OPSD PowerDesk: Day-Ahead Forecasting and Anomaly Detection

Saipranav Reddy (SE23UARI025)
Varun Pandrangi (SE23UCSE137)
Bharat Reddy (SE23UARI020)

November 27, 2025

Introduction

Reliable short-term electricity load forecasting is crucial for system operators, market participants, and policymakers because it directly affects security of supply and economic dispatch decisions in European power systems. This report investigates day-ahead load forecasting and anomaly detection using hourly demand data from the Open Power System Data (OPSD) time-series package, focusing on three large European countries and comparing a classical SARIMA model with a GRU neural network for both accuracy and robustness.

The central research question is: *How well can SARIMA and GRU models forecast day-ahead electricity load for Germany, France, and Spain, and how can their residuals be exploited for systematic anomaly detection and online model adaptation?* Answering this question is important because accurate forecasts and timely detection of structural breaks, such as COVID-19 lockdowns or sudden behavioral shifts, enable better reserve sizing, risk management, and data quality control in operational tools built on OPSD data.

Data and STL Decomposition

The analysis uses the OPSD time-series data package, which provides cleaned hourly electricity consumption (load) for many European countries, compiled from the ENTSO-E Transparency Platform and national sources. For this project, three countries are selected: Germany (DE), France (FR), and Spain (ES), each with roughly five years of hourly load data between 2015 and 2020, giving on the order of 50,000 observations per country.

Table 1: Countries and time ranges used

Code	Country	Data Points	Time Range
DE	Germany	$\approx$ 50,000	2015–2020
FR	France	$\approx$ 50,000	2015–2020
ES	Spain	$\approx$ 50,000	2015–2020

To understand structure in the load series, STL (Seasonal-Trend decomposition using Loess) is applied with a period of 24 hours, decomposing each series into trend, seasonal, and

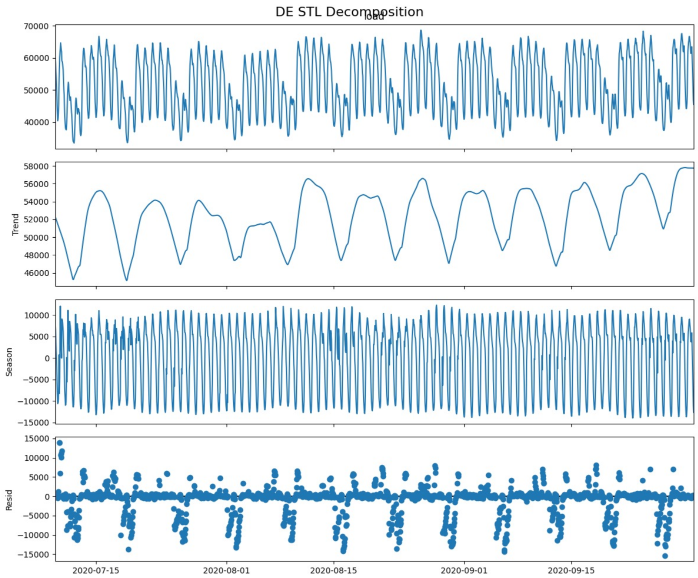

Figure 1: STL decomposition of German hourly load with daily

remainder components. STL uses local regression (LOESS) smoothing in iterative inner and outer loops to alternately refine the seasonal and trend components while down-weighting outliers, which makes it robust to gradual structural changes and extreme values in power-system time series.

For Germany, the STL components reveal strong daily seasonality with pronounced business-hour peaks (approximately 8–18h) and a gradual downward trend in the later years, consistent with efficiency improvements and evolving demand patterns in large industrial systems. For France, the seasonal component shows higher winter loads and clear weekend troughs, reflecting extensive use of electric heating and strong working-week cycles. For Spain, the decomposition indicates moderate daily seasonality with midday peaks and relatively lower overall variability compared to Germany and France, which aligns with its climate and different industrial and residential load mix.

Order Selection

Model order selection for SARIMA starts from differencing decisions and inspection of autocorrelation structures. Non-seasonal differencing order d is kept at 0 or 1 as needed for stationarity, while seasonal differencing with order $D = 1$ and period $s = 24$ is used to remove daily seasonality before fitting autoregressive (AR) and moving-average (MA) components.

After differencing, autocorrelation function (ACF) and partial autocorrelation function (PACF) plots are computed for the residual series. For Germany, the ACF shows significant spikes at multiples of 24 hours and slowly decaying seasonal correlations, while the PACF has a few strong low-order lags, indicating the need for both non-seasonal and seasonal AR and MA terms.

For Germany, a grid search over SARIMA parameters is performed and models are ranked by information criteria such as AIC and BIC, which balance fit quality and parameter count. The five best configurations by BIC are reported in Table 2; the chosen specification aims to

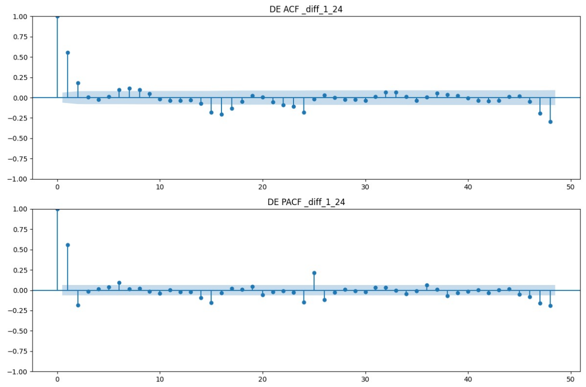

Figure 2: ACF and PACF for Germany after seasonal differencing

capture key dynamics with minimal complexity.

Table 2: Top 5 SARIMA configurations for Germany by BIC

Rank	Order (p, d, q)	Seasonal $(P, D, Q, 24)$	AIC	BIC
1	(2, 0, 1)	(1, 1, 1, 24)	10071.76	10098.36
2	(1, 0, 1)	(1, 1, 1, 24)	10085.42	10105.89
3	(2, 0, 2)	(1, 1, 1, 24)	10073.21	10106.01
4	(1, 0, 2)	(1, 1, 1, 24)	10082.55	10109.15
5	(2, 1, 1)	(1, 1, 1, 24)	10089.33	10115.93

Across all three countries, the final SARIMA specifications converge to similar seasonal structures with daily period $s = 24$ and seasonal differencing $D = 1$, differing mainly in the low-order non-seasonal AR and MA parts. The selected orders are summarized in Table 3.

Table 3: Final SARIMA orders

Country	Order (p, d, q)	Seasonal (P, D, Q, s)
DE	(2, 0, 1)	(1, 1, 1, 24)
FR	(2, 0, 2)	(1, 1, 1, 24)
ES	(1, 0, 2)	(1, 1, 1, 24)

Forecast Results

Two models are evaluated for day-ahead forecasting: a SARIMA model that produces point forecasts with prediction intervals, and a GRU-based recurrent neural network implemented in PyTorch with 168 hours of input history and 24-hour direct multi-horizon output. Forecast quality is assessed using several standard metrics: MASE, sMAPE, MSE, RMSE, MAPE, and coverage of 80 percent prediction intervals in the SARIMA case.

On the development set (10 percent validation split at the end of each time series), GRU models generally achieve slightly lower sMAPE and MSE than SARIMA, while SARIMA provides calibrated uncertainty estimates with coverage in a reasonable range around the nominal level. Table 4 reports detailed numbers for each country and model.

Table 4: Development-set metrics (10% of data)

Country	Model	MASE	sMAPE (%)	MSE	RMSE	MAPE (%)	80% Cov.
DE	SARIMA	0.608	5.24	1.44e7	3796	5.31	86.6%
DE	GRU	0.626	4.91	1.34e7	3662	4.81	–
FR	SARIMA	0.828	4.60	1.23e7	3508	4.56	72.7%
FR	GRU	0.753	4.09	8.86e6	2977	4.02	–
ES	SARIMA	0.722	4.68	3.39e6	1841	4.76	76.2%
ES	GRU	0.650	3.97	2.39e6	1545	3.89	–

Final test-set results on a separate 10 percent holdout confirm that all models achieve MASE below 1.0, meaning that they outperform a seasonal naive baseline that repeats the value from one day earlier. Table 5 shows that GRU tends to be slightly more accurate in Germany and France, while SARIMA and GRU are very close in Spain, where the load series is smoother.

Table 5: Test-set metrics (10% of data)

Country	Model	MASE	sMAPE (%)	MSE	RMSE	MAPE (%)	80% Cov.
DE	SARIMA	0.539	5.19	1.22e7	3496	5.27	90.4%
DE	GRU	0.524	4.65	1.21e7	3474	4.50	–
FR	SARIMA	0.705	4.84	1.20e7	3469	4.98	80.6%
FR	GRU	0.569	3.97	8.44e6	2906	4.09	–
ES	SARIMA	0.637	4.52	2.46e6	1567	4.58	78.8%
ES	GRU	0.664	4.63	2.43e6	1560	4.64	–

For operators, this means that GRU offers modest improvements in point accuracy, especially in systems with strong weather-driven variability, but SARIMA remains valuable when calibrated prediction intervals are needed for risk-aware decision making. The overall performance across metrics indicates that both approaches are suitable for deployment in a day-ahead forecasting module of OPSD PowerDesk.

