CONFORMAL INFERENCE FOR ENERGY PRICE FORECASTING

Frederik V. Svane (s224766)
Marcus P. L. Christoffersen (s224750)

ABSTRACT

The abstract should contain about 100 to 150 words summarizing the problem, approach, and key findings.

1. INTRODUCTION

Energy price forecasting, like other investment domains, benefits immensely from uncertainty estimates for proper risk management. Conformal prediction is able to provide these prediction intervals, with guaranteed coverage. Setting a miscoverage rate of $\alpha = 0.1$ means we expect 90% of true values to fall within our intervals. However, standard conformal methods assume exchangeable data, which is violated in time series, where observations are sequentially dependent/inexchangeable.

Extensions to standard conformal inference, such as Adaptive Conformal Inference (ACI) and Online Gradient Descent (OGD), have been developed to address this problem of assumed exchangeability, by dynamically adjusting intervals based on binary coverage feedback. Error-Quantified Conformal Inference (ECI) [1] extends these methods further, by incorporating the *magnitude* of prediction errors, instead of mere binary miss/cover feedback.

In this project, we apply these conformal methods to English electricity spot price forecasting using 2022 hourly data with weather and consumption features. We implement two forecasting models, SARIMAX and LSTM, and compare OGD and ECI for constructing prediction intervals. This allows us to explore how online conformal prediction performs on real energy data from the 2022 energy crisis, where the market experienced increased volatility and distribution shifts.

2. METHODS

2.1. Conformal Prediction

Conformal prediction constructs prediction intervals by comparing new observations to a calibration set. Given a point prediction $\hat{y}_t$ and a true value y_t, we define a nonconformity score $s_t = |y_t - \hat{y}_t|$ measuring how "unusual" the observation is. A prediction interval is then constructed as $[\hat{y}_t - q_t, \hat{y}_t + q_t]$, where q_t is a threshold chosen such that the interval covers the true value with probability $1 - \alpha$.

Standard conformal prediction assumes exchangeability: the joint distribution of observations is invariant to permutation. This assumption fails for time series data, where observations are chronologically dependent (ie. the next spot price depends on the current spot price).

Additionally, the underlying distribution may shift over time, meaning the statistical properties of prices in summer differ from those during winter, and likewise for crisis periods compared to stable periods. So a calibration set from the past may not represent future conditions, causing coverage guarantees to break down.

2.2. Online Conformal Methods

Online conformal methods address non-exchangeability by updating the threshold q_t dynamically based on observed coverage. The Online Gradient Descent (OGD) method uses the update rule

$$q_{t+1} = q_t + \eta \cdot (\text{err}_t - \alpha) \tