

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. Reference herein to any social initiative (including but not limited to Diversity, Equity, and Inclusion (DEI); Community Benefits Plans (CBP); Justice 40; etc.) is made by the Author independent of any current requirement by the United States Government and does not constitute or imply endorsement, recommendation, or support by the United States Government or any agency thereof.

Artificial Intelligence and Post-Deployment Model Drift in Grid-based Applications

INL/RPT-26-94128
Revision 0

OCTOBER 2026

Lauren Ortiz, Patience Yockey, Bradley Marx,
Remy Stolworthy, Zachary Dalton, Jeremy Jones,
and Juliana Ocampo Giraldo

Idaho National Laboratory

Alex Jenkins

AVEVA



DISCLAIMER

This information was prepared as an account of work sponsored by an agency of the U.S. Government. Neither the U.S. Government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness, of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. References herein to any specific commercial product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the U.S. Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the U.S. Government or any agency thereof.

Artificial Intelligence and Post-Deployment Model Drift in Grid-based Applications

**Lauren Ortiz, Patience Yockey, Bradley Marx,
Remy Stolworthy, Zachary Dalton, Jeremy Jones,
and Juliana Ocampo Giraldo
Idaho National Laboratory
Alex Jenkins
AVEVA**

October 2026

**Idaho National Laboratory
Idaho Falls, Idaho 83415**

<http://www.inl.gov>

**Prepared for the
U.S. Department of Energy
Office of Electricity
Under DOE Idaho Operations Office
Contract DE-AC07-05ID14517**

Page intentionally left blank

ABSTRACT, SUMMARY, FOREWORD, AND ACKNOWLEDGMENTS

Artificial intelligence (AI) and machine learning (ML) are increasingly integrated in grid planning, operations, and asset management as part of the broader digital transformation of the United States (U.S.) electric sector. Utilities are adopting AI-enabled analytics to support such functions as anomaly detection, predictive maintenance, forecasting, outage management, and decision support.^a When responsibly deployed, these applications can provide measurable benefits—improving reliability, reducing unplanned outages, and more efficiently using infrastructure—particularly as grid complexity grows due to electrification, distributed energy assets, and large AI-driven data-center loads. However, while many AI/ML models perform well at deployment, their validity cannot be assumed over time. Changes in load composition, asset behavior, weather patterns, and operating conditions can cause deployed models to diverge from real-world system behavior, a phenomenon known as model drift. When this post-deployment model drift goes unmeasured, operators risk relying on degraded AI/ML models in grid applications where accuracy directly affects economic efficiency and operational reliability.

This whitepaper presents a quantitative, model-agnostic methodology for measuring AI/ML model drift and estimating model lifetime using Kullback–Leibler (KL) divergence. It is important to note that KL divergence is one of several approaches to drift detection; its selection here is motivated by the specific operational constraints of grid operational technology (OT) environments, and its value is realized in combination with complementary mechanisms, specifically, cumulative-sum (CUMSUM) sequential accumulation and population stability index (PSI)-derived thresholds, rather than as a standalone signal. By comparing the evolving statistical relationship between predicted and observed values, the approach identifies low-, moderate-, and high-drift conditions using thresholds informed by PSI guidance. The method provides operators with objective, data-driven indicators for when AI/ML models remain reliable and when retraining or replacement is required. The methodology is demonstrated using a long-term load-forecasting case study for the Texas South-Central (SCENT) region, showing how initially accurate AI/ML models progressively degrade as system conditions evolve and how the KL-divergence signal distinguishes structural degradation from normal, expected forecasting error. This methodology was tested using the AVEVA Predictive Analytics Operational Profile Topology Information and Classification System algorithm (a commercially available algorithm designed for industrial use cases), and common open-source algorithms, including seasonal autoregressive integrated moving average, linear regression, and multilayer perceptron.

^a M. Culler, P. Yockey, R. Stolworthy E. Stewart, “Adoption of AI in the Utility T&D Sector: Use Cases, Consequence, Assessment, and Benefits,” Idaho National Laboratory, August 2025, <https://www.osti.gov/servlets/purl/2997112>.

The methodology presented in this whitepaper—using KL divergence combined with CUMSUM sequential analysis and PSI thresholding—represents one viable, structured approach to measuring AI/ML model drift in grid-based applications. It is not intended to be, nor should it be interpreted as, the exclusive or universally superior method for drift detection. A range of complementary and alternative techniques exist, including Wasserstein distance, Kolmogorov-Smirnov statistics, adaptive windowing methods (ADWIN), process-control-based detectors (drift-detection method [DDM] and early drift-detection method [EDDM]), and hybrid approaches combining statistical monitoring with physics-based validation. The authors encourage practitioners to evaluate multiple methods against their specific operational context, data characteristics, risk tolerance, and retraining capacity.

KL divergence is presented here because it satisfies a specific set of operational requirements common in electric-grid OT environments, detailed in Section 3.3, and because its combination with CUMSUM and PSI thresholds produces a tractable, interpretable, and model-agnostic life-cycle management framework. The limitations of this approach are explicitly addressed in Section 5.

The research was conducted under the Artificial Intelligence Management and Research for Advanced Networked Testbed Hub (AMARANTH) program within the Center for Security and Digital Energy Technologies, funded by the U.S. Department of Energy's (DOE's) Office of Electricity and performed by Idaho National Laboratory. The work supports DOE's grid-modernization and resilience objectives by enabling trustworthy, risk-aware AI deployment in critical infrastructure. By treating AI/ML models as life-cycle-dependent operational assets, this approach helps utilities and decision-makers reduce risk, improve transparency, and sustain reliable grid operations as AI adoption expands. This work is complementary to, but distinct from, the Operational Resilience for Critical Hardware and Infrastructure under Drift (ORCHID) program, which addresses understanding why drift happens, supply-chain security in AI/ML systems, and adversarial-drift scenarios. The AMARANTH methodology focuses on life-cycle management and operational-drift thresholds under normal operating conditions; readers seeking treatment of adversarial scenarios, including how a compromised sensor stream could produce KL-divergence patterns that mimic benign drift, are directed to the ORCHID program.

CONTENTS

1.	INTRODUCTION.....	1
1.1.	Types of Artificial Intelligence	2
2.	BACKGROUND.....	3
2.1.	The Risk of Unmeasured AI/ML Model Drift and Validity Degradation in Power System Applications	3
2.2.	AI and ML Task Domains in the Electric Sector	5
3.	QUANTIFYING AI/ML MODEL LIFETIME and POST-DEPLOYMENT DRIFT	6
3.1.	Required Inputs and Assumptions	6
3.1.1.	Model and Data Requirements.....	6
3.1.2.	Platform and Model Scope: AVEVA Predictive Analytics Software.....	7
3.2.	Data Preparation and Distribution Modeling	7
3.3.	AI/ML Model-Drift Quantification.....	8
3.3.1.	Drift-Metric Selection	8
3.3.2.	Relationship Between KL-Divergence Thresholds and Operational Reality.....	9
3.3.3.	Time-Resolved Drift Calculation.....	10
3.4.	Cumulative Drift Analysis and Thresholding	10
3.5.	AI/ML-Model Lifetime Estimation.....	11
4.	USE CASES: IMPACTS OF UNMEASURED AI/ML MODEL DRIFT.....	13
4.1.	Load Forecasting.....	13
4.2.	Predictive Maintenance.....	14
5.	LIMITATIONS	15
6.	CONCLUSION	16
7.	REFERENCES.....	17
	Appendix A Introduction to Machine Learning and Artificial Intelligence.....	19
	Appendix B Kullback-Leibler Divergence on Different Architectures	20

FIGURES

Figure 1. Actual vs. forecasted power usage in the SCENT region (2017–2025), with background shading indicating low- (green), moderate- (yellow), and high-drift (red) periods.	4
Figure 2. Daily KL divergence for AVEVA Operational Profile Topology Information and Classification System on SCENT load data (2017–2025).	4
Figure 3. Weekly KL divergence for AVEVA Operational Profile Topology Information and Classification System on SCENT load data (2017–2025).	4
Figure 4. Flow chart demonstrating the method used to determine model drift and the latest point at which a model should be retrained.	11
Figure B-1. KL-divergence and drift zones for SARIMA load forecasting with weekly aggregation.	20
Figure B-2. KL-divergence and drift zones for linear-regression load forecasting with weekly aggregation.	21
Figure B-3. KL-divergence and drift zones for MLP load forecasting with weekly aggregation.	21

TABLES

Table 1. Summary of algorithm drift performance on the ERCOT dataset using weekly aggregation.	11
---	----

ACRONYMS

ADWIN	adaptive windowing method
AI	Artificial intelligence
AMARANTH	Artificial Intelligence Management and Research for Advanced Networked Testbed Hub
CSDet	Center for Security and Digital Energy Technologies
CUMSUM	cumulative-sum
DDM	drift-detection method
DL	Deep learning
DOE	Department of Energy
EDDM	early drift-detection method
ERCOT	Electric Reliability Council of Texas
GenAI	Generative AI
KL	Kullback-Leibler
KS	Kolmogorov-Smirnov
MAPE	mean absolute percentage error
ML	machine learning
MLP	multilayer perceptron
OE	Office of Electricity
OPTiCS	Operational Profile Topology Information and Classification System
ORCHID	Operational Resilience for Critical Hardware and Infrastructure under Drift
OT	operational technology
PSI	population stability index
SARIMA	seasonal autoregressive integrated moving average
SCENT	South-Central

Page intentionally left blank

Artificial Intelligence and Post-Deployment Model Drift in Grid-based Applications

1. INTRODUCTION

As the United States (U.S.) electric sector undergoes rapid transformation, artificial intelligence (AI) and machine learning (ML) are emerging as strategic enablers of operational resilience and efficiency. AI is no longer a future consideration, but a present top-level focus: AI is viewed as a cost-effective tool to optimize grid operations and improve energy efficiency within the electric sector.¹

This white paper is designed to support decision making by offering a structured methodology for evaluating AI/ML integration into critical infrastructure. It addresses not only the technical capabilities of AI systems but also the organizational and operational risks associated with their deployment. Additionally, the paper explores how to quantify AI/ML model lifetime and drift, assess post-deployment trustworthiness, and align AI applications with long-term grid-modernization goals. By grounding the discussion in proven methodologies and real-world use cases, this paper aims to equip technical teams with the insights needed to guide responsible AI adoption.

On the scope and limitations of the methodology presented herein. The Kullback–Leibler (KL) divergence methodology, paired with cumulative-sum (CUMSUM) sequential accumulation and population stability index (PSI)-derived thresholds, represents one viable, structured approach to measuring AI/ML model drift in grid-based applications. It is not intended to be, nor should it be interpreted as, the exclusive or universally superior method for drift detection. A range of complementary and alternative techniques exist, including Wasserstein distance, Kolmogorov-Smirnov (KS) statistics, adaptive windowing methods (ADWIN), process-control-based detectors (drift-detection method [DDM], and early drift-detection method [EDDM]), and hybrid approaches combining statistical monitoring with physics-based validation. The authors encourage practitioners to evaluate multiple methods against their specific operational context, data characteristics, risk tolerance, and retraining capacity.

KL-divergence is presented here because it satisfies a specific set of operational requirements common in electric-grid operational technology (OT) environments, detailed in Section 3.3, and because its combination with CUMSUM and PSI thresholds produces a tractable, interpretable, and model-agnostic life-cycle management framework. Critically, drift measurement using this or any other single metric is a necessary but not sufficient condition for comprehensive AI/ML model assurance. A full performance assurance program should also include pointwise error metrics (mean absolute percentage error [MAPE], root mean square error [RMSE]), calibration checks to verify that confidence intervals are accurate, feature-importance stability monitoring to detect shifts in those input variables the model relies upon, data-quality monitoring to ensure inputs arrive cleanly and on time, and operational-outcome tracking to assess whether decisions informed by the model are producing better results. This paper does not develop a complete multimetric framework; it presents drift measurement as one structured component of that broader ecosystem. The limitations of the KL-divergence approach specifically are addressed in Section 5.

This research was conducted under the Center for Security and Digital Energy Technologies (CSDET) Artificial Intelligence Management and Research for Advanced Networked Testbed Hub (AMARANTH) program funded by the U.S. Department of Energy (DOE) Office of Electricity (OE) and performed by Idaho National Laboratory (INL). The work supports DOE OE’s mission to strengthen grid resilience and operational reliability by advancing secure, trustworthy applications of AI and ML for the electric sector. By integrating technical rigor with risk-aware deployment frameworks, this effort directly complements national grid-modernization initiatives and a utility’s decision-making priorities. Additionally, this work employed INL’s High Performance Computing resources at the Collaborative Computing Center, supported by the DOE Office of Nuclear Energy and the Nuclear Science User Facilities. As part of this effort, AVEVA, a global industrial software company specializing in engineering, operational, and performance-intelligence solutions for asset-intensive industries, including energy and utilities, provided key access to their AVEVA Predictive Analytics (AVEVA PA) software platform, enabling INL to demonstrate the proposed AI/ML drift methodology on a commercially deployed AI/ML system and reinforcing the value of industry partnerships in advancing secure AI solutions for the grid.

Note on terminology: Throughout this paper, “AI/ML model” refers exclusively to ML (statistical or neural-network-based) systems trained on historical data to generate predictions or classifications. This is distinct from physics-based grid-simulation models (e.g., power-flow solvers, transient tools), which operate on first-principles equations and physics rather than learned patterns. The term “model” alone is avoided to prevent ambiguity.

1.1. Types of Artificial Intelligence

AI and ML are not new concepts; however, the increase in mainstream popularity of AI and ML leads to conflation of their definitions. AI is the broadest category and includes code and instructions that enable computers to emulate human behaviors and learn from pattern recognition. ML is a subset of AI that uses algorithms to detect patterns in large datasets. Deep learning, a subset of ML, uses neural networks on large datasets to detect more-complex patterns and extract unique features from the large dataset. Generative AI is a subset of deep learning that uses transformers to encode and decode data and involves the generation of new content after the generative AI has been trained on large datasets. For a deeper dive on the different types of AI, terminology, and definitions, see Appendix A.

2. BACKGROUND

Over the last few years, AI and ML have experienced significant interest and growth in usage. Over 80% of respondents in a 2025 McKinsey study see AI as a means to drive business growth and innovation within their organization.² AI and ML are tools for solving complex problems and improving efficiency across many industries. As these technologies continue to evolve, AI/ML applications in the electric sector are expanding, offering new ways to optimize operations, improve reliability, and support smarter infrastructure. The following sections provide background on AI development and terminology as well as describe how different types of AI and ML are currently applied within the electric sector. This context informs later discussions on deployment considerations, trust, and the need for quantitative assessment of AI/ML performance over time.

2.1. The Risk of Unmeasured AI/ML Model Drift and Validity Degradation in Power System Applications

The development of AI/ML products at high-technology-readiness (i.e., TRL 7–9) for the electric sector, combined with increased government and industry investment in grid-focused AI/ML, accelerates the adoption of these AI/ML models across electric-grid planning and operations.¹ With the adoption of AI in grid applications, ranging from operational decision-support tools to asset-health monitoring, the concept of AI/ML model drift becomes operationally significant. Part of any AI/ML model's robustness is its ability to generalize changes in the underlying physical system, such as equipment upgrades, load growth, or shifting weather patterns. However, even a well-designed AI/ML model, trained on representative historical data, will eventually diverge from current-system behavior as conditions evolve, requiring retraining or controls to maintain reliability.

The distinction between forecasting error and AI/ML model drift must be established prior to proceeding; these are related but fundamentally different phenomena:

Forecasting error is the expected, day-to-day variability between an AI/ML model's prediction and actual observed values. Error is inherent to any probabilistic system and is driven by factors outside the AI/ML model's control, such as unanticipated weather events, sudden load shifts, equipment outages, etc. Grid operators routinely manage this variability through reserve margins and operational adjustments.

AI/ML model drift is a systematic, cumulative increase in error over time as the AI/ML model's training distribution diverges from current operating conditions. Unlike episodic forecast error, which produces temporary spikes in divergence that decay, structural drift produces a persistent upward trend in cumulative divergence. KL-divergence accumulated with CUMSUM is used to detect this upward trend. The KL-divergence methodology is specifically a leading indicator of degradation through detecting distributional shifts before short-term forecasting error measures such as MAPE increases, providing an advance warning ahead of operational impact. However, it is important to note that KL divergence is most effective when used in conjunction with these complementary metrics: MAPE or other pointwise error measures provide operational grounding, CUMSUM provides the temporal accumulation needed to separate signal from noise, and PSI-derived thresholds provide calibrated decision boundaries. No single metric should be relied upon in isolation.

The distinction between forecasting error and AI/ML model drift is exhibited in the Texas South-Central (SCENT) case study described in Section 3, where the divergence shown in Figure 1 and the KL drift signal in Figure 2 and Figure 3 are not directly attributable to any single operational event. The cumulative, monotonically increasing nature of the KL CUMSUM signal persists across multiple years as the signature of structural AI/ML-model drift instead of episodic forecasting noise.

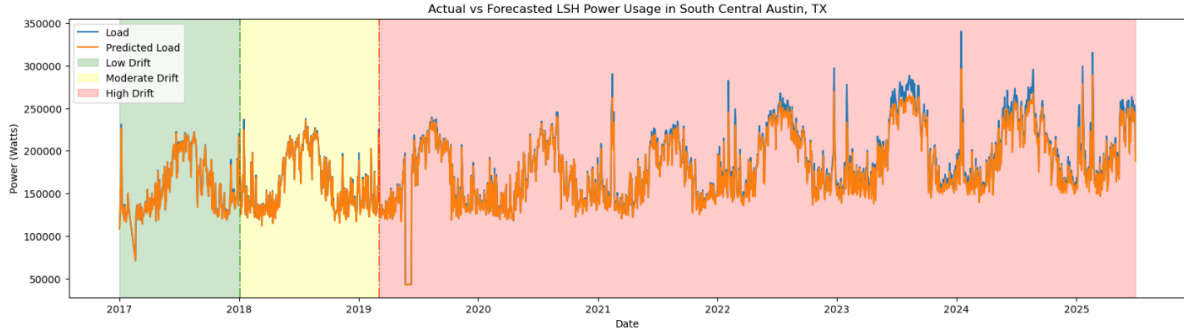


Figure 1. Actual vs. forecasted power usage in the SCENT region (2017–2025), with background shading indicating low- (green), moderate- (yellow), and high-drift (red) periods.

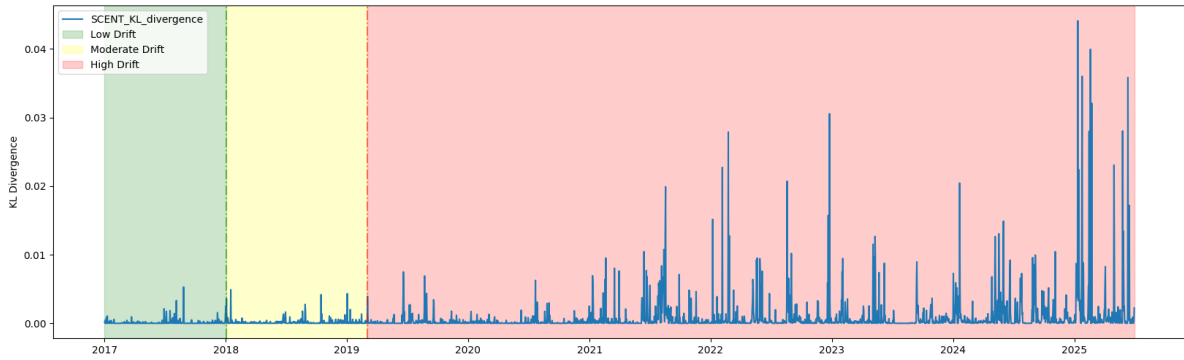


Figure 2. Daily KL divergence for AVEVA Operational Profile Topology Information and Classification System on SCENT load data (2017–2025). Background shading indicates low- (green), moderate- (yellow), and high-drift (red) zones. Individual daily divergence values are typically in the range of 0.001 to 0.05; it is the running sum of KL divergence that crosses the 0.05 and 0.125 thresholds.

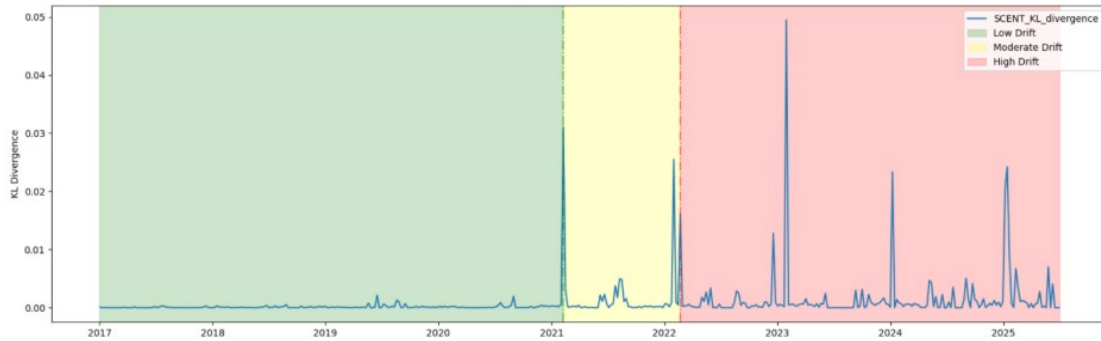


Figure 3. Weekly KL divergence for AVEVA Operational Profile Topology Information and Classification System on SCENT load data (2017–2025). Drift thresholds are reached approximately 2–3 years later than at daily resolution, providing a longer operational window before retraining is required.

AI/ML model drift can be driven by changes to the physical system, such as equipment upgrades or major failures, and retraining AI/ML models allows them to adapt to these changes. Measuring drift is a crucial piece of understanding a model's performance in a production environment and provides operators with assurance of reliable performance and visibility into the model's lifespan to plan when retraining needs to occur. It is important to emphasize, however, that drift measurement is a necessary but not sufficient condition for operational assurance. Operators should treat quantitative drift metrics as one component of a broader performance-monitoring program that includes pointwise-accuracy tracking, data-quality monitoring, and operational-outcome assessment. A quantifiable definition of drift allows the model's current performance to be measured against operator-defined limits, but those limits and the monitoring framework around them must be deliberately designed for each deployment context.

When AI/ML model lifetime has not been measured or quantified, an AI/ML model may be inadvertently used in production beyond its lifespan, causing the AI/ML model to produce inaccurate results. Conversely, the AI/ML model may be retrained or retired and replaced far before it reaches the end of its lifespan, leading to early and unnecessary investment in model replacement. Without quantified AI/ML model drift, operators have no reference for whether an AI/ML model has drifted out of tolerance.

2.2. AI and ML Task Domains in the Electric Sector

From a grid perspective, when AI/ML models drift or lose validity, their outputs may no longer accurately represent grid conditions. Degraded AI/ML models can lead to missed indicators of asset degradation or failure; incorrect forecasts of load, generation, or system behavior; and/or improperly prioritized maintenance or restoration actions.

AI and ML usage in the electric sector is increasingly considered a way to solve operational challenges and increase efficiency across different use cases.¹ AI and ML are estimated to be able to enable up to 175 gigawatts of additional line transmission capacity across existing lines through optimization of transmission and demand profiles.³ Electric utilities currently use, or plan to use, AI and ML models to support real-time operations, asset management, maintenance, and business and customer operations. As of 2024, up to 75% of electric utilities report using AI/ML for maintenance and asset operation.³

Prior INL CSDET analysis of the adoption of AI in the electric sector identified four key task domains: detection, prediction, optimization and control, and business-customer applications.¹ For the methodology outlined in this whitepaper, we focus on two of these task domains: detection and prediction.

- **Detection** applications identify abnormal system behavior that may indicate faults, asset degradation, or cyber-physical anomalies, using operational data such as supervisory control and data acquisition, phasor measurement units, advanced metering infrastructure, and sensor streams to enhance situational awareness.
- **Prediction** applications forecast future system states, including load, generation, outages, and asset health, supporting proactive maintenance, planning, and grid development.

As AI/ML implementation scales within the electric sector, utilities, vendors, and stakeholders are increasingly interested in the deployment and post-deployment trustworthiness of AI/ML models for long-term usage. Requiring quantitative assessments to increase trust and transparency of AI/ML models that assist with grid functions is crucial to the long-term viability of AI and ML in the electric sector. These assessments also allow engineers and AI/ML developers to communicate and develop a common language to build out controls that ensure safe, reliable, and performant AI-integrated systems.

3. QUANTIFYING AI/ML MODEL LIFETIME and POST-DEPLOYMENT DRIFT

The following proof-of-concept methodology provides a systematic approach to quantifying AI/ML model drift and estimating AI/ML model lifetime for electric grid applications. As AI/ML models increasingly support grid planning, forecasting, and operational decision making, operators require objective, data-driven methods to determine when an AI/ML model remains reliable and when retraining or replacement will be necessary.

It bears repeating that the methodology described here is one structured approach among several viable options. Practitioners working in environments with different data characteristics, model types, or retraining architectures should evaluate whether this approach or a hybrid that incorporates it alongside other drift metrics is the best fit for their context.

The proposed approach (beginning in Section 3.1) evaluates divergence between AI/ML model predictions and observed system behavior across different points in time using KL-divergence statistical metrics (explained in Section 3.1.1). KL divergence was selected because it is AI/ML model-agnostic, computationally lightweight in terms of its theoretical and structural properties (requiring closed-form computation over discretized distributions without iterative optimization), applicable to multivariate and noisy datasets, and requires no knowledge of the underlying parametric distribution, properties which are essential for deploying in OT grid environments. It should be noted that formal benchmarking of KL divergence against alternative methods on representative OT hardware would be needed to substantiate computational performance claims quantitatively, and this is identified as an area for future empirical work. By measuring how predicted and true-value distributions evolve relative to one another, the methodology identifies the onset and progression of AI/ML-model drift in grid operations. Drift severity is categorized using predefined thresholds—low, moderate, and high drift—enabling asset owners and operators to assess AI/ML model health continuously and estimate a practical AI/ML model lifespan. The resulting outputs support proactive life-cycle management, reduce the risk from degraded AI/ML model performance, and improve trust in AI-assisted grid operations.

3.1. Required Inputs and Assumptions

3.1.1. Model and Data Requirements

This methodology requires three primary inputs: a trained AI/ML model (in this case, AVEVA PA), the dataset used to train that model (SCENT forecasting data from 2006 to 2016), and an out-of-sample test dataset that reflects future operational conditions (SCENT forecasting data from 2017 to 2025). The test dataset must include time-stamped observations to support chronological drift analysis. Ground-truth values are required to enable direct comparison between predicted and observed system behavior.

The SCENT region is one of the eight zones defined by the Electric Reliability Council of Texas (ERCOT) for load-forecasting purposes. It encompasses the greater San Antonio metropolitan areas and surrounding communities. The SCENT zone is representative of a load profile with high summer air-conditioning demand, seasonal weather variability, and growing data-center loads. The Austin-San Antonio corridor's specific characteristics are not universally representative of U.S. utility service territories, and findings from this case study should be generalized with caution. Validation across additional regions with different load profiles is identified as a necessary next step before this methodology can be recommended for broad adoption. The SCENT dataset contains hourly observations of load in megawatt-hours from 2006 to 2025. Paired with weather data for the greater San Antonio region—including temperature, humidity, weather conditions, wind speed, and precipitation—a load-forecasting model can be constructed to predict load with respect to weather. The 2006–2016 period is used for AI/ML training while the 2017–2025 period is used for out-of-sample drift measurement.

3.1.2. Platform and Model Scope: AVEVA Predictive Analytics Software

This methodology is demonstrated using the AVEVA PA software.⁴ AVEVA PA provides AI-driven predictive maintenance and asset-health monitoring capabilities that are directly relevant to the electric sector's operational challenges. This software uses ML, statistical analysis, and pattern recognition to identify anomalies in historical and real-time operational data, forecast time to failure, and support proactive maintenance decisions before traditional alarms trigger alerts.

AVEVA PA uses the Operational Profile Topology Information and Classification System (OPTiCS) ML algorithm, created and patented by AVEVA, and other open-source AI/ML models to enable users to detect subtle deviations from expected behavior and estimate when an asset is likely to fail, helping to reduce unplanned downtime and improve reliability and performance.⁴ This methodology was validated across multiple forecasting approaches, including AVEVA proprietary OPTiCS algorithm and locality-sensitive hashing, as well as multilayer perceptron (MLP), and linear regression (see Appendix B).^{5,6,7,8}

Because the software continuously monitors assets over time and integrates with enterprise data sources, it exemplifies the operational environments where AI/ML model drift can emerge as underlying physical systems evolve. By demonstrating the proposed KL divergence-based drift-measurement approach on AI/ML models implemented within AVEVA PA, this work shows how operators can objectively assess when predictive AI/ML models no longer represent evolving system states and require retraining or replacement to maintain accuracy and fidelity.

3.2. Data Preparation and Distribution Modeling

To enable statistically valid drift measurement, true and predicted values are transformed into comparable probability distributions. Probability distributions are fitted exclusively using the training dataset to avoid data leakage. These fitted distributions are then used to normalize and transform both predicted and true values in the test dataset. This process preserves the statistical characteristics learned during training while ensuring that future data are consistently evaluated. By separating distribution fitting from drift measurement, the methodology ensures that the AI/ML model maintains fidelity to real-world deployment conditions, where future system states are unknown.

Figure 1 shows the actual versus forecasted energy consumption for the SCENT region from 2017 to 2025. Over this period, the actual and forecasted energy consumption progressively diverge. From 2017 to 2018 there was low AI/ML model drift with minimal difference between the predicted and actual load. From 2018 to early 2019, moderate drift emerged. From early 2019 onward, high drift occurred and grew as time progressed. As noted previously, this divergence is not attributable to any single operational event. The SCENT region experienced significant structural load changes during this period, including rapid data-center growth in the Austin-San Antonio corridor and evolving industrial load patterns, which the AI/ML model trained on 2006–2016 data could not anticipate.⁹

Over-reliance on a drifted AI/ML model would cause grid-asset owners and operators to systematically underestimate grid consumption, potentially resulting in costly last-minute energy purchases or unplanned reserve deployment. Quantitative drift indicators allow operators to identify this degradation before operational impact occurs. The model drift demonstrated in Figure 1 can be addressed with consistent retraining on more-recent data, and an understanding of when the model has begun to drift and to what extent drift has occurred allows organizations to maximize their period of model utility before retraining is necessary.

3.3. AI/ML Model-Drift Quantification

3.3.1. Drift-Metric Selection

Drift metrics and modeling is a well-studied area within applied ML.¹⁰ In grid applications, particularly within OT or OT-adjacent environments, safety, reliability, and performance are key to maintaining operability and providing meaningful operator actions.¹¹ Drift is a key metric for understanding the reliability of an AI/ML model in deployment, with safety and economic implications if drift occurs and goes unaccounted for.

As OT environments vary depending on the company, regulatory standards, and utility size, drift can occur through a variety of AI/ML use cases, requiring the drift-detection algorithm to be flexible and computationally lightweight. Two categories exist for drift metrics: active, which monitors AI/ML performance and alerts when drift has occurred, and passive, which updates the AI/ML model at a constant rate to reduce the chances of drift occurring. Due to potential resource constraints in OT environments, active drift metrics are considered more viable in grid deployments. Following an extensive review of active drift metrics, it was determined the drift-detection method must satisfy the following characteristics:^{10,12,}

- Able to handle multivariate and noisy datasets
- Usable in semi-supervised and unsupervised environments
- Able to detect slow and gradual changes in drift
- Applicable across different applications
- Consuming low amounts of resources
- Highly interpretable to non-AI/ML experts.

Active drift-detection methods generally fall into four categories: time-window based, statistical process control, statistical, and sequential analysis. Time-window-based methods (e.g., ADWIN) split data into subwindows and compare averages.¹³ These are effective for univariate supervised tasks but are resource-heavy and not suitable for noisy multivariate data.¹⁴ Process-control-based methods (e.g., DDM and EDDM) track error distances and deviations. These work for classification tasks but struggle with gradual drift and other AI/ML use cases.¹⁵

These constraints narrowed the drift-metric space to statistical and sequential analysis methods. Of the tested approaches, it was determined that KL divergence, combined with PSI thresholding and CUMSUM sequential analysis, provided the most-consistent model-agnostic method of determining model drift. Alternative metrics were evaluated and rejected for the following reasons:

- The two-sample KS statistic is nonparametric but is defined for single continuous variables. Applying it to the mixed categorical and continuous data typical of grid OT environments requires separate tests for each feature and correction for multiple comparisons. The KL-based approach avoids this added complexity.
- Wasserstein distance can be computed directly from data, but there is no established thresholding guidance for it comparable to PSI, and its computation for multivariate data is more expensive than KL divergence over discretized distributions. Ongoing research is exploring hybrid Wasserstein-KL approaches for robustness calculation in supply-chain security contexts that are outside the scope of this AMARANTH life-cycle management methodology.

KL divergence is a relative entropy measure that captures distributional differences between predicted values $P(y_i)$ and observed distribution $Q(\mu_i(x_i))$:¹⁶

$$D_{KL}(P(y_i)||Q(\mu_i(x_i))) = P(y_i) \ln \frac{P(y_i)}{Q(\mu_i(x_i))} - P(y_i) + Q(\mu_i(x_i))$$

Unlike pointwise error metrics such as MAPE, KL divergence reflects changes in the underlying statistical structure of AI/ML model outputs, making it well-suited for detecting gradual degradation in model validity.¹⁷ This leading-indicator property is only realized when KL divergence is paired with CUMSUM accumulation, the raw per-timestep KL values are too small and noisy to act as standalone alerts. The combination of the two mechanisms is what produces an actionable signal. If it is assumed under normal drift conditions where $P(y)$ and $Q(\mu(x))$ are symmetric at every value of i , PSI can be used to derive thresholds at which drift occurs as the first point found through CUMSUM where $\sum_i^n D_{KL}(P(y_i)||Q(\mu_i(x_i))) > 0.05$ for drift to occur at point n implying drift has occurred.⁴

PSI is a symmetric metric defined as:⁴

$$PSI = D_{KL}(P(y)||Q(\mu(x))) + D_{KL}(Q(\mu(x)||P(y)))$$

where the approximate symmetry between $P(y)$ and $Q(\mu(x))$ holds when the AI/ML model is performing well and the predictions are centered on actual values yielding $PSI \approx 2D_{KL}(P(y)||Q(\mu(x)))$. This allows PSI-derived thresholds (where moderate drift occurs $0.1 < PSI < 0.25$ and a high drift occurs $PSI \geq 0.25$) to be translated into KL-divergence space, producing the thresholds of 0.05 for moderate drift and 0.125 for high drift.

KL divergence itself is inherently asymmetric, where $D_{KL}(P(y)||Q(\mu(x))) \neq D_{KL}(Q(\mu(x)||P(y)))$ and the direction of the comparison matters. The methodology described computes $D_{KL}(P(y)||Q(\mu(x)))$, predicted relative to observation to capture how much information is lost using the AI/ML model's predictions to the approximate true distribution. KL divergence asymmetry is not a limitation but a feature in this case, where the symmetry assumption breaks down under drift as the two directional KL divergences differ from one another and the asymmetry itself is a signal of directional degradation (over-versus under-forecasting) that PSI would obscure by averaging. This sensitivity to directional bias is particularly valuable in applications such as load forecasting where under-forecasting has different operational consequences than over-forecasting.

