RIFT: Reliability of LLM and Physics Forecasters across Time-Horizons in Coupled PDE Systems

Kuntal Thakur [1] **Ayan Banerjee** [1] **Christian Stoddard** [2] **Majid M. Sadeghi** [2] **Sandeep Gupta** [1]

Abstract

Transformer-based time series models achieve strong short-term forecasts, but their reliability over multi-year horizons in PDE-governed physical systems is untested. We introduce RIFT, a systematic comparison of learning-based (Time-LLM with GPT-2 and BERT backbones), physics-based (MODFLOW 6 + GWT, ODE), and hybrid forecasters for subsurface thermal and discharge prediction, spanning 1 month to 10 years across a proprietary mine dataset (1-year ground truth) and USGS Kennecott (3-year). We find a horizon-dependent crossover: Time-LLM dominates short horizons, BERT degrades 4× by 5 months while GPT-2 holds through 1 year, and physics solvers dominate beyond. Hybrid residual learning fails on mine data and only marginally helps on USGS. We give a closed-form predictor for the crossover, $h^* \sim \min(L_{\text{lookback}}, \tau_{\text{phys}})$, recovering the empirical ≈1-year transition on both systems.

1. Introduction

Large language models repurposed for time series forecasting have achieved remarkable short-term accuracy across domains from finance to weather (Zhou et al., 2021; Nie et al., 2023; Das et al., 2024; Jin et al., 2024). By treating temporal sequences as token streams, these models leverage pretrained representations to capture complex patterns without domain-specific engineering. But a fundamental question remains open: **can LLM-based forecasters maintain their advantage over multi-year horizons in physical systems governed by coupled partial differential equations?**

We investigate this question in the context of subsurface thermal forecasting for deep mining operations, where accurate 10-year temperature projections drive large-capital cooling infrastructure planning (McPherson, 1993). Subsurface thermal evolution couples groundwater flow (governed by Darcy's law) with heat transport (advection-diffusion), creating a system where local pumping operations affect temperature distributions kilometers away over multi-year timescales. Traditional machine learning approaches treat this as a univariate time series problem. Physics-based models instead solve coupled PDEs:

$$\frac{\partial}{\partial t}(S_s h) = \nabla \cdot (K \nabla h) - Q_{\text{pump}} \tag{1}$$

$$\frac{\partial}{\partial t}\left[(\phi \rho_w c_w + (1-\phi)\rho_s c_s)\, T\right] = \nabla \cdot (\lambda_{\text{eff}} \nabla T) - \rho_w c_w \nabla \cdot (\mathbf{q} T) \tag{2}$$

where Equation 1 governs hydraulic head evolution and Equation 2 couples fluid flow ($\mathbf{q}$) to thermal transport. Temperature changes depend not just on past values, but on spatial flow patterns, geological structure, and geothermal gradients, information absent from univariate discharge data.

This work presents the first systematic comparison of Time-LLM (Jin et al., 2024) with GPT-2 and BERT backbones against physics-based PDE solvers (MODFLOW 6 + GWT) for subsurface thermal forecasting, spanning horizons from 1 month to 10 years across two datasets. We ask two questions: (1) at what forecast horizon do physics-based models overtake learning-based approaches for coupled subsurface systems? and (2) can combining physics solvers with LLM residual learning outperform either approach alone?

Beyond the specific application, this study contributes to the evaluation of foundation-model forecasters by identifying when pattern-extrapolating models cease to be reliable in physical-world settings, and by giving a closed-form predictor for that transition.

Contributions.

- We give a closed-form predictor for the LLM-to-physics crossover, $h^* \sim \min(L_{\text{lookback}}, \tau_{\text{phys}})$, derived from the lookback window and the dominant physical timescale, and verify it on both thermal diffusion ($\tau_{\text{th}} \sim L^2/(4\pi\alpha)$) and aquifer recession ($\tau \approx 1000$ days) systems.

[1]IMPACT Lab, Arizona State University, Tempe, Arizona, USA [2]Resolution Copper, USA. Correspondence to: Kuntal Thakur <kthakur9@asu.edu>.

Accepted at the ICML 2026 Workshop on Forecasting as a New Frontier of Intelligence, Seoul, South Korea. Copyright 2026 by the author(s).

- GPT-2 stays stable through 1 year while BERT degrades 4× at 5 months, isolating autoregressive structure as the long-horizon factor.
- MODFLOW+LLM residual learning underperforms MODFLOW alone at every horizon, showing well-calibrated physics leaves no learnable residual.
- The crossover replicates on public USGS Kennecott data across a different conservation law (mass vs. energy) and solver class (lumped ODE vs. 3D PDE).

2. Related Work

Our work bridges transformer-based time series forecasting (Zhou et al., 2021; Nie et al., 2023; Jin et al., 2024), physics-informed ML (Raissi et al., 2019; Jia et al., 2021; Willard et al., 2022), and subsurface thermal modeling (Harbaugh, 2005; Langevin et al., 2022). Foundation-model forecasters (Jin et al., 2024; Ansari et al., 2024; Das et al., 2024) are benchmarked almost exclusively on sub-year, repeating-pattern domains, and Tan et al. (2024) show pretrained LLM weights are not what drives their accuracy. Physics-informed ML typically treats the neural net as a surrogate for a PDE solver; RIFT inverts the question and asks when the solver must be kept, and against foundation forecasters in particular. ML for subsurface systems remains thin (Sun et al., 2020; Tao et al., 2022). Appendix F gives the full positioning table.

3. Method

We compare three distinct forecasting paradigms: (1) Learning-only approaches using transformer-based language models, (2) Physics-only methods using numerical PDE solvers, and (3) Hybrid approaches that combine physics baselines with learned residual corrections. This framing connects to a broader challenge in deploying AI for safety-critical physical systems. Just as LLM-based planners for autonomous medical devices require explicit physics verification to ensure safe operation (Banerjee & Gupta, 2025), forecasting coupled physical processes in mining operations demands that predictions respect governing conservation laws. When prediction failure translates to large-capital misdesigned infrastructure, understanding exactly when data-driven models can be trusted becomes essential.

3.1. Problem Formulation

Given a time series of mine water discharge temperatures $\mathbf{T} = \{T_1, T_2, \ldots, T_n\}$ observed at timestamps $\{t_1, t_2, \ldots, t_n\}$, our objective is to predict future temperatures $\{\hat{T}_{n+1}, \ldots, \hat{T}_{n+H}\}$ over a forecast horizon H ranging from 1 month to 10 years.

