* performance as an integrated system.

Concretely, benchmark design should move beyond single-site reconstruction. One class of tests should evaluate whether a model trained in several fields can adapt to a new site with limited calibration while preserving the quality of the uncertainty. A second class should evaluate whether belief-state policies remain robust under hidden shifts in permeability structure, fracture connectivity, or sensing sparsity. A third class should test whether discovered correction terms remain stable across resampling, data subsets, and nearby sites, indicating that they capture repeatable mechanisms rather than incidental fits.

A named, concrete benchmark with standardized metrics and shared datasets would give the community a common target, accelerating progress across all three research directions.

Trustworthiness in high-stakes deployment. A natural concern about a generative reframing of geothermal AI is whether such models can be made trustworthy enough for safety-critical operational use, where induced seismicity, pressure excursions, and thermal short-circuiting are partially irreversible. Trustworthiness here does not arise from any single safeguard but from four commitments intrinsic to the closed loop. *First*, physical constraints—conservation laws, thermodynamic monotonicity, and stress envelopes—should be built into generative dynamics rather than imposed as post-hoc filters [3], so that every sampled trajectory is feasible by construction. *Second*, the value of a trajectory ensemble lies in calibration rather than diversity: a model that produces many visibly different futures is not trustworthy unless its predicted distribution achieves nominal coverage of held-out outcomes (CRPS, reliability curves), with conformal prediction [49] offering one route to finite-sample coverage guarantees. *Third*, belief-state policies must operate under safety envelopes—e.g., CVaR bounds on induced seismicity and pressure excursions [72], combined with the minimax formulation in Eq. (3)—and route irreversible actions through human-in-the-loop oversight; the Geysers example in Section 2.1 is precisely the regime where bounding the tail matters more than optimizing the mean. *Fourth*, the residual-as-signal principle makes the system auditable: persistent mismatch is surfaced for equation discovery rather than absorbed into opaque weights, and GeoGenBench is designed to falsify the framing itself if it fails to deliver better calibration, transfer, and reusable insight than deterministic baselines. In high-stakes domains, falsifiability is itself a trustworthiness property.

A falsifiability commitment. GeoGenBench is designed not only to measure progress but to test whether the closed-loop framing itself is correct. If systems that win on GeoGenBench do not also yield better transfer, better uncertainty, and more reusable physical insight than systems optimized under the old single-site paradigm, then the perspective advanced in this paper should be revised.

6. BROADER IMPLICATIONS

This shift is not merely of academic interest; it changes what geothermal computation is expected to deliver. The implications are scientific, engineering, and economic, but

all follow from the same claim: once prediction, intervention, and discovery are treated as a coupled process, geothermal systems should no longer be evaluated as static modeling exercises but as adaptive knowledge systems.

Scientific implications. Because persistent residual becomes a signal rather than noise, a new mode of scientific exploration opens in geophysics. Rather than hypothesizing governing equations *a priori* and fitting parameters, we can *discover* site-specific physics directly from data and compare these discoveries across geological settings. This data-to-equation approach can reveal previously unknown coupling mechanisms, identify geological conditions under which standard models systematically fail, and accelerate the development of next-generation geothermal physics. More broadly, it establishes a model paradigm for AI-driven scientific discovery in other subsurface domains, including carbon storage, groundwater management, and mineral extraction.

Engineering and system implications. Because decision support becomes feedback-coupled rather than scenario-based, generative world models and belief-state policies enable a new class of *geothermal digital twins*. These are not static simulations of one scenario, but uncertainty-aware environments in which operators can compare plausible futures, inspect trade-offs, and update decisions as observations arrive. The architecture of geothermal decision support shifts from a sequential workflow—run simulations, inspect outputs, manually adjust—to a closed-loop process in which modeling, intervention, and refinement continuously inform one another.

Economic and deployment implications. Because transferable physical structure reduces startup cost, the most immediate economic impact is faster deployment of new geothermal projects. If backbone models, reusable uncertainty representations, and partially transferable control principles carry knowledge from one site to the next, new deployments may require less bespoke calibration and shorter model-development cycles before operational analysis can begin. The importance of this shift is not that it removes site-specific work, but that it changes how much prior knowledge can be carried from one deployment to the next. This matters for scaling geothermal fast enough to meet ambitious deployment targets such as the DOE's Enhanced Geothermal Shot and longer-horizon U.S. geothermal expansion goals [52, 53].