Readers should note that the individual timestep KL-divergence values in Figure 2 and Figure 3 are instantaneous measurements, typically in the range of 0.001 to 0.005 during the low-drift period. These KL-divergence values cumulate via CUMSUM to reach the 0.05 and 0.125 threshold over time.

3.3.2. Relationship Between KL-Divergence Thresholds and Operational Reality

The drift thresholds used in this methodology should be understood as starting points, not universal constants. The appropriate threshold for a given deployment depends on the operator's risk tolerance, the consequence of AI/ML-model errors in that application, and the cost and downtime associated with retraining. An operator with high retraining capacity and low tolerance for forecast error might choose tighter thresholds. If retraining is expensive and forecasting errors are routinely absorbed into large reserve margins, an operator might choose higher thresholds. This methodology provides the measurement framework; the threshold calibration is the operator's responsibility based on the specific operational context.

3.3.3. Time-Resolved Drift Calculation

KL divergence is calculated at each timestamp in the test dataset between the transformed predicted and true values. The timestamps can be adjusted to different temporal resolutions, such as daily, weekly, or monthly, depending on what temporal resolution best represents the dataset and the resolution of the input dataset. Finer (e.g., daily) resolution will accumulate divergence faster and reach drift thresholds sooner than coarser (e.g., weekly) resolutions, allowing operators to tune the sensitivity of drift detection to their retraining capacity and risk tolerance. The SCENT dataset has hourly resolution, aggregated to daily and weekly for the drift calculations shown in Figure 2 and Figure 3. Sub-hourly analysis would require higher-frequency operational data. Each timestep produces a single divergence value, where lower values indicate closer alignment between AI/ML model predictions and observed behavior. This produces a time-resolved drift signal that captures both sudden shifts and gradual divergence.

3.4. Cumulative Drift Analysis and Thresholding

To translate individual divergence measurements into actionable insights, KL-divergence values are accumulated chronologically across the test period (2017–2025) using CUMSUM. Predefined drift thresholds categorize AI/ML model behavior into low (< 0.05), moderate (between 0.05 and 0.125), and adoption (≥ 0.125).⁴ These thresholds are adapted from PSI guidance and translated into the KL-divergence space as described in Section 3.3.

KL-divergence for the AVEVA OPTiCS model reached 0.05 in 2018 and 0.125 in 2019 cumulatively, as shown in Figure 2 above. Figure 3 shows the weekly equivalent, where the moderate-drift threshold is reached in 2021, and the high-drift threshold in 2022.

The figures above demonstrate the impact that temporal resolution can have on drift measurement, as evidenced by the greatly increased amount of time the model spent in the low-drift region when using a coarse (weekly) resolution compared to a fine (daily) resolution (approximately 4 years vs. 1 year, respectively). This emphasizes the importance of choosing a proper temporal resolution for a given use case. Coarser temporal resolutions limit the frequency of updated predictions, but can resist drift for longer, thereby lowering overall model-maintenance workload. Conversely, a finer temporal resolution can provide more frequent updates but will begin to drift sooner. Thresholds should be calibrated at the resolution used in production.

Another factor that affects drift was demonstrated during this analysis, which is the choice of the AI/ML algorithm itself. With this methodology, it is possible to construct several versions of an AI/ML model, all using the same underlying training and test datasets, and compare them for their resilience to drift. An operator can include this metric as part of the criteria for deciding which AI/ML models to promote to a production environment. For example, Table 1 summarizes the drift periods (aggregated weekly) for the four algorithm types tested on the ERCOT dataset. AVEVA OPTiCS, designed for use in industrial scenarios, resisted drift the longest, lasting approximately 5 years before reaching the high-drift threshold. The open-source algorithms lasted approximately 1 to 1.5 years before reaching the high-drift threshold. Note that this table shows only values from this particular test case, and care should be taken before generalizing them. Nevertheless, it is intended to illustrate how a comparative analysis might be performed during a testing phase, prior to deciding which AI/ML to deploy in production. For further details on this comparative analysis, please see Appendix B.

Table 1. Summary of algorithm drift performance on the ERCOT dataset using weekly aggregation.

Algorithm	Dataset Start	Beginning of Moderate Drift	Beginning of High Drift	Low-Drift Duration (Weeks)	Moderate Drift Duration (Weeks)	Total Duration Before High Drift (Weeks)
AVEVA PA OPTiCS	1/1/2017	2/14/2021	2/26/2022	215	54	269
Seasonal autoregressive integrated moving average	1/1/2017	4/24/2017	1/7/2018	16	37	53
Linear regression	1/1/2017	5/1/2017	1/7/2018	17	36	53
Multilayer Perceptron	1/1/2017	5/1/2017	5/20/2018	17	55	72

3.5. AI/ML-Model Lifetime Estimation

The transition into the high-drift zone defines the estimated end of AI/ML model operational lifetime. Beyond this point, the AI/ML model is expected to exhibit elevated risk of performance degradation and reduced representational fidelity. This threshold provides operators with a clear, quantitative signal for retraining or replacement decisions functioning as a condition-based maintenance interval for AI/ML models (Figure 4).

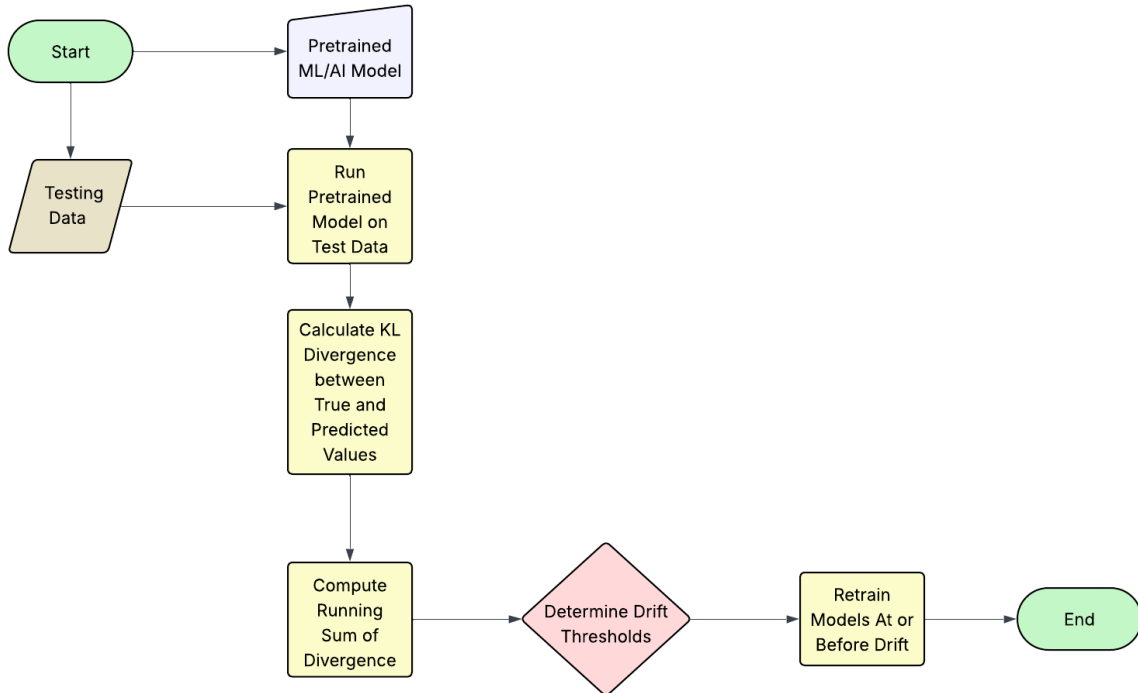


Figure 4. Flow chart demonstrating the method used to determine model drift and the latest point at which a model should be retrained.

In practice, the moderate-drift zone serves as an action window; the model is still performing adequately, but operators should initiate retraining planning so that a retrained model is ready before the high-drift threshold is crossed, minimizing operational risks. Operators should treat the moderate-drift signal as a prompt to begin retraining planning, rather than as a guarantee of any specific lead time. The roughly 1-year window observed between moderate- and high-drift thresholds in the SCENT case study should not be treated as a fixed or universally predictable planning horizon. Sudden structural changes—a large load facility coming online, a grid topology change, or an extreme weather event—can compress this window substantially and unpredictably.

It is also important to note that the paper does not prescribe a specific retraining protocol. The appropriate approach—rolling window versus full historical retraining, required volume of new data, validation requirements before returning a retrained model to production—depends on the operator’s data environment, model architecture, and consequence tolerance. The retraining protocol design for grid OT environments is a meaningful area for follow-on work, and operators should consult relevant operational AI governance guidance when developing their retraining processes.

4. USE CASES: IMPACTS OF UNMEASURED AI/ML MODEL DRIFT

As AI/ML models become more-deeply embedded in electric-grid planning and operations, their performance over time increasingly determines operational reliability, economic efficiency, and operational decision quality. While many grid AI/ML applications demonstrate strong accuracy at deployment, changes in system conditions, asset behavior, and data distributions can cause AI/ML models to degrade in ways that are not immediately visible to operators. The following use cases illustrate how unmeasured AI/ML-model drift can introduce operational and economic risk and demonstrate where quantitative assessment of AI/ML-model lifetime and drift provides actionable value.

Load forecasting and predictive maintenance are examined as representative applications in which post-deployment AI/ML-model fidelity directly affects decision quality and grid performance. Both function primarily as decision-support tools that inform human operators and influence real-time dispatch decisions, maintenance prioritization, and capital planning, but do not directly actuate equipment or operate autonomously. The economic and operational significance of these functions makes sustained accuracy important where persistent inaccuracy from undetected AI/ML model drift can erode the efficiency gains that motivated AI adoption.

4.1. Load Forecasting

Accurate load forecasting is critical to maintaining balance between electricity supply and demand. As established in Section 2.2, there is an important operational distinction between the day-to-day forecast error that grid operators routinely manage and the systematic cumulative degradation caused by AI/ML-model drift. This section describes how the KL-divergence methodology can be applied to detect and manage AI/ML-model drift.

Applying the KL-divergence methodology to a load-forecasting context follows the general workflow described in Figure 4. In the SCENT case study, the moderate-drift threshold was crossed in 2018 at a daily resolution, providing operators approximately 1 year of advance warning before the high-drift threshold was crossed in 2019. This lead time represents a window during which retraining can be planned and executed without emergency disruption to operations.

The operational significance of these drift thresholds can be anchored directly to ERCOT's published forecasting standards. ERCOT's 2017 "Long-Term Hourly Peak Demand and Energy Forecast" targets forecast errors within approximately 1–3% MAPE under normal operating conditions, with deviations beyond that range treated as material to planning and reserve decisions.¹⁸ During the low-drift period (2017–2018), hourly MAPE for the AVEVA OPTiCS model remained within 1–3%, consistent with acceptable operational performance under ERCOT's planning framework. The KL-divergence CUMSUM signal, however, began accumulating meaningfully from the start of the test period, crossing the moderate-drift threshold (0.05) in 2018. MAPE did not rise to operationally notable levels until approximately 2019, averaging 1.53% on a sustained basis until that point. This sequencing illustrates the practical value of distributional monitoring: a utility relying solely on rolling MAPE monitoring would have had no quantitative signal to initiate retraining planning until performance had already degraded whereas the KL CUMSUM signal provided advance warning while the model was still producing acceptable outputs.

It should be noted that this leading-indicator relationship—where KL-divergence crosses drift thresholds before MAPE degrades to operationally significant levels—is demonstrated here for one model on one dataset and one regional load profile. This relationship may not hold in all deployment contexts, particularly for classification tasks, anomaly-detection applications, or datasets with substantially different noise characteristics from SCENT's. The leading-indicator property of KL-divergence should be treated as an empirical finding in this case study that motivates further validation across deployment contexts, rather than an established general property.

Changes in the grid environment further exacerbate the forecasting challenge. Increased data-center load from generative AI, evolving consumption patterns, weather variability, and structural changes to the grid can cause historical training data to become unrepresentative of current conditions. Without a quantitative method to measure divergence from an AI/ML model's original baseline, operators lack objective indicators for when retraining or replacement is required.

4.2. Predictive Maintenance

AI/ML models are increasingly used to support predictive maintenance by identifying early indicators of equipment degradation.⁹ When functioning as intended, these AI/ML models enable proactive maintenance, extend asset lifetimes, and reduce unplanned outages. Their effectiveness, however, depends on continued alignment between AI/ML model behavior and evolving asset conditions.

Applying the KL-divergence methodology to predictive maintenance follows the same workflow as the load-forecasting case (Figure 4) where the AI/ML model's training-period prediction distributions are fitted, KL divergence is computed between training-era and production-era outputs at each timestep, and cumulative divergence is tracked against the 0.05 and 0.125 thresholds. For asset-health applications, the quantity compared is typically an anomaly score or remaining-useful-life estimate rather than a load value, and practitioners should be aware that these outputs have distributional properties that differ fundamentally from load data. Equipment-failure distributions are typically right-skewed and bounded in ways that load data are not, and the PSI symmetry assumption underlying the threshold derivation is less likely to hold for remaining-useful-life estimates or anomaly scores. The methodology therefore requires further validation before application to predictive-maintenance use cases, and threshold calibration in particular may need to be revisited for non-Gaussian output distributions.

If a predictive maintenance AI/ML model drifts undetected, subtle signs of equipment degradation may no longer be detected, allowing equipment to fail with little to no warning, resulting in service interruptions and emergency repairs. Conversely, drift may cause an AI/ML model to falsely indicate degradation, prompting unnecessary maintenance interventions that consume workforce resources, increase operational costs, and reduce confidence in AI-assisted diagnostics. Quantitative drift-measurement methods such as KL divergence provide a mechanism to identify when AI/ML model outputs no longer reliably reflect asset health, allowing operators to recalibrate, retrain, or retire AI/ML models before decision quality is compromised. In predictive-maintenance contexts especially, the KL-divergence signal should be interpreted alongside domain-specific validation signals, such as comparison against known equipment-failure histories or physics-based health baselines since the consequences of false positives and false negatives in maintenance decisions may differ significantly across asset types and grid criticality levels.

5. LIMITATIONS

A central limitation of the presented methodology is that KL divergence does not diagnose the cause of distributional drift, whether it be concept or data drift. The KL divergence-based signal measures only the magnitude and trajectory of divergence between predicted and observed value distributions.

AMARANTH's KL-divergence method does not differentiate whether this divergence is due to:

- Degradation of the AI/ML model itself (data drift)
- Structural changes in the underlying physical system (concept drift)
- A combination of both data and concept drift.

The SCENT case study exhibits the limitation of the KL-divergence drift method directly, where the observed divergence was partly driven by structural load growth, including rapid data-center expansion in the Austin-San Antonio corridor and evolving industrial load patterns—changes not represented in the 2006–2016 training data. In such cases, KL-divergence correctly identifies that the model is no longer statistically aligned with current-system behavior, but it cannot determine why.

Distinguishing between data drift and concept drift has operational significance because the appropriate response differs. Data drift—where the model's learned relationships remain valid, but the input distribution has shifted—may be addressable through retraining on recent data with the existing model architecture. Concept drift—where the physical system relationship has fundamentally changed—may require more-substantial intervention, such as feature reengineering, model redesign, or the incorporation of new explanatory variables that capture the changed system structure. The KL-divergence signal alone cannot direct operators to the appropriate response; distinguishing between these cases requires additional system-level context, physics-based modeling, domain expertise, and external diagnostics that are outside the scope of this life-cycle-focused framework. Operators are encouraged to treat a rising KL CUMSUM signal as a trigger for deeper investigation, not solely as a retraining prompt.

KL-divergence acts as a drift indicator, reliably and computationally efficient. However, it must be interpreted appropriately: the methodology detects distributional divergence regardless of cause, and its value lies in signaling that the model has exceeded its statistical-validity window and should be retrained or replaced. Determining the cause of that drift requires complementary analyses.

The methodology presented rests on a single dataset and geographic region. The Appendix B comparison of linear regression and MLP adds model-agnostic validation, but both models use the same SCENT dataset. Validating this approach on additional regional datasets with different load characteristics beyond the SCENT dataset tested here is identified as a priority for future work.

Because KL divergence is never negative, the cumulative signal can only grow. Any model will eventually cross a fixed threshold if monitored long enough, even without structural drift. Threshold crossing times therefore depend on the temporal resolution and the length of the monitoring period, as seen in the daily versus weekly differences in above. The PSI-derived thresholds should be viewed as calibrated to this case study rather than derived for cumulative use. Future work should focus on separating baseline accumulation and seasonal variation from true drift. Options include CUMSUM formulations that reset with a drift allowance, seasonally adjusted baselines, and calibration against a stationary reference dataset.

6. CONCLUSION

As AI/ML models become more-deeply integrated into electric-grid planning, forecasting, and asset management, their continued reliability cannot be assumed over time. Changes in load composition, asset condition, operating regimes, weather patterns, and data-generation sources can cause AI/ML models trained on historical data to diverge from current-system behavior. When this divergence goes unmeasured, operators risk relying on degraded decision-support tools, with implications for economic efficiency and operational reliability.

This whitepaper presented a quantitative, AI/ML model-agnostic methodology for measuring post-deployment AI/ML model drift and estimating model lifetime using KL divergence. This approach is one structured option among a broader landscape of drift-detection methods. KL divergence was selected over alternative drift metrics, including KS statistics, Wasserstein distance, ADWIN, and DDM/EDDM, because it satisfied the operational requirements of grid OT environments, handling multivariate and categorical data without requiring knowledge of the underlying parametric distribution, operating effectively in both semi-supervised and unsupervised environments. KL divergence has the ability to detect the gradual distributional shift that pointwise error metrics miss while it remains sufficiently computationally lightweight for resource-constrained OT systems. These properties make it applicable across a large range of grid AI use cases, such as load forecasting and predictive maintenance, without modification. Applying this model-drift methodology to the AVEVA PA AI/ML model demonstrates its relevance to commercially deployed systems.

A core contribution of this work is the explicit distinction between normal forecasting error and AI/ML model drift. The KL-divergence CUMSUM signal is a leading indicator of structural degradation, providing operators advance warning to plan retraining proactively. In the SCENT case study, this advanced-warning window was approximately 1 year at daily resolution, representing meaningful operational value over approaches that rely solely on rolling MAPE monitoring. Operators should note, however, that this interval is not a fixed guarantee; structural changes can compress it unpredictably, and the moderate-drift threshold should prompt immediate retraining planning rather than a measured wait.

The drift thresholds presented (0.05 for moderate drift, 0.125 for high drift) are grounded in PSI guidance translated into KL-divergence space and validated against the SCENT operational data. They are best understood as calibrated starting points; operators should adjust thresholds to reflect their specific risk tolerance, retraining capacity, and operational consequences of AI/ML model error. Similarly, drift measurement should be situated within a broader performance assurance program that includes pointwise error tracking, calibration checks, feature stability monitoring, and operational-outcome assessment; KL divergence is one necessary component of that program, not a sufficient condition for assurance on its own.

More broadly, quantitative drift and lifetime metrics support trustworthy AI deployment by enabling transparency, consistency, and operational accountability. As AI/ML adoption continues to scale across the electric sector, integrating drift measurement into standard operational practices will be essential to ensuring that AI-enabled systems remain reliable, resilient, and aligned with real-world grid conditions. Future work should prioritize validation of this methodology on additional regional datasets with diverse load profiles, formal benchmarking of computational performance on representative OT hardware, and development of retraining protocols suited to grid OT deployment environments.

7. REFERENCES

- 1 U.S. Department of Energy, “Artificial Intelligence for Energy,” accessed December 08, 2025, <https://www.energy.gov/topics/artificial-intelligence-energy>.
- 2 A. Singla, A. Sukharevsky, B. Hall, L. Yee, M. Chui and T. Balakrishnan, “The state of AI in 2025: Agents, innovation, and transformation,” Mckinsey & Company, November 5, 2025, accessed February 17, 2026, <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai>.
- 3 IEA, “Energy and AI: AI for energy optimisation and innovation,” accessed December 08, 2025, <https://www.iea.org/reports/energy-and-ai/ai-for-energy-optimisation-and-innovation>.
- 4 AVEVA, “AVEVA Predictive Analytics,” accessed January 27, 2026, <https://www.aveva.com/en/products/predictive-analytics>.
- 5 H. Al-Chalabi et al., “Time Series Forecasting using ARIMA Model,” in ADVCOMP 2018: The Twelfth International Conference on Advanced Engineering Computing and Applications in Sciences, 2018.
- 6 W. Wu and B. Li, “Hierarchical Locality Sensitive Hashing for Structured Data: A Survey,” accessed January 27, 2026, <https://arxiv.org/pdf/2204.11209>.
- 7 ERCOT, “Supply and Demand,” Electric Reliability Council of Texas, accessed: January 27, 2026, <https://www.ercot.com/gridmktinfo/dashboards/supplyanddemand>.
- 8 AVEVA, “Tata Power uses AVEVA Predictive Analytics software,” accessed: September 1, 2026, https://www.aveva.com/content/dam/aveva/documents/perspectives/success-stories/SuccessStory_TataPower.pdf.coredownload.inline.pdf.
- 9 A. Mok, “Utilities are tiptoeing into AI as climate change and data center growth add stress to the energy grid,” Business Insider, accessed January 28, 2026, <https://www.businessinsider.com/utilities-modernize-energy-grid-generative-ai-predictive-maintenance-2025-7>.
- 10 M. Han, Z. Chen, M. Li, H. Wu and X. Zhang, “A survey of active and passive concept drift handling methods,” *Computational Intelligence* 38, no. 4 (2022): 1492–1535, <https://onlinelibrary.wiley.com/doi/10.1111/coin.12520>.
- 11 B. Lampe, “Cyber-Informed Engineering (CIE) – Engineered Controls Database and Use,” Idaho National Laboratory, September 2025, <https://www.osti.gov/servlets/purl/3006995>.
- 12 F. Bayram, B. S. Ahmed, and A. Kassler, “From concept drift to model degradation: An overview on performance-aware drift detectors,” *Knowledge-Based Systems* 245, Issue C (2022), <https://dl.acm.org/doi/10.1016/j.knosys.2022.108632>.
- 13 S. Ashok, S. Ezhumalai, T. Patwa, “Remediating data drifts and re-establishing ML models,” *Procedia Computer Science* 218 (2023): 799–809, <https://www.sciencedirect.com/science/article/pii/S1877050923000601>.
- 14 River, “ADWIN,” accessed: September 1, 2026, <https://riverml.xyz/dev/api/drift/ADWIN/>.
- 15 River, “EDDM,” accessed: September 1, 2026, <https://riverml.xyz/0.13.0/api/drift/EDDM/>.
- 16 S. Boyd and L. Vandenberghe, *Convex Optimization*, (Cambridge University Press, 2004).

-
- 17 S. Basterrech and M. Wozniak, “Tracking changes using Kullback-Leibler divergence for the continual learning,” (2022), <https://arxiv.org/abs/2210.04865>.
 - 18 ERCOT, “2017 Long-Term Hourly Peak Demand and Energy Forecast,” Electric Reliability Council of Texas, December 2016, https://www.ercot.com/files/docs/2016/12/13/2017_Long-Term_Hourly_Peak_Demand_and_Energy_Forecast.pdf.

Appendix A

Introduction to Machine Learning and Artificial Intelligence

Even though artificial intelligence (AI) capabilities have recently garnered more attention, they are not new concepts. AI can be traced back to the 1950s when Alan Turing asked the question “can machines think?” and developed the concept of a Turing test of an AI system’s ability to fool a human interrogator into thinking it is human through question-answer responses.¹ Although the Turing test is commonly conducted through interrogation and question-answer responses, the idea of a digital system simulating or conducting common human tasks and behaviors with high fidelity is arguably the root idea behind the Turing test.

AI as a field of study does not have one simple definition because these “tools are capable of a wide range of tasks and outputs.”¹ Multiple executive orders, standards, and research papers aim to define AI; however, at its core, AI functions to emulate human behavior to cause varying degrees of impact on the physical world, passing a Turing test through emulation.² Via this definition, the ideal AI can simulate all potential human behaviors and tasks. Although AI is not an all-seeing oracle and cannot solve problems unknown to humans, AI is able to process larger datasets and identify more-complicated patterns than the human mind due to the increased amount of memory that AI has at its disposal. Like a compiler (or interpreter), AI is a distillation of existing human knowledge into patterns and mathematical formulae. The types of AI relevant to this whitepaper are described below:

- **Machine Learning (ML)** itself is a subset of AI that uses large datasets for pattern recognition through supervised (training data includes a target) and unsupervised (training data does not include a target) learning methods. Examples of ML include logistic or linear regression, support vector machines, decision trees and random forest algorithms, K-nearest neighbors and K-means, and naïve Bayes to learn patterns through large datasets. These algorithms have higher interpretability, explainability, and performance on smaller datasets than other subsets of AI and have uses such as anomaly detection, predictive maintenance, and system optimization.
- **Deep learning (DL)** is a subset of ML where neural networks are layered to extract hidden patterns and extract high-level features from a raw dataset. Examples of DL algorithms include convolutional neural networks, recurrent neural networks, and deep neural networks. DL is useful in handling highly complex and massive datasets where most DL algorithms are highly adaptable to changes in post-deployment applications. Example use cases include automated inspection and system control and complex event predictions.
- **Generative AI (GenAI)** is a specific subset of DL that generates new content such as text, images or code and can create outputs based on learned patterns that relate to a given input based on massive amounts of input. GenAI can do associative reasoning tasks and data synthesis with use cases such as creation or updating of system documentation, scenario planning, and simulation of plausible datasets. Examples of GenAI algorithms include generative adversarial networks, variational autoencoders, transformer models, and autoregressive models.

¹ A. M. Turing, “Computing Machinery and Intelligence,” in *Parsing the Turing Test*, R. Epstein, G. Roberts, and G. Beber, Eds., Dordrecht, (Springer Netherlands, 2009), 23–65. doi: 10.1007/978-1-4020-6710-5_3.

² NASA, “What is Artificial Intelligence,” National Aeronautics and Space Administration, accessed October 29, 2025, <https://www.nasa.gov/what-is-artificial-intelligence/>.

Appendix B

Kullback-Leibler Divergence on Different Architectures

Kullback-Leibler (KL) divergence can be used across a variety of different architectures from traditional machine-learning (ML) algorithms to deep-learning (DL) algorithms, where the architecture and the amount of data vary significantly. As an example, the graphs shown below provide examples of model drift between a statistical model (seasonal autoregressive integrated moving average [SARIMA]), a traditional ML algorithm (linear regression), and a DL model (multilayer perceptron [MLP]) determined using KL divergence in which low drift occurs and where KL divergence is below or at 0.05, medium drift occurs between 0.05 and 0.125, and high drift occurs at or above a KL-divergence value of 0.125, all using the Texas South-Central (SCENT) dataset and weekly resolution used to develop Figure 3. It can be seen that the cumulative KL-divergence of the linear regression and SARIMA models reached the drift thresholds sooner than the MLP model under the same resolution, dataset, and conditions. KL-divergence can be used to determine the resilience of one model versus the other to drift and inform the artificial intelligence (AI)-developer of the best choice for models given the organization's risk tolerance and retraining capability. When comparing SARIMA, linear regression, MLP, and AVEVA Predictive Analytics (AVEVA PA) Operational Profile Topology Information and Classification System (OPTiCS), it can be shown visually through Figure 3, Figure B-1, Figure B-2, and Figure B-3 that AVEVA PA OPTiCS was more resilient to drift than SARIMA, linear regression, and MLP models where drift occurred later and had a larger length of time in the moderate-drift zone.

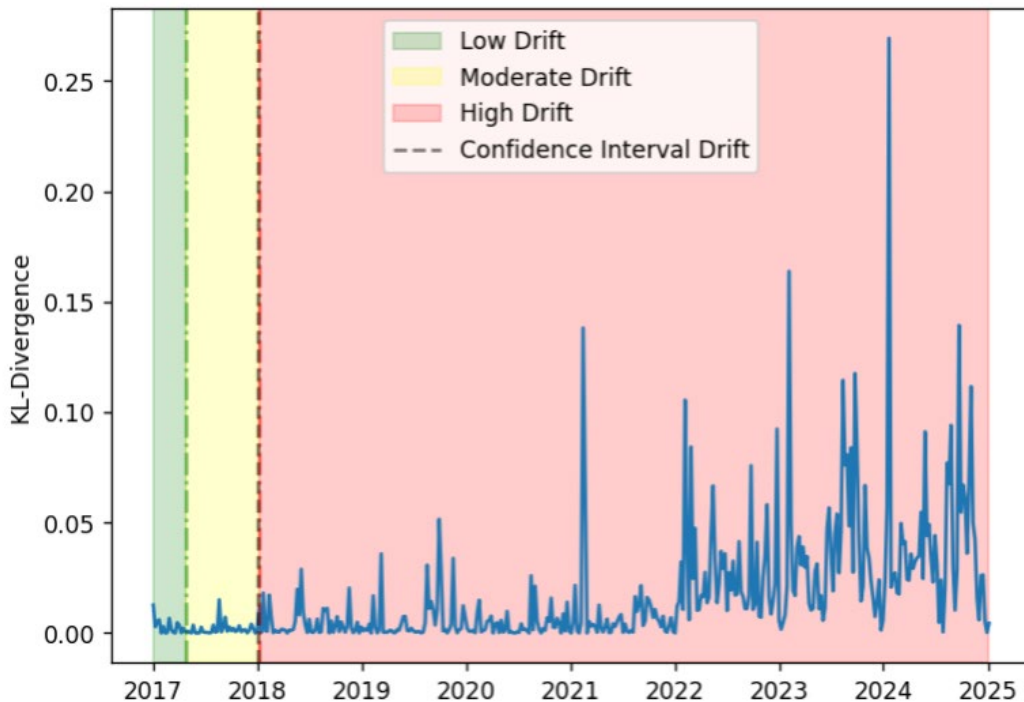


Figure B-1. KL-divergence and drift zones for SARIMA load forecasting with weekly aggregation.

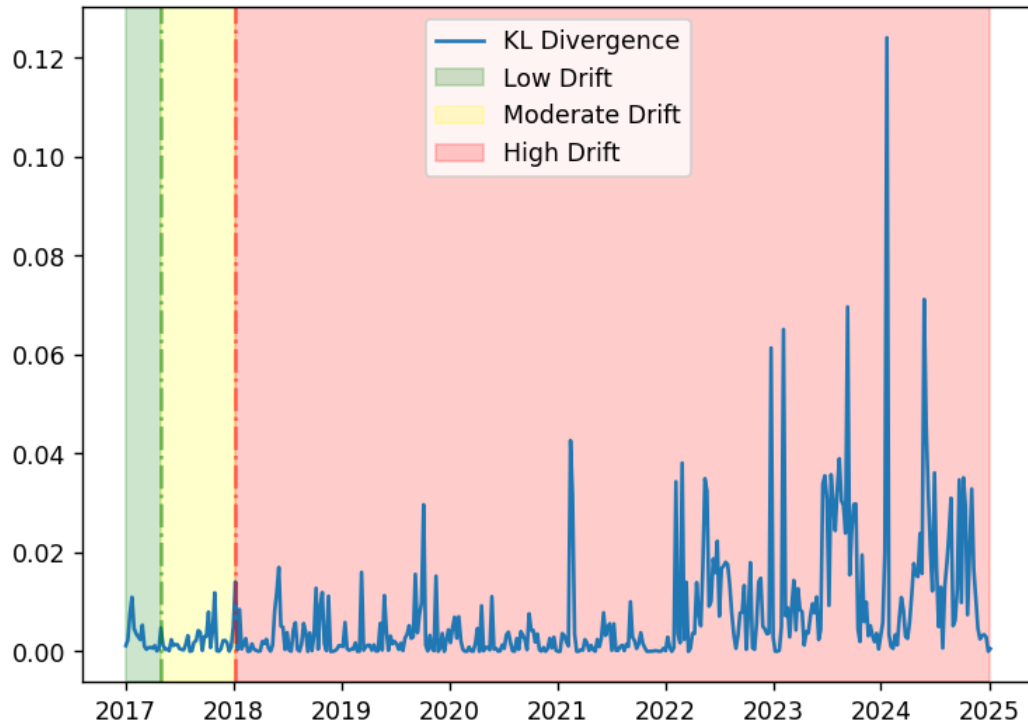


Figure B-2. KL-divergence and drift zones for linear-regression load forecasting with weekly aggregation.

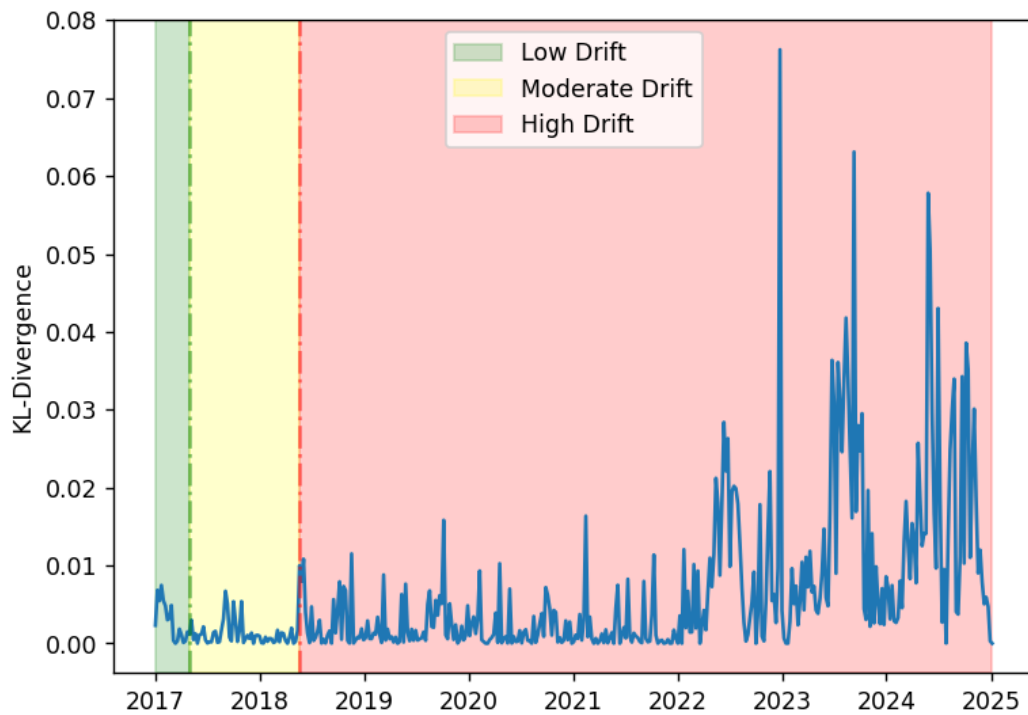


Figure B-3. KL-divergence and drift zones for MLP load forecasting with weekly aggregation.

Linear regression and MLP models represent opposite ends of the AI/ML complexity spectrum and were selected deliberately to demonstrate that the KL-divergence drift methodology is model-agnostic. Linear regression is the simplest supervised-learning approach in that it fits a linear relationship between input features (weather variables, calendar indicators) and the target output (load in MWh), producing a highly interpretable model with relatively few parameters. Its simplicity means it generalizes conservatively and will not overfit the training period, but it also cannot capture nonlinear seasonal interactions or complex load patterns. The MLP, by contrast, is a DL architecture consisting of multiple layers of interconnected nodes that learn nonlinear feature representations from the training data. MLPs can capture more-complex input-output relationships and typically achieve lower training-period error on datasets with nonlinear structure, such as the temperature-load relationship in summer-peaking regions like SCENT.