Anomaly Detection

Anomaly detection is applied to the forecast residuals to highlight time periods where observed load behaves unexpectedly relative to recent history and model predictions. Two complementary methods are used: a rolling z-score based on a 336-hour (14-day) window and a CUSUM procedure that tracks cumulative deviations from the expected residual mean.

The rolling z-score standardizes residuals using a moving mean and standard deviation, and potential anomalies are those with absolute z-score above a chosen threshold, here inspected around $|z| \geq 3.0$. CUSUM is parameterized with $k = 0.5$ and $h = 5.0$ to detect persistent mean shifts that may indicate events such as lockdowns, data jumps, or step changes in demand.

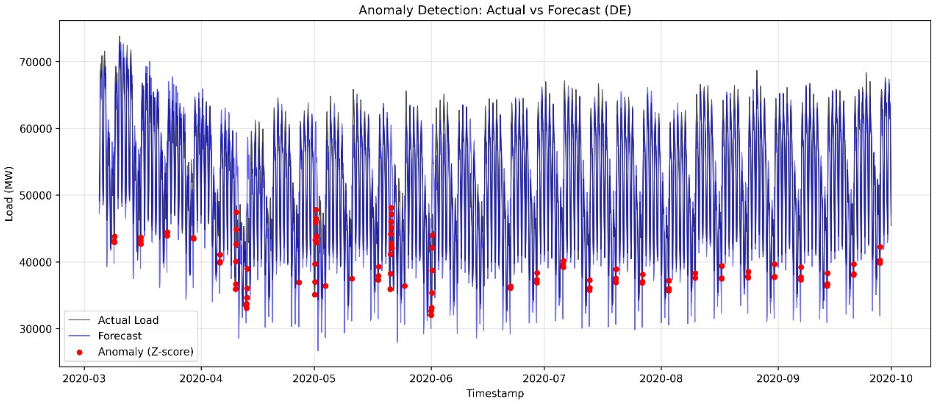

Figure 3: Example anomaly detection window for Germany: actual vs. SARIMA forecast with flagged anomalies in red.

Table 6 lists the ten largest rolling z-scores for Germany during the test period. Several of the largest deviations cluster in March and April 2020, coinciding with the first COVID-19 lockdown and associated changes in industrial and residential consumption, and some occur during summer weekends where load patterns differ from the historical norm.

Table 6: Top 10 rolling z-score anomalies for Germany

Rank	Timestamp	Actual (MW)	Pred. (MW)	Z-Score	Flagged
1	2020-07-25 00:00	38,467	29,863	2.97	No
2	2020-03-28 00:00	42,406	34,532	2.94	No
3	2020-08-01 00:00	38,943	30,872	2.89	No
4	2020-04-04 00:00	42,387	34,034	2.89	No
5	2020-09-19 00:00	39,985	32,155	2.78	No
6	2020-03-21 00:00	44,513	36,365	2.77	No
7	2020-09-27 23:00	40,414	31,914	2.76	No
8	2020-03-29 23:00	43,930	36,140	2.74	No
9	2020-03-14 00:00	46,875	38,024	2.74	No
10	2020-08-08 00:00	38,602	30,907	2.72	No

Figure **??** illustrates a typical anomaly window, showing actual load, SARIMA prediction, prediction intervals, and residual z-scores. Figure **??** shows the CUSUM chart of standardized residuals, which exhibits a pronounced upward excursion during early 2020 that then stabilizes after refits, consistent with the pattern of COVID-19 related demand shocks.

The results suggest that weekend midnight hours often show the largest pointwise deviations because the model underestimates load in those specific contexts, and that COVID-19 triggered sustained positive residuals before the online adaptation loop re-aligned the model. No residuals in the final test window exceed the strict $|z| \geq 3.0$ flagging threshold, indicating that rolling normalization combined with regular refits gives a reasonably adaptive and robust anomaly filter.

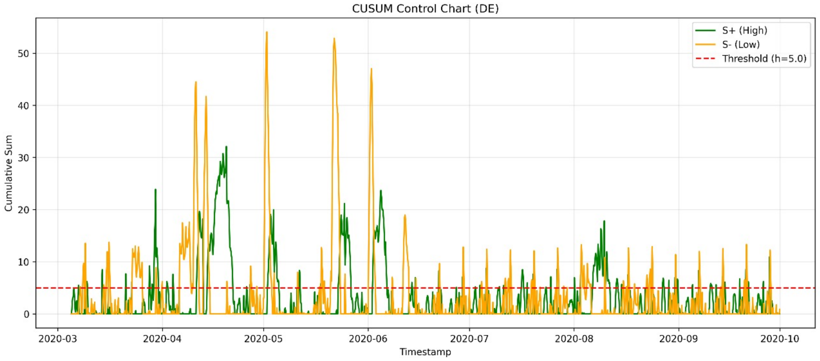

Figure 4: CUSUM chart for German residuals highlighting persistent positive drift during early 2020.