7. CONCLUDING REMARKS

The next phase of progress in computational geothermal science will likely not come from faster surrogates alone. It will come from changing the computational framing of the problem. We have argued that geothermal is not best understood as deterministic simulation followed by downstream optimization, but as a closed-loop problem in which plausible subsurface futures must be inferred, interventions must be chosen under partial observability, and persistent mismatch must be turned into improved physical understanding. Under this view, simulation, control, and calibration are no longer separate modules. They become coupled parts of one computational loop. This does not make physics-based simulation obsolete, nor does it imply that geothermal can be reduced to generic machine learning. Rather, it suggests that the next generation of geothermal AI should be built around hybrid systems that generate plausible futures, adapt decisions as information evolves, and accumulate transferable scientific knowledge across sites. The residual-as-signal principle is the conceptual center of this agenda. It says that the most informative geothermal errors are not merely discrepancies to be minimized after the fact; they are structured traces of what the current representation fails to capture. If that principle is correct, then the real opportunity is not simply to solve familiar tasks more efficiently. It is to redesign the loop by which geothermal systems are modeled, operated, and scientifically understood. This claim is not specific to geothermal energy. Carbon storage, groundwater management, and nuclear waste disposal share the same defining features: partial observability, multi-physics dynamics, sparse data, and the entanglement of universal laws with site-specific geology. If this adaptive and integrated reframing succeeds in geothermal, it can provide a template for AI-driven scientific discovery across subsurface systems. Recent advances in generative modeling, scientific machine learning, and sequential decision making make this shift newly plausible. The most important step is revision of the abstraction itself. If that revision succeeds, geothermal can become a model domain for AI-driven scientific discovery in partially observed, intervention-sensitive physical systems.

8. ACKNOWLEDGEMENTS

Kunpeng Liu is supported by NSF 2550105, NSF 2550106, and NSF 2242812. Nori Nakata and Guodong Chen are supported by the U.S.Department of Energy under Award Number DE-AC02-5CH11231.

9. REFERENCES

[1] H. Bai, G. Chen, W. Ying, X. Wang, N. Gong, S. Dong, G. Pedrielli, H. Wang, H. Chen, and Y. Fu. Brownian bridge augmented surrogate simulation and injection planning for geological co2 storage. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 14459–14466, 2026.

[2] H. Bai, H. Wang, N. Gong, X. Wang, W. Ying, H. Chen, and Y. Fu. Supply chain optimization via generative simulation and iterative decision policies. In *2025 Winter Simulation Conference (WSC)*, pages 558–569. IEEE, 2025.

[3] J.-H. Bastek, W. Sun, and D. Kochmann. Physics-informed diffusion models. In *International Conference on Learning Representations*, volume 2025, pages 3360–3385, 2025.

[4] Z. Bi and N. Nakata. Learning injection–seismicity coupling for probabilistic multi-horizon forecasting in geothermal systems. 2026.

[5] Z. Bi, N. Nakata, R. Nakata, P. Ren, X. Wu, and M. W. Mahoney. Advancing data-driven broadband seismic wavefield simulation with multiconditional diffusion model. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–9, 2025.

[6] L. Biggio, T. Bendinelli, A. Neitz, A. Lucchi, and G. Parascandolo. Neural symbolic regression that scales. In *International Conference on Machine Learning (ICML)*, 2021.

[7] E. Bjarkason, A. Yeh, J. O'Sullivan, A. Croucher, and M. O'Sullivan. Non-uniqueness of geothermal natural-state simulations. 11 2019.

[8] D. Blackwell, M. Richards, Z. Frone, J. Batir, A. Ruzo, R. Dingwall, and M. Williams. Temperature-at-depth maps for the conterminous us and geothermal resource estimates. Southern Methodist University Geothermal Laboratory, Dallas, TX (United States), 10 2011.

[9] H. Cao, H. Bai, and Y. Fu. Structured memory and role-aware decision making for supply chain transportation. In *2025 IEEE International Conference on Big Data (BigData)*, pages 1837–1846. IEEE, 2025.

[10] H. Cao, J. Zhang, K. Liu, D. Wang, F. Xia, H. Chen, X. Hu, and Y. Fu. Sim2act: Robust simulation-to-decision learning via adversarial calibration and group-relative perturbation. *arXiv preprint arXiv:2603.09053*, 2026.

[11] D. K. Chandra, P. Wang, J. Leopold, and Y. Fu. Collective representation learning on spatiotemporal heterogeneous information networks. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 319–328, 2019.

[12] G. Chen, J. J. Jiao, Q. Liu, Z. Wang, and Y. Jin. Machine-learning-accelerated multi-objective design of fractured geothermal systems. *Nexus*, 1(4), 2024.

[13] Y. Chen, D. Voskov, and A. Daniilidis. Open-source simulation study for direct use geothermal systems. 02 2024.

[14] M. Cranmer. Interpretable machine learning for science with pysr and symbolicregression.jl. 2023.

[15] W. Fan, K. Liu, H. Liu, H. Zhu, H. Xiong, and Y. Fu. Feature and instance joint selection: A reinforcement learning perspective. *arXiv preprint arXiv:2205.07867*, 2022.

[16] S. Finsterle. Practical notes on local data-worth analysis. *Water Resources Research*, 51(12):9904–9924, 2015.

[17] Y. Fu, D. Wang, W. Ying, X. Wang, X. Zhang, H. Liu, and J. Pei. Autonomous data agents: A new opportunity for smart data. *arXiv preprint arXiv:2509.18710*, 2025.

[18] R. Gambini, D. W. Waters, F. Sansone, and V. Memmo. Risk and uncertainty in geothermal projects: Characteristics, challenges and application of the novel reverse enthalpy methodology. *Energies*, 18(15), 2025.

[19] W. Gao, J. Gao, Q. Han, H. Pan, and K. Liu. Graph random walk with feature-label space alignment: A multi-label feature selection method. *arXiv preprint arXiv:2505.23228*, 2025.

[20] W. Gao, Z. Man, Z. He, Y. Tang, J. Gao, and K. Liu. Two-stage feature generation with transformer and reinforcement learning. *arXiv preprint arXiv:2505.21978*, 2025.

[21] G. Grubac, W. El-Rabaa, A. Bonneville, I. Ben-Fayed, R. A. Gonzalez, G. Gullickson, and O. Perez. Implementation of the world's first greater than 300 c propped egs reservoir.

[22] R. Harsuko, Z. Bi, and N. Nakata. Gaia: Geothermal analytics and intelligent agent. *arXiv preprint arXiv:2511.03852*, 2025.

[23] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[24] R. Horne, A. Genter, M. McClure, W. Ellsworth, J. Norbeck, and E. Schill. Enhanced geothermal systems for clean firm energy generation. *Nature reviews clean technology*, 1(2):148–160, 2025.

[25] H. Hoteit, X. He, B. Yan, and V. Vahrenkamp. Uncertainty quantification and optimization method applied to time-continuous geothermal energy extraction. *Geothermics*, 110:102675, 2023.

[26] J. Jello and T. Baser. Utilization of existing hydrocarbon wells for geothermal system development: A review. *Applied Energy*, 348:121456, 2023.

[27] P.-A. Kamienny, S. d'Ascoli, G. Lample, and F. Charton. End-to-end symbolic regression with transformers. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

[28] H. Kurniawati. Partially observable markov decision processes (pomdps) and robotics, 2021.

[29] Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhatt, A. Stuart, and A. Anandkumar. Fourier neural operator for parametric partial differential equations. *International Conference on Learning Representations (ICLR)*, 2021.

[30] Z. Liang, J. Yu, R. Thibaut, F. Zheng, C. Hoiland, and A. Pollack. Surrogate modeling for geothermal systems: Accelerating optimization, history matching, and uncertainty quantification.

[31] C. Liu, X. Xu, and D. Hu. Multiobjective reinforcement learning: A comprehensive overview. volume 45, pages 385–398, 2015.

[32] R. Liu, S. J. Quan, Z.-R. Peng, Z. Yao, H. Wang, Z. Chen, K. Liu, Y. Fu, and D. Wang. City editing: Hierarchical agentic execution for dependency-aware urban geospatial modification, 2026.

[33] R. Liu, R. Xie, Z. Yao, Y. Fu, and D. Wang. Continuous optimization for feature selection with permutation-invariant embedding and policy-guided search. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1857–1866, 2025.

[34] R. Liu, T. Zhe, Y. Fu, F. Xia, T. Senator, and D. Wang. Permutation-invariant representation learning for robust and privacy-preserving feature selection, 2026.

[35] R. Liu, T. Zhe, Z.-R. Peng, N. Catbas, X. Ye, D. Wang, and Y. Fu. Urban planning in the age of agentic ai: Emerging paradigms and prospects. *ACM SIGKDD Explorations Newsletter*, 27(2):35–42, 2026.

