FFG
Forschung wirkt.
Bundesministerium
Klimaschutz, Umwelt,
Energie, Mobilität,
Innovation und Technologie
AI for Green – Ausschreibung 2023

This project is funded by the FFG.
www.ffg.at

transpAIrent.energy

Transparent AI Forecasts for Green Energy in Austria

Project number: 910239

D3.1 Algorithm review and implementation documentation

D3.2 Algorithm validation and result documentation

30.11.2025

Authors	**Organisation**
Peter WIDHALM	AIT
Verena ALTON	AIT

Funding Disclaimer

This project is being carried out as part of the 2023 “AI for Green” call for proposals from the Federal Ministry for Innovation, Mobility and Infrastructure (BMIMI). The processing is carried out on behalf of the BMIMI by the Austrian Research Promotion Agency (FFG). The project is funded as part of the topic of digital technologies, an initiative of the BMIMI, under the grant agreement number FO999910239. The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the texts, or in the decision to publish the results.

Table of Contents

Funding Disclaimer 1
Executive Summary 4
1. Introduction 5
1.1. State of the Art 5
1.1.1. Deterministic Models 5
1.1.2. Generative Models 6
2. Methodology 7
2.1. Baseline models 7
2.1.1. SARIMAX 7
2.1.2. Long Short-Term Memory (LSTM) 7
2.1.3. Multivariate Linear Regression 8
2.2. Generative Adversarial Networks (GAN) 8
2.2.1. Conditional Time Series GAN (GAN_0) 8
2.2.2. Modified Conditional Time Series GAN (GAN_1) 9
2.2.3. Combined LSTM and Conditional Time Series GAN (GAN_2) 9
2.3. Temporal Fusion Transformer (TFT) 9
2.4. Implementation 10
2.4.1. Technical Details 10
2.4.2. Features 10
2.4.3. Forecasting Setup 10
2.4.4. Data Preprocessing 10
3. Results and Discussion 11
3.1. Qualitative Assessment 11
3.1.1. Day-ahead Prices 11
3.1.2. CO_2-intensity 12
3.1.3. Balancing Energy Activations 12
3.2. Quantitative Assessment 13
3.2.1. Day-ahead prices 14
3.2.2. CO_2-intensity 15
3.2.3. Positive aFRR activations 16
3.2.4. Negative aFRR activations 17
3.3. Case Study: Day-ahead Price Prediction in the Energy Crisis 17
3.3.1. Metrics Results 19
3.3.2. Simplified Optimization Approach 20
3.4. Temporal Fusion Transformer 23
4. Conclusion 24
4.1. Forecasting Results 24

4.2. Case Study: Day-Ahead Price Forecasting During the Energy Crisis 25
4.3. Key Findings 25
References 26
Appendix 28
Data 28
Variable Name 28
Description 28
Source 28

Executive Summary

The increasing volatility and complexity of modern electricity systems, driven by the growing share of renewable energy sources, dynamic market conditions, and weather dependencies, pose significant challenges for energy system forecasting and operation. This document combines Deliverables 3.1 and 3.2 of the *transpAIrent.energy* project: Deliverable 3.1 documents the review, implementation, and evaluation of forecasting algorithms developed within Work Package 3, which focuses on probabilistic prediction of key energy system variables in Austria, while Deliverable 3.2 presents the results of the algorithm validation using real-world data.

Objectives and Scope

Deliverable 3.1 presents the methodological foundations and implementation details for forecasting three core target variables:

- Day-ahead electricity prices
- CO_2-intensities of electricity generation
- Balancing energy activations (aFRR)

These variables are critical for ensuring grid stability, supporting transparent market participation, and enabling the optimized operation of flexible energy assets such as biomass plants.

Methodological Approach

We implemented a diverse set of forecasting models, including both deterministic and generative approaches:

- **Baseline Models**: SARIMAX, LSTM, Multivariate Linear Regression, and Previous Day baseline.
- **Three GAN-based architectures**: (1) Conditional Time Series GAN (GAN_0), (2) a modified GAN with improved training stability (GAN_1), and (3) a hybrid ensemble combining LSTM and GAN components (GAN_2).
- **Transformer-Based Model**: Temporal Fusion Transformer (TFT) for complex temporal forecasting.

Each model was trained on real-world data and evaluated using probabilistic metrics (CRPS, NLL, Coverage Probability, Sharpness) and a simplified optimization framework that mimics real-life energy trading decisions.

Results

- **Forecasting Performance**: GAN-based models demonstrated strong performance across most metrics, particularly in capturing uncertainty and adapting to volatile conditions. GAN_1 and GAN_2 were especially effective in providing well-calibrated probabilistic forecasts.
- **Case Study of Day-ahead Price Predictions in Energy Crisis Scenario**: A comparative analysis showed that GAN models significantly outperformed deterministic baselines when trained on normal data but tested on crisis-period data. Their ability to model non-linear, non-stationary patterns proved advantageous in extreme market conditions.
- **Optimization Relevance**: By applying forecasts in a simplified buy-low/sell-high trading model, we illustrated the operational utility of probabilistic predictions. GAN models consistently yielded higher and more stable revenues under crisis scenarios, while deterministic models often failed to adapt.

1. Introduction

The transition to a low-carbon and sustainable energy system requires the widespread integration of renewable energy sources such as photovoltaic and wind power. While essential for decarbonization, these technologies introduce significant variability into electricity generation, increasing the complexity and volatility of modern power systems. At the same time, fluctuating electricity demand, dynamic market conditions, and weather dependencies further increase operational uncertainty.

To maintain grid stability and ensure that electricity supply and demand are continuously balanced, accurate forecasting of key energy system variables is critical. Such forecasts enable a wide range of benefits: improved grid reliability, more efficient energy trading strategies, reductions in greenhouse gas emissions, and optimized operation of flexible energy assets.

This deliverable focuses on three forecasting targets that are essential for the Austrian electricity system:

- Day-ahead electricity prices, which influence bidding strategies and market participation.
- CO_2 intensities of electricity production, which are relevant for environmental impact assessment and emissions reporting.
- Automatic Frequency Restoration Reserve (aFRR) activations, which reflect real-time grid balancing requirements and system flexibility needs.

Forecasting these variables provides actionable insights for various stakeholders, particularly operators of flexible assets such as biomass plants. Unlike variable renewables, biomass facilities can store fuel and adjust output based on expected market and system conditions. However, their operational planning requires foresight, due to constraints such as fuel conversion time and regulatory commitments. Probabilistic forecasts support such forward-looking decision-making by quantifying the likelihood of different future outcomes.

Recent advances in artificial intelligence (AI) have opened new possibilities for forecasting in the energy sector. In particular, generative AI models go beyond traditional point forecasts by generating entire distributions of possible future values. This probabilistic perspective captures the inherent uncertainty of energy systems, offering more robust inputs for both operational and strategic decision-making. (Chai et al., 2024; Lago et al., 2021; Maji et al., 2023)

In this deliverable, we present a review and practical documentation of the forecasting algorithms developed within Work Package 3 of the *transpAIrent.energy* project. The focus lies on the implementation and evaluation of advanced generative AI models for forecasting day-ahead electricity prices, CO_2-intensities, and aFRR activations. We assess their performance both quantitatively (through statistical metrics) and qualitatively (via visual assessment and a practical use case), with the aim of advancing transparent, accessible, and high-quality forecasting tools for the Austrian energy market.

1.1. State of the Art

This section provides an overview of relevant forecasting models and techniques applicable to probabilistic time series forecasting of energy market variables.

1.1.1. Deterministic Models

A variety of deterministic models have been established for time series forecasting. Among the most common are the Autoregressive Integrated Moving Average (ARIMA) models, including extensions such as ARMAX (which incorporates exogenous variables) and SARIMAX (which accounts for seasonality). These models are valued for their interpretability and solid baseline performance. However, they often struggle to capture the highly non-linear interactions and complex dynamics that characterize modern electricity markets. Nevertheless, they continue to serve as important benchmarks in both academic research and real-world applications. (Shah et al., 2022)

Deep Neural Networks (DNNs) have been widely applied to capture complex, non-linear patterns in data. However, they are not inherently designed for sequential data like time series, which limits their suitability for electricity market forecasting tasks. (Lago et al., 2021; Tschora et al., 2022)

Long Short-Term Memory (LSTM) networks are specifically designed to handle sequential data and can effectively model long-term dependencies within time series. This makes them more appropriate than DNNs for forecasting electricity market variables. LSTM networks are therefore commonly used as a reference model in comparative studies. (Tschora et al., 2022)

Hybrid models that combine different machine learning techniques have demonstrated improved performance in many cases. For example, combining Convolutional Neural Networks (CNNs) with LSTM networks has been shown to enhance predictive accuracy by capturing both spatial and temporal features. Ensemble methods also stand out for their robustness and accuracy, as they aggregate predictions from multiple models, typically through techniques such as stacking or bagging. These methods are widely used in state-of-the-art forecasting systems. (Lago et al., 2021)

1.1.2. Generative Models

Recent advancements have led to the development of generative models tailored for probabilistic forecasting in energy markets. One notable example is the Weak Innovation AutoEncoder-based Generative Probabilistic Forecasting (WIAE-GPF) model. This approach leverages the Wiener-Kallianpur innovation representation to generate realistic future samples of multivariate time series data, including electricity prices. By incorporating a Variational Autoencoder (VAE), WIAE-GPF learns a latent representation of historical price dynamics and generates diverse sets of possible future price trajectories. This method has demonstrated superior performance compared to both classical statistical methods and other machine learning approaches on real-world electricity market datasets. (Wang et al., 2024)

Another promising approach is the Conditional Time Series GAN (CTSGAN) from (Lu et al., 2022), which extends the Time Series GAN model (Smith & Smith, 2020) by incorporating conditional inputs and scenario-based forecasting capabilities. Built on the Wasserstein GAN framework (Arjovsky et al., 2017), CTSGAN can generate multiple plausible future scenarios for day-ahead electricity prices, effectively capturing the uncertainty inherent in electricity markets. The model has been validated in case studies from the Australian National Electricity Market (NEM) and has shown robust performance across various market conditions. In this project, both the original CTSGAN model and two modified versions have been implemented and tested on Austrian market data to assess their predictive accuracy and operational relevance.

The Temporal Fusion Transformer (TFT) architecture, proposed by (Lim et al., 2020) and building upon the class of Transformer models, combines recurrent layers with self-attention layers to provide interpretable results and capture long-term dynamics. The authors have shown superior performance on several real-world datasets compared to baseline models. In this project, the TFT model was implemented as another generative AI model and compared against the GAN-based models.

2. Methodology

Three GAN-based models and one Transformer-based model were implemented, together with three established baseline models allowing for a rigorous comparison of the performance and quality of the forecasts. The following section provides an overview of the different models and their implementation.

2.1. Baseline models

Three different baseline models were implemented, which present well-established methods for time series forecasting and serve as benchmark for the later proposed generative AI models.

2.1.1. SARIMAX

The SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous variables) model serves as a baseline approach and is widely recognized in time series forecasting due to its balance between interpretability and effectiveness.

The SARIMAX model builds upon the traditional ARIMA model by combining three key components:

- Autoregression (AR): Captures dependencies between current values and previous values in the time series.
- Integration (I): Applies differencing to handle non-stationarity, which is common in energy market time series like electricity prices.
- Moving Average (MA): Accounts for the impact of past forecast errors on current predictions.

To improve the model's ability to reflect real-world electricity markets, SARIMAX extends ARIMA by incorporating seasonal patterns (e.g., daily or weekly cycles) and exogenous variables that influence the time series. Examples of exogenous variables relevant to energy forecasting include weather conditions, demand forecasts, and fuel prices.

By explicitly integrating these external factors, the SARIMAX model enhances its predictive accuracy compared to simpler ARIMA or SARIMA models. Its straightforward implementation and interpretability make it a solid baseline reference in our project before introducing more advanced machine learning and generative approaches. (Stier, 2001)

2.1.2. Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks are a specialized type of recurrent neural network (RNN) designed by (Hochreiter & Schmidhuber, 1997) to address the challenge of learning long-term dependencies in sequential data, such as time series of electricity prices. Unlike standard RNNs, which struggle with issues like vanishing gradients, LSTMs incorporate internal gating mechanisms that allow them to effectively retain or discard information over time.

An LSTM cell processes each time step by combining the current input with information from previous time steps. Through three main gates—forget, input, and output—it decides how much past information should be carried forward and how much new information should be incorporated. This internal state update allows the LSTM to capture both short-term and long-term trends in the data.

In time series forecasting, an LSTM model receives a historical sequence of data points and processes them step by step, generating a hidden representation at each time step. These hidden states can then be mapped to the target variable (e.g., future electricity price) using a simple output layer. During training, the model learns to minimize the difference between predicted and actual values using standard loss functions like Mean Squared Error (MSE). Training is performed with backpropagation through time, which is well-suited for sequential data.

Thanks to their ability to model complex temporal relationships and capture both short- and long-term dependencies, LSTMs are particularly suitable as a baseline model for electricity market forecasting tasks.

2.1.3. Multivariate Linear Regression

The multivariate linear regression model generalizes the traditional linear regression to produce probabilistic forecasts instead of point forecasts. It captures the linear relationships between a set of input variables and the target variable by estimating a set of regression coefficients that define the strength and direction of each relationship in form of a conditional covariance matrix assuming a multivariate normal distribution. This probabilistic perspective not only allows the model to generate point forecasts but also to quantify the associated prediction uncertainty. (Helwig, 2017)

2.2. Generative Adversarial Networks (GAN)

In this section, we introduce one class of generative AI model that was used in our study, the Generative Adversarial Network (GAN). GANs are a class of deep learning models that learn to generate realistic synthetic data through an adversarial training process between two neural networks: a generator and a discriminator.

In a standard GAN, the generator network takes random noise as input and produces synthetic samples, while the discriminator network tries to distinguish between real and generated data. During training, the generator learns to produce increasingly realistic samples that can deceive the discriminator. Figure 1 shows the architecture of a standard GAN with the networks in blue and input/output data in orange. Despite their innovative design, GANs can face challenges like unstable training dynamics, vanishing gradients, and mode collapse (where the generator produces only a limited variety of outputs). (Goodfellow et al., 2014)

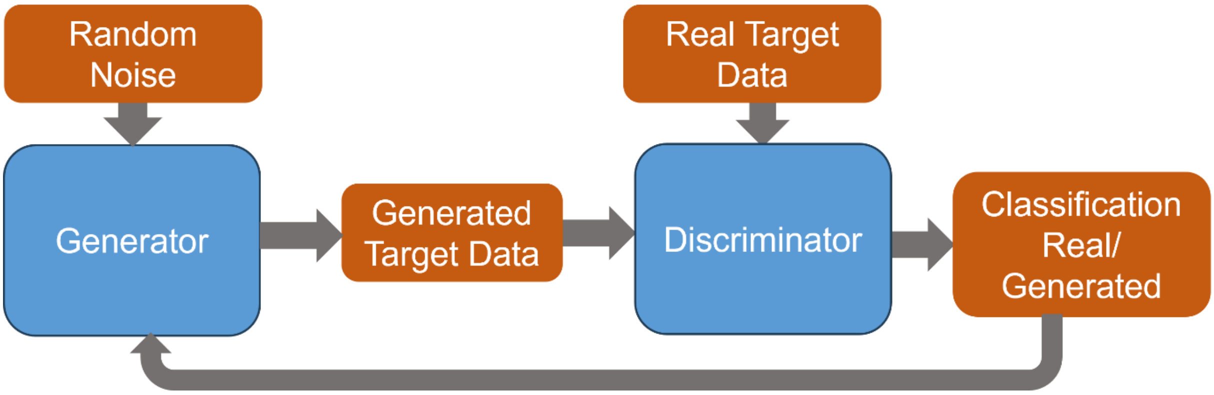

Figure 1: Architecture of the Generative Adversarial Network (GAN)

2.2.1. Conditional Time Series GAN (GAN_0)

Building on the GAN framework, we implemented a Conditional Time Series GAN (CTSGAN) based on (Lu et al., 2022), which integrates exogenous variables (e.g., historical prices, weather data) into the model to generate forecasts conditioned on relevant features. This allows the GAN to produce forecasts that reflect the influence of external factors, making the forecasts more accurate and realistic. To capture temporal dependencies, we incorporated LSTM layers into the GAN architecture, ensuring that the model could learn both short- and long-term relationships in the time series data. Furthermore, a low-dimensional latent space is introduced for the adversarial training, for which two additional networks are incorporated. The embedding network transforms features and target data into output in the latent space, while the recovery network should recover this output and transform it back into the target data. Figure 2 shows the architecture of the Conditional Time Series GAN.

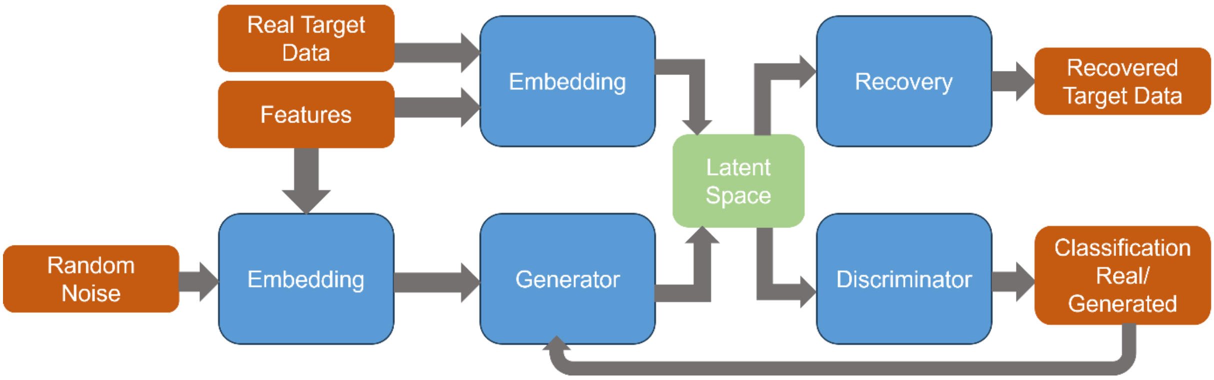

Figure 2: Architecture of the Conditional Time Series GAN from (Lu et al., 2022)

2.2.2. Modified Conditional Time Series GAN (GAN_1)

We further enhanced the CTSGAN by modifying its architecture to improve training stability and reduce mode collapse. In this version, noise is integrated differently, and a gradient penalty technique is applied to enforce smooth training. This helps the model produce more diverse and stable outputs, which is particularly important when forecasting volatile electricity prices.

2.2.3. Combined LSTM and Conditional Time Series GAN (GAN_2)

To use the strengths of both deterministic and generative approaches, we developed a hybrid model (GAN_2). In this setup, a deterministic model (in this case the LSTM baseline) is first used to generate point forecasts. The residuals (the differences between the actual and predicted values) are then modelled using a GAN. This two-step process allows the GAN to focus on capturing the variability and uncertainty not explained by the deterministic model, leading to improved probabilistic forecasts.

2.3. Temporal Fusion Transformer (TFT)

Transformer models are another class of generative AI models that show promising results especially for complex sequential data. The Temporal Fusion Transformer (TFT) is a deep learning architecture specifically designed for forecasting tasks that involve complex time series data with multiple variables. Developed to combine high predictive accuracy with interpretability, the TFT integrates several advanced neural network components to capture both short-term and long-term temporal dependencies, as well as relationships between different input variables.

The TFT uses attention mechanisms to dynamically weigh the importance of different time steps and input features. This allows the model to focus on the most relevant parts of the input sequence for each prediction. Additionally, the TFT includes gating mechanisms and variable selection networks, which enhance its ability to handle noisy and irrelevant inputs by learning to emphasize only the most important variables at each point in time.

Another strength of the TFT is its interpretability. By analyzing the attention weights and variable importance scores, the model can provide insights into which variables and time steps were most influential in driving the predictions, helping to enhance the transparency. (Lim et al., 2020)

2.4. Implementation

2.4.1. Technical Details

The programming language Python 3 (Python, 2025) was used to implement all models investigated in this work package. The data processing relied heavily on the library pandas (pandas, 2025). Statsmodels (Josef Perktold et al., 2024) and pmdarima (pmdarima, 2025) where applied for the SARIMAX baseline. The LSTM baseline and GAN models were developed with Tensorflow (Tensorflow, 2025) and Keras (Keras, 2025), while the Temporal Fusion Transformer was realised with PyTorch (‘PyTorch’, 2025). The multivariate linear regression baseline was built with functions of the library NumPy (NumPy, 2025). The plotting of the data as well as the results was conducted with Plotly (Plotly, 2025) and Matplotlib (Matplotlib, 2025).

2.4.2. Features

All implemented models except for the “Previous Day” baseline are conditional, i.e., they use exogenous factors for their prediction, so called features. These features can include temporal information (e.g., day of the week, month), weather data (e.g., temperature, solar surface radiation), and energy-related data (e.g., load forecast, electricity generation). However, not all available features are needed or even helpful for the models. Therefore, feature selection is the standard approach in modelling and machine learning tasks to select the relevant features for the models. In this work, for each target variable a correlation analysis between all available features and the target variable was conducted, followed by a lasso regression for a preselection of relevant features. With these preselected features a forward-feature selection (FFS) was conducted with the best baseline model, the LSTM. The same identified relevant features where then used for all models, to ensure a fair comparison.

2.4.3. Forecasting Setup

A whole forecasting day (96 time steps) is predicted at a time instead of iteratively predicting one time step (15 minutes) per forecast. Regarding the features, it was differentiated between forecasted features (e.g., day-ahead wind and solar production, load forecast, weather forecast) and historical (non-forecasted) features (electricity generation per type, target variable of previous day). Figure 3 exemplifies the schematic of the forecasting setup for the target variable “day-ahead price”, which remains the same for all target variables.

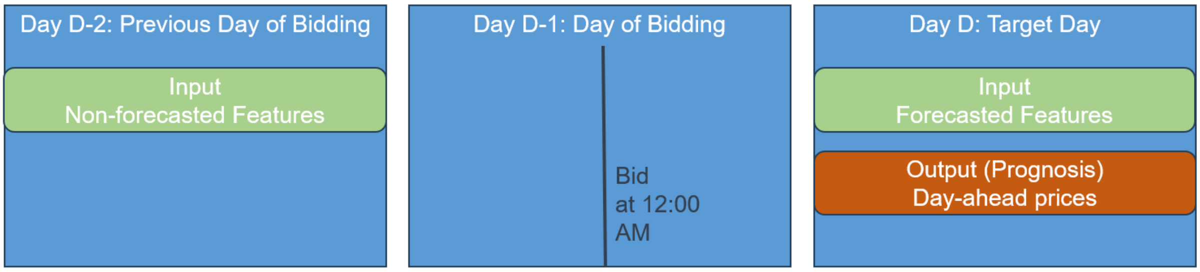

Figure 3: Forecasting setup for the target variable “day-ahead prices”

2.4.4. Data Preprocessing

Cutting Peaks: As proposed in literature (Lu et al., 2022), we cut peaks in the training target data as we assume that outliers confuse the models more than provide valuable input. Therefore, peaks are defined as outliers in the top and bottom 2,5 percentiles of the data. To avoid loosing too many data points, the outliers are not removed but replaced with the top and bottom caps respectively.

