*Proceedings of Machine Learning Research* **318** (2026)
The 39th Canadian Conference on Artificial Intelligence

Canadian AI 2026
25-29 May 2026,Vancouver,BC

# Physics-Guided Diffusion Models for Production Forecasting with Limited Well Data

Temesgen Mikael Abraha*
Yves Lucet
Department of Computer Science, I. K. Barber Faculty of Science
University of British Columbia, Kelowna, BC, Canada

**Abstract**

Accurate forecasting of oil and gas well production from limited early-time data is critical for reservoir management and timely land cleanup. We present Physics-SIMS-TS, a novel framework that adapts Self-Improving Diffusion Models with Synthetic Data (SIMS) for time series forecasting and integrates physics-based decline-curve constraints. Given only the first 20% of a well's production history, our model predicts the remaining 80%. We first adapt the SIMS methodology, originally developed for image generation, to time series by training a conditional diffusion model with negative guidance that steers generation away from synthetic artifacts. Experiments show that domain-aware data augmentation (Inverse Distance Weighting) outperforms generic generative approaches (TimeGAN, TimeVAE) by 1.7×, demonstrating that incorporating domain-specific knowledge improves forecasting performance. Building on this insight, we introduce Physics-SIMS-TS, which integrates Arps decline-curve dynamics through gradient guidance during sampling and monotonicity projection via isotonic regression. Experiments on 16,216 gas wells from British Columbia, Canada, spanning multiple geological formations, demonstrate that Physics-SIMS-TS achieves 1.9–6.2× lower prediction error than traditional machine learning baselines across all dataset sizes, with the largest improvements on small datasets where physics constraints most effectively regularize the learning problem.

**Keywords:** Diffusion models, time series forecasting, production decline curves, physics-informed machine learning, oil and gas, self-improving models

## 1. Introduction

Forecasting hydrocarbon production from unconventional reservoirs remains a persistent challenge in the field of petroleum engineering. Economic decisions regarding well completion, field development, and reserve estimation depend critically on accurate predictions of future production rates. However, reliable forecasts traditionally require years of production history [1], yet business decisions often demand predictions from as little as 6–12 months of data.

Among these decisions, forecasts determine when a well reaches its economic limit and should be scheduled for abandonment and site reclamation. Wells without reliable production forecasts risk being left idle indefinitely, eventually becoming orphaned, with no solvent party responsible for decommissioning and cleanup. This is a growing challenge across North America: British Columbia currently has over 870 orphan well sites awaiting restoration [2], Alberta's orphan inventory has risen to over 3,300 wells with cleanup costs exceeding $1 billion [3], and the United States has at least 120,000 documented orphaned wells with an estimated 310,000–800,000 more remaining undocumented [4].

The classical approach to production forecasting relies on Arps decline-curve analysis, which models production rate $q(t)$ through the hyperbolic equation

$$q(t) = \frac{q_i}{(1 + b \cdot D_i \cdot t)^{1/b}}, \tag{1.1}$$

* temesgen.abraha@ubc.ca

where $q_i$ is the initial rate, $D_i$ is the initial decline rate, and $b$ is the decline exponent. Although the model is physically grounded, fitting these parameters from limited early data is ill-posed, particularly in tight formations where wells remain in transient flow with $b > 1$.

Recent advances in deep generative models, particularly diffusion models [5], have shown strong results in image and time-series generation [6, 7]. However, naïve application of synthetic data augmentation can lead to model autophagous disorder, a progressive degradation that occurs when models train on their own outputs [8]. The Self-Improving diffusion Models with Synthetic data (SIMS) [8] framework addresses this through *negative guidance*, steering generation away from synthetic artifacts rather than toward them.

In this work, we make three contributions:

(1) We adapt SIMS to time series forecasting (SIMS-TS), demonstrating effective production curve generation conditioned on early production data.
(2) Through comparative experiments with data augmentation methods, we show that domain-aware approaches such as Inverse Distance Weighting (IDW) outperform domain-agnostic generative models (TimeGAN, TimeVAE), motivating physics integration.
(3) We introduce Physics-SIMS-TS, which incorporates Arps decline physics through gradient guidance and monotonicity projection, achieving 1.9–6.2× mean squared error (MSE) improvement over baselines across dataset sizes from 51 to 16,216 wells.

## 2. Related Work

Arps' empirical decline equations [1] remain the industry standard for production forecasting. Extensions such as the Stretched Exponential Decline Model [9] and the Duong model [10] address transient flow in unconventional reservoirs. Tadjer et al. [11] demonstrated that these models require modification for unconventional wells where complex flow regimes make traditional analysis problematic. Physics-informed machine learning approaches have also been explored: Han et al. [12] proposed a physics-informed neural network architecture for large-scale systems with sparse well data, and Bi et al. [13] integrated flow theory directly into loss functions for production prediction.

Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) represent two foundational approaches for time-series synthesis. Yoon et al. [14] introduced TimeGAN, combining adversarial training with supervised autoregressive control across four network components. Desai et al. [15] proposed TimeVAE with interpretable temporal decomposition into level, trend, and seasonal components, offering more stable training than GAN-based methods. The TSGBench evaluation framework [16] provides systematic comparison across ten generative methods, finding that VAE-based approaches perform well in distance-based metrics while TimeGAN ranks lower than expected despite its widespread use as a baseline.

Diffusion models achieve state-of-the-art results in time series generation through iterative denoising. Yuan and Qiao [6] introduced Diffusion-TS with encoder-decoder transformers and disentangled seasonal-trend decomposition. Kollovieh et al. [7] proposed TSDiff with self-guidance mechanisms for unconditionally trained diffusion models, demonstrating that synthetic data training can outperform real data training on certain benchmarks. Tashiro et al. [17] presented CSDI, conditioning on observed data through two-dimensional attention for imputation and forecasting tasks. These architectures provide the foundation upon which our SIMS-TS adaptation builds.

Wen et al. [18] provided a comprehensive taxonomy of time series augmentation, categorizing approaches into time domain, frequency domain, and time-frequency domain transformations. Iwana and Uchida [19] evaluated twelve augmentation techniques across 128 datasets, establishing empirical baselines. For spatially distributed data, Inverse Distance

Weighting [20], which estimates values at unsampled locations as weighted averages of nearby measurements, offers a domain-aware alternative. Despite the availability of these methods, data augmentation using generative models for oil and gas production remains largely unexplored, with recent reviews [21] framing it as a future direction rather than an established practice.

Model Autophagy Disorder (MAD) [22], also termed model collapse [23], describes the progressive degradation when generative models train on their own synthetic outputs. Alemohammad et al. [8] introduced SIMS to address this through negative guidance, steering generation away from synthetic artifacts during inference. SIMS maintained stable performance across 100 generations on image benchmarks without degradation. However, to the best of our knowledge, SIMS has not been adapted to time series domains, nor has negative guidance been combined with physics-informed constraints for industrial forecasting applications.

## 3. Methodology

### 3.1. SIMS-TS

Given a normalized production curve $\mathbf{x} \in \mathbb{R}^{100}$ representing monthly production over the well lifetime, we observe only the first 20 points $\mathbf{c} = \mathbf{x}_{1:20}$ and predict the remaining 80 points $\mathbf{y} = \mathbf{x}_{21:100}$. We employ a conditional denoising diffusion model where the forward process gradually adds Gaussian noise

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0,\ (1-\bar{\alpha}_t)\mathbf{I}) \tag{3.1}$$

with cumulative noise schedule $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$. The denoising network $\epsilon_\theta(\mathbf{x}_t, t, \mathbf{c})$ is a Transformer with Adaptive Layer Normalization (AdaLN) that injects both the diffusion timestep $t$ and conditioning signal $\mathbf{c}$ through additive embeddings. Training minimizes the standard score-matching objective

$$\mathcal{L} = \mathbb{E}_{t,\mathbf{x}_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t, \mathbf{c})\|^2\right]. \tag{3.2}$$

The key insight from SIMS [8] is that synthetic data should provide *negative* guidance. We train a base model $\epsilon_{\theta_r}$ on real production curves, generate synthetic curves, and train an auxiliary model $\epsilon_{\theta_s}$ (initialized from base model weights) on these synthetic samples. At inference, predictions are combined as

$$\tilde{\epsilon} = (1+\omega)\cdot\epsilon_{\theta_r}(\mathbf{x}_t, t, \mathbf{c}) - \omega\cdot\epsilon_{\theta_s}(\mathbf{x}_t, t, \mathbf{c}), \tag{3.3}$$

where $\omega \geq 0$ controls guidance strength. This amplifies the direction from synthetic toward real data distributions.

### 3.2. Physics-SIMS-TS

Production curves approximately follow Arps decline dynamics Equation (1.1) and exhibit monotonically decreasing behavior. We incorporate these constraints through three physics-based loss terms applied during the reverse diffusion process. In the following, $k$ indexes production time steps to distinguish from the diffusion timestep $t$.

