Technical Report: Multi-Horizon CO_2 Forecasting Using Deep Learning

Aleksandra Kiuberis

23 februari 2026

1 Problem Definition

The objective of this assignment is to build a forecasting model that predicts the Dutch electricity CO_2 emission factor for the next 7 days (168 hourly values), based on renewable energy production data:

$$\hat{y}_{t+1:t+168}$$

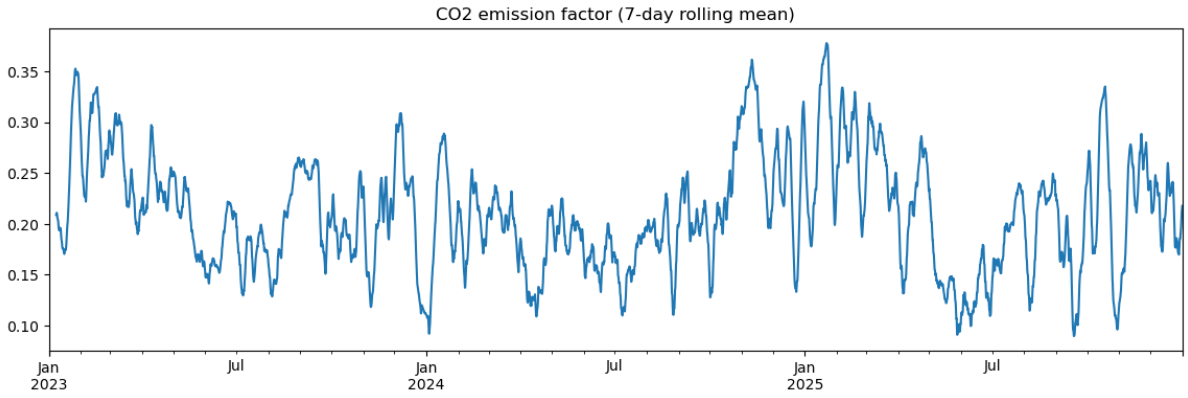
where y_t denotes the hourly CO_2 emission factor (kg CO_2/kWh). This is formulated as a multi-horizon time series forecasting task with exogenous variables: hourly frequency, 168-step prediction horizon, strong daily and weekly seasonality.

1.1 CO_2 data analysis:

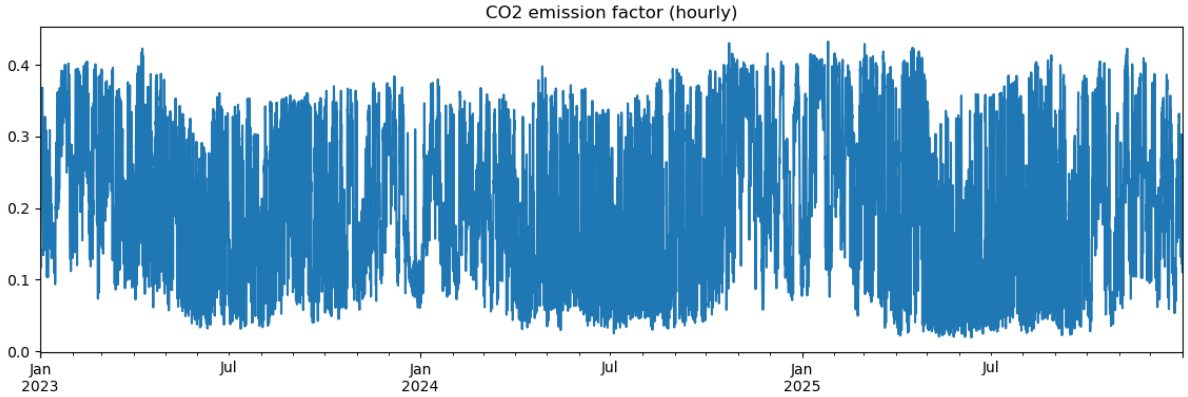
Before model development, an exploratory data analysis (EDA) was conducted to understand the temporal structure, variability, and statistical properties of the hourly CO_2 emission factor. The raw time series exhibits strong short-term volatility, with values ranging approximately between 0.03 and 0.42 kg CO_2/kWh. However, the 7-day rolling mean reveals smoother medium-term dynamics and clear regime shifts, indicating the presence of non-stationarity and seasonal effects 1. The analysis of average CO_2 intensity by hour 2 of day shows a pronounced diurnal cycle: lower emission factors are typically observed during late morning and midday hours, while higher values occur during evening and nighttime periods. This pattern likely reflects variations in electricity demand and generation mix throughout the day. Similarly, the aggregation by day of week suggests systematic weekday–weekend differences, indicating that calendar effects should be explicitly modeled 3.

The heatmap visualization (hour vs. date) confirms the presence of strong daily seasonality and evolving yearly trends 4. Horizontal banding indicates consistent intraday structure, while gradual shifts in color intensity over time suggest medium-term structural changes in the generation mix. Finally, the distribution of CO_2 values is clearly non-Gaussian and slightly multi-modal, implying that linear models may struggle to fully capture the underlying dynamics without nonlinear feature transformations.

Overall, the EDA indicates that the forecasting task involves multiple interacting temporal patterns: high-frequency noise, strong daily seasonality, weekly structure, and medium-term trend shifts. These observations justify the use of models capable of handling nonlinear relationships and multi-horizon dependencies, such as gradient boosting or deep learning architectures.



Figuur 1: 7-day rolling mean of the hourly CO_2 emission factor, highlighting medium-term trends and seasonal dynamics.



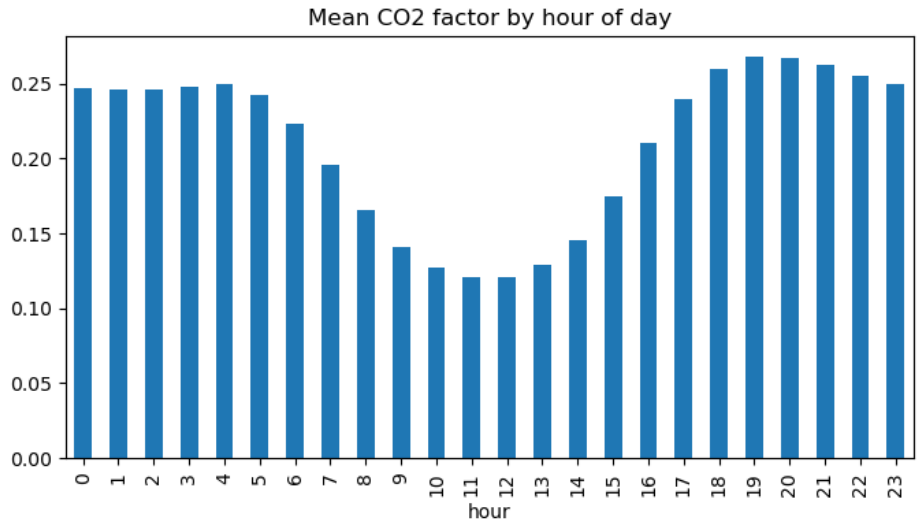
Figuur 2: Hourly CO_2 emission factor over the full study period (2023–2025), showing high-frequency variability and evolving seasonal patterns.

1.2 Evaluation Metrics

The following metrics were computed on the test set: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Absolute Percentage Error (MAPE) and MAE by horizon buckets (1-24, 25-72 and 73-168 hours). Due to sensitivity near low CO_2 values, MAE and RMSE are considered more reliable than MAPE.

2 Available Data:

Historical data was collected from the Dutch NED portal/API and includes: CO_2 emission factor, solar production, wind (onshore and offshore), biomass, waste, fossil gas, coal, nuclear. Additional features such as calendar encodings and optional weather forecasts were also incorporated. The dataset is recorded at an hourly frequency. To preserve the temporal structure and prevent information leakage, the data were split strictly in chronological order. All observations prior to 01-01-2025 were used for training, while the period from 01-01-2025 to 01-01-2026 was reserved for out-of-sample evaluation. No random shuffling was applied at



Figuur 3: Average CO_2 emission factor by hour of day. The pronounced diurnal pattern indicates strong intraday seasonality, with lower values during midday and higher values in evening and nighttime hours.

any stage.

2.1 Data Engineering & Quality Control

The data pipeline was designed with production-grade reliability in mind, with a focus on temporal consistency and leakage prevention.