Standardization: We assume that the target data follows a normal distribution and use normally distributed noise as input to the generative models. To ensure an effective learning process, we standardise the training data rather than applying Min-Max normalisation as is often done in the literature (e.g. (Lu et al., 2022)).

3. Results and Discussion

The algorithms were tested on the three different variables of the Austrian electricity market: day-ahead electricity prices, CO_2-intensity of the grid and aFRR activations. This section presents the results and performance evaluations of the different models and their respective forecasts of these variables. First, some forecasting days will be presented for qualitative evaluation. Different metrics were assessed and will be presented for quantitative comparison of the performance. Lastly, a special case study for the day-ahead electricity prices investigating the effect of the energy crisis is presented.

The performance of the Temporal Fusion Transformer (TFT) is analysed separately using day-ahead prices, demonstrating the special ability of the TFT to explain the predictions.

3.1. Qualitative Assessment

The following plots will present a comparison of the forecasts from the implemented models. The x-axis presents the time steps (we have 96 time steps in one forecasting day), the y-axis are the predicted values (prices, CO_2-intensity, balancing energy activations). The orange dotted line presents the real values while the blue line(s) are the predictions. The models in the top row are the baselines (SARIMAX, LSTM and multivariate linear regression) and the models in the bottom row the GAN models (GAN_0, GAN_1 and LSTM-cTSGAN ensemble model). Due to convergence difficulties, the multivariate linear regression baseline is replaced by the Values of Previous Day baseline for the variables CO_2-intensity and aFRR activations.

3.1.1. Day-ahead Prices

The first target variable investigated are the day-ahead electricity prices. Data points from 2019 until mid-2024 were gathered from the ENTSO-E transparency platform (ENTSO-E, 2025).

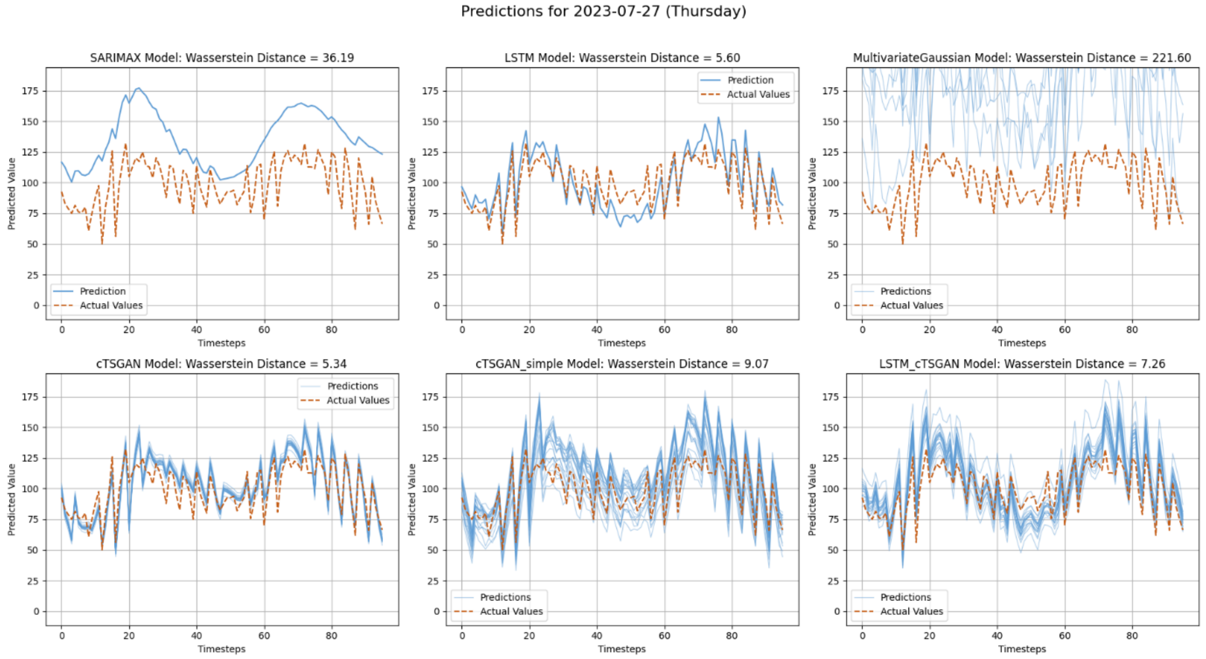

Figure 4: Forecasting day example of day-ahead electricity prices

Comparing the probabilistic models in Figure 4, the multivariate linear regression baseline has by far the highest variance, while the GAN_0 has a very small variance. This is probably a symptom of mode collapse in the GAN model training, a well-known challenge of GANs

(Gulrajani et al., 2017). For GAN_1 and GAN_2 mechanisms were implemented that help to reduce mode collapse (gradient penalty, fewer adversarial training epochs).

3.1.2. CO_2-intensity

The next target variable is the CO_2-intensity of the grid, which depends on the amount of electricity produced from different generation types. This data was gathered from Electricity Maps (*Methodology | Electricity Maps*, 2025) for the years 2021-2024.

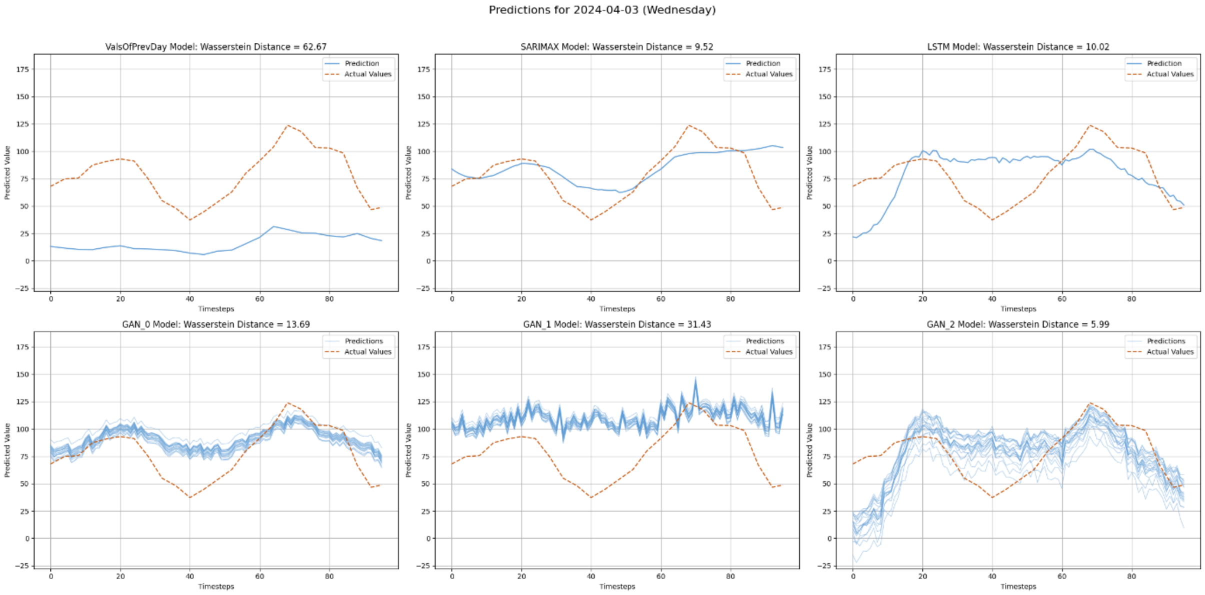

Figure 5: Forecasting day example of the CO_2-intensity of the electricity grid

From the visual assessment in Figure 5 it can be assumed that the SARIMAX baseline as well as GAN_0 and GAN_2 achieve to capture the curve of the variable the best, where GAN_2 has a significantly higher variance and therefore uncertainty estimation.

3.1.3. Balancing Energy Activations

For the forecasting of balancing energy activations, the positive as well as negative aFRR activation data was gathered from APG transparency (*Transparency*, 2025) from 2018 until mid-2024. Since positive and negative activations are separated in the data sets and cannot be combined to a single time series because of simultaneous activations, they are treated as separate target variables for the forecasting task.

The preceding correlation analysis (including lasso regression) revealed very low correlations between the balancing energy activations and the given features (weather data, temporal features, etc.), especially compared to the other target variables.