The *Arps ODE residual* penalizes deviations from the expected decline rate $D(k) = D_i/(1+b\cdot D_i\cdot k)$

$$\mathcal{L}_{\text{Arps}} = \frac{1}{K}\sum_{k=1}^{K}\left|\frac{d\hat{q}_k}{dk} + D(k)\cdot\hat{q}_k\right|^2, \tag{3.4}$$

where temporal derivatives are computed via central finite differences and $K$ is the production sequence length. The *monotonicity loss* penalizes any increase in production

$$\mathcal{L}_{\text{mono}} = \frac{1}{K-1}\sum_{k=1}^{K-1} \max(0,\ \hat{q}_{k+1} - \hat{q}_k)^2, \tag{3.5}$$

and the *boundary loss* enforces non-negativity, $\mathcal{L}_{\text{bound}} = \frac{1}{K}\sum_{k=1}^{K} \max(0, -\hat{q}_k)^2$.

At each denoising step, we first compute the SIMS-guided noise prediction $\tilde{\epsilon}$ via Equation (3.3), then estimate clean data $\hat{\mathbf{x}}_0$ from the current noisy state $\mathbf{x}_t$. The combined physics loss $\mathcal{L}_{\text{physics}} = \lambda_1\mathcal{L}_{\text{Arps}} + \lambda_2\mathcal{L}_{\text{mono}} + \lambda_3\mathcal{L}_{\text{bound}}$ is evaluated on $\hat{\mathbf{x}}_0$, and the noise prediction is modified via gradient guidance

$$\tilde{\epsilon} \leftarrow \tilde{\epsilon} + \eta(t)\cdot \nabla_{\hat{\mathbf{x}}_0}\mathcal{L}_{\text{physics}}(\hat{\mathbf{x}}_0), \tag{3.6}$$

where the gradient is computed with respect to the estimated clean data and clipped to unit norm for stability. We apply adaptive guidance strength

$$\eta(t) = \eta_0 \cdot \left(1 - \frac{t}{T}\right)^{0.5}, \tag{3.7}$$

which increases physics influence as samples become cleaner in later diffusion steps (lower $t$), since physics constraints are more meaningful on less noisy estimates.

As a final post-processing step, we project generated samples onto the space of monotonically decreasing sequences using isotonic regression via the Pool Adjacent Violators (PAV) algorithm [24], providing a hard guarantee of physical plausibility.

## 4. Experiments

### 4.1. Experimental Setup

We compare three approaches: (1) *IDW+Random Forest*, a baseline using 5× Inverse Distance Weighting augmentation [20] with IDW power parameter $p = 3$ and Random Forest regression; (2) *SIMS-TS*, our adapted SIMS with guidance strength $\omega \in \{0.0, 0.1, 0.3, 0.5\}$; and (3) *Physics-SIMS-TS*, combining SIMS with physics guidance $\eta_0 \in \{0.0, 0.05, 0.1\}$, loss weights $\lambda_1$=1.0, $\lambda_2$=2.0, $\lambda_3$=1.0, and monotonicity projection. For clusters with fewer than 2,000 wells, we fine-tune the base model on cluster-specific data (50 epochs, learning rate $10^{-5}$) before applying SIMS. All diffusion sampling uses DDIM [25] with 50 steps. We report the best configuration per method for each cluster. To establish that domain-aware methods outperform generic generative models, we also conduct a preliminary augmentation comparison on a 201-well Montney subset (160 train, 41 test), evaluating IDW, TimeGAN [14], and TimeVAE [15] across augmentation factors from 1.5× to 20× using both Random Forest (RF) and Gradient Boosting (GB) regressors.

The denoising network is a Transformer with hidden dimension 256, 6 layers, 4 attention heads, and sinusoidal timestep embeddings combined additively with a learned conditional encoder for the observed 20-point input (7.3M parameters). Adaptive Layer Normalization injects both the diffusion timestep and conditioning signal. The noise schedule uses 1,000 cosine-scheduled timesteps with smoothing parameter $s$=0.008.

We train the base model with AdamW (learning rate $10^{-4}$, weight decay 0.01) for 300 epochs with 15-epoch linear warmup and gradient clipping at norm 1.0. For cluster-specific fine-tuning (clusters with fewer than 2,000 wells), we reduce the learning rate to $10^{-5}$ and train for 50 epochs. The auxiliary model for SIMS is initialized from base model weights and fine-tuned on 200 generated synthetic curves for 50 epochs.

For IDW-based augmentation, we use power parameter $p$=3, up to 8 nearest neighbors, a 70/30 mix ratio between the reference and IDW-interpolated curves, and 5% Gaussian noise scaled by mean curve amplitude.