For learning-based approaches, the model learns a mapping $f_\theta : \mathbf{T}_{1:n} \rightarrow \hat{\mathbf{T}}_{n+1:n+H}$ parameterized by θ. For physics-based approaches, predictions emerge from solving the coupled PDE system (Equations 1-2) given geological parameters and boundary conditions. This distinction is fundamental: learning-based methods treat the problem as statistical pattern extrapolation, while physics-based methods encode conservation of mass and energy as hard constraints. We preserve this realistic deployment asymmetry: physics models receive geological parameters and boundary conditions unavailable to the univariate LLM.

3.2. Learning-Only Approaches

3.2.1. Time-LLM Architecture

Time-LLM (Jin et al., 2024) reprograms frozen large language models for time series forecasting by learning an input transformation that maps temporal patches to the LLM's token embedding space via cross-attention with text prototype embeddings. We evaluate two backbone variants: **Time-LLM (GPT-2)** using the 124M parameter autoregressive model (Radford et al., 2019), and **Time-LLM (BERT)** using BERT-base (Devlin et al., 2019) (110M parameters) with bidirectional attention. Both backbones remain frozen; only reprogramming and output projection layers are learned (architecture details in Appendix H).

3.2.2. Statistical and Deep Learning Baselines

We include standard forecasting baselines: **Naive Persistence** (predicts $\hat{T}_{t+h} = T_t$), **ARIMA** with automatic order selection (Hyndman & Khandakar, 2008), **Prophet** (Taylor & Letham, 2018) for additive decomposition, **LSTM** (Hochreiter & Schmidhuber, 1997) as the standard deep learning baseline, and **Physics-LSTM** (Jia et al., 2021) with physics-informed regularization:

$$\mathcal{L}_{\text{PI-LSTM}} = \mathcal{L}_{\text{MSE}} + \lambda \cdot \mathcal{L}_{\text{physics}} \tag{3}$$

where $\mathcal{L}_{\text{physics}}$ penalizes predictions violating thermal bounds derived from geothermal gradients (detailed in Appendix G).

3.3. Physics-Based Approaches

We solve the coupled PDE system using MODFLOW 6 + GWT for the proprietary site (3D grid, 6 hydrogeologic layers) and a lumped-parameter Boussinesq ODE for USGS Kennecott (Appendix J).

3.4. Hybrid Approach: Physics + LLM Residual Learning

The hybrid approach decomposes forecasting into a physics-based baseline and a learned residual correction $\hat{T}(t) = T_{\text{physics}}(t) + \Delta T_{\text{LLM}}(t)$. We train Time-LLM on residuals

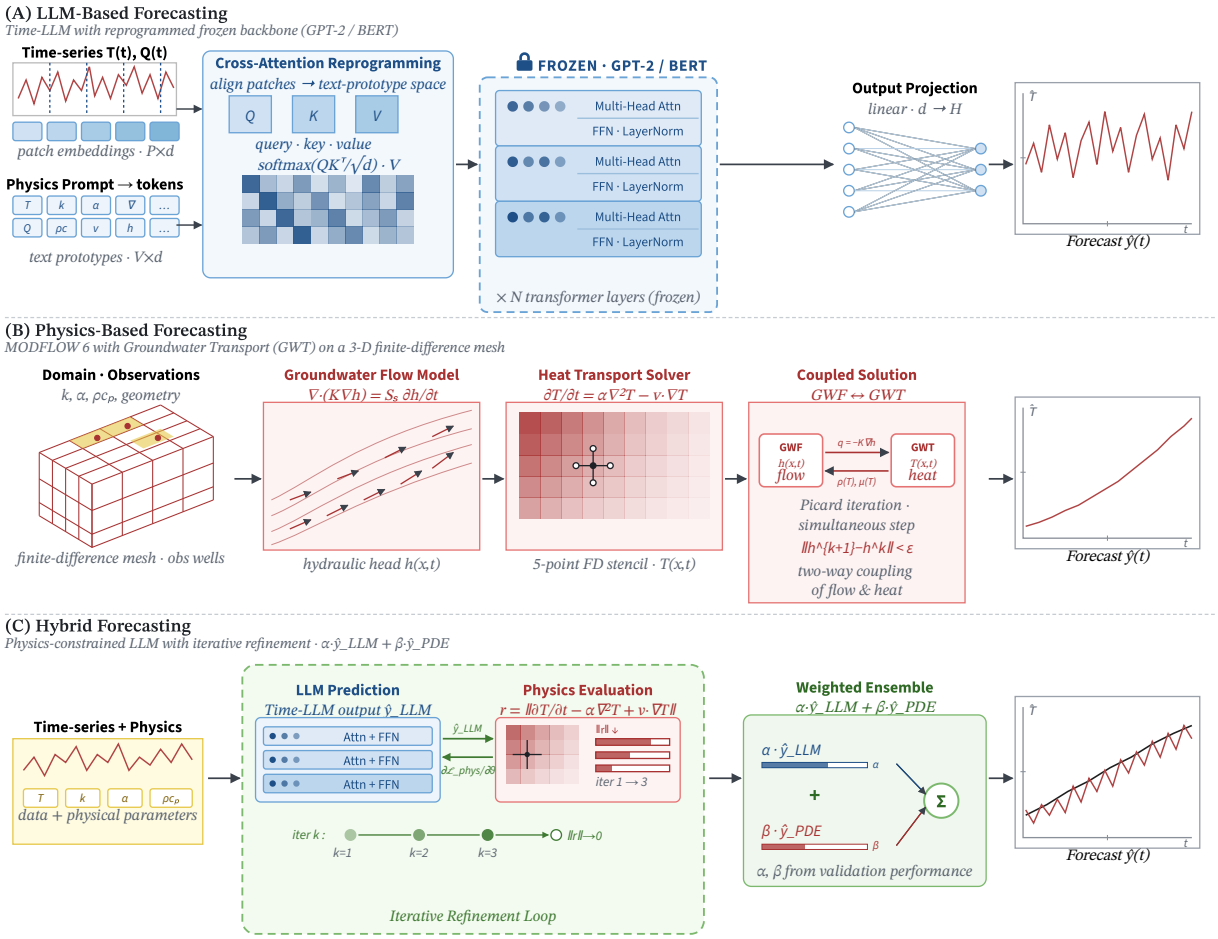

Figure 1. **Comparison of forecasting paradigms.** (A) Learning-only Time-LLM with patch embedding, cross-attention reprogramming, and frozen GPT-2/BERT backbone. (B) Physics-only MODFLOW 6 with GWT solving coupled groundwater flow and heat transport. (C) Hybrid approach combining LLM prediction with physics evaluation in a weighted ensemble.

from the physics model and weight LLM and physics outputs using validation performance. We additionally evaluate **PhysicsLoss-LLM**, which adds thermal-bound penalties to the Time-LLM objective (Appendix I).

