ENERGYX: AN AGENTIC FRAMEWORK FOR EXPLAINABLE ENERGY FORECASTING AND ANOMALY DETECTION

Sanyat Fale
Dept of Mathematics and Computing
Indian Institute of Science
Bangalore, India
sanyatvinod@iisc.ac.in

Naman Srivastava
Robert Bosch Centre For Cyber Physical Systems
Indian Institute of Science
Bangalore, India
snaman@iisc.ac.in

Pandarasamy Arjunan
Robert Bosch Centre For Cyber Physical Systems
Indian Institute of Science
Bangalore, India
samy@iisc.ac.in

ABSTRACT

Energy time series forecasting and anomaly detection are critical for grid stability, demand response, and operational reliability. However, deploying effective solutions remains challenging due to complex model selection, hyperparameter tuning, and limited explainability, which restrict accessibility for domain users. We present EnergyX, an agentic framework that translates natural language user intent into executable forecasting and anomaly detection workflows. EnergyX performs automated data profiling, model selection with cross-validated performance-based aggregation, and generates explanations grounded in statistical diagnostics, feature attribution methods (e.g., SHAP, feature importance), and counterfactual analysis. The system integrates decomposition, lag analysis, residual diagnostics, and anomaly consensus scoring to provide transparent and actionable insights. Through forecasting and anomaly detection scenarios, we demonstrate that EnergyX enables accurate, explainable, and user-steerable time series analysis, highlighting the potential of LLM-driven agentic systems for accessible energy analytics.

Track: Research

1 INTRODUCTION

Energy consumption forecasting and anomaly detection are fundamental to modern energy systems International Energy Agency (2024); U.S. Energy Information Administration (2023), enabling efficient resource allocation, grid stability, and operational decision-making. Accurate forecasts support demand planning, while anomaly detection helps identify equipment failures, inefficiencies, and unexpected consumption patterns. However, building reliable forecasting pipelines remains challenging due to dataset variability, model selection complexity, and the expertise required to interpret model outputs.

A wide range of forecasting models exist, including statistical approaches (ARIMA Tunnicliffe Wilson (2016), ETS Hyndman et al. (2008)), tree-based methods (Random Forest Breiman (2001), Gradient Boosting Ke et al. (2017); Chen & Guestrin (2016)), and neural architectures (N-BEATS Oreshkin et al. (2020), Transformers Zhou et al. (2021); Lim et al. (2020)). While these models can achieve strong performance, their effectiveness varies across datasets due to differences in temporal structure, noise, and feature availability. Ensemble approaches Bates & Granger (1969); Dietterich (2000) improve robustness by combining multiple models, but selecting appropriate models, configuring them, and determining aggregation strategies introduces significant overhead.

Recent agentic systems such as TimeSeriesScientist Zhao et al. (2025) and TimeCopilot Garza & Rosillo (2025) use Large Language Models (LLMs) to translate natural language queries into forecasting workflows and automate model orchestration. However, these systems rely on predefined pipelines and provide limited transparency into model behavior and decision-making. Existing explainability methods such as SHAP Lundberg & Lee (2017); Lundberg et al. (2020) and feature importance Breiman (2001) provide partial insights, but lack integrated analysis of temporal patterns such as seasonality, lag relationships, residual behavior, and anomaly agreement. These limitations create barriers for domain users seeking to deploy and trust forecasting systems. Users must manually navigate model selection, interpret results, and integrate multiple diagnostic tools to understand model behavior. The lack of unified, explainable, and user-steerable systems reduces accessibility and limits real-world adoption. Addressing these challenges requires systems that can automate forecasting workflows while providing transparent, interpretable, and actionable insights grounded in statistical evidence.

We propose EnergyX, an agentic framework that translates natural language user intent into executable energy forecasting and anomaly detection workflows. EnergyX integrates automated model selection, performance-based prediction aggregation, and multi-modal explainability grounded in statistical diagnostics, feature attribution, and counterfactual analysis Wang et al. (2023). This enables accurate, transparent, and user-steerable time series analysis.

Contributions: Our key contributions are:

- An agentic, human-in-the-loop framework that converts natural language queries into executable energy forecasting and anomaly detection workflows.
- A cross-validation-based model aggregation strategy that combines multiple model families using performance-derived weights.
- A unified explainability framework integrating feature attribution, seasonal decomposition, lag correlation analysis, residual diagnostics, and anomaly agreement scoring.
- Support for counterfactual forecasting and structured reporting for actionable decision-making.

2 ENERGYX: AGENTIC FRAMEWORK DESIGN

EnergyX implements an agentic architecture in which a Large Language Model (LLM) orchestrates time series forecasting, anomaly detection, and counterfactual analysis through iterative tool invocation. The system operates in two sequential phases: (1) planning and (2) execution, as shown in Figure 1.

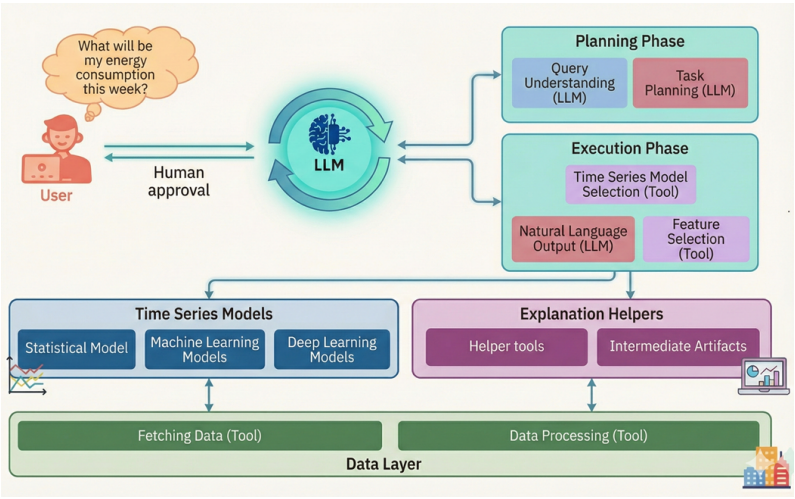

Figure 1: Overview of EnergyX System Architecture

Planning Phase: The process begins with automated data profiling, which analyzes temporal structure, statistical properties, stationarity, and seasonality, and produces a structured summary. Only this statistical profile, not raw data, is provided to the LLM. Given the profile and the user's natural language query, the LLM generates a structured task plan specifying the task type (forecasting, anomaly detection, or counterfactual analysis), prediction horizon, feature configuration, candidate models, focused columns and explainability requirements. This plan is presented to the user

through an interactive interface, allowing review and modification before execution, ensuring human oversight and control.
Execution Phase: After approval, the LLM enters an iterative tool-calling loop, where it dynamically selects analytical tools based on intermediate results. Each tool returns structured outputs such as validation metrics, predictions or diagnostic statistics, which inform subsequent decisions. All intermediate artifacts (like cross-validation scores, feature importance, and diagnostic statistics) accumulate in a shared session dictionary consumed by the ensemble combiner, explanation and report generators. For forecasting, the agent trains candidate models using rolling-origin cross-validation and combines predictions using performance-based aggregation weights. For anomaly detection, multiple detection methods are applied and aggregated through consensus scoring. For counterfactual analysis, the agent evaluates feature perturbations or identifies minimal feature changes required to achieve target outcomes. After tool completion, the LLM produces a grounded final analysis citing exact metrics from the shared session state. Explainability is embedded throughout via temporal decomposition, statistical diagnostics, feature attribution, and counterfactual reasoning. This two-phase design separates user-editable planning from LLM-driven execution, avoiding rigid pipelines while preserving human-in-the-loop control and adaptive behavior.The analytical tools used by the agent and their functions are summarized in Appendix 6.

2.1 EXPLAINABILITY

Explainability is essential for understanding prediction reliability, identifying key drivers, and supporting operational decision-making in energy systems. EnergyX integrates multiple complementary analyses that characterize temporal structure, quantify feature influence, explain anomalies, and evaluate hypothetical scenarios.

Temporal Structure and Statistical Diagnostics: EnergyX analyzes the temporal structure of the series using seasonal-trend decomposition Hyndman & Khandakar (2008) and lag correlation analysis. Decomposition Cleveland et al. (1990) separates the signal into trend, seasonal, and residual components, revealing long-term changes, periodic patterns etc. Strong trend and seasonal components indicate predictable behavior, while high residual variance signals inherent uncertainty. Lag correlations identify persistence and recurring cycles, such as daily or weekly consumption patterns.
Feature Attribution and Driver Identification: To explain what drives model predictions, EnergyX computes feature attribution using SHAP values Lundberg et al. (2020), tree-based feature importance Breiman (2001); Ke et al. (2017), and feature-target correlations. These provide complementary perspectives: SHAP values quantify directional feature contributions, feature importance measures model reliance on features. Together with correlation, these analyses distinguish consistent drivers from context-dependent or weak influences, helping users understand which factors most affect energy consumption.

Anomaly Explanation through Contextual Evidence: When anomalies are detected, EnergyX explains them using statistical severity measures, agreement across detection methods, and local temporal context. Statistical scores quantify anomaly extremeness, while multi-method agreement increases confidence. Context windows showing target and feature behavior before and after each anomaly help distinguish isolated measurement errors from feature-driven or system-level events. Detection methods and parameters are detailed in Appendix 7.

Counterfactual Analysis for Actionable Insight: Counterfactual analysis enables users to understand how feature changes influence predictions and explore actionable scenarios.

In *forward analysis* ("what-if" reasoning), users modify input features and observe the resulting prediction change. For example, increasing temperature by 2°C may result in higher predicted energy consumption, revealing sensitivity to weather conditions.
In *inverse analysis* ("target-seeking" reasoning), users specify a desired outcome and the system identifies feature changes that could achieve it. For example, reducing peak consumption by 20% may require lowering specific operational features within feasible ranges. Counterfactual generation details are provided in Appendix 7. Together, these complementary methods provide transparent, grounded, and actionable explanations that help users interpret predictions, understand anomalies, and evaluate potential interventions.

3 EXPERIMENTAL EVALUATION

The orchestrator LLM used is **llama-3.3-70b** for all evaluation purposes. For implementation details refer to Table 3, 4 and 5.We demonstrate EnergyX's capabilities through representative user interactions covering univariate forecasting, multivariate forecasting, and anomaly detection on the ASHRAE dataset Miller et al. (2020) in the figure 2.

Example 1: Forecast query
Query: Forecast electricity consumption for the next 4 days.
Output: Mean = 215 $\pm$ 22.9 kWh/hr. Best model: LightGBM (MAPE 11.26%). Aggregation: inverse-MAPE weighted.

Example 2: Forecast with explanation
Query: Forecast next 5 days using air temperature, cloud coverage, and hour.
Output: Mean = 264.62 $\pm$ 10.11 kWh/hr. Best model: RandomForest (MAPE 8.74%).
Explanation: RandomForest slightly outperforms LightGBM (MAPE 8.74% vs 8.86%) with lower variance, indicating better accuracy and stability. For both models, hour is the dominant driver, confirming strong daily seasonality, while air temperature shows weak linear correlation but meaningful non-linear influence. Lag analysis reveals strong persistence (lag-1=0.94) and daily seasonality (lag-24=0.83), supported by decomposition showing high seasonal strength (0.77) and moderate trend. Overall, forecasts are primarily governed by time-of-day patterns with secondary weather effect.

Example 3: Detect anomalies with explanation
Query: Detect anomalies in electricity and explain.
Output: Four anomalies were detected with moderate inter-method agreement (3.8/7), indicating uncertainty but consistent statistical severity. The anomalies are extreme relative to the distribution (mean z-score = 247.11 ± 97.59) and lie far beyond normal IQR bounds (28.50–472.16), with values exceeding the 95th (398.96) and approaching the 99th percentile (428.17), confirming rarity. Context analysis shows a large spike at index 0 (571.0) and sharp drops at indices 553 (70.645), 808 (69.621), and 809 (0.0). These abrupt deviations suggest potential measurement errors, sensor faults, or external system events rather than normal process variation.

Figure 2: Example natural language queries and grounded agent outputs. Additional examples in Appendix 7.

We compare the MAE and MAPE% metrics of forecasting on the ETT dataset, with TimeSeriesScientist Zhao et al. (2025), on forecast horizons of 96 and 192 of all 4 variants. All datasets divided into 25 slices, 512 time steps per slice, results are averaged across the 25 slices.

Table 1: Comparison on ETT Zhou et al. (2021) datasets (MAE / MAPE%). Lower is better.

Dataset-H	EnergyX (Ours)	TimeSeriesScientist
	↓ MAE / MAPE	↓ MAE / MAPE
ETTh1-96	2.165 / 45.73	1.81 / 13.9
ETTh1-192	2.24 / 40.67	2.05 / 31.0
ETTh2-96	4.791 / 21.78	4.50 / 18.9
ETTh2-192	5.432 / 47.46	4.47 / 12.8
ETTm1-96	1.775 / 22.20	1.68 / 15.7
ETTm1-192	1.989 / 41.76	1.89 / 19.9
ETTm2-96	3.890 / 44.09	3.63 / 40.5
ETTm2-192	5.89 / 35.79	4.77 / 30.5

From Table 1, EnergyX achieves reasonable forecasting performance relative to TimeSeriesScientist while maintaining a general-purpose, model-agnostic design. The observed performance reflects the capabilities of the underlying forecasting components and can be further improved by incorporating more advanced or specialized models within the framework.

Beyond forecasting accuracy, EnergyX's primary contribution lies in unifying forecasting, anomaly detection, and counterfactual analysis into a single, user-steerable workflow. As shown in Table 8, the system reliably executes diverse tasks from natural language queries, achieving 100% success on forecasting, anomaly detection, and report generation, with slightly lower performance on complex counterfactual queries (80–84%), which are mitigated through the human-in-the-loop planning step.

Importantly, the LLM operates only at the task specification stage, contributing minimal overhead ($< 3\%$ runtime), while model training dominates computation.

These results highlight a key trade-off: EnergyX prioritizes usability, interpretability, and task generality over single-model optimization. This enables domain users to interact with time series data through natural language while obtaining transparent, multi-level explanations grounded in statistical diagnostics and model behavior. At the same time, the modular design allows seamless integration of stronger forecasting models, making the framework extensible without redesigning the system pipeline.

A current limitation is the reliance on preprocessed inputs, which may restrict usability in real-world deployments. Additionally, while explainability is effective at the individual time-series level, extending it to fleet-level or cross-entity analysis remains an open challenge.

4 Conclusion

We presented EnergyX, an agentic framework for accessible and explainable energy forecasting and anomaly detection. By combining LLM-based task specification, automated ensemble modeling, and integrated explainability, EnergyX translates natural language intent into executable workflows and actionable insights.

Experimental results demonstrate that EnergyX provides reliable forecasting performance while enabling unified, multi-task time series analysis with transparent explanations. The framework emphasizes usability and interpretability, making it suitable for domain users who lack expertise in deploying complex models.

Future work will focus on integrating automated data preprocessing, expanding the pool of forecasting models, evaluating on labeled anomaly detection benchmarks such as LEAD Gulati & Arjunan (2022), and extending explainability to fleet-level analysis for improved scalability and real-world applicability.

References

J. M. Bates and C. W. J. Granger. The combination of forecasts. *OR*, 20(4):451–468, 1969. ISSN 14732858. URL http://www.jstor.org/stable/3008764.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794. ACM, August 2016. doi: 10.1145/2939672.2939785. URL http://dx.doi.org/10.1145/2939672.2939785.

Robert B Cleveland, William S Cleveland, Jean E McRae, and Irma Terpenning. Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1):3–73, 1990. URL https://math.unm.edu/~lil/Stat581/STL.pdf.

Thomas G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, pp. 1–15, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg. ISBN 978-3-540-45014-6.

Azul Garza and Renée Rosillo. Timecopilot, 2025. URL https://arxiv.org/abs/2509.00616.

Manoj Gulati and Pandarasamy Arjunan. Lead1.0: A large-scale annotated dataset for energy anomaly detection in commercial buildings. *arXiv preprint arXiv:2203.17256*, 2022.

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

Rob J. Hyndman, Anne B. Koehler, J. Keith Ord, and Ralph D. Snyder. *Forecasting with Exponential Smoothing: The State Space Approach*. Springer Series in Statistics. Springer, Berlin, Heidelberg, 2008. doi: 10.1007/978-3-540-71918-2.

International Energy Agency. Electricity 2024: Analysis and forecast to 2026. Technical report, International Energy Agency, Paris, France, 2024. URL https://www.iea.org/reports/electricity-2024. Licence: CC BY 4.0.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. 12 2017.

Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting, 2020. URL https://arxiv.org/abs/1912.09363.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020. doi: 10.1038/s42256-019-0138-9. URL https://doi.org/10.1038/s42256-019-0138-9.

Clayton Miller, Anjukan Kathirgamanathan, Bianca Picchetti, Pandarasamy Arjunan, June Young Park, Zoltan Nagy, Paul Raftery, Brodie W. Hobson, Zixiao Shi, and Forrest Meggers. The building data genome project 2, energy meter data from the ashrae great energy predictor iii competition. *Scientific Data*, 7(1):368, 2020. doi: 10.1038/s41597-020-00712-x. URL https://doi.org/10.1038/s41597-020-00712-x.

Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting, 2020. URL https://arxiv.org/abs/1905.10437.

Granville Tunnicliffe Wilson. Time series analysis: Forecasting and control,5th edition, by george e. p. box, gwilym m. jenkins, gregory c. reinsel and greta m. ljung, 2015. published by john wiley and sons inc., hoboken, new jersey, pp. 712. isbn: 978-1-118-67502-1. *Journal of Time Series Analysis*, 37:n/a–n/a, 03 2016. doi: 10.1111/jtsa.12194.

U.S. Energy Information Administration. International energy outlook 2023. Technical report, U.S. Energy Information Administration, 2023. URL https://www.eia.gov/outlooks/ieo/. Accessed: 2026-02-14.

Zhendong Wang, Ioanna Miliou, Isak Samsten, and Panagiotis Papapetrou. Counterfactual explanations for time series forecasting. In *2023 IEEE International Conference on Data Mining (ICDM)*, pp. 1397–1402. IEEE, 2023.

Haokun Zhao, Xiang Zhang, Jiaqi Wei, Yiwei Xu, Yuting He, Siqi Sun, and Chenyu You. Timeseriesscientist: A general-purpose ai agent for time series analysis, 2025. URL https://arxiv.org/abs/2510.01538.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):11106–11115, May 2021. doi: 10.1609/aaai.v35i12.17325. URL https://ojs.aaai.org/index.php/AAAI/article/view/17325.

A APPENDIX

A.1 PLOTS FROM ASHRAE DATASET TESTING:

(Data for building_id 15, site_id 0 has been isolated with its corresponding weather data)