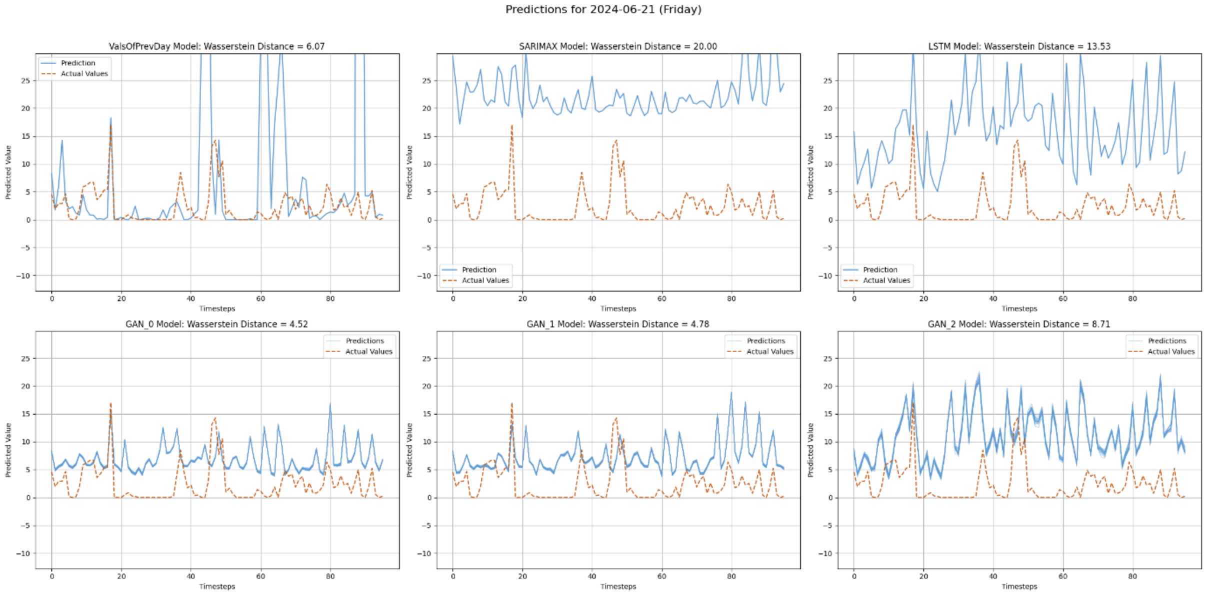

Figure 6: Forecasting day example of positive aFRR activations

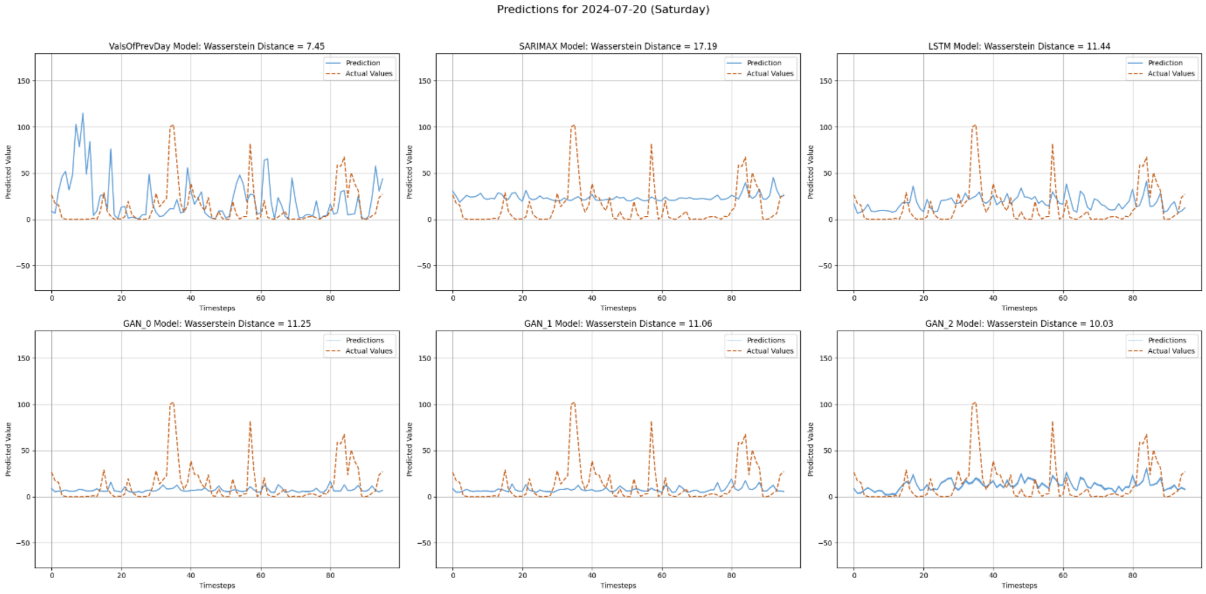

Figure 7: Forecasting day example of negative aFRR activations

Visual evaluations in Figure 6 and Figure 7 show that it is very difficult for all models to accurately determine all activation peaks. In future developments, it might therefore be more useful to model total activations in 4-hour intervals instead of forecasting them in 15-minute intervals.

3.2. Quantitative Assessment

To achieve a rigorous evaluation of the predictions, four different metrics were used.

The **Continuous Ranked Probability Score** (CRPS) can be used for deterministic as well as probabilistic forecasts, as it generalizes the Mean Absolute Error (MAE) for distributions. The cumulative distribution function (CDF) of the predictions is measured against the real CDF. A lower CRPS indicates a better probabilistic forecast, meaning the predicted distribution places more mass near the observed value while still reflecting the inherent uncertainty in the data. CRPS rewards distributions that balance sharpness and reliability.

The following metrics can only be applied for probabilistic forecasts and not for the deterministic baselines.

The **Negative Log-Likelihood** (NLL) measures how well a probabilistic prediction explains the real data. The calculation of the NLL depends on the assumed distribution of the predictions, which is often considered Gaussian with mean μ=0 and variance σ^2=1. A lower NLL means that the predicted distributions assign high probability density to the actual observed values, implying better calibration of both mean and variance (uncertainty).

The **Coverage Probability** does not account for the whole predicted distribution but only for the lower and upper bounds of a prediction. With that, the amount of real values that are in-between these bounds is calculated. An ideal probabilistic forecast should have an empirical coverage probability close to a desired nominal confidence level. If the coverage is too low, the model underestimates uncertainty, leading to overconfident predictions. Conversely, very high coverage can be a sign of underconfident predictions, meaning the uncertainty bounds are too wide and provide limited practical value.

Sharpness evaluates the concentration of the predicted distribution, independent of its accuracy. A sharper prediction has a narrower spread, meaning high confidence in the forecast and low estimated uncertainty. However, sharpness is a trade-off between precision and coverage: a model that is too sharp but miscalibrated may result in insufficient coverage probability. In contrast, a forecast with a high spread (high sharpness) has a not very useful precision. An optimal probabilistic forecast achieves both high sharpness and proper calibration, meaning it provides precise predictions while maintaining the correct uncertainty levels.

3.2.1. Day-ahead prices

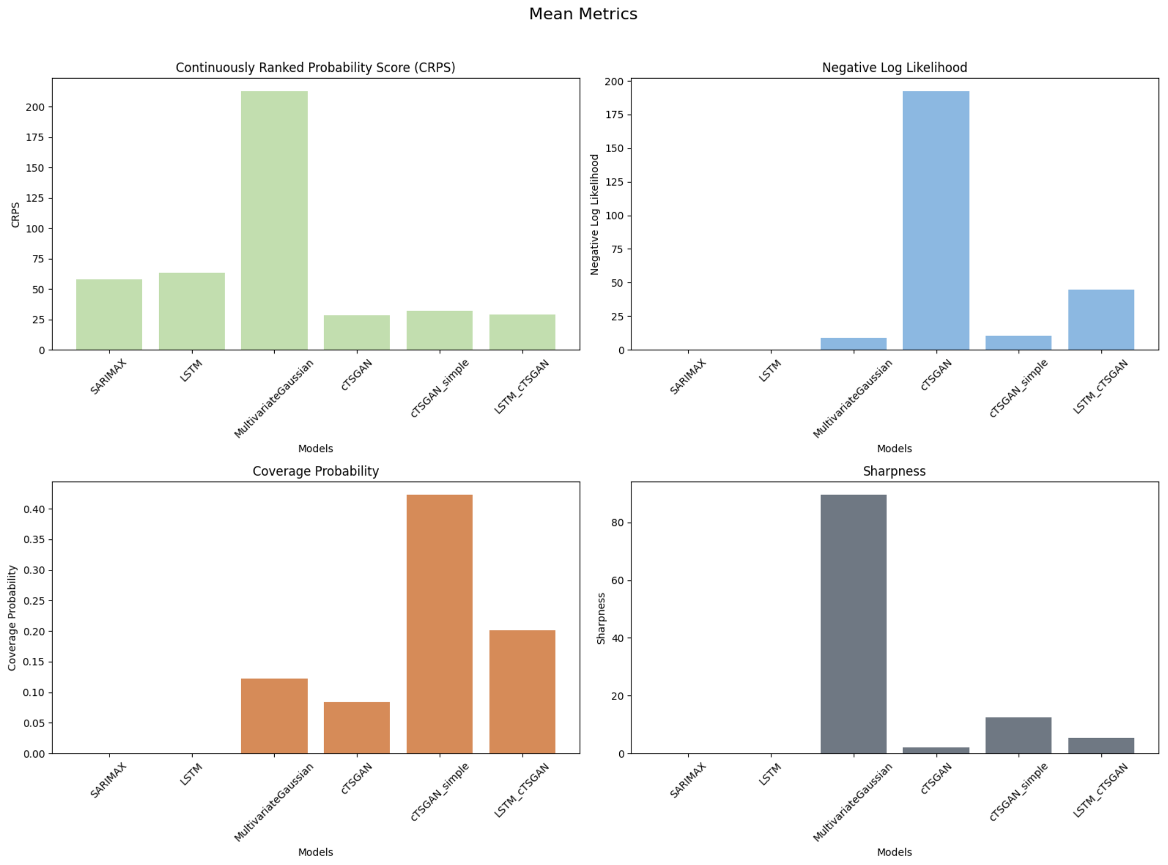

Figure 8: Evaluation metrics for day-ahead prices averaged over all test data

For the day-ahead prices (Figure 8), the GAN-based models achieve the best CRPS score followed by the SARIMAX and LSTM baselines, with the Multivariate Linear Regression model performing the worst. However, the Multivariate Linear Regression baseline achieves the best Negative Log Likelihood score, closely followed by GAN_1, probably because of the higher variance and therefore uncertainty estimation from these models. GAN_1 has by far the best Coverage Probability with a comparatively low sharpness, which points to a well calibrated forecast. GAN_2 is on the second place on both Coverage Probability and Sharpness, while the Multivariate Linear Regression has a very high Sharpness and therefore low precision.

3.2.2. CO_2-intensity

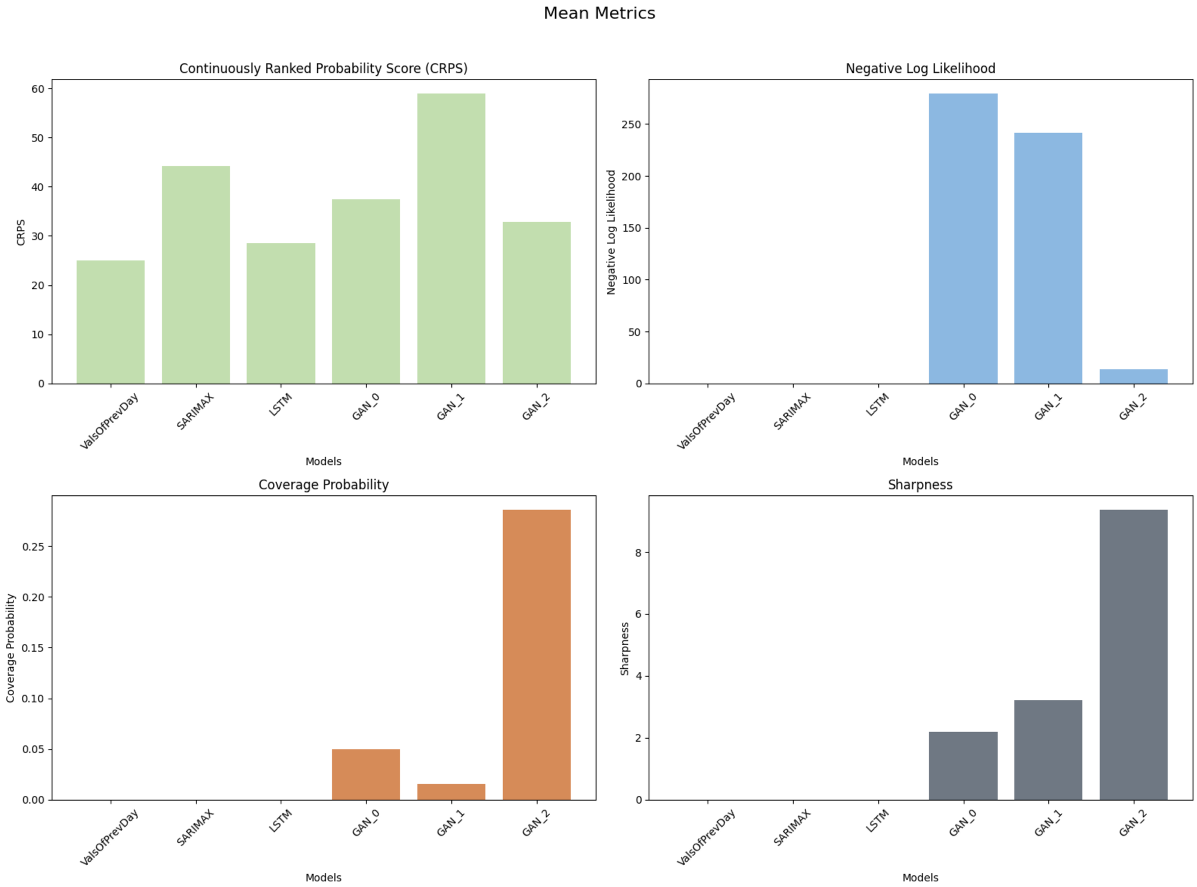

Figure 9: Evaluation metrics for CO_2-intensity of the grid averaged over all test data

Looking at the results for the variable "CO_2-intensity of the grid" in Figure 9, the very simple Previous Day baseline achieves the best CRPS score, followed closely by the LSTM baseline and GAN_2 model. GAN_2 performs also the best regarding the NLL and Coverage Probability compared to the other probabilistic models. However, the Sharpness score is comparatively high, pointing at a rather high variance and therefore low precision.