3.5. Data, Metrics, and Protocol

We use proprietary daily discharge temperatures from a deep mining operation and 10 years of daily USGS mine discharge data from Kennecott Drain near Magna, Utah (Site 10172650) (U.S. Geological Survey, 2024); preprocessing details appear in Appendix K.

Validated-horizon MAE/MSE are reported at 1 month, 3 months, 5 months, and 1 year for the proprietary mine data, and at 1 and 3 years for USGS. Mine-temperature metrics are standardized; USGS discharge metrics use ft^3/s. For 10-year extrapolation, where no ground truth is available, we report physical-consistency metrics instead of accuracy. Stochastic models run 3 times with different seeds (mean ± std reported). **Reproducibility.** The proprietary mine dataset cannot be released for confidentiality reasons; we document the preprocessing pipeline and MODFLOW configuration (Appendix I) to support replication on comparable discharge-temperature datasets.

4. Results

We evaluate all models on (1) proprietary thermal discharge from an active deep mine (Table 1) and (2) public USGS mine discharge from Kennecott Drain, Utah (Table 3), with units and metric choices as described in §3.5.

4.1. Short to Medium Horizon Performance (1 Month to 1 Year)

At short horizons (1–3 months), Time-LLM variants dominate all baselines (Table 1). BERT gives the best 1-month MAE (0.075) and remains strong at 3 months, while GPT-2 follows closely. MODFLOW+GWT underperforms at short

Table 1. Benchmark results on proprietary mine temperature forecasting. Metrics computed on standardized data (zero mean, unit variance). Best results in **bold**, second-best underlined. Mean ± std over 3 runs where applicable.

		1-Month		3-Month		5-Month		1-Year	
Model	**Type**	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Naive	Baseline	0.232	0.066	0.304	0.099	0.296	0.093	0.274	0.094
ARIMA	Statistical	0.245	0.073	0.317	0.107	0.309	0.101	0.287	0.091
Prophet	Statistical	1.093	1.207	1.109	1.246	1.175	1.399	1.341	1.910
LSTM	Deep Learning	0.250±.039	0.079±.025	0.197±.016	0.045±.006	0.183±.005	0.042±.002	0.935±.017	2.260±.234
Physics-LSTM	Physics-DL	0.246±.047	0.074±.026	0.225±.008	0.057±.005	**0.117**±.007	**0.018**±.003	0.924±.010	2.303±.242
Time-LLM (BERT)	LLM	**0.075**±.019	**0.008**±.004	**0.105**±.018	**0.013**±.004	0.434±.009	0.746±.133	0.424±.038	0.783±.125
Time-LLM (GPT-2)	LLM	0.148±.017	0.027±.007	0.113±.013	0.015±.003	0.184±.091	0.103±.105	**0.187**±.088	0.088±.076
MODFLOW+GWT	PDE Solver	0.211	0.054	0.273	0.079	0.266	0.075	0.248	**0.068**
PhysicsLoss-LLM (BERT)	Physics-LLM	0.124±.025	0.019±.009	0.109±.012	0.014±.004	0.238±.107	0.194±.141	0.320±.029	0.302±.011
PhysicsLoss-LLM (GPT-2)	Physics-LLM	0.137±.033	0.025±.012	0.126±.038	0.020±.012	0.180±.026	0.061±.014	0.226±.052	0.130±.100
MODFLOW+LLM (BERT)	Hybrid	0.248±.008	0.076±.005	0.316±.016	0.111±.010	0.383±.013	0.169±.012	0.329±.002	0.162±.007
MODFLOW+LLM (GPT-2)	Hybrid	0.261±.007	0.085±.006	0.324±.013	0.114±.007	0.341±.046	0.132±.032	0.317±.019	0.142±.030

horizons, reflecting the PDE solver's difficulty in capturing fine-grained daily variation from numerical discretization alone.

A critical transition emerges at 5 months: BERT degrades sharply, while GPT-2 remains stable, indicating that autoregressive structure provides more robust extrapolation than bidirectional encoding. At 1 year, GPT-2 achieves the best MAE, but MODFLOW+GWT achieves the lowest MSE, producing fewer large deviations. Hybrid MODFLOW+LLM approaches do not improve over MODFLOW+GWT alone at any horizon.

Predicting the crossover. Time-LLM uses a 512-step lookback (≈1.4 years of daily data), bounding the temporal patterns it can extract. Thermal drawdown operates on the diffusion timescale $\tau_{\text{th}} \sim L^2/(4\pi\alpha)$, where L is the characteristic drawdown length scale and α the thermal diffusivity; for our site geometry this is on the order of years. We therefore predict the crossover horizon h^* as $h^* \sim \min(L_{\text{lookback}}, \tau_{\text{phys}})$, where L_{lookback} is the model's lookback window (here ≈1.4 years) and τ_{phys} the dominant physical timescale (τ_{th} for the thermal system). This is consistent with the empirical ≈1-year transition. The same prediction holds for USGS via the aquifer recession time $\tau \approx 1000$ days (Appendix D).

4.2. USGS Discharge Forecasting

Appendix E reports full results on the public USGS Kennecott Drain dataset, extending evaluation to 3-year horizons. At 1 year, ODE+LLM (GPT-2) achieves best performance (MAE = 22.69), narrowly outperforming pure Physics-ODE (MAE = 23.11). Statistical methods remain competitive, while LSTM-based approaches exhibit high variance.

At 3 years, physics-based ODE achieves lowest MAE (25.41), with hybrid ODE+LLM trailing closely (MAE = 25.67). Time-LLM variants degrade notably (GPT-2 MAE = 31.94; BERT MAE = 34.22), confirming the crossover observed in proprietary data: **physics-based approaches become essential as horizons extend beyond temporal patterns in the training window.**

4.3. Long-Horizon Consistency Metrics (10 Years)

Beyond validated horizons, we report consistency with physical expectations rather than accuracy. In 10-year forecasts, all models remain within the VRT safety envelope, but Time-LLM variants are near-flat while MODFLOW+GWT captures cooling drift from sustained pumping (Table 2; Appendix C). Thus, infrastructure planning based on the LLM forecast would miss the long-term thermal trajectory.

Table 2. 10-year consistency metrics.

Model	**Drift (°F/yr)**	**Var. (°F)**
Time-LLM (GPT-2)	≈0	<0.20
Time-LLM (BERT)	≈0	<0.20
MODFLOW+GWT	−1.36	0.48

5. Conclusion

RIFT shows LLM forecasters help at short horizons but lose reliability once forecasts exceed the dominant physical timescale. Across both datasets, long-horizon LLM forecasts flatten while physics-based dynamics remain plausible; conservation-law structure, not solver complexity, drives reliable extrapolation.

References

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Pineda Arango, S., Kapoor, S., et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Bakker, M., Post, V., Langevin, C. D., Hughes, J. D., White, J. T., Starn, J. J., and Fienen, M. N. Scripting MODFLOW model development using Python and FloPy. *Groundwater*, 54(5):733–739, 2016.

Banerjee, A. and Gupta, S. K. Personalized open world plan generation for safety-critical human centered autonomous systems: A case study on artificial pancreas. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025.

Bluhm, S., Moreby, R., von Glehn, F., and Pascoe, C. Life-of-mine ventilation and refrigeration planning for resolution copper mine. *Journal of the Southern African Institute of Mining and Metallurgy*, 114(6):497–503, 2014. URL https://scielo.org.za/pdf/jsaimm/v114n6/16.pdf.

Das, A., Kong, W., Sen, R., and Zhou, Y. A decoder-only foundation model for time-series forecasting. In *International Conference on Machine Learning*, 2024.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

Harbaugh, A. W. MODFLOW-2005, the U.S. Geological Survey modular ground-water model: The ground-water flow process. Techniques and Methods 6-A16, U.S. Geological Survey, 2005.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

Hyndman, R. J. and Khandakar, Y. Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 27(3):1–22, 2008.

Jia, X., Willard, J., Karpatne, A., Read, J. S., Zwart, J. A., Steinbach, M., and Kumar, V. Physics-guided machine learning for scientific discovery: An application in simulating lake temperature profiles. *ACM/IMS Transactions on Data Science*, 2(3):1–26, 2021.

Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J. Y., Shi, X., Chen, P.-Y., Liang, Y., Li, Y.-F., Pan, S., and Wen, Q. Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations*, 2024.

Kharazmi, E., Zhang, Z., and Karniadakis, G. E. hp-VPINNs: Variational physics-informed neural networks with domain decomposition. *Computer Methods in Applied Mechanics and Engineering*, 374:113547, 2021.

Langevin, C. D., Hughes, J. D., Banta, E. R., Niswonger, R. G., Panday, S., and Provost, A. M. Documentation for the MODFLOW 6 groundwater flow model. Techniques and Methods 6-A55, U.S. Geological Survey, 2017.

Langevin, C. D., Hughes, J. D., Banta, E. R., Provost, A. M., Niswonger, R. G., and Panday, S. MODFLOW 6 modular hydrologic model version 6.4.1. Technical report, U.S. Geological Survey, 2022. Software Release, https://doi.org/10.5066/P9FL1JCC.

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., and Anandkumar, A. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

Lu, L., Pestourie, R., Yao, W., Wang, Z., Verdugo, F., and Johnson, S. G. Physics-informed neural networks with hard constraints for inverse design. *SIAM Journal on Scientific Computing*, 43(6):B1105–B1132, 2021.

McPherson, M. J. *Subsurface Ventilation and Environmental Engineering*. Chapman & Hall, London, 1993.

Meng, X. and Karniadakis, G. E. A composite neural network that learns from multi-fidelity data: Application to function approximation and inverse PDE problems. *Journal of Computational Physics*, 401:109020, 2020.

Nie, Y., Nguyen, N. H., Sinthong, P., and Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. 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.

Sun, L., Gao, H., Pan, S., and Wang, J.-X. Surrogate modeling for fluid flows based on physics-constrained deep learning without simulation data. *Computer Methods in Applied Mechanics and Engineering*, 361:112732, 2020.

Tan, M., Merrill, M. A., Gupta, V., Althoff, T., and Hartvigsen, T. Are language models actually useful for time series forecasting? In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Tao, H., Hameed, M. M., Marhoon, H. A., Zounemat-Kermani, M., Heddam, S., et al. Groundwater level prediction using machine learning models: A comprehensive review. *Neurocomputing*, 489:271–308, 2022. DOI: 10.1016/j.neucom.2022.03.014.

Taylor, S. J. and Letham, B. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018.

U.S. Geological Survey. USGS water data for the nation: Site 10172650, Kennecott drain near Magna, UT. https://waterdata.usgs.gov/nwis, 2024. Accessed: 2025-01-15.

Willard, J., Jia, X., Xu, S., Steinbach, M., and Kumar, V. Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Computing Surveys*, 55(4):1–37, 2022.

WSP USA. Resolution copper groundwater flow model report. Technical Report Project No. 31400968, WSP USA, Greenwood Village, Colorado, February 2019. URL https://www.resolutionmineeis.us/sites/default/files/references/wsp-groundwater-flow-model-2019.pdf. Prepared for the Resolution Copper Project and Land Exchange EIS; U.S. Forest Service public record.

Wu, H., Xu, J., Wang, J., and Long, M. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22419–22430, 2021.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., and Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12): 11106–11115, 2021.