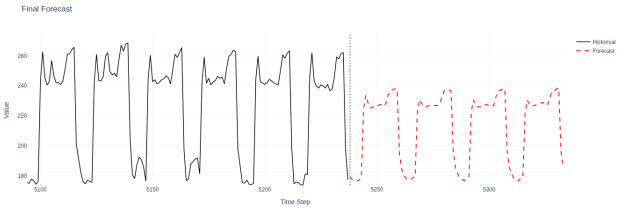

Figure 3: Univariate forecast prediction

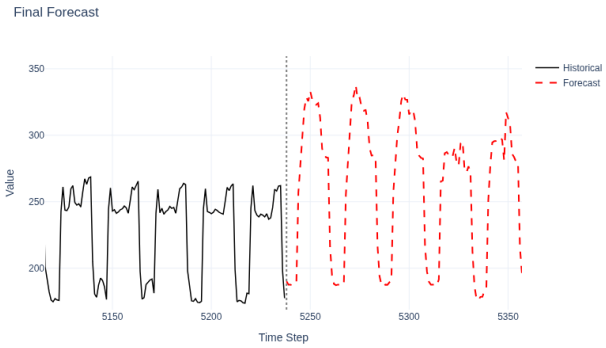

Figure 4: Multivariate forecast prediction

The Firgures 5 and 6 represents the anomalies detected in the historical energy consumption data from the meter_types ”electricity” and ”chilled_water”respectively.

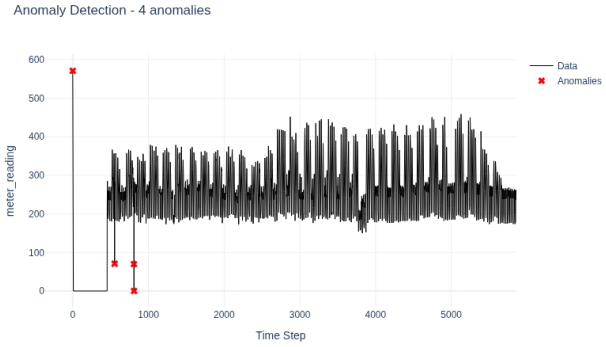

Figure 5: Anomaly Detection for meter_type = electricity, on ASHRAE

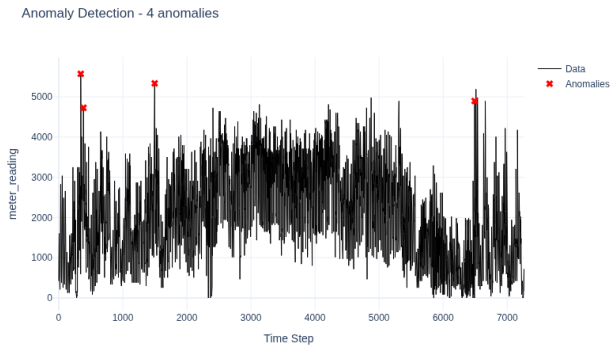

Figure 6: Anomaly Detection for meter_type = chilled_water, on ASHRAE

A.2 EXPERIMENTAL SETUP:

Table 2: Model inventory used by EnergyX

CATEGORY	MODELS
Forecasting	Naive, SeasonalNaive, ARIMA (auto), ETS, RandomForest, LightGBM, N-Beats
Anomaly	Z-score, Modified Z-score (MAD), Rolling statistics, IQR, STL decomposition, Isolation Forest, DBSCAN,

Table 3: System-Level Configuration

Component	Detail
LLM	LLaMA-3.3-70B via Cerebras Cloud (OpenAI-compatible API); local fallback: LLaMA 3.2:3B via Ollama
Two-temperature pattern	Routing/tool selection: T=0.1; report synthesis: T=0.7
Structured output	Typed plans with JSON parsing fallback (3 retries, exponential backoff)
Tool calling	Native tool_calls; regex-based text adapter for fallback
Random seed	42 (models and hyperparameter sampling)
Max agent iterations	25
State management	Pydantic models *(UserTaskPlan, DataProfile)*; shared session dict accumulates cross-tool results

Table 4: Forecasting Models and Hyperparameter Search

Model	Search Space
SeasonalNaive	seasonal_period $\in \{7, 12, 24, 30\}$
ETS	trend $\in \{\text{add}, \text{mul}, \text{none}\}$, seasonal $\in \{\text{add}, \text{mul}, \text{none}\}$
RandomForest	n_estimators $\in \{50, 100, 200\}$, max_depth $\in \{5, 10, 15\}$, n_lags $\in \{6, 12, 24\}$
LightGBM	num_leaves $\in \{15, 31, 63\}$, learning_rate $\in \{0.01, 0.1, 0.3\}$, n_estimators $\in \{50, 100, 200\}$, n_lags $\in \{6, 12, 24\}$
N-BEATS	n_lags $\in \{12, 24, 48\}$, hidden_size $\in \{32, 64, 128\}$, epochs $\in \{30, 50, 100\}$
ARIMA	auto_arima (stepwise, $p \leq 5$, $d \leq 2$, $q \leq 5$, non-seasonal, 720-step context cap)