ML Anomaly Classifier

To reduce false positives from simple z-score thresholding, a logistic regression classifier is trained to distinguish true anomalies from normal points using silver labels built from residual magnitude and prediction interval information. Positive labels are assigned to observations with $|z| \geq 3.5$ or those outside the SARIMA prediction interval with $|z| \geq 2.5$, while negative labels correspond to $|z| < 1.0$ and residuals inside the interval.

The feature set includes lagged loads at 24 and 48 hours, a rolling 24-hour mean and standard deviation, categorical hour-of-day and day-of-week indicators, and local forecast error context via recent residuals. On a held-out test set of 88 samples with 44 anomalies, the classifier achieves a PR-AUC of 0.822 and an F1 score of 0.649 at 80 percent precision, which indicates effective prioritization of the most critical anomalies while limiting false alarms.

Inspecting the fitted coefficients shows that recent forecast errors (for example residuals at lag 1 and lag 24) are the most informative features because persistent local error patterns strongly signal a structural deviation rather than random noise. Rolling standard deviation and calendar features add value by capturing volatility regimes and typical low-load hours where deviations are more surprising, leading to a cleaner anomaly list than using z-scores alone.

Live Simulation and Online Adaptation

A live simulation is implemented by feeding historical data sequentially through an online forecasting loop for Germany with 2,000 consecutive hours. The chosen adaptation strategy is rolling SARIMA with a 120-day sliding training window and two trigger types: scheduled daily refits at 00:00 and additional refits when an exponentially weighted moving average of absolute z-scores exceeds the empirical 95th percentile of its historical distribution.

Table 7 shows two example update events around March 2020, comparing rolling 7-day MASE and 7-day prediction interval coverage before and after refitting. In both cases, refits substantially reduce MASE (better accuracy) but temporarily lower coverage as intervals tighten around the new regime before stabilizing in subsequent cycles.

Over the full simulation, the online system performs 132 model updates, of which 84 are scheduled daily refits and 48 are drift-triggered events. The average refit duration is

Table 7: Before/after statistics for selected SARIMA updates (Germany)

Update Date	Reason	7d MASE Before	7d MASE After	7d Cov. Before	7d Cov. After
2020-03-12	Scheduled	0.588	0.466	81.5%	53.6%
2020-03-17	Scheduled	0.517	0.470	86.3%	52.4%

approximately 70–80 seconds, which is acceptable for daily adaptation in an offline planning context but would need optimization or parallelization to support more frequent updates or many concurrent regions.

Limitations

First, the GRU model provides only point forecasts; obtaining calibrated prediction intervals would require probabilistic modeling, ensembles, or techniques such as Monte Carlo dropout, which increase computational cost and implementation complexity. Second, the live simulation uses historical data in a pseudo-real-time fashion but does not include a full streaming architecture, so real deployments would still need robust pipelines for data ingestion, monitoring, and fallback strategies.

Third, SARIMA refitting is relatively expensive, taking around one minute per update, so very frequent adaptation or scaling to many countries would require substantial compute resources or lighter-weight incremental models. Fourth, the models do not incorporate explicit exogenous variables such as temperature, holidays, or policy interventions, even though these are known to be important drivers of electricity demand and would likely reduce systematic biases during extreme events.

Finally, only a single adaptation strategy based on rolling SARIMA is implemented, while alternative or complementary strategies such as online fine-tuning of GRU models or hybrid statistical–machine-learning approaches could provide faster adaptation but would require careful regularization and access to hardware accelerators.

Appendix: File Locations

Table 8 lists the main artifacts produced by this project and their locations in the repository.

Table 8: Key project artifacts and locations

Artifact	Path
STL plots	`outputs/plots/<CC>_stl_decomposition.png`
ACF/PACF plots	`outputs/plots/<CC>_acf_pacf_diff_1_24.png`
Forecast results (SARIMA)	`outputs/<CC>_forecasts_test.csv`
Forecast results (GRU)	`outputs/<CC>_forecasts_test_gru.csv`
Anomaly data	`outputs/<CC>_anomalies.csv`
ML anomaly evaluation	`outputs/anomaly_ml_eval.json`
Live simulation log (DE)	`outputs/DE_online_updates.csv`
Metrics summary	`outputs/metrics_summary.csv`

https://github.com/VarunPandrangi/OPSD$_{powerdesk.git}$