3.2.3. Positive aFRR activations

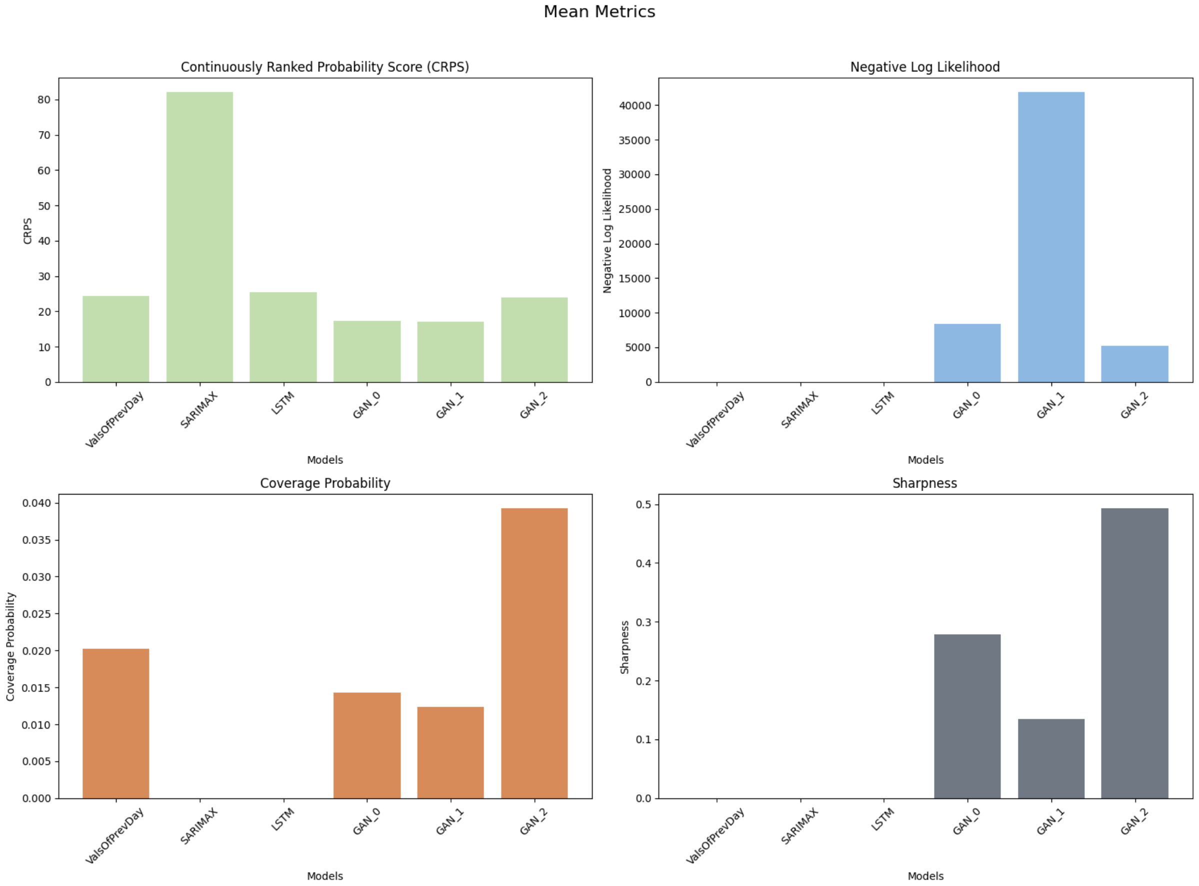

Figure 10: Evaluation metrics for positive aFRR activations averaged over all test data

The evaluation for the positive aFRR activations in Figure 10 shows that GAN_0 and GAN_1 achieve the best CRPS score, with the Previous Day baseline, LSTM and GAN_2 in the mid-tier, and the SARIMAX model performing worst. However, concerning the NLL, GAN_2 performs best closely followed by GAN_0 and with distinct gap to GAN_1. GAN_2 also performs best regarding the coverage probability with a mildly higher sharpness/variance than the other two GAN models.

3.2.4. Negative aFRR activations

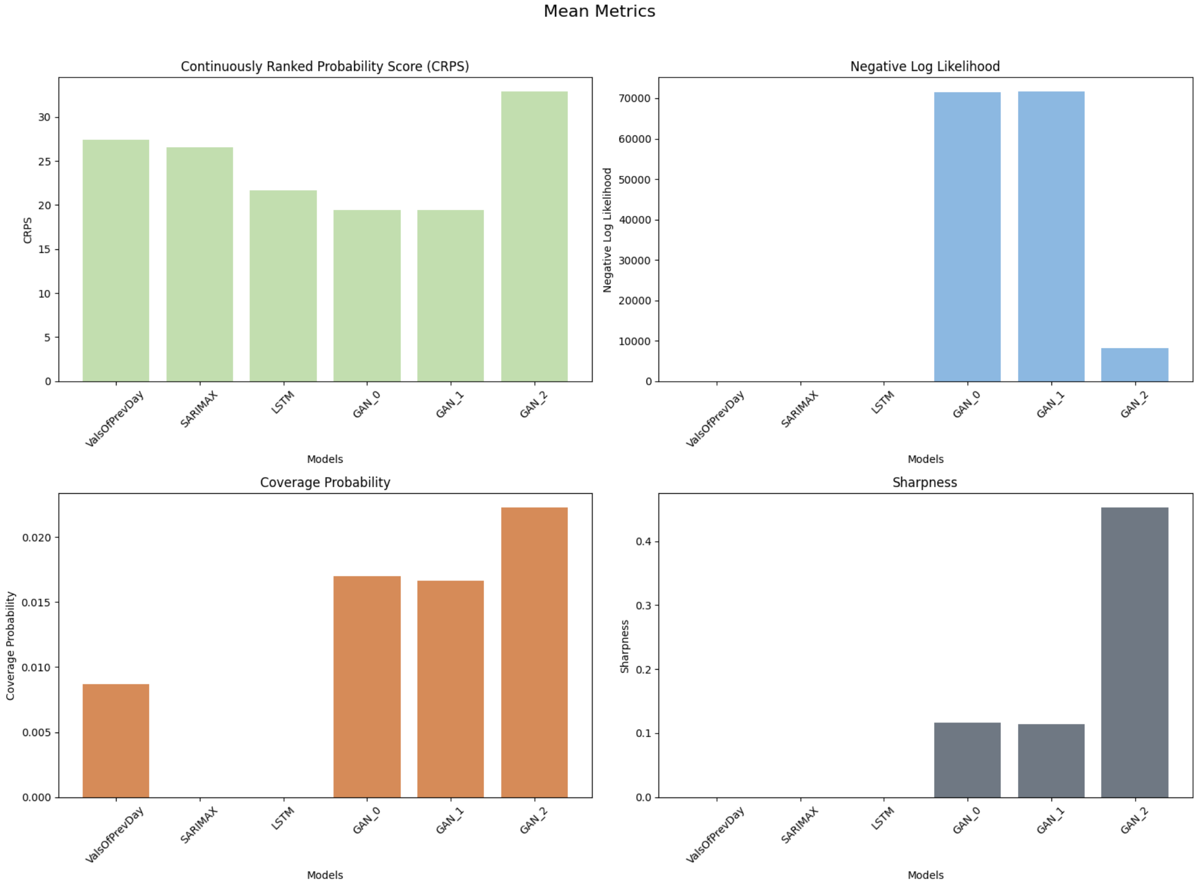

Figure 11: Evaluation metrics for negative aFRR activations averaged over all test data

For the negative aFRR activations, Figure 11 reveals that the GAN_0 and GAN_1 models perform the best regarding CRPS, with the baselines in the mid-tier and GAN_2 achieving the worst score. However, GAN_2 performs best in the NLL calculation and Coverage Probability, but with a comparatively high Sharpness and therefore possible lack of precision.

3.3. Case Study: Day-ahead Price Prediction in the Energy Crisis

Looking at the data of day-ahead prices in the given time range, a clear distortion can be seen from around September 2021 to January 2023: the energy crisis (see orange data points in Figure 12). The global energy crisis was the result of a combination of geopolitical tensions, supply chain disruptions, and overall uncertainties. A key factor has been the significant reduction in fossil fuel imports due to geopolitical conflicts, particularly the Russian-Ukrainian war, which resulted in a rather sudden decline in natural gas supplies to Europe. This caused high volatility and uncertainty in energy markets, leading to extreme price fluctuations in electricity markets, where prices reflect short-term supply and demand conditions. Moreover, policy interventions, such as price caps and emergency market regulations, have influenced market behaviour, sometimes distorting price signals. The resulting uncertainty and increased costs lead to significant challenges for grid operators, market participants, and policymakers.

(Förster et al., 2024)

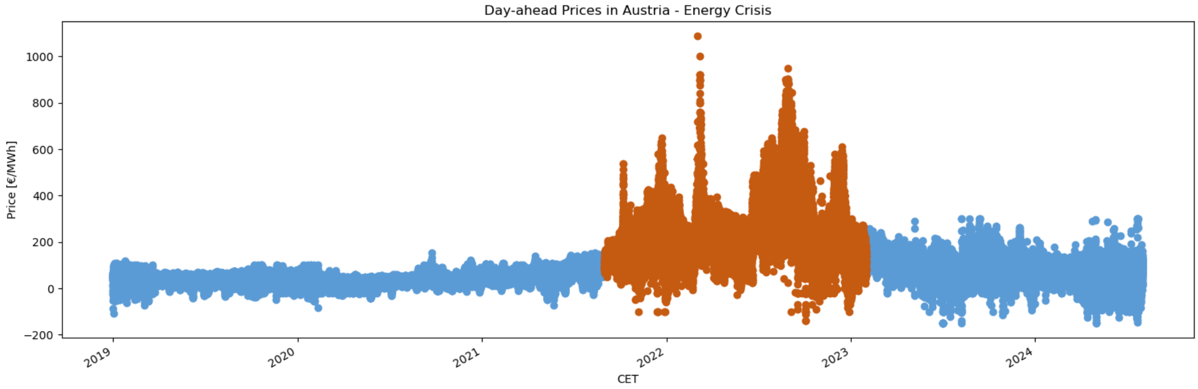

Figure 12: Day-ahead prices in Austria with energy crisis marked in orange.

As these price distortions cannot be explained by regular features like weather or electricity demand, it suggests to remove the data for training to avoid confusing the models. However, the energy crisis also presents an interesting case study as it provides an extreme situation with a lot of uncertainty, which is inherently intriguing for probabilistic forecasting. Therefore, it was decided to investigate two different case studies: one with the energy crisis completely removed from the data and a standard train-test split, and one where the non-crisis data is used for training and the crisis-data is used for testing.

Normal mode: In the case study *normal mode*, all data of the time period of the energy crisis (here defined as September 2021 to January 2023) is removed to avoid confusing the models with distortions that cannot be explained by the given features. The train-test-split is illustrated in Figure 13 with the training data in blue and the test data in dark gray. It should be noted that although not as extreme as in the energy crisis, the price variance after the energy crisis is still different and significantly higher than before the energy crisis. Therefore, the bulk of the training data, that is dated before the energy crisis, is still inherently different and not really representative for the test data.