API ingestion includes retry logic and pagination handling to ensure completeness under rate limits and transient failures. All series are resampled to a strict hourly resolution and aligned to a single master UTC index to eliminate silent time drift between sources, a common source of subtle modeling errors in energy data.

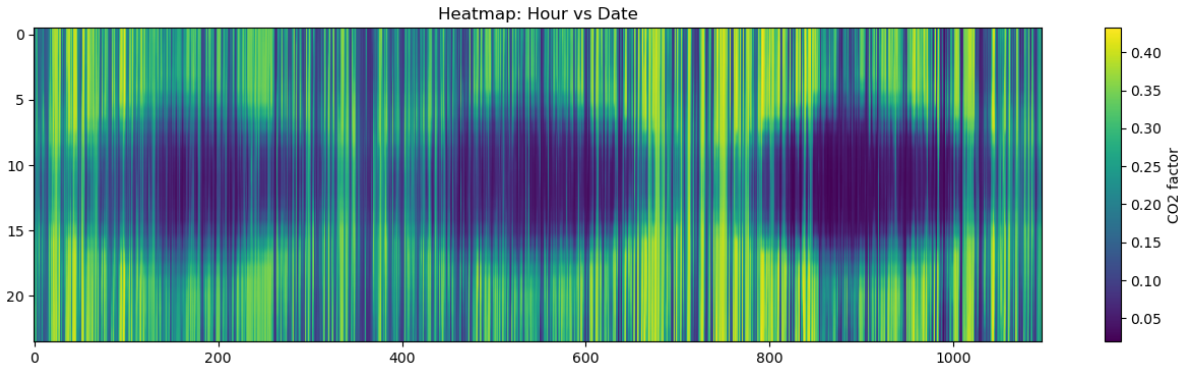
Missing values are handled conservatively and in a domain-aware manner. Short gaps in the CO_2 factor and wind production are interpolated within a limited horizon to preserve local continuity without distorting longer trends; corresponding missingness indicators are retained. Solar production values are set to zero where missing, reflecting a domain assumption rather than statistical imputation.

3 Modeling:

3.1 Feature Engineering:

Feature construction was guided by the seasonality and nonlinearity observed during exploratory analysis.

Lag features: For tree-based models, lagged versions of the CO_2 emission factor and renewable production were generated at multiple temporal offsets (1, 24, 168 and 336 hours). These lags enable the model to capture short-term dynamics as well as daily and weekly dependencies.



Figuur 4: Heatmap of hourly CO_2 emission factor values over time (hour of day vs. date), illustrating intraday seasonality and temporal variability.

Calendar features: Hour of day, day of week, and month were encoded. Cyclical encoding using sine and cosine transformations was applied to hour-of-day to preserve circular structure.

Rolling statistics: Rolling means and short-term smoothing windows were optionally included to provide local trend information while avoiding look-ahead bias.

Missingness indicators: Binary flags were retained for interpolated values to allow the model to account for imputation effects.

Scaling: Neural network models were trained using normalized targets and covariates, whereas tree-based models were trained on raw scale features.

All features were constructed strictly using past information relative to the prediction origin to prevent leakage.

3.2 Baseline: Seasonal Naive (t-24)

The Seasonal Naive (t-24) model was selected as a baseline because the CO_2 emission factor exhibits strong daily seasonality in hourly data. In this setting, the forecast for a given hour is simply equal to the observed value at the same hour on the previous day.

$$\hat{y}_{t+h} = y_{t+h-24}$$

This makes it a transparent and robust reference model that captures recurring diurnal patterns without introducing any learned parameters or risk of overfitting. Using a Seasonal Naive (t-24) baseline allows us to evaluate whether more advanced models (e.g., gradient boosting or transformer-based architectures) genuinely extract additional predictive signal beyond simple daily repetition. If a complex model cannot outperform this baseline, it suggests that the data are dominated by seasonal structure or that the modeling approach requires refinement.

3.3 Light Gradient-Boosting Model(LGBM):

LightGBM (LGBM) is a gradient boosting framework based on decision trees, widely used for tabular regression tasks due to its efficiency and strong predictive performance. In the context of multi-horizon time series forecasting, two modeling strategies can be applied. In

the single-model approach, one LightGBM regressor is trained to predict all forecast horizons simultaneously (e.g., 1-168 hours ahead), typically by including horizon-specific features or structuring the target accordingly. This approach is computationally efficient, easier to maintain in production, and ensures consistent behavior across horizons, but may struggle to optimally capture horizon-specific dynamics. In contrast, the direct multi-model approach trains 168 separate LightGBM models, each dedicated to a specific forecast horizon. This allows each model to specialize in the statistical structure of its respective lead time (e.g., short-term vs. long-term dependencies), often improving accuracy at the cost of increased training time, model management complexity, and deployment overhead. The choice between these strategies reflects a trade-off between scalability and horizon-specific optimization.

Model	MAE (H=1)	MAE (H=24)	MAE (H=168)
1 model - LGBM	0.0199	0.0654	0.0788
168 models - LGBM	0.0132	0.0686	0.0793

Tabel 1: N-HiTS performance (test set)

3.4 Neural Hierarchical Interpolation for Time Series Forecasting (NHiTS)

N-HiTS (Neural Hierarchical Interpolation for Time Series) is a deep learning architecture designed for multi-horizon time series forecasting. Unlike recurrent (LSTM) or attention-based (Transformer) models, N-HiTS relies on a stack of fully connected neural networks that iteratively decompose the input signal into hierarchical components. Each block operates on a progressively refined residual of the time series: it produces a coarse forecast at a lower temporal resolution, interpolates it to the full forecast horizon, and passes the remaining residual to the next block. The final prediction is obtained as the sum of forecasts from all hierarchical stacks. This multi-resolution structure allows N-HiTS to efficiently capture both long-term trends and short-term fluctuations while maintaining relatively low computational complexity. In practice, the model has shown strong performance on long forecasting horizons (e.g., 168 steps ahead) and is particularly suitable for structured, regularly sampled time series such as energy demand or CO_2 intensity forecasting, where stability, scalability, and inference speed are important considerations. N-HiTS was evaluated as an alternative deep learning architecture.

Model	MAE	RMSE	MAPE
N-HiTS	0.0529	0.0665	39.71%
TFT	0.0364	0.0458	27.69%

Tabel 2: N-HiTS&TFT performances (test set)

Model	MAE 1-24h:	MAE 25-72h:	MAE 73-168h:
N-HiTS	0.0427	0.0528	0.0555
TFT	0.0354	0.0364	0.0366

Tabel 3: N-HiTS&TFT performances (test set)

3.5 Temporal Fusion Transformer (TFT)

Temporal Fusion Transformer (TFT) is a deep learning architecture specifically designed for multi-horizon time series forecasting. It combines recurrent layers (for local sequential patterns), attention mechanisms (for long-range dependencies), and gating components that dynamically select relevant features. Unlike classical models, TFT can simultaneously use static features (e.g., region identifiers), known future inputs (e.g., calendar variables, weather forecasts), and observed historical variables (e.g., past CO_2 values). Its variable selection networks help the model automatically determine which features are important at each time step, improving interpretability. The attention layer further enables insight into which historical periods most influence a given forecast. As a result, TFT is particularly useful for complex forecasting tasks, such as energy demand or CO_2 intensity prediction, where nonlinear relationships, multiple exogenous drivers, and long forecast horizons must be modeled jointly and coherently.

Key architectural components:

- Encoder–Decoder structure for sequence-to-sequence learning
- Gating mechanisms to suppress irrelevant signals
- Variable Selection Networks
- Multi-head temporal attention
- Support for known future covariates

3.5.1 Temporal Fusion Transformer: Implementation Details

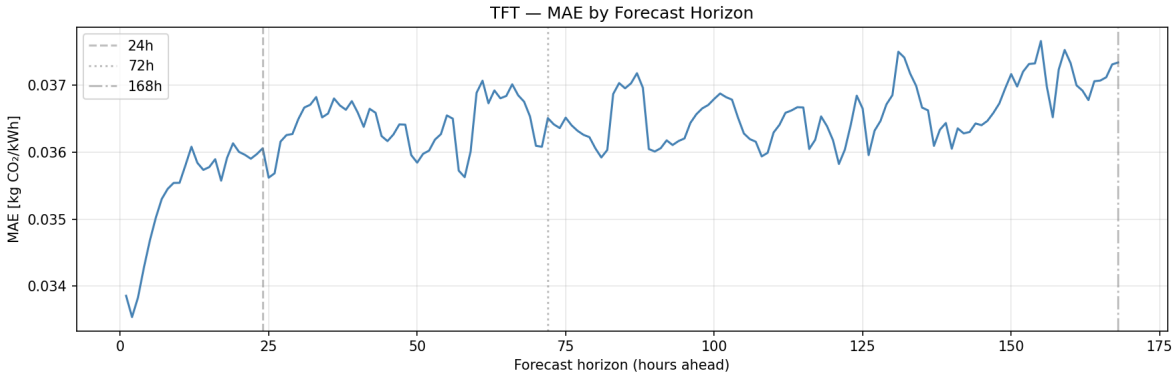
The TFT model was implemented using PyTorch Forecasting in a sequence-to-sequence configuration. Key configuration parameters:

- Encoder length: 168 hours
- Prediction length: 168 hours
- Hidden size: 64
- Dropout: 0.1
- Loss function: MAE
- Optimizer: Adamw
- Early stopping with validation monitoring

The model uses static covariates, known future inputs (calendar features), and observed historical variables (past CO_2 and renewable production). Attention layers allow dynamic weighting of historical time steps, while gating mechanisms suppress irrelevant features. Training was performed with mini-batch gradient descent and gradient clipping to ensure numerical stability.

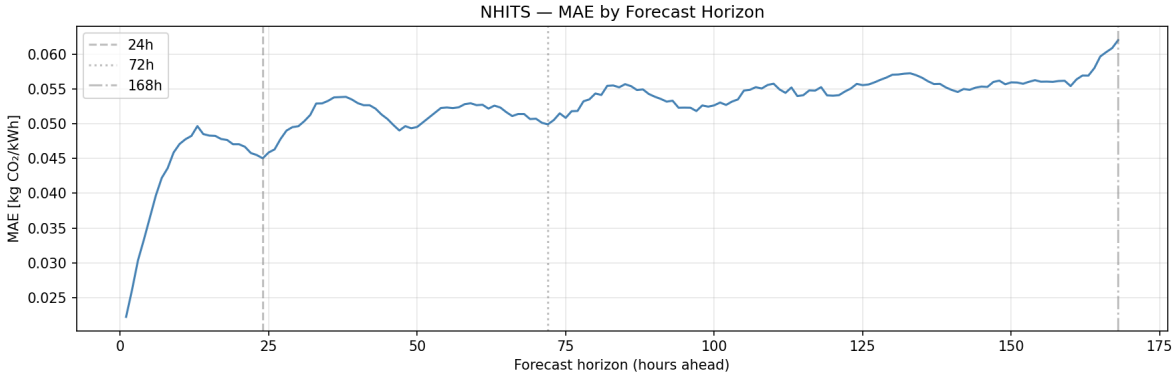
4 Error Analysis

Worst-performing windows were stored for inspection. Error spikes typically correspond to rapid shifts in generation mix or unusual operational states.



Figuur 5: Mean Absolute Error (MAE) as a function of forecast horizon for the TFT model over a 168-hour prediction window.

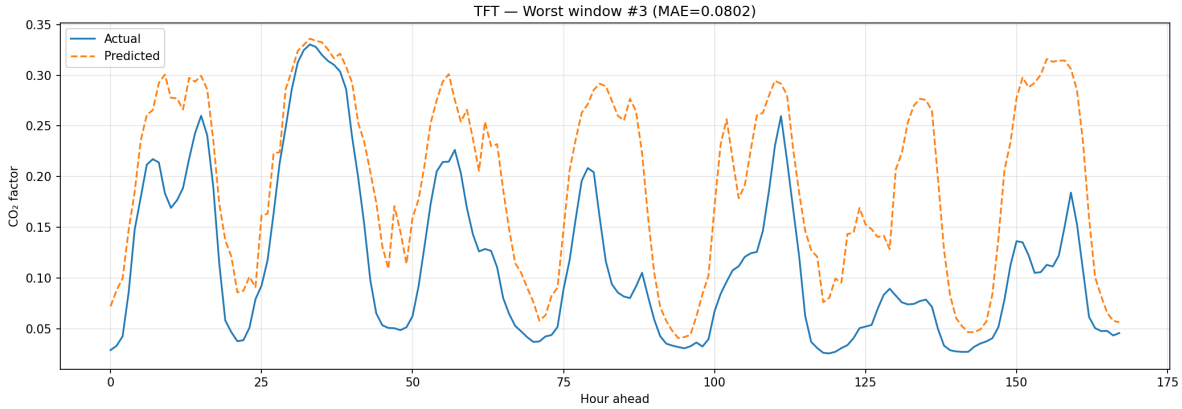
Figure5 shows the MAE as a function of forecast horizon for the Temporal Fusion Transformer. The error remains remarkably stable across the full 168-hour prediction window, indicating that the model generalizes well to long-term forecasts. Unlike simpler autoregressive approaches, the degradation in accuracy with increasing horizon is minimal, which highlights the advantage of sequence-to-sequence attention-based modeling.



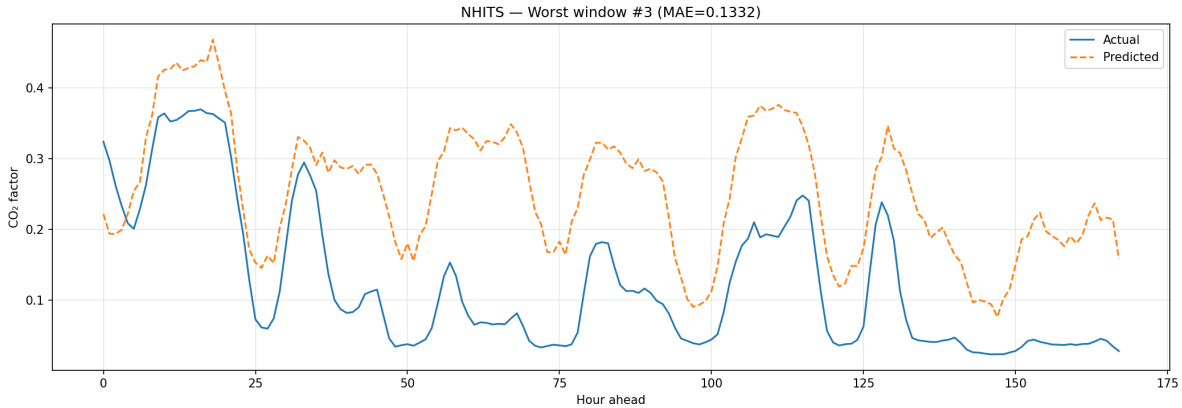
Figuur 6: Mean Absolute Error (MAE) as a function of forecast horizon for the N-HiTS model over a 168-hour prediction window.

Figure6 presents the MAE by forecast horizon for N-HiTS. The model exhibits a noticeable increase in error during the first 24 hours and a gradual degradation toward longer horizons. Compared to TFT, the error growth is more pronounced, suggesting that N-HiTS struggles more with long-range dependencies in this dataset.

To better understand model robustness, I analyzed the worst-performing prediction window for both models 7,8. N-HiTS exhibits systematic over prediction and fails to accurately



Figuur 7: Worst-performing 168-hour prediction window for the Temporal Fusion Transformer (MAE=0.0802).



Figuur 8: Worst-performing 168-hour prediction window for the N-HiTS model (MAE=0.1332).

capture rapid CO_2 factor fluctuations. In contrast, TFT preserves the overall temporal structure more effectively and maintains lower peak errors, even in challenging periods.

5 Conclusion:

This study evaluated multiple approaches for multi-horizon CO_2 emission factor forecasting. While daily seasonality explains a significant portion of variance, tree-based models demonstrate that nonlinear relationships with renewable production provide additional predictive signal. Deep learning architectures further improve long-horizon stability. Among all evaluated models, the Temporal Fusion Transformer achieved the best overall performance (MAE = 0.0364 kg CO_2/kWh), maintaining stable accuracy across the full 168-hour forecast window. Unlike tree-based approaches, TFT demonstrates robustness to regime shifts and better captures complex temporal interactions.

These results suggest that attention-based sequence-to-sequence modeling provides a mea-

ningful advantage in structured energy forecasting tasks involving multiple exogenous drivers and long prediction horizons.

However, the current setup does not explicitly incorporate renewable production forecasts and does not include probabilistic calibration. Future work should address two-stage forecasting pipelines and drift-aware monitoring to enhance production readiness.