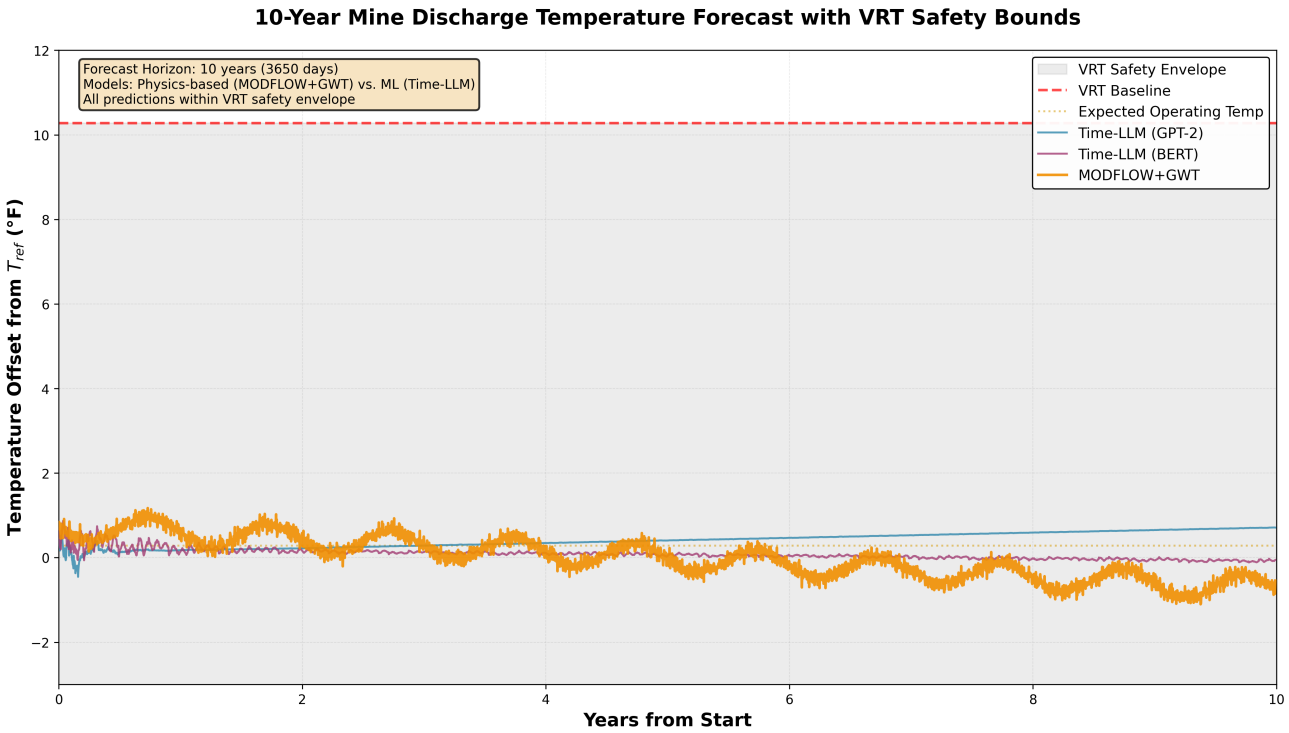

Figure 2. 10-year mine discharge temperature forecasts. Time-LLM variants produce near-flat projections while MODFLOW+GWT captures gradual cooling drift (-1.36°F/yr) from sustained pumping. All predictions remain within the VRT safety envelope (+10.3°F upper bound, off-scale).

A. Appendix

B. Limitations and Scope

Ground truth constrains what we can claim. The proprietary mine series is validated only to 1 year and USGS Kennecott to 3 years, so the 10-year forecasts are assessed for physical plausibility, not accuracy. The mine dataset cannot be released, so only the USGS results are fully reproducible. We evaluate two LLM backbones (GPT-2 and BERT) on univariate targets at one site per dataset, leaving generalization across architectures, geologies, and multivariate coupling untested. The crossover predictor $h^* \sim \min(L_{\text{lookback}}, \tau_{\text{phys}})$ recovers the empirical transition but is an order-of-magnitude timescale argument rather than a calibrated bound, and the physical timescales depend on site parameters that vary between settings.

Scope and extensions. The scope is univariate discharge, two backbones (GPT-2, BERT), and horizons up to 3-year ground truth; see Appendix M.

C. 10-Year Forecast Visualization

D. Crossover Timescale Explanation

Time-LLM uses a 512-step lookback (≈1.4 years of daily data), bounding the temporal patterns it can extract. Thermal drawdown from sustained pumping operates on the diffusion timescale $\tau_{\text{th}} \sim L^2/(4\pi\alpha)$, where L is the characteristic flow-path length and α the effective thermal diffusivity; for our site geometry this yields τ_{th} on the order of years, comparable to the lookback. Once the forecast horizon exceeds τ_{th}, the dominant physical mode lies outside the LLM's evidence window and pattern matching cannot recover it. The empirical crossover at ≈1 year is consistent with $h^* \sim \min(L_{\text{lookback}}, \tau_{\text{phys}})$. The same argument explains the USGS crossover via the aquifer recession time $\tau \approx 1000$ days (Appendix J).

E. USGS Benchmark Results

F. Extended Related Work

F.1. Transformer-Based Time Series Forecasting

Transformers have been extensively adapted for time series. Informer (Zhou et al., 2021) introduced ProbSparse self-attention, Autoformer (Wu et al., 2021) incorporated

Table 3. Benchmark results on USGS mine discharge forecasting (ft^3/s). Best results in **bold**, second-best underlined. Mean ± std over 3 runs where applicable.

		1-Year		3-Year	
Model	**Type**	MAE	MSE	MAE	MSE
Naive	Baseline	34.93	2043.29	41.38	2584.73
ARIMA	Statistical	24.13	1034.09	26.87	1318.54
Prophet	Statistical	26.18	1000.27	26.71	**917.34**
LSTM	Deep Learning	51.46±19.44	3661.50±2012.09	39.83±6.75	2391.83±861.65
Physics-LSTM	Physics-DL	49.93±16.95	3273.26±1615.64	32.98±2.78	1581.41±227.26
Time-LLM (BERT)	LLM	29.56±2.00	1396.04±147.31	34.22±1.34	1873.78±128.67
Time-LLM (GPT-2)	LLM	26.22±1.53	1146.74±55.40	31.94±2.15	1699.17±179.64
Physics-ODE	Physics	<u>23.11</u>	<u>929.19</u>	**25.41**	<u>1162.56</u>
ODE+LLM (BERT)	Physics-LLM	23.54±0.27	952.46±10.70	<u>25.67</u>±1.40	1180.11±107.98
ODE+LLM (GPT-2)	Physics-LLM	**22.69**±0.64	**911.24**±11.39	26.15±0.06	1180.69±5.59

decomposition blocks, and PatchTST (Nie et al., 2023) demonstrated that patching improves long-horizon forecasting. Foundation models have emerged as general-purpose forecasters: Time-LLM (Jin et al., 2024) reprograms frozen LLMs via learned input transformations, Chronos (Ansari et al., 2024) uses quantization-based tokenization, and TimesFM (Das et al., 2024) trains decoder-only architectures on 100 billion time points, achieving strong zero-shot performance on standard benchmarks.