Expanding-window CV (3 folds); minimum train size $\max(50, 2H)$. If grid > 10, uniformly sample 10 (seed=42). Best model selected by lowest mean MAPE.

Table 5: Ensemble Strategy and Anomaly Detection

Component	Detail
Forecast ensemble	Adaptive selection between inverse-MAPE weighting (top-5) and trim-mean averaging (top-3), choosing the lower estimated CV MAPE. Inverse-MAPE uses $w_i = (1/\text{MAPE}_i)/\sum_j (1/\text{MAPE}_j)$
Statistical detectors (5)	Z-score (σ=3.0), MAD (σ=3.5, 5% contamination cap), rolling stats (window=24, k=3.0), IQR (k=1.5), STL residuals (σ=3.0)
ML detectors (2)	Isolation Forest (100 trees, contamination=0.02), DBSCAN (ε=0.5, min_samples=5; BallTree subsampling for $n > 50$k)
Voting rule	7-method majority ensemble with min_votes=3
Multivariate mode	IsolationForest + DBSCAN on full [target|features]; statistical methods per-channel on target + top-K variance features ($K \leq 9$), union voting

A.3 Tool Descriptions

Table 6: Analytical tools used by EnergyX

TOOL CATEGORY	DESCRIPTION
Data Profiling	Extracts temporal structure, statistical summaries, stationarity, and seasonality
Model Training	Trains forecasting models using rolling-origin cross-validation and hyperparameter search
Forecast Aggregation	Combines model predictions using cross-validation–derived performance weights
Anomaly Detection	Identifies anomalies using statistical, distribution-based, and machine learning methods with consensus scoring
Explainability	Computes feature attribution (SHAP, feature importance), decomposition, lag correlation, and residual diagnostics
Counterfactual Analysis	Estimates forecast changes under hypothetical feature modifications
Report Generation	Synthesizes predictions, diagnostics, and explanations into structured reports

Table 7: Anomaly Explainability Helpers

TOOL CATEGORY	DESCRIPTION
Statistical Quantification	Computes anomaly severity using Z-score, modified Z-score (MAD based), percentile rank, and IQR bounds to quantify how extreme each observation is relative to historical behavior.
Method Agreement (Consensus Scoring)	Calculates the fraction of detection methods that flag each timestep. High agreement indicates high-confidence anomalies, while low agreement highlights uncertain or method-specific detections.
Contextual Feature Evidence	Provides a 5-timestep window of the target and all features around each anomaly. This enables interpretation of whether anomalies arise from isolated target deviations, feature-driven changes, or lagged feature effects.

A.5 RESULTS

Category	Avg Success (%)
Univariate Forecast	100.0
Multivariate Forecast	100.0
Anomaly Detection	100.0
Anomaly Explained	84.0
Forecast Explained	100.0
Counterfactual Forward	84.0
Counterfactual Inverse	80.0
Filtered Query	100.0
Forecast + Anomaly	100.0
Report Generation	100.0

Table 8: Category-wise average success over five runs (each category contains five queries).

Example 1: Forecast query
Query: Forecast electricity consumption for the next 4 days.
Output: Mean = 215 $\pm$ 22.9 kWh/hr. Best model: LightGBM (MAPE 11.26%). Aggregation: inverse-MAPE weighted.

Example 2: Forecast with explanation
Query: Forecast next 5 days using air temperature, cloud coverage, and hour.
Output: Mean = 264.62 $\pm$ 10.11 kWh/hr. Best model: RandomForest (MAPE 8.74%).
Explanation: RandomForest slightly outperforms LightGBM (MAPE 8.74% vs 8.86%) with lower variance, indicating better accuracy and stability. For both models, hour is the dominant driver, confirming strong daily seasonality, while air temperature shows weak linear correlation but meaningful non-linear influence. Lag analysis reveals strong persistence (lag-1=0.94) and daily seasonality (lag-24=0.83), supported by decomposition showing high seasonal strength (0.77) and moderate trend. Overall, forecasts are primarily governed by time-of-day patterns with secondary weather effect.

Example 3: Anomaly detection query
Query: Detect anomalies in consumption through chilled water.
Output: The anomaly detection task has been completed. A total of 4 anomalies were found in the dataset. The method counts are as follows: ZScore (2), ModifiedZScore (0), RollingStats (15), IQR (1), STL (167), IsolationForest (136), and DBSCAN (0). The minimum votes required for an anomaly to be confirmed is 3, and the average agreement among the methods is 3.5. The total number of data points in the dataset is 7269. The task was completed successfully.

Example 4: Anomaly detection with explanation
Query: Detect anomalies in electricity and explain.
Output: Four anomalies were detected with moderate inter-method agreement (3.8/7), indicating uncertainty but consistent statistical severity. The anomalies are extreme relative to the distribution (mean z-score = 247.11 ± 97.59) and lie far beyond normal IQR bounds (28.50–472.16), with values exceeding the 95th (398.96) and approaching the 99th percentile (428.17), confirming rarity. Context analysis shows a large spike at index 0 (571.0) and sharp drops at indices 553 (70.645), 808 (69.621), and 809 (0.0). These abrupt deviations suggest potential measurement errors, sensor faults, or external system events rather than normal process variation.

Example 5: Counterfactual Inverse
Query: Under what conditions would the oil temperature drop to 9 degrees?
Output: The current prediction (16.33) exceeds the target (9.0), and optimization reduces it to 10.69 but cannot fully reach the target, suggesting feature limitations. Achieving the closest feasible value requires small increases in LULL (+1.7%) and MUFL (+3.3%), and a substantial decrease in LUFL (16.7%), the primary driver. While LULL and MUFL remain within historical ranges, the large LUFL reduction may be impractical. The remaining gap indicates the target is likely unattainable with the available features, implying influence from unobserved factors.

Example 6: Counterfactual Forward
Query: What happens to oil temperature if HUFL increases by 2 for the next 24 hours? The baseline forecast mean is 13.32, while the counterfactual forecast mean is 12.87. The mean impact of the intervention is -0.45, and the max impact is 1.26. This change means that the target variable is expected to decrease by 0.45 on average, with a maximum decrease of 1.26, when the HUFL feature is increased by 2.0. However, the model MAPE is 38.7%, which indicates high uncertainty in the counterfactual estimates.

A.4 Counterfactual Analysis

We support two complementary counterfactual modes using the best multivariate model selected via cross-validation.

A.1 Forward Counterfactual (What-if). Given user-defined feature interventions, future feature trajectories are first forecast univariately, perturbed by the specified deltas, and passed through the multivariate predictor to estimate outcome change Δy relative to baseline.

A.2 Inverse Counterfactual (Target-seeking). To reach a desired target y^*, we solve a constrained optimization over the top-3 SHAP-ranked features, minimizing $(\mathcal{M}(\mathbf{x}) - y^*)^2$ within historical bounds using Nelder–Mead, returning minimal feasible feature adjustments.

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} (\mathcal{M}(\mathbf{x}) - y^*)^2 \quad \text{s.t.} \quad x_i^{\min} \le x_i \le x_i^{\max},\ i \in \mathcal{F}_{\text{top-3}},$$

Predictions are accompanied by model MAPE for uncertainty context; warnings are issued if optimization error exceeds 10%. Limitations include additive forward assumptions, independent feature optimization, extrapolation risk, and possible local minima.

Figure 7: Example natural language queries and grounded agent outputs.