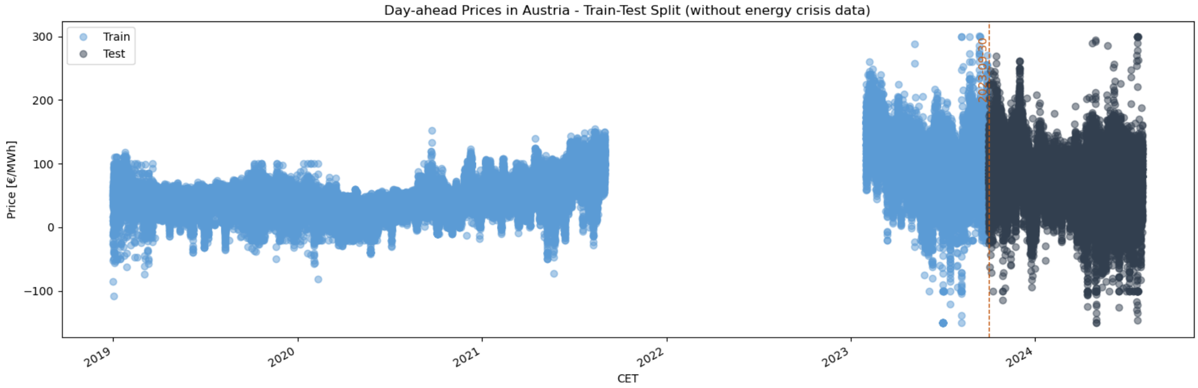

Figure 13: Train-test split of day-ahead prices data set in normal mode *(without energy crisis).*

Crisis mode: As mentioned above, the energy crisis presents an interesting case study to investigate the performance of forecasting models in extreme situations with lots of uncertainty. Therefore, all non-crisis data was used for the training, also the time after the energy crisis to have a sufficient amount of data points. The crisis-data was then used to test the models (see Figure 14 for the train-test-split).

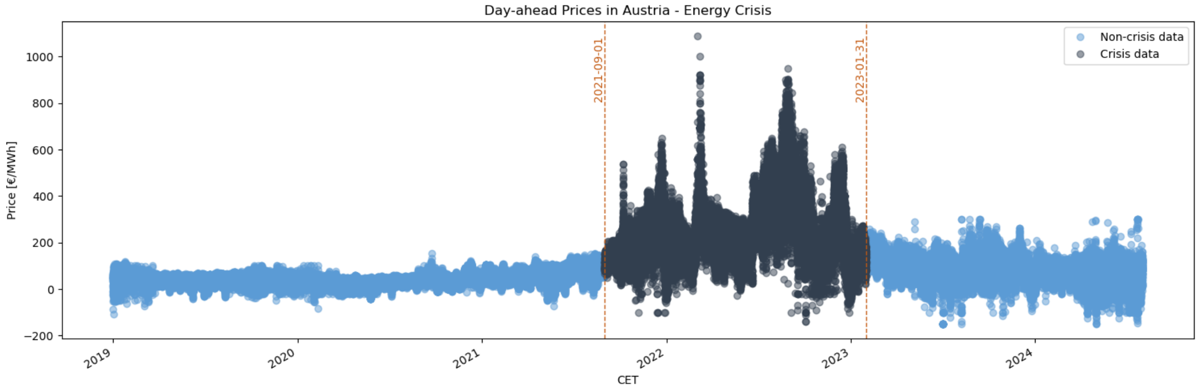

Figure 14: Train-test split of day-ahead prices data set in crisis mode *(energy crisis as test data).*

3.3.1. Metrics Results

As an overall result comparing both modes it can be seen that in the metrics, all models perform significantly worse in *crisis mode* than in *normal mode*, which is expected due to the test data being very different from the training data in *crisis mode*. However, comparing the models against each other in each of the modes shows interesting results.

Continuous Ranked Probability Score (CRPS)

The CRPS results are presented in Figure 15 and reveal notable patterns in both modes. In *normal mode*, where market conditions are stable, the LSTM baseline achieves the lowest CRPS, indicating the highest forecasting accuracy. The GAN-based models exhibit slightly higher scores, performing competitively but not outperforming the LSTM. The multivariate linear regression model has the second-highest CRPS, while the SARIMAX model yields the worst performance, suggesting that its linear assumptions are insufficient to capture electricity price dynamics.

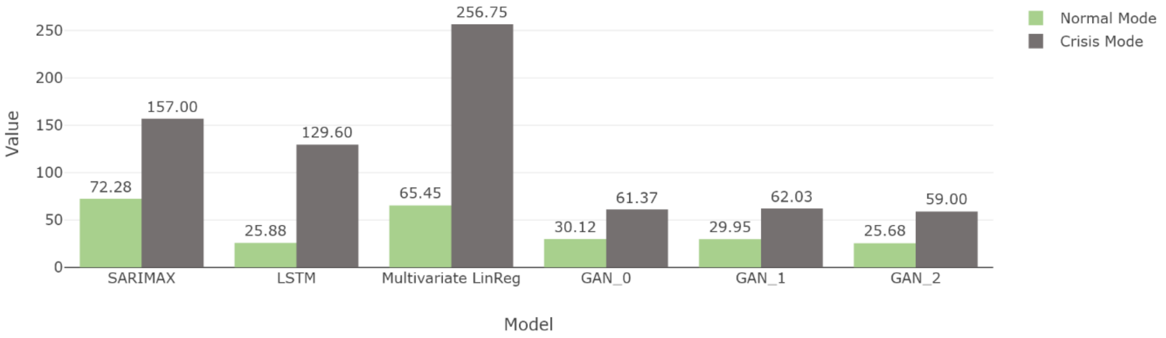

Figure 15: CRPS comparing normal mode *with* crisis mode.

In *crisis mode*, where the test data includes extreme market fluctuations from the energy crisis, the ranking of the models looks different. The three GAN-based models obtain the lowest CRPS values, which demonstrates that they are better able to capture the increased uncertainty during volatile market periods. This suggests that GANs effectively model the complex and non-stationary behaviour of electricity prices under extreme conditions. In contrast, the LSTM baseline, which performed best under normal conditions, now exhibits a higher CRPS, indicating reduced robustness in handling sudden price shifts. The SARIMAX and linear regression models perform even worse, with the highest CRPS values, supporting the hypothesis that they are rather limited to adapt to crisis-induced deviations from historical patterns.

These results highlight the trade-off between model stability and adaptability. While traditional time series models may perform adequately under stable conditions, probabilistic generative models such as GANs can provide a significant advantage when forecasting electricity prices in highly uncertain environments.

Negative Log-Likelihood (NLL)

The Negative Log-Likelihood (NLL) results highlight more pronounced differences between the two forecasting modes, as can be observed in Figure 16. Across both settings, the multivariate linear regression baseline achieves the lowest (best) NLL, meaning it produces relatively well-calibrated probabilistic forecasts. Conversely, the GAN_0 exhibits the highest (worst) NLL, suggesting that its predictive distributions frequently assign low probability density to the actual observed values and penalties for the low variance are applied

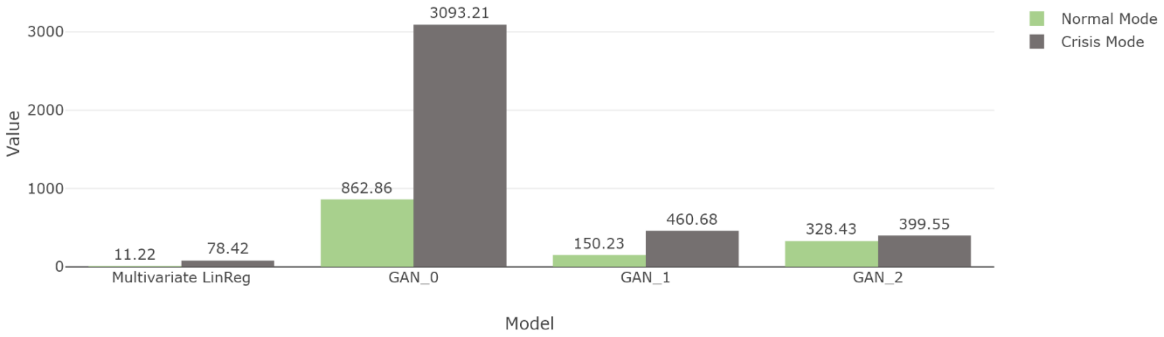

Figure 16: NLL comparing normal mode *with* crisis mode.

For the other GAN models, performance varies significantly between modes. In *normal mode*, the GAN_1 has a slightly higher NLL than the GAN_2 model, indicating that it may be slightly less well-calibrated in stable market conditions. However, in *crisis mode*, the GAN_1 model significantly improves, achieving an NLL value close to that of the best-performing multivariate linear regression. Meanwhile, the GAN_2 model exhibits a substantial increase in NLL, making it the second worst-performing model after the GAN_0 in this setting.

For day-ahead electricity price forecasting, a lower NLL indicates a model's ability to assign high probability to actual prices. That the GAN models have significantly worse performance regarding the NLL as opposed to the CRPS is probably due to their low variance, which is strongly punished in the NLL calculation.

3.3.2. Simplified Optimization Approach

To assess the practical value of the developed forecasting models (particularly for day-ahead electricity prices), we implemented a simplified optimization framework. While full-scale stochastic optimization can be computationally intensive, our goal was to evaluate model performance in a realistic yet tractable way. This was done by exploring how well forecasts could guide operational decisions under both typical conditions and stress scenarios (e.g., energy crisis situations).

The core idea of the optimization task is to identify the time steps in a forecast where a market participant could buy low and sell high, based on predicted values. Using a forecasted time series, we search for the hours with the lowest and highest predicted prices. The difference between these two prices represents the theoretical revenue potential. This potential revenue is then evaluated using the actual market values at those selected time steps. This allows us to assess how well the forecast guided the decision, i.e., whether the predicted "best times" to buy and sell would have translated into real profit. Figure 17 exemplifies the optimization approach for one forecasting day.

This approach is applicable to both deterministic and probabilistic forecasts. For deterministic models, we use the point forecasts directly. For probabilistic models, we work with the expected values or sample distributions provided by the generative models.

Probabilistic Optimization Strategies

To fully utilize the uncertainty information provided by generative models, we implemented several enhanced strategies tailored to probabilistic forecasts:

- **Pessimistic Strategy (Risk-Averse):** Instead of optimizing based on the average (expected) difference between high and low prices, this method focuses on a conservative percentile (e.g., the 10th percentile). This allows the strategy to prioritize scenarios where the downside risk is minimized, offering a more cautious, risk-averse approach.
- **Score-Based Optimization**: In this strategy, each potential buy/sell combination is scored using a weighted trade-off between expected revenue and variance (i.e., uncertainty).
- **Probability of Positive Revenue**: Here, we evaluate how likely it is that a chosen buy/sell pair will generate a positive return across all forecast samples. We select the time steps that not only offer a high expected return but also have a high probability of success. This strategy is particularly useful for risk-averse decision-makers who prefer reliability over high but uncertain gains.

By comparing results across these settings, we assessed not only the models' forecasting accuracy but also their robustness and value in decision-making during unexpected market disruptions.

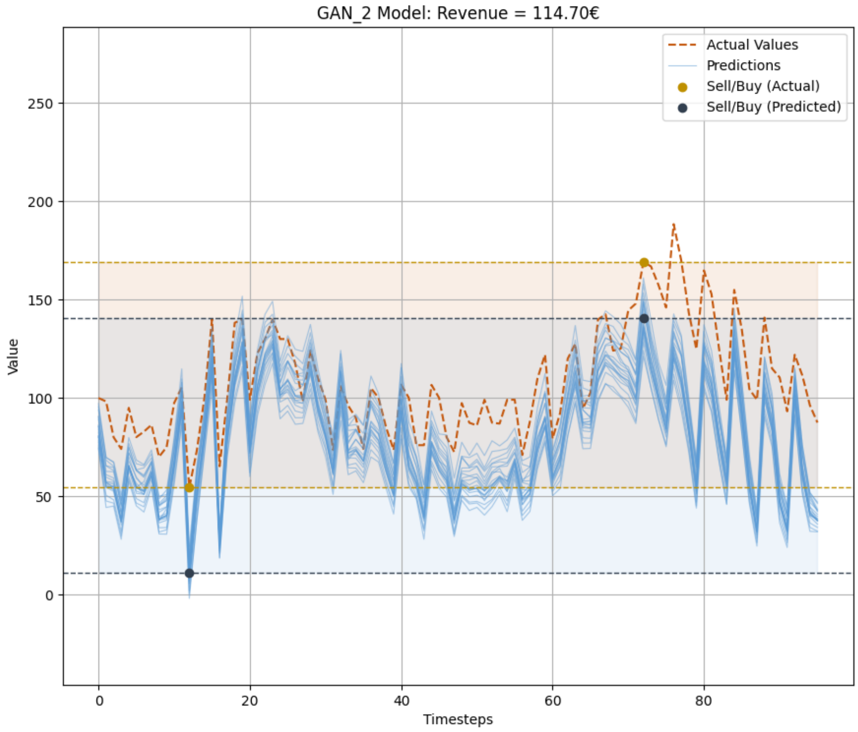

Figure 17: Example of one forecasting day with optimisation: the blue-grey points are the two chosen time steps with respective heights of the forecasted prices, the orange points are the same time steps but with the height of the actual prices. The coloured area between the points is the respective achieved revenue with those chosen time steps, where the blue-grey area is again the forecasted revenue and the orange area the actual revenue.