However, Tan et al. (2024) show that removing LLM backbones does not degrade performance, suggesting patching and attention structure, rather than pretrained language knowledge, drive results. More fundamentally, existing evaluations focus on horizons under one year and domains where patterns repeat predictably. RIFT extends this critique by showing that even where LLMs provide value at short horizons, they fail when forecasting physical systems governed by coupled PDEs over multi-year timescales.

F.2. Physics-Informed Machine Learning

Physics-informed neural networks (PINNs) embed governing equations directly into training objectives (Raissi et al., 2019), spawning a rich literature on variational formulations (Kharazmi et al., 2021), hard-constraint architectures (Lu et al., 2021), and multi-fidelity extensions (Meng & Karniadakis, 2020). In environmental modeling, physics-informed LSTMs (Jia et al., 2021) penalize conservation-law violations during training, while neural operators (Li et al., 2021) learn parametric PDE solution maps. Hybrid strategies that couple mechanistic simulators with neural residual corrections (Willard et al., 2022) further blur the boundary between learning and solving.

These approaches generally treat the neural network as a surrogate for the PDE solver. RIFT asks a different question: given a fixed forecast task, when does a purely data-driven model suffice and when must one use the classical numerical solver it would replace? By benchmarking end-to-end learning against validated PDE solvers across systematically varying horizons, we identify a practical crossover beyond which soft physics constraints fail to prevent long-horizon degradation.

F.3. Subsurface Flow and Heat Transport

MODFLOW (Harbaugh, 2005; Langevin et al., 2017) is the standard for saturated groundwater flow simulation, with MODFLOW 6 + GWT (Langevin et al., 2022) providing integrated flow and heat transport solvers. ML applications to subsurface systems remain limited (Sun et al., 2020; Tao et al., 2022). No prior work systematically compares transformer architectures against physics-based solvers for thermal forecasting across horizons spanning months to decades.

Table 4. Positioning relative to prior work. Our study uniquely combines long-horizon evaluation (10+ years), physics-based baselines, and subsurface thermal application.

Study	**Horizon**	**Physics**	**Subsurface**
Informer (Zhou et al., 2021)	<1 year	✗	✗
Time-LLM (Jin et al., 2024)	<1 year	✗	✗
Chronos (Ansari et al., 2024)	<1 year	✗	✗
TimesFM (Das et al., 2024)	<1 year	✗	✗
Tan et al. (Tan et al., 2024)	<1 year	✗	✗
PINNs (Raissi et al., 2019)	Variable	Surrogate	✗
PI-LSTM (Jia et al., 2021)	<1 year	Loss reg.	Lakes
RIFT (Ours)	**10 years**	**PDE solver**	**Mining**

G. Baseline Model Details

G.1. Physics-LSTM Regularization

Physics-LSTM augments the standard LSTM architecture with physics-informed regularization that constrains predictions to respect thermal bounds derived from the geothermal gradient at the extraction depth:

$$\mathcal{L}_{\text{physics}} = \frac{1}{H}\sum_{h=1}^{H} \max\left(0,\ \hat{T}_{n+h} - T_{\text{VRT}}(z)\right)^2 + \max\left(0,\ T_{\text{min}} - \hat{T}_{n+h}\right)^2 \tag{4}$$

This soft constraint approach mirrors the physics-informed loss regularization used in lake temperature modeling (Jia et al., 2021), adapted for subsurface thermal systems where virgin rock temperature provides a natural upper bound. The penalty weight λ in Equation 3 is tuned via validation performance (range tested: $\lambda \in [0.001, 0.1]$).

G.2. Baseline Architectures

LSTM Configuration: Two-layer long short-term memory network with 128 hidden units per layer, trained with MSE loss and Adam optimizer (learning rate 10^{-3}, batch size 32). Dropout rate 0.2 applied between layers.

ARIMA: We use the `auto.arima` implementation from the `forecast` R package with stepwise search over orders (p, d, q) where $p \in [0, 5]$, $d \in [0, 2]$, $q \in [0, 5]$. Model selection via AIC minimization with seasonal component detection.

Prophet: Additive decomposition model with piecewise linear trend, Fourier-based seasonality (10 terms for annual cycle), and automatic changepoint detection. We use default hyperparameters with changepoint prior scale 0.05.

H. Time-LLM Architecture Details

Given an input time series $\mathbf{x} \in \mathbb{R}^L$ of lookback length L, Time-LLM applies:

$$\mathbf{z} = \text{Reprogram}(\mathbf{x}) = \mathbf{W}_p \cdot \text{Patch}(\mathbf{x}) + \mathbf{E}_{\text{pos}} \tag{5}$$

where $\text{Patch}(\cdot)$ segments the input into non-overlapping patches of length P, $\mathbf{W}_p \in \mathbb{R}^{d_{\text{model}} \times P}$ projects each patch to the LLM embedding dimension, and $\mathbf{E}_{\text{pos}}$ adds learned positional information. Patch embeddings are aligned with text embeddings through cross-attention:

$$\mathbf{z}' = \text{CrossAttn}(\mathbf{z}, \mathbf{E}_{\text{text}}, \mathbf{E}_{\text{text}}) \tag{6}$$

where $\mathbf{E}_{\text{text}} \in \mathbb{R}^{V \times d_{\text{model}}}$ are text prototype embeddings. A physics-aware prompt prefix provides domain context, the frozen LLM backbone processes the concatenated sequence, and a learned output projection $\mathbf{W}_o$ maps final hidden states to forecast values.

I. Hybrid Model Details

The residual-learning hybrid computes $r_t = T_{\text{observed}}(t) - T_{\text{physics}}(t)$ and trains Time-LLM on $\{r_1, \ldots, r_n\}$ rather than raw temperatures. The final ensemble weights LLM and physics outputs based on validation performance.

PhysicsLoss-LLM trains directly on raw temperatures but incorporates thermal bound penalties into the Time-LLM training objective:

$$\mathcal{L}_{\text{PhysicsLoss-LLM}} = \mathcal{L}_{\text{MSE}} + \lambda \cdot \mathcal{L}_{\text{physics}} \tag{7}$$

This tests whether physics constraints in the loss function can compensate for the absence of explicit PDE solving.