[36] R. Liu, T. Zhe, D. Wang, Z. Yao, K. Liu, Y. Fu, H. Liu, and J. Pei. Agentos: From application silos to a natural language-driven data ecosystem, 2026.

[37] L. Lu, P. Jin, G. Pang, Z. Zhang, and G. E. Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021.

[38] E. L. Majer and J. E. Peterson. The impact of injection on seismicity at the geysers, california geothermal field. *International Journal of Rock Mechanics and Mining Sciences*, 44(8):1079–1090, 2007.

[39] N. Nakata, H. Chang, S. Wu, Z. Bi, L. Chen, F. Soom, H. Gao, A. Titov, and S. Dadi. Fracture characterization and stress communication revealed by induced seismicity at the cape modern geothermal field. *Utah, GRC Transactions*, 49(2025):1191–1202, 2025.

[40] M. I. of Technology. *The Future of Geothermal Energy: Impact of Enhanced Geothermal Systems (EGS) on the United States in the 21st Century : an Assessment*. The Future of Geothermal Energy: Impact of Enhanced Geothermal Systems (EGS) on the United States in the 21st Century : an Assessment by an MIT-led Interdisciplinary Panel. Massachusetts Institute of Technology, 2006.

[41] H. Oh, K. Beckers, and K. McCabe. Geospatial characterization of low-temperature heating and cooling demand in residential, commercial, manufacturing, agricultural, and data center sectors for potential geothermal applications in the united states. *Renewable and Sustainable Energy Reviews*, 206:114875, 2024.

[42] P. Olasolo, M. Juárez, M. Morales, S. D´Amico, and I. Liarte. Enhanced geothermal systems (egs): A review. *Renewable and Sustainable Energy Reviews*, 56:133–144, 2016.

[43] J. Pathak, S. Subramanian, P. Harrington, S. Raja, A. Chattopadhyay, M. Mardani, T. Kurth, D. Hall, Z. Li, K. Azizzadenesheli, P. Hassanzadeh, K. Kashinath, and A. Anandkumar. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators, 2022.

[44] K. Pruess, C. Oldenburg, and G. Moridis. TOUGH2: A general-purpose numerical simulator for multiphase fluid and heat flow. *Lawrence Berkeley National Laboratory Report*, 1999. LBNL-43134.

[45] H. Rahimian and S. Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

[46] P. Ren, R. Nakata, M. Lacour, I. Naiman, N. Nakata, J. Song, Z. Bi, O. A. Malik, D. Morozov, O. Azencot, et al. Learning earthquake ground motions via conditional generative modeling. *Nature Communications*, 2026.

[47] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

[48] J. Rutqvist and O. Stephansson. The role of hydromechanical coupling in fractured rock engineering. *Hydrogeology Journal*, 11(1):7–40, 2003.

[49] G. Shafer and V. Vovk. A tutorial on conformal prediction. *Journal of machine learning research*, 9(3), 2008.

[50] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[51] J. Taron, D. Elsworth, and K.-B. Min. Numerical simulation of thermal-hydrologic-mechanical-chemical processes in deformable, fractured porous media. *International Journal of Rock Mechanics and Mining Sciences*, 46(5):842–854, 2009.

[52] U.S. Department of Energy. Doe launches new energy earthshot to slash the cost of geothermal power, September 2022. Enhanced Geothermal Shot aims to reduce EGS cost by 90% to $45/MWh by 2035.

[53] U.S. Department of Energy. Geovision: Harnessing the heat beneath our feet. `https://www.energy.gov/eere/geothermal/geovision`, 2024. U.S. geothermal market analysis and projections.

[54] Utah FORGE. Utah FORGE: Frontier observatory for research in geothermal energy, n.d. Accessed March 22, 2026.

[55] A. Vignesh Malarkkan, X. Wang, K. Liu, D. Zhang, and Y. Fu. Causally-guided diffusion for stable feature selection. *arXiv e-prints*, pages arXiv–2603, 2026.

[56] X. Wang, H. Cao, K. Liu, and Y. Fu. Dataforge: Agentic platform for autonomous data engineering. *arXiv preprint arXiv:2511.06185*, 2025.

[57] X. Wang, D. Wang, W. Ying, H. Bai, N. Gong, S. Dong, K. Liu, and Y. Fu. Efficient post-training refinement of latent reasoning in large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 33692–33700, 2026.