Normal Mode

In *normal mode*, all models perform similarly well regarding the optimisation revenues. The GAN_2 model achieves in average the highest revenue of around 111€, followed by the LSTM baseline at 106€ and the two other GAN models at 96-99€. The SARIMAX baseline obtains on average 95€ and the multivariate linear regression baseline 85-89€ depending on the optimisation strategy.

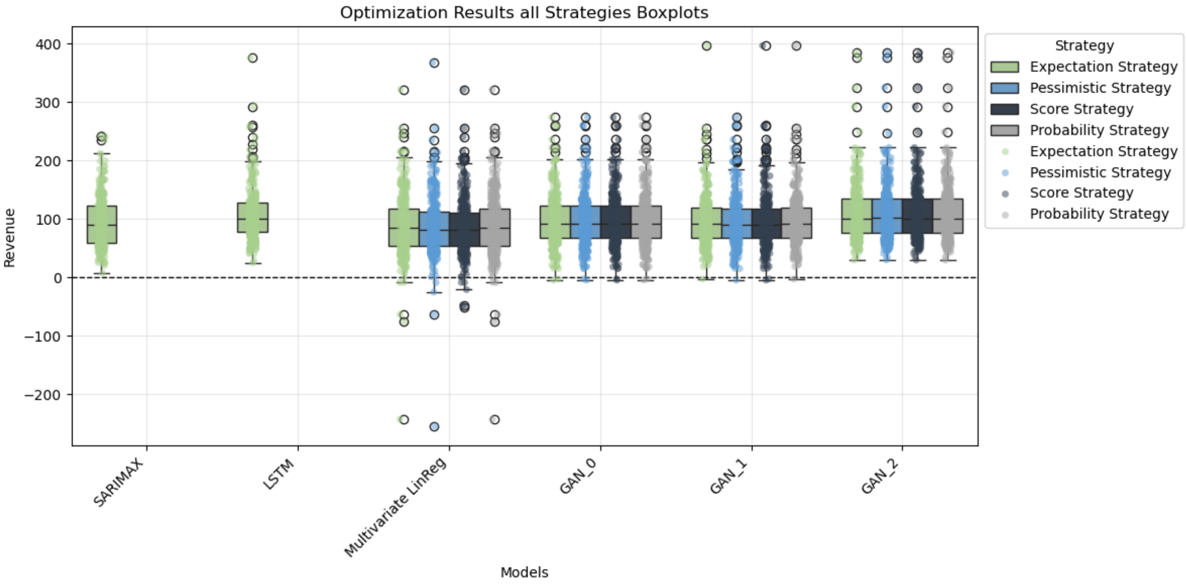

Figure 18: Achieved revenues in box-plots comparing optimisation strategies in normal mode.

Differences between the strategies are rather small in the averages. Looking at the daily results, there are some differences but no clear trend suggesting that one strategy is significantly better or worse than the others. The magnitude of differences in the strategies seems to correlate with the magnitude of variance of the forecasting distributions of the models, which is also intuitively understandable. Therefore, the multivariate linear regression baseline shows more difference between strategies than the GAN_0 with its very low variance.

Looking at the distributions of the revenues in Figure 18, not only the averages are of interest from an economical perspective but also how often the models incur losses (negative revenues). Here it can be seen that the two deterministic baselines SARIMAX and LSTM as well as the GAN_2 model never obtain losses, while the other two GAN models incur a few losses over all test data days. The multivariate linear regression model accounts for the highest amount of losses.

Crisis Mode

In *crisis mode*, the results differ again significantly, as can be seen in Figure 19. Here, all three GAN models achieve far higher revenues than the baselines. The GAN_0 performs the best with an average revenue of 192€, closely followed by the GAN_1 (170-171€) and the former best-performer, the GAN_2, with 157€. The baselines are all significantly worse: the LSTM achieves 125€ on average, the multivariate linear regression 31-32€ and the SARIMAX model has a negative average revenue with -52€.

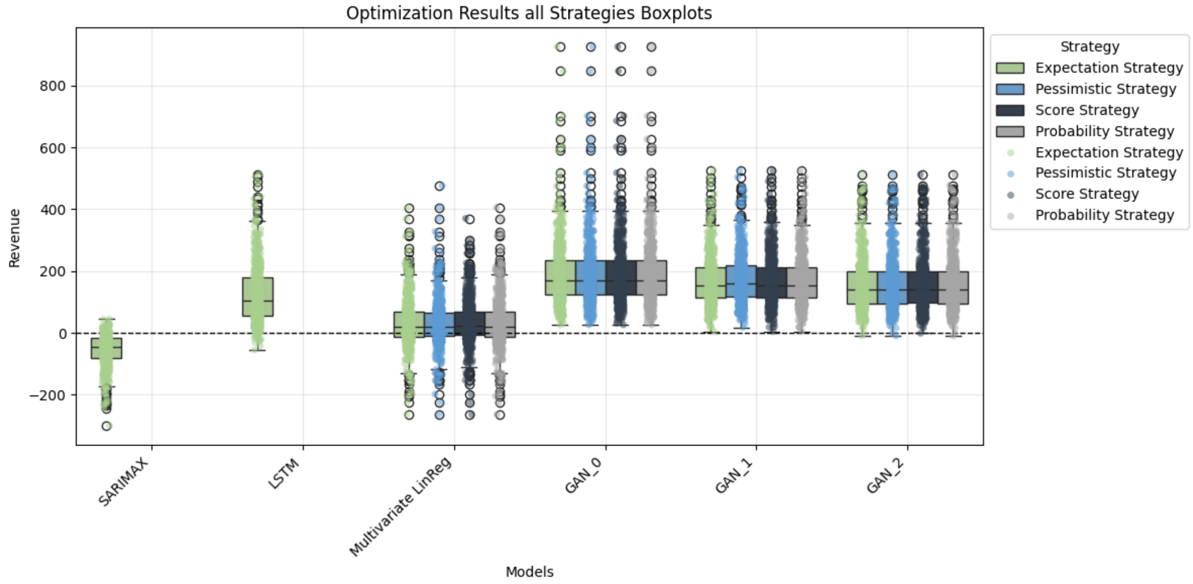

Figure 19: Achieved revenues in box-plots comparing optimisation strategies in crisis mode.

The box-plots of the revenues reveals that the GAN models make either zero losses (GAN_0 and GAN_1) or only very few (GAN_2 model), while the baselines perform significantly worse, achieving losses very often. These results suggest that the GAN models are indeed capable of capturing uncertainties and performing well in extreme situations.

3.4. Temporal Fusion Transformer

To validate the Temporal Fusion Transformer (TFT) and compare it to the other models, the model was trained on day-ahead prices from 2019 to 2022 and tested on data from 2023 to 2025. The following metrics were computed: CRPS, Coverage Probability and Sharpness value. On the test dataset, the CRPS is 30.4, which is similar to that of the GAN-based models. At 37.8%, the coverage probability is close to the best value achieved by the GAN-based models. The sharpness (standard deviation of the predicted distribution) stands at 43.2, which is higher than that of the GAN models, reflecting greater model uncertainty.

While the performance proved to be similar to that of the other generative AI models, the strength of the TFT lies in its ability to leverage its attention weights and Variable Selection Networks (VSN) to answer three core questions:

1. Which historical variables were important? (Importance of Encoder Variables)
2. Which known future variables were important? (Importance of Decoder Variables)
3. Which points in time in the past does the model look at to predict the future? (Temporal Attention).

This additional explanatory information is provided not only as a global statistic across the entire test set but for each individual forecasting sample.

Figure x illustrates an exemplary TFT electricity price forecast for the next 36 hours, including the distribution and expected value. The temporal attention weighting is mapped on a second axis, visualizing the influence of past observations on the forecast.

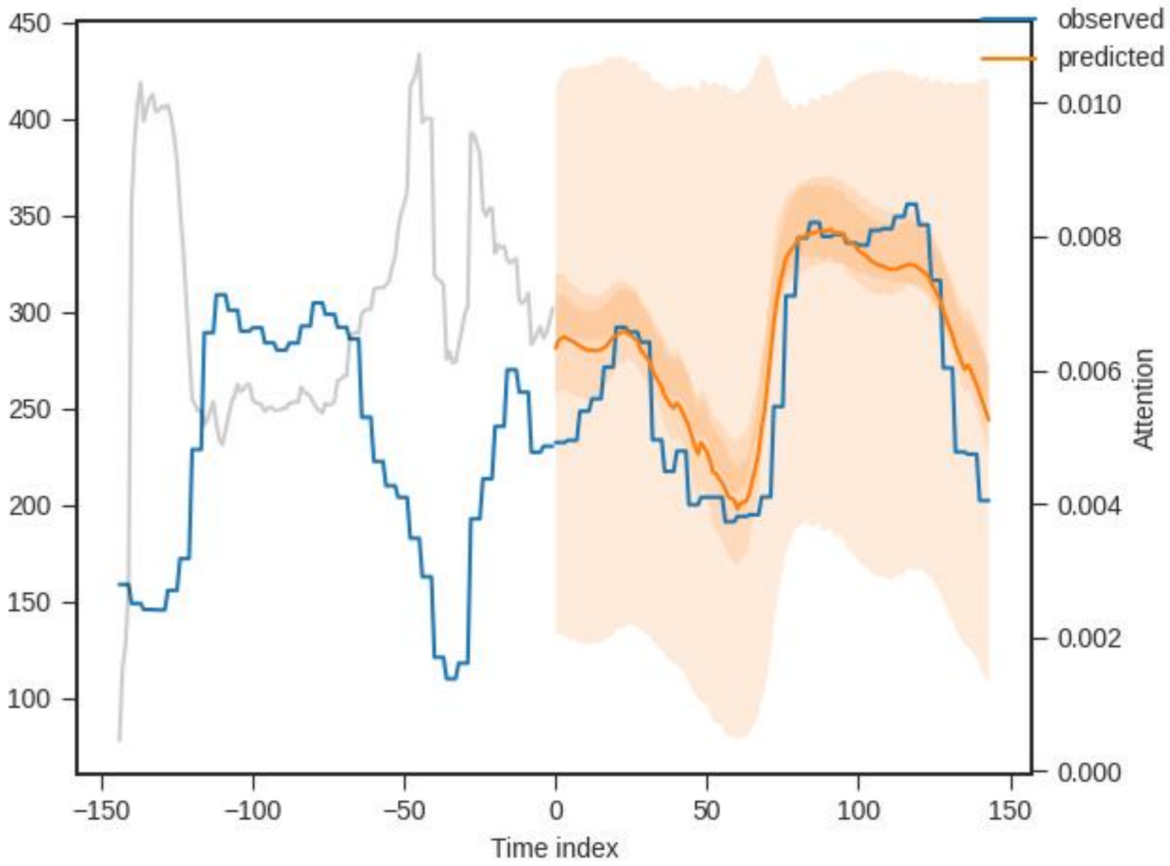

Figure 20 TFT electricity price forecast with probabilistic bands and temporal attention weights. The primary axis displays the predicted price distribution (quantiles) and the expected value against the actual price. The secondary axis illustrates the temporal attention scores, highlighting the model's focus on local minima and sharp price changes in the historical 36-hour window.