J. Physics Model Equations

J.1. Thermal Transport Model (Proprietary Data)

Discharge temperature at a deep mine reflects the integrated thermal state along the flow path from extraction depth to surface. We solve the coupled PDE system (Equations 1 –2) using MODFLOW 6 + GWT (Langevin et al., 2022) via FloPy (Bakker et al., 2016).

J.1.1. Domain and Layer Properties

The subsurface is discretized into a $20 \times 20 \times 6$ grid (200 m horizontal spacing, total domain 4 km $\times$ 4 km $\times$ deep (>1 km) extraction depth). Each layer corresponds to a distinct hydrogeologic unit with independently assigned hydraulic conductivity ($K = 10^{-8}$ to 10^{-6} m/s) and porosity (0.04–0.15), broadly consistent with the publicly documented hydrogeologic characterization of the site (WSP USA, 2019). Thermal conductivity (1.95–6.55 W/(m·K), reaching 4.8–6.6 in quartz-rich rock) and rock heat capacity follow the publicly documented site thermal characterization (Bluhm et al., 2014). Vertical hydraulic conductivity is set to 10% of horizontal in all layers.

J.1.2. Geothermal Baseline

The virgin rock temperature (VRT) at extraction depth z establishes the initial thermal field:

$$T_{\text{VRT}}(z) = T_{\text{surface}} + \Gamma \cdot z \tag{8}$$

where T_{surface} is the mean annual surface temperature and

Γ [°C/m] is the site geothermal gradient determined from borehole temperature logs, consistent with the publicly documented site value (Bluhm et al., 2014)

J.1.3. Heat-as-Solute Formulation

MODFLOW 6 GWT solves the advection-dispersion equation for a generic solute. We map thermal transport onto this framework by treating temperature as concentration. Thermal retardation accounts for heat storage in rock versus fluid via the linear sorption model with bulk density $\rho_b = (1-\phi)\rho_s$ and distribution coefficient $K_d = c_s/(\rho_w c_w)$, giving retardation factor:

$$R = 1 + \frac{\rho_b K_d}{\phi} \tag{9}$$

Thermal diffusion maps to molecular diffusion coefficient $D_{\text{eff},k} = \lambda_k/(\rho_w c_w)$ for layer k, with longitudinal and transverse dispersivities of 10 m and 1 m respectively.

J.1.4. Boundary Conditions and Pumping

Constant-head boundaries maintain lateral hydraulic equilibrium. Two extraction wells pump at a site-typical combined rate, with the majority of flow drawn from a primary shaft and the remainder from a secondary shaft. Extraction wells remove water at the local cell temperature via the SSM (Source-Sink Mixing) package. The weighted discharge temperature is:

$$T_{\text{discharge}} = w_A\, T_{\text{Shaft A}} + w_B\, T_{\text{Shaft B}} \tag{10}$$

with $w_A + w_B = 1$ set from site pumping records.

J.1.5. Temporal Discretization

Stress periods use daily resolution for the first 60 days, weekly for days 60–365, and monthly thereafter, balancing accuracy with computational cost.

J.1.6. Calibration

A single constant offset ΔT aligns model predictions with observations:

$$T_{\text{cal}}(t) = T_{\text{MODFLOW}}(t) + \Delta T, \quad \Delta T = \bar{T}_{\text{obs}} - \bar{T}_{\text{model}} \tag{11}$$

where $\bar{T}_{\text{obs}}$ and $\bar{T}_{\text{model}}$ are training-period means. This compensates for unresolved sub-grid heterogeneity while preserving all temporal dynamics from the PDE solution. Hydraulic parameters are consistent with the publicly documented hydrogeologic characterization and mine dewatering records for the site (WSP USA, 2019); thermal parameters and the geothermal gradient follow the publicly documented site thermal characterization (Bluhm et al., 2014).

J.2. Discharge Dynamics Model (USGS Kennecott)

The USGS dataset records daily water discharge $Q(t)$ [ft^3/s] from a mine drainage tunnel. Unlike the thermal system, no site-specific hydrogeological parameters (hydraulic conductivity, aquifer geometry, porosity) are available. We therefore solve a lumped-parameter ODE derived from integrating the groundwater flow PDE over the aquifer domain.

J.2.1. Governing Equation

The governing equation treats the aquifer as a linear reservoir with seasonal forcing:

$$\frac{dQ}{dt} = -\frac{Q - Q_{\text{base}}}{\tau} + R(t) \tag{12}$$

where τ [days] is the aquifer response time (analogous to the ratio of storage to transmissivity, $\tau \sim S \cdot L^2/T$, where S is storativity, L is aquifer length, and T is transmissivity), Q_{base} [ft^3/s] is the long-term equilibrium discharge, and $R(t)$ is a seasonal recharge function capturing snowmelt and precipitation inputs.

This equation is the temporal component of the 1D Boussinesq equation for unconfined groundwater flow, where spatial integration over the aquifer length L reduces the PDE to an ODE in discharge.

J.2.2. Analytical Solution

The analytical solution is:

$$Q(t) = (Q_0 - Q_{\text{base}}) \cdot e^{-t/\tau} + Q_{\text{base}} + R(t) \tag{13}$$

where Q_0 is the initial discharge. The hydraulic diffusivity D [m^2/day] relates to the response time via $D = L^2/(\pi^2\tau)$, providing a physical interpretation even without direct aquifer measurements.

J.2.3. Calibration Components

Three components are calibrated from the training data:

(i) Recession parameters. The decay time τ and baseline Q_{base} are fit by minimizing prediction error on the detrended discharge:

$$(\tau^*, Q^*_{\text{base}}) = \underset{\tau, Q_{\text{base}}}{\arg\min} \sum_t \left(Q_{\text{obs}}(t) - Q_{\text{pred}}(t; \tau, Q_{\text{base}})\right)^2 + \lambda_{\text{reg}}(\tau - \tau_{\text{prior}})^2 \tag{14}$$

with regularization constraining τ to physically plausible ranges (100 to 8000 days for fractured rock aquifers). The prior $\tau_{\text{prior}} = 1000$ days represents typical recession timescales for regional mine drainage systems.