[58] X. Wang, D. Wang, W. Ying, R. Xie, H. Chen, and Y. Fu. Knockoff-guided feature selection via a single pre-trained reinforced agent. *IEEE Transactions on Big Data*, 12(2):625–642, 2026.

[59] Z. Wang, J. Zhang, X. Zhang, K. Liu, P. Wang, and Y. Zhou. Diversity-oriented data augmentation with large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22265–22283, 2025.

[60] G. Wen, Z. Li, K. Azizzadenesheli, A. Anandkumar, and S. M. Benson. U-FNO: An enhanced Fourier neural operator-based deep-learning model for multiphase flow. *Advances in Water Resources*, 163:104180, 2022.

[61] M. C. White and N. Nakata. Induced seismicity and geothermal energy production in the salton sea geothermal field, california. *Scientific Reports*, 15(1):1638, 2025.

[62] M. Xiao, D. Wang, M. Wu, K. Liu, H. Xiong, Y. Zhou, and Y. Fu. Traceable group-wise self-optimizing feature transformation learning: A dual optimization perspective. *ACM Transactions on Knowledge Discovery from Data*, 18(4):1–22, 2024.

[63] J. Xu, Y. Tian, P. Ma, D. Rus, S. Sueda, and W. Matusik. Prediction-guided multi-objective reinforcement learning for continuous robot control. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[64] W. Ying, H. Bai, N. Gong, X. Wang, S. Dong, H. Chen, and Y. Fu. Bridging the domain gap in equation distillation with reinforcement feedback. *arXiv preprint arXiv:2505.15572*, 2025.

[65] W. Ying, H. Bai, K. Liu, and Y. Fu. Topology-aware reinforcement feature space reconstruction for graph data. *ACM Transactions on Knowledge Discovery from Data*, 20(1):1–22, 2025.

[66] W. Ying, C. Wei, N. Gong, X. Wang, H. Bai, A. V. Malarkkan, S. Dong, D. Wang, D. Zhang, and Y. Fu. A survey on data-centric ai: Tabular learning from reinforcement learning and generative ai perspective, 2025.

[67] W. Ying, J. Zhang, H. Bai, N. Gong, X. Wang, K. Liu, C. K. Reddy, and Y. Fu. Data-efficient symbolic regression via foundation model distillation. *arXiv preprint arXiv:2508.19487*, 2025.

[68] J. Zhang, F. Mo, T. C. Weerasooriya, X. Ye, D. Wang, Y. Fu, and K. Liu. Blind spot navigation in large language model reasoning with thought space explorer. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 3691–3707, 2026.

[69] J. Zhang, X. Wang, Y. Jin, C. Chen, X. Zhang, and K. Liu. Prototypical reward network for data-efficient rlhf. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13871–13884, 2024.

[70] J. Zhang, X. Wang, F. Mo, Y. Zhou, W. Gao, and K. Liu. Entropy-based exploration conduction for multi-step reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3895–3906, 2025.

[71] J. Zhang, X. Wang, W. Ren, L. Jiang, D. Wang, and K. Liu. Ratt: A thought structure for coherent and correct llm reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26733–26741, 2025.

[72] J. Zhang, H. Xie, X. Zhang, and K. Liu. Enhancing risk assessment in transformers with loss-at-risk functions. In *2024 IEEE International Conference on Knowledge Graph (ICKG)*, pages 477–484. IEEE, 2024.

[73] X. Zhang, J. Zhang, F. Mo, D. K. Chandra, Y.-Z. Chen, F. Xie, and K. Liu. Retrieval-augmented feature generation for domain-specific classification. In *2025 IEEE International Conference on Data Mining (ICDM)*, pages 943–952. IEEE, 2025.

[74] X. Zhang, J. Zhang, F. Mo, D. Wang, Y. Fu, and K. Liu. Leka: Llm-enhanced knowledge augmentation. *arXiv preprint arXiv:2501.17802*, 2025.

[75] X. Zhang, J. Zhang, B. Rekabdar, Y. Zhou, P. Wang, and K. Liu. Dynamic and adaptive feature generation with llm. *arXiv preprint arXiv:2406.03505*, 2024.

[76] T. Zhe, R. Liu, F. Memar, X. Luo, W. Fan, X. Ye, Z. Peng, and D. Wang. Constraint-aware route recommendation from natural language via hierarchical llm agents, 2025.