All experiments were conducted on a workstation with an AMD Ryzen Threadripper PRO 7955WX (16 cores), 256 GB RAM, and two NVIDIA RTX 6000 Ada Generation GPUs (48 GB VRAM each); a single GPU was used for all training and inference. Base model training on 12,972 wells (300 epochs) completed in approximately 24 minutes ($\sim$5 seconds per epoch). Cluster-specific fine-tuning (50 epochs) completed in under 1 minute per cluster. At inference on the full 16,216-well test set, IDW+Random Forest required 161 seconds (dominated by 5$\times$ augmentation and RF fitting), while SIMS-TS and Physics-SIMS-TS required 25 and 38 seconds respectively using 50-step DDIM sampling on GPU. For small clusters (fewer than 300 wells), all three methods completed inference in under 4 seconds.

### 4.2. Dataset

We use production data from 16,216 gas wells in British Columbia, Canada, obtained from the British Columbia Energy Regulator (BCER, formerly the BC Oil and Gas Commission), the Crown corporation responsible for regulating oil and gas activities in the province [26]. Wells span multiple geological formations including Montney, Doig, Charlie Lake, Baldonnel, Debolt, Golata, and Wilrich. Each production curve is normalized to 100 monthly values after per-curve min-max scaling. We organize wells into nine formation-based clusters ranging from 51 to 16,216 wells, enabling evaluation across a wide range of dataset sizes. The Montney formation appears at two scales: a 201-well development subset used for data augmentation experiments, and the complete 5,819-well formation set, enabling evaluation of within-formation scaling behavior. All error metrics reported in this paper are computed on these normalized curves (values in $[0, 1]$); an MSE of 0.01 corresponds to an RMSE of 0.10, or roughly 10% of the full production range.

### 4.3. Evaluation Metrics

We report Mean Squared Error (MSE) as the primary metric, supplemented by Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). MSE and RMSE are standard choices in production forecasting [11, 21] because they disproportionately penalize large prediction errors, which is desirable in reservoir management where severely inaccurate forecasts carry outsized economic consequences. RMSE additionally expresses error in the original data units, improving interpretability [27]. MAE provides a complementary perspective by weighting all errors equally regardless of magnitude, offering robustness to outliers [28]. Reporting both squared-error and absolute-error metrics enables distinguishing methods that reduce typical errors from those that specifically mitigate catastrophic mispredictions.

## 5. Results and Discussion

Augmentation Comparison. Figure 1 shows that IDW achieves the largest MSE reduction for both regressors: 6.0% for RF (at 5$\times$) and 13.2% for GB (at 10$\times$), compared to 4.0%/7.1% for TimeGAN and 1.0%/3.9% for TimeVAE. The average improvement across both models is 9.6% for IDW versus 5.5% for TimeGAN and 2.5% for TimeVAE, representing a 1.7$\times$ and 3.8$\times$ advantage, respectively.

Table 1 reveals that IDW maintains stable performance across augmentation factors (optimal at 5–10$\times$), while TimeGAN and TimeVAE degrade at higher factors: TimeGAN MSE increases from 0.0066 at 1.5$\times$ to 0.0079 at 20$\times$ for RF. This pattern suggests that generative models introduce increasingly unrealistic samples at higher augmentation rates, whereas IDW spatial interpolation preserves the structure of real production data. Gradient Boosting benefits more from augmentation than Random Forest across all methods,

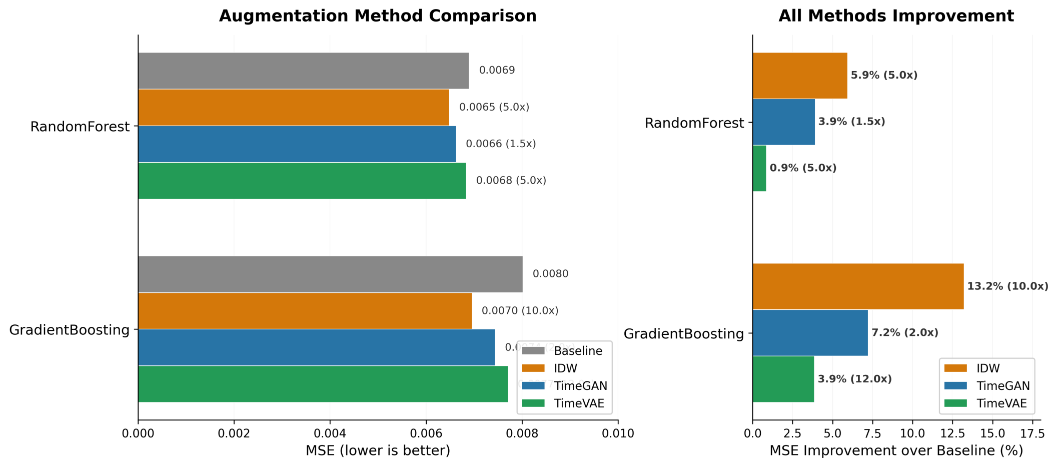


*Figure 1.* Data augmentation comparison on 201 Montney wells. **Left:** MSE across methods for Random Forest and Gradient Boosting, with augmentation factor in parentheses. **Right:** MSE improvement over the no-augmentation baseline for all three methods. IDW consistently outperforms TimeGAN and TimeVAE across both models.