Alongside the day of the week, the most heavily weighted variables include electricity demand over the last 36 hours in Germany and Austria. The most critical known information regarding the future period to be forecasted was the load forecast for Austria, followed by season, time of day, and the load forecast for Germany. Temporal attention primarily focuses on local minima and steep ramps of the electricity price within the last 36 hours.

4. Conclusion

This deliverable presents the development, implementation, and evaluation of various forecasting models targeting three critical energy system variables: day-ahead electricity prices, CO_2-intensities, and balancing energy activations (aFRR). The focus was on assessing the performance of both deterministic baseline models and advanced generative approaches, particularly GAN-based architectures.

4.1. Forecasting Results

Across all three target variables, the generative models, especially the improved GAN variants, demonstrated strong potential for probabilistic forecasting in the energy domain:

- **Day-ahead Electricity Prices**: GAN-based models performed best in terms of probabilistic accuracy (e.g., CRPS and coverage probability), with GAN_1 showing a particularly well-calibrated distribution. Although the multivariate linear regression model achieved strong results in Negative Log-Likelihood (NLL), it also showed high variance and low sharpness, indicating poor precision.

- **CO_2-Intensity of the Grid**: Surprisingly, the very simple Previous Day baseline performed competitively, but GAN_2 achieved the best NLL and coverage probability among the probabilistic models. While its high variance reduced sharpness, the model's uncertainty representation adds value in operational settings.
- **Balancing Energy Activations**: Due to low correlations between features and target variables, this task proved especially difficult. However, GAN models generally outperformed the SARIMAX and Previous Day baselines. GAN_2 provided the most balanced performance across CRPS, NLL, and coverage probability, indicating its robustness even with noisy and weakly correlated inputs.

4.2. Case Study: Day-Ahead Price Forecasting During the Energy Crisis

A dedicated case study was conducted to evaluate model robustness under crisis conditions, simulating a scenario where models were trained on historical, stable data and tested on periods characterized by extreme market volatility.

In normal market conditions, deterministic models like LSTM performed well and achieved high average optimization revenues. GAN-based models were competitive, with GAN_2 leading among them. All models exhibited relatively similar performance, reflecting the predictable nature of the data.

In crisis conditions, the GAN models clearly outperformed all baselines. GAN_0 and GAN_1, in particular, achieved the highest average revenues and rarely incurred losses. Their ability to model uncertainty and adapt to unexpected dynamics allowed for superior performance in volatile environments. In contrast, traditional models like SARIMAX not only produced poor forecasts but also led to frequent and substantial financial losses.

4.3. Key Findings

- **Forecasting Accuracy**: GAN-based models consistently achieved strong performance across key probabilistic metrics, demonstrating their value in modelling uncertainty for energy system variables.
- **Robustness in Volatile Conditions**: The crisis scenario highlighted a clear advantage for generative models, which adapted better to market irregularities and extreme conditions than deterministic models.
- **Operational Relevance**: By integrating probabilistic forecasts into a simplified optimization framework, we demonstrated how such models can inform real-world decisions, such as energy dispatch or trading strategies.
- **Explainability**: Unlike blackbox-models, the TFT provides granular insights into variable importance and temporal attention for each individual forecast. This feature facilitates a deeper understanding of the time-dependent correlations driving the electricity price dynamics and the core mechanics behind the model's predictions.
- **Limitations and Outlook**: While generative models offer clear benefits, challenges such as training stability and mode collapse remain. Further improvements, such as ensemble strategies and model calibration, could enhance performance even further. Additionally, the proposed aggregation approach for balancing energy activations (e.g., forecasting over 4-hour intervals) could improve forecast reliability for this particularly noisy target variable.

This concludes the algorithmic documentation for Deliverables 3.1 and 3.2. The insights gained from this work provide a solid foundation for continued development and real-world deployment of probabilistic forecasting tools within the *transpAIrent.energy* platform.

References

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein Generative Adversarial Networks. *Proceedings of the 34th International Conference on Machine Learning*, 214–223. https://proceedings.mlr.press/v70/arjovsky17a.html

Chai, S., Li, Q., Abedin, M. Z., & Lucey, B. M. (2024). Forecasting electricity prices from the state-of-the-art modeling technology and the price determinant perspectives. *Research in International Business and Finance*, *67*, 102132. https://doi.org/10.1016/j.ribaf.2023.102132

ENTSO-E. (2025). *ENTSO-E Transparency Platform*. https://transparency.entsoe.eu/

Förster, R., Harding, S., & Buhl, H. U. (2024). Unleashing the economic and ecological potential of energy flexibility: Attractiveness of real-time electricity tariffs in energy crises. *Energy Policy*, *185*, 113975. https://doi.org/10.1016/j.enpol.2023.113975

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). *Generative Adversarial Networks* (No. arXiv:1406.2661). arXiv. https://doi.org/10.48550/arXiv.1406.2661

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. (2017). *Improved Training of Wasserstein GANs* (No. arXiv:1704.00028). arXiv. https://doi.org/10.48550/arXiv.1704.00028

Helwig, N. E. (2017). *Mutlivariate Linear Regression*. University of Minnesota, Twin Cities. http://users.stat.umn.edu/~helwig/notes/mvlr-Notes.pdf

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Josef Perktold, Skipper Seabold, Kevin Sheppard, ChadFulton, Kerby Shedden, jbrockmendel, j-grana6, Peter Quackenbush, Vincent Arel-Bundock, Wes McKinney, Ian Langmore, Bart Baker, Ralf Gommers, yogabonito, s-scherrer, Yauhen Zhurko, Matthew Brett, Enrico Giampieri, yl565, … Yaroslav Halchenko. (2024). *statsmodels/statsmodels: Release 0.14.2* (Version v0.14.2) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.593847

Keras. (2025). *Keras: Deep Learning for humans*. https://keras.io/

Lago, J., Marcjasz, G., De Schutter, B., & Weron, R. (2021). Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy*, *293*, 116983. https://doi.org/10.1016/j.apenergy.2021.116983

Lim, B., Arik, S. O., Loeff, N., & Pfister, T. (2020). *Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting* (No. arXiv:1912.09363). arXiv. https://doi.org/10.48550/arXiv.1912.09363

Lu, X., Qiu, J., Lei, G., & Zhu, J. (2022). Scenarios modelling for forecasting day-ahead electricity prices: Case studies in Australia. *Applied Energy*, *308*, 118296. https://doi.org/10.1016/j.apenergy.2021.118296

Maji, D., Shenoy, P., & Sitaraman, R. K. (2023). Multi-Day Forecasting of Electric Grid Carbon Intensity Using Machine Learning. *ACM SIGEnergy Energy Informatics Review*, *3*(2), 19–33. https://doi.org/10.1145/3607114.3607117

Matplotlib. (2025). *Matplotlib—Visualization with Python*. https://matplotlib.org/

Methodology | Electricity Maps. (2025). https://www.electricitymaps.com/methodology

NumPy. (2025). *NumPy*. https://numpy.org/

pandas. (2025). *pandas—Python Data Analysis Library*. https://pandas.pydata.org/

Plotly. (2025). *Plotly*. https://plotly.com/python/

pmdarima. (2025). *Pmdarima: ARIMA estimators for Python—Pmdarima 2.0.4 documentation*. https://alkaline-ml.com/pmdarima/

Python. (2025, April 24). *Welcome to Python.org*. Python.Org. https://www.python.org/

PyTorch. (2025). *PyTorch*. https://pytorch.org/

Shah, I., Iftikhar, H., & Ali, S. (2022). Modeling and Forecasting Electricity Demand and Prices: A Comparison of Alternative Approaches. *Journal of Mathematics*, *2022*(1), 3581037. https://doi.org/10.1155/2022/3581037

Smith, K. E., & Smith, A. O. (2020). *Conditional GAN for timeseries generation* (No. arXiv:2006.16477). arXiv. https://doi.org/10.48550/arXiv.2006.16477

Stier, W. (2001). *Methoden der Zeitreihenanalyse*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-56709-4

Tensorflow. (2025). *TensorFlow*. Tensorflow. https://www.tensorflow.org/

Transparency. (2025). Austrian Power Grid. https://markt.apg.at/en/transparency/

Tschora, L., Pierre, E., Plantevit, M., & Robardet, C. (2022). Electricity price forecasting on the day-ahead market using machine learning. *Applied Energy*, *313*, 118752. https://doi.org/10.1016/j.apenergy.2022.118752

Wang, X., Zhao, Q., & Tong, L. (2024). *Probabilistic Forecasting of Real-Time Electricity Market Signals via Interpretable Generative AI* (No. arXiv:2403.05743). arXiv. https://doi.org/10.48550/arXiv.2403.05743

Appendix

Data

Variable Name	Description	Source
day_ahead_prices_60min	Day-ahead prices of Austria in 60min time steps (from EPEX)	ENTSO-E
day_ahead_prices_15min	Day-ahead prices of Austria in 15min time steps (from EXAA)	ENTSO-E
AT-day_ahead_Solar	Day-ahead prognosis of electricity generation from solar in Austria	ENTSO-E
AT-day_ahead_Wind Onshore	Day-ahead prognosis of electricity generation from onshore wind plants in Austria	ENTSO-E
AT-Biomass_Actual Aggregated	Electricity generation from biomass in Austria	ENTSO-E
AT-Fossil Gas_Actual Aggregated	Electricity generation from fossil gas in Austria	ENTSO-E
AT-Fossil Hard coal_Actual Aggregated	Electricity generation from fossil hard coal in Austria	ENTSO-E
AT-…_Actual Aggregated	Electricity generation from … in Austria	ENTSO-E
AT-…_Actual Aggregated_shifted_2D	Electricity generation from … in Austria from 2 days ago	ENTSO-E
AT-Forecasted Load	Day-ahead load forecast in Austria	ENTSO-E
AT-Actual Load	Actual load in Austria	ENTSO-E
power_afrr_neg_act_at	Activations of negative aFRR (Automatic Frequency Restoration Reserves) in Austria	APG
power_afrr_pos_act_at	Activations of positive aFRR (Automatic Frequency Restoration Reserves) in Austria	APG
AT-export-CH	Energy exports from Austria to Switzerland	ENTSO-E
Country1-export-country2	Energy exports from country 1 to country 2	ENTSO-E
Country1-import-country2	Energy imports from country 1 to country 2	ENTSO-E
hour_x	Hour of the day from 0 to 23	-
Morning	Time between (incl.) 6:00 and (excl.) 10:00	-
Midday	Time between (incl.) 10:00 and (excl.) 14:00	-
Afternoon	Time between (incl.) 14:00 and (excl.) 16:00	-

Evening	Time between (incl.) 16:00 and (excl.) 20:00	-
Night	Time between (incl.) 20:00 and (excl.) 6:00	-
Spring	March-May	-
Summer	June-August	-
Autumn	September-November	-
Winter	December-February	-

Project coordinator:

Klara Maggauer, M. Sc., B. Sc.

AIT Austrian Institute of Technology GmbH

Center for Energy

Giefinggasse 4

A-1210 Vienna

E-mail: klara.maggauer@ait.ac.at