(ii) Seasonal recharge. Monthly mean residuals between observed discharge and the recession baseline define the seasonal recharge pattern $R(t)$:

$$R(m) = \frac{1}{N_m} \sum_{t \in \text{month } m} [Q_{\text{obs}}(t) - Q_{\text{recession}}(t)] \tag{15}$$

where N_m is the number of days in month m, and $m \in \{1, 2, \ldots, 12\}$. Recharge values are bounded to $|R(m)| \leq 2\sigma_Q$ to prevent overfitting to anomalous months.

(iii) Short-term autoregressive correction. A 7-day AR model on detrended, deseasonalized residuals captures persistence in day-to-day fluctuations:

$$\epsilon(t) = \sum_{k=1}^{7} \phi_k \epsilon(t-k) + \eta(t) \tag{16}$$

where ϕ_k are AR coefficients fit by least squares and $\eta(t)$ is white noise. The AR contribution decays exponentially with forecast horizon (half-life $\approx$ 60 days), ensuring that long-horizon predictions rely entirely on the physics components rather than extrapolated statistical patterns:

$$w_{\text{AR}}(h) = \exp(-h/60) \tag{17}$$

where h is the forecast horizon in days.

K. Data Preprocessing Details

K.1. Missing Value Imputation

For gaps ≤ 7 days, we use linear interpolation:

$$T(t_i) = T(t_{i-1}) + \frac{T(t_{i+k}) - T(t_{i-1})}{k+1} \tag{18}$$

where k is the gap length in days. For gaps > 7 days, the surrounding window (gap $\pm$ 14 days) is excluded from both training and evaluation to prevent spurious patterns from interpolation artifacts.

K.2. Outlier Detection and Removal

Values exceeding 3 standard deviations from a 30-day rolling mean are flagged as outliers:

$$\text{outlier}(t) = |T(t) - \mu_{\text{rolling}}(t)| > 3\sigma_{\text{rolling}}(t) \tag{19}$$

Flagged values are replaced via cubic spline interpolation using the nearest 5 non-outlier points on each side. Outlier detection is applied independently to each monitoring point before aggregation.

K.3. Normalization

Zero-mean, unit-variance standardization using statistics computed exclusively on training data:

$$\mu_{\text{train}} = \frac{1}{N_{\text{train}}} \sum_{t \in \text{train}} T(t) \tag{20}$$

$$\sigma_{\text{train}} = \sqrt{\frac{1}{N_{\text{train}}} \sum_{t \in \text{train}} (T(t) - \mu_{\text{train}})^2} \tag{21}$$

$$T_{\text{norm}}(t) = \frac{T(t) - \mu_{\text{train}}}{\sigma_{\text{train}}} \tag{22}$$

Test-set predictions are inverse-transformed for reporting in original units: $\hat{T}(t) = \sigma_{\text{train}} \cdot \hat{T}_{\text{norm}}(t) + \mu_{\text{train}}$.

K.4. Train/Validation/Test Splitting

Chronological 70%/15%/15% split preserving temporal ordering:

- Training: Days 1 to $\lfloor 0.70N \rfloor$
- Validation: Days $\lfloor 0.70N \rfloor + 1$ to $\lfloor 0.85N \rfloor$
- Test: Days $\lfloor 0.85N \rfloor + 1$ to N

where N is total number of days. No future information leaks into training or validation sets. For multi-step forecasting, we use a sliding window approach with no overlap between train/validation/test forecast origins.

L. Training Configuration Details

L.1. Time-LLM Hyperparameters

- **Lookback window**: 512 time steps
- **Patch length**: $P = 16$
- **Patch stride**: 8 (50% overlap)
- **Learning rate**: Initial 10^{-3} with cosine annealing
- **Batch size**: 16
- **Optimizer**: AdamW with weight decay 10^{-4}
- **Early stopping**: Patience of 10 epochs on validation MAE
- **Max epochs**: 100
- **Text prototype vocabulary**: Top 100 most frequent words from domain-specific corpus (mining/geology terms)
- **Physics prompt length**: 64 tokens

L.2. Physics-Aware Prompt Template

The physics-aware prompt prefix for Time-LLM includes domain context:

For the USGS discharge dataset, the prompt is adapted:

> *"Groundwater discharge from mine drainage tunnel. Aquifer recession dynamics. Seasonal recharge from snowmelt and precipitation. Storage-release governed by hydraulic properties. Conservation of mass controls discharge rates."*

L.3. Computational Resources

All experiments conducted on:

- **GPU**: Single NVIDIA RTX 6000 Ada Generation (48GB VRAM)
- **CPU**: AMD EPYC 7763 (64 cores)
- **RAM**: 512GB DDR4
- **Training time**: 2-4 hours per Time-LLM variant (depending on horizon)

M. Future Work

The scope statement in the main paper identifies three axes - input dimensionality, model family, and horizon length - along which RIFT can be extended. We discuss each here, together with two further directions (spatial generalization, tighter hybrid coupling) that follow naturally from the crossover analysis.

Multivariate inputs. The current evaluation treats discharge as a univariate stream. Operational records include pumping rates, ambient temperature, precipitation, and discharge from neighboring wells. Conditioning Time-LLM on these covariates could plausibly extend the short-horizon LLM advantage and sharpen the crossover. Whether the closed-form predictor $h^* \sim \min(L_{\text{lookback}}, \tau_{\text{phys}})$ continues to hold under richer inputs is an open empirical question.

Additional foundation backbones. We tested GPT-2 and BERT as the two canonical autoregressive/bidirectional choices. A broader sweep across modern foundation forecasters - Chronos (Ansari et al., 2024), TimesFM (Das et al., 2024), Lag-Llama, and Moirai - would test whether the crossover is universal across model family and pretraining corpus, or whether some backbones erode the gap.

Long-horizon validation. Ground truth on the proprietary site extends to 1 year and on USGS Kennecott to 3 years. Continued monitoring will yield multi-year holdouts that directly validate the 10-year regime. In the interim, conservation-law residuals and the VRT safety envelope serve as soft validators: any forecast that violates the upper bound is provably wrong regardless of held-out observations.

Spatial generalization. RIFT evaluates two sites with distinct conservation structure (energy vs. mass). Replicating the analysis across additional mine operations - varying depth, lithology, and pumping regime - would establish whether the crossover horizon scales with τ_{phys} across geological settings.

Tighter hybrid coupling. Our hybrid models use ensemble residual stacking. End-to-end architectures that backpropagate through a differentiable PDE solver (Willard et al., 2022) may close the gap that residual stacking leaves open, particularly when physics parameters themselves are uncertain.