*Table 1.* MSE across augmentation factors. Best result per method shown in **bold**.

| Model | Method | 1.5× | 2× | 3× | 5× | 8× | 10× | 12× | 20× |
|---|---|---|---|---|---|---|---|---|---|
| RF | IDW | 0.0071 | 0.0067 | 0.0070 | **0.0065** | 0.0067 | 0.0068 | 0.0067 | 0.0066 |
| | TimeGAN | **0.0066** | 0.0068 | 0.0070 | 0.0069 | 0.0075 | 0.0071 | 0.0077 | 0.0079 |
| | TimeVAE | 0.0069 | 0.0071 | 0.0069 | **0.0068** | 0.0075 | 0.0072 | 0.0077 | 0.0077 |
| GB | IDW | 0.0080 | 0.0078 | 0.0073 | 0.0075 | 0.0072 | **0.0070** | 0.0074 | 0.0075 |
| | TimeGAN | 0.0075 | **0.0074** | 0.0077 | 0.0075 | 0.0076 | 0.0077 | 0.0078 | 0.0077 |
| | TimeVAE | 0.0078 | 0.0078 | 0.0079 | 0.0077 | 0.0078 | 0.0078 | **0.0077** | 0.0080 |

*Note:* Baseline MSE (no augmentation): RF = 0.0069, GB = 0.0080.

likely because its sequential error-correction mechanism is better able to exploit augmented training samples.

These results establish a key insight: domain-aware augmentation outperforms domain-agnostic generative approaches for production forecasting. This motivates incorporating even stronger domain knowledge, specifically physics-based decline constraints, directly into the generative model.

We evaluate Physics-SIMS-TS against SIMS-TS and IDW+RF across nine formation-based clusters spanning three orders of magnitude in dataset size.

Table 2 shows that Physics-SIMS-TS achieves the lowest MSE and RMSE on all nine clusters, with MSE improvements ranging from 1.9× (Montney, 5,819 wells) to 6.2× (Golata, 53 wells) over the IDW+RF baseline, averaging 3.4× across all clusters. For MAE, Physics-SIMS-TS is best on five clusters while SIMS-TS is marginally better on four; this difference arises because physics constraints primarily reduce large prediction errors that are penalized more heavily by squared-error metrics (MSE, RMSE) than by absolute-error metrics (MAE). Standard SIMS-TS also substantially outperforms IDW+RF on every cluster across all metrics, confirming that diffusion-based generation is superior to traditional machine learning with augmentation even without physics constraints. Physics-SIMS-TS further improves upon SIMS-TS on all clusters for MSE and RMSE, demonstrating the additive benefit of physics guidance.

*Table 2.* Comparison of IDW+Random Forest (baseline), SIMS-TS, and Physics-SIMS-TS across geological formation clusters. Best value per cluster per metric shown in **bold**. Gain measures Physics-SIMS-TS MSE improvement over the baseline. The Montney formation appears at two scales: a 201-well development subset and the complete 5,819-well formation set.

| | | MSE (↓) | | | MAE (↓) | | | RMSE (↓) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Formation | Wells | IDW+RF | SIMS | P-SIMS | IDW+RF | SIMS | P-SIMS | IDW+RF | SIMS | P-SIMS | Gain |
| Wilrich | 51 | 0.0110 | 0.0053 | **0.0051** | 0.0698 | **0.0370** | 0.0379 | 0.1049 | 0.0730 | **0.0712** | 2.2× |
| Golata | 53 | 0.0376 | 0.0066 | **0.0060** | 0.1483 | 0.0534 | **0.0533** | 0.1938 | 0.0812 | **0.0777** | 6.2× |
| Debolt | 188 | 0.0300 | 0.0078 | **0.0064** | 0.1247 | 0.0584 | **0.0554** | 0.1733 | 0.0885 | **0.0798** | 4.7× |
| Montney | 201 | 0.0111 | 0.0049 | **0.0041** | 0.0651 | **0.0433** | 0.0439 | 0.1052 | 0.0698 | **0.0643** | 2.7× |
| Baldonnel | 293 | 0.0218 | 0.0053 | **0.0045** | 0.1022 | 0.0464 | **0.0463** | 0.1476 | 0.0730 | **0.0672** | 4.8× |
| Charlie Lake | 690 | 0.0223 | 0.0073 | **0.0065** | 0.1043 | 0.0564 | **0.0553** | 0.1493 | 0.0853 | **0.0804** | 3.5× |
| Doig | 1,390 | 0.0150 | 0.0066 | **0.0063** | 0.0816 | **0.0515** | 0.0519 | 0.1223 | 0.0815 | **0.0794** | 2.4× |
| Montney | 5,819 | 0.0127 | 0.0084 | **0.0068** | 0.0773 | 0.0519 | **0.0500** | 0.1128 | 0.0914 | **0.0826** | 1.9× |
| All Wells | 16,216 | 0.0173 | 0.0084 | **0.0075** | 0.0883 | **0.0555** | 0.0556 | 0.1317 | 0.0914 | **0.0865** | 2.3× |

*P-SIMS wins: MSE 9/9, RMSE 9/9, MAE 5/9. Average MSE gain: 3.4×*

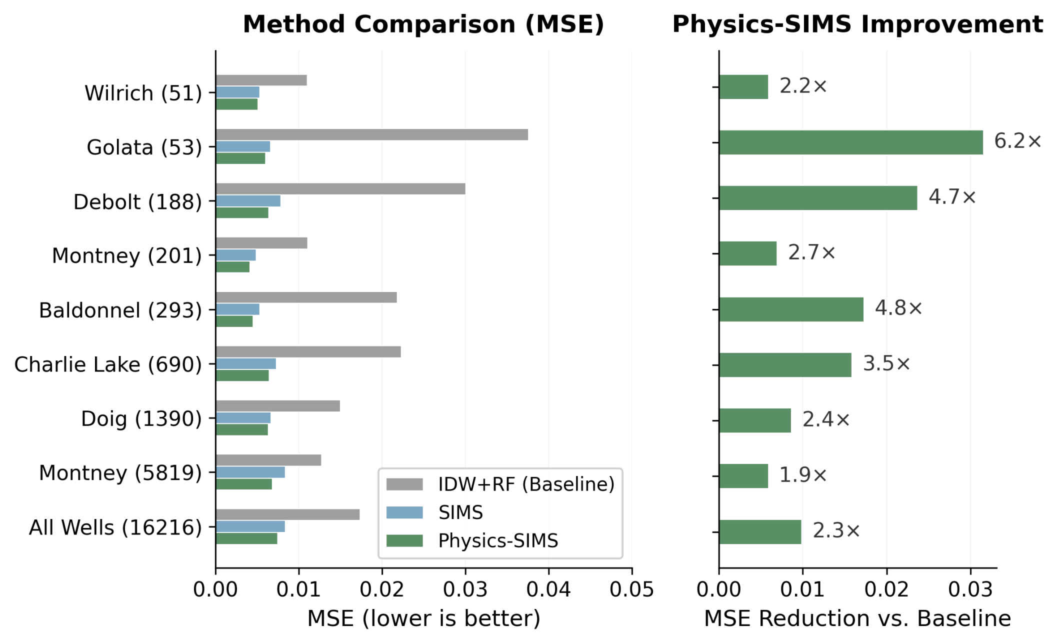


*Figure 2.* Comparison across all nine clusters. **Left:** MSE for each method per cluster (lower is better). IDW+RF (gray) consistently has the highest error, while Physics-SIMS-TS (green) achieves the lowest. **Right:** Absolute MSE reduction and improvement factor of Physics-SIMS-TS over the IDW+RF baseline. The largest gains occur on small clusters where physics constraints provide the strongest regularization.

Figure 2 visualizes these results. The left panel shows that the MSE gap between methods is largest on small, geologically distinct formations: Golata (53 wells, 6.2×), Baldonnel (293 wells, 4.8×), and Debolt (188 wells, 4.7×). The right panel highlights that absolute MSE improvement is also largest on these clusters, confirming that physics constraints provide the greatest benefit where data is most limited. Notably, even on the full 16,216-well dataset, Physics-SIMS-TS reduces MSE by 2.3× (from 0.0173 to 0.0075), demonstrating that physics constraints remain beneficial even with abundant training data.

**Discussion.**

The consistent superiority of Physics-SIMS-TS reflects a coherent progression: spatial domain knowledge (IDW) outperforms generic generation (TimeGAN/TimeVAE), and temporal physics knowledge (Arps constraints) outperforms spatial knowledge alone. Production forecasting from 20% of the timeline is inherently ill-posed, as many trajectories are consistent with early observations. Physics constraints reduce this ambiguity by encoding known decline behavior, and monotonicity projection eliminates physically implausible predictions where production increases.

The strongest improvements on small clusters (51–293 wells) arise because physics constraints serve as a regularizer. With limited data, the diffusion model cannot learn formation-specific decline patterns from examples alone; physics knowledge compensates by constraining the hypothesis space to physically plausible trajectories. With abundant data (5,819–16,216 wells), the model can partially learn these patterns directly, so the marginal benefit of physics constraints is smaller but still meaningful.

## 6. Conclusion

We presented Physics-SIMS-TS, a framework combining self-improving diffusion models with physics-based decline-curve constraints for oil and gas production forecasting. By adapting the SIMS negative guidance mechanism to time series and incorporating Arps decline dynamics through gradient guidance and monotonicity projection, we achieve 1.9–6.2× MSE improvement over traditional baselines across dataset sizes from 51 to 16,216 wells. Our data augmentation experiments further demonstrate that domain-specific knowledge provides stronger inductive bias than generic generative approaches, with the largest benefits on small datasets where physics constraints most effectively regularize the learning problem.

A limitation of our approach is the monotonic decline assumption, which does not hold for wells with workovers, recompletions, or other interventions that temporarily increase production. Additionally, for small clusters, we fine-tune a base model trained on all BC formations, introducing a cross-formation transfer assumption. Our evaluation is also restricted to a single geographic domain (British Columbia gas wells); validating Physics-SIMS-TS on production data from other basins and play types (e.g., the Permian Basin or Bakken Formation) would provide stronger evidence of generalizability. Similarly, evaluation on standard time series benchmarks such as the ETT, Weather, and Exchange datasets would clarify how the framework performs outside the production forecasting domain. The physics guidance mechanism is currently heuristic; future work could develop theoretical analysis of the interaction between gradient guidance and the diffusion sampling process to better understand convergence properties and guidance strength selection. Finally, comparison against additional baselines, including state-of-the-art transformer-based forecasting models and zero-shot foundation models, would further contextualize the contribution. Future work could also incorporate event-conditional models to handle non-monotonic production profiles.

## Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) [Alliance Grant #ALLRP 567562-2021 sponsored by British Columbia Energy Regulator (BC-ER)]. Part of the research was performed in the Computer-Aided Convex Analysis (CA2) laboratory funded by a Leaders Opportunity Fund [LOF, John R. Evans Leaders Fund, Funding for research infrastructure] from the Canada Foundation for Innovation (CFI) and by the British Columbia Knowledge Development Fund (BCKDF).

The authors thank BC-ER, especially Logan Gray and Jason Gregg, for providing data and technical expertise. The authors used AI-based tools (Claude, Anthropic) for grammar editing and code debugging. The authors take full responsibility for all content in this paper.

## References


[1] J. J. Arps. "Analysis of Decline Curves". In: *Transactions of the AIME* 160.01 (1945), pp. 228–247. DOI: 10.2118/945228-G.

[2] BC Energy Regulator. *Orphan Sites*. Accessed: 2025-02-07. 2025. URL: https://www.bc-er.ca/what-we-regulate/oil-gas/orphan-sites/.

[3] Orphan Well Association. *Annual Report 2024/2025*. Tech. rep. Calgary, AB, Canada: Orphan Well Association, 2025. URL: https://www.orphanwell.ca/.

[4] M. Kang, J. Boutot, A. McVay, B. Stackhouse, P. Suska, B. Pavilonis, L. M. Golston, S. Mao, and A. R. Brandt. "Documented Orphaned Oil and Gas Wells Across the United States". In: *Environmental Science & Technology* 57.3 (2023), pp. 1089–1098. DOI: 10.1021/acs.est.2c03268.

[5] J. Ho, A. Jain, and P. Abbeel. "Denoising Diffusion Probabilistic Models". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 6840–6851.

[6] X. Yuan and Y. Qiao. "Diffusion-TS: Interpretable Diffusion for General Time Series Generation". In: *The Twelfth International Conference on Learning Representations (ICLR)*. 2024. URL: https://openreview.net/forum?id=4h1apFjO99.

[7] M. Kollovieh, A. F. Ansari, M. Bohlke-Schneider, J. Zschiegner, H. Wang, and Y. Wang. "Predict, Refine, Synthesize: Self-Guiding Diffusion Models for Probabilistic Time Series Forecasting". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 36. 2023, pp. 28341–28364.

[8] S. Alemohammad, A. I. Humayun, S. Agarwal, J. Collomosse, and R. Baraniuk. *Self-Improving Diffusion Models with Synthetic Data*. 2024. arXiv: 2408.16333 [cs.LG].

[9] P. P. Valkó. "Assigning Value to Stimulation in the Barnett Shale: A Simultaneous Analysis of 7000 Plus Production Histories and Well Completion Records". In: *SPE Hydraulic Fracturing Technology Conference*. Society of Petroleum Engineers. 2009. DOI: 10.2118/119369-MS.

[10] A. N. Duong. "Rate-Decline Analysis for Fracture-Dominated Shale Reservoirs". In: *SPE Reservoir Evaluation & Engineering* 14.03 (2011), pp. 377–387.

[11] A. Tadjer, A. Hong, and R. B. Bratvold. "Machine Learning Based Decline Curve Analysis for Short-Term Oil Production Forecast". In: *Energy Exploration & Exploitation* 39.5 (2021), pp. 1747–1769.

[12] J.-X. Han, L. Xue, Y.-S. Wei, Y.-D. Qi, J.-L. Wang, Y.-T. Liu, and Y.-Q. Zhang. "Physics-Informed Neural Network-Based Petroleum Reservoir Simulation with Sparse Data Using Domain Decomposition". In: *Petroleum Science* 20.6 (2023), pp. 3450–3460. DOI: 10.1016/j.petsci.2023.10.019.

[13] J. Bi, J. Li, K. Wu, Z. Chen, S. Chen, L. Jiang, D. Feng, and P. Deng. "A Physics-Informed Spatial-Temporal Neural Network for Reservoir Simulation and Uncertainty Quantification". In: *SPE Journal* 29.4 (2024), pp. 2026–2043.

[14] J. Yoon, D. Jarrett, and M. van der Schaar. "Time-Series Generative Adversarial Networks". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 32. 2019.

[15] A. Desai, C. Freeman, Z. Wang, and I. Beaver. *TimeVAE: A Variational Auto-Encoder for Multivariate Time Series Generation*. 2021. arXiv: 2111.08095 [cs.LG].

[16] Y. Ang, Q. Huang, Y. Bao, A. K. Tung, and Z. Huang. "TSGBench: Time Series Generation Benchmark". In: *Proceedings of the VLDB Endowment* 17.3 (2024), pp. 305–318.

[17] Y. Tashiro, J. Song, Y. Song, and S. Ermon. "CSDI: Conditional Score-Based Diffusion Models for Probabilistic Time Series Imputation". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 34. 2021, pp. 24804–24816.

[18] Q. Wen, L. Sun, F. Yang, X. Song, J. Gao, X. Wang, and H. Xu. "Time Series Data Augmentation for Deep Learning: A Survey". In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*. 2021, pp. 4653–4660.

[19] B. K. Iwana and S. Uchida. "An Empirical Survey of Data Augmentation for Time Series Classification with Neural Networks". In: *PLOS ONE* 16.7 (2021), e0254841.

[20] D. Shepard. "A Two-Dimensional Interpolation Function for Irregularly-Spaced Data". In: *Proceedings of the 1968 23rd ACM National Conference*. New York, NY, USA: ACM, 1968, pp. 517–524. DOI: 10.1145/800186.810616.

[21] D. Fan, S. Lai, H. Sun, Y. Yang, C. Yang, N. Fan, and M. Wang. "Review of Machine Learning Methods for Steady State Capacity and Transient Production Forecasting in Oil and Gas Reservoir". In: *Energies* 18.4 (2025), p. 842. DOI: 10.3390/en18040842.

[22] S. Alemohammad, J. Casco-Rodriguez, L. Luzi, A. I. Humayun, H. Babaei, D. LeJeune, A. Siahkoohi, and R. G. Baraniuk. "Self-Consuming Generative Models Go MAD". In: *The Twelfth International Conference on Learning Representations (ICLR)*. 2024.

[23] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal. "AI Models Collapse When Trained on Recursively Generated Data". In: *Nature* 631.8022 (2024), pp. 755–759.

[24] R. E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk. *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. New York: Wiley, 1972.

[25] J. Song, C. Meng, and S. Ermon. "Denoising Diffusion Implicit Models". In: *The Ninth International Conference on Learning Representations (ICLR)*. 2021. URL: https://openreview.net/forum?id=St1giarCHLP.

[26] BC Oil and Gas Commission. *Resource Assessment of the Montney Formation*. Tech. rep. BC Oil and Gas Commission, 2020.

[27] T. Chai and R. R. Draxler. "Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? Arguments Against Avoiding RMSE in the Literature". In: *Geoscientific Model Development* 7.3 (2014), pp. 1247–1250. DOI: 10.5194/gmd-7-1247-2014.

[28] R. J. Hyndman and A. B. Koehler. "Another Look at Measures of Forecast Accuracy". In: *International Journal of Forecasting* 22.4 (2006), pp. 679–688. DOI: 10.1016/j.ijforecast.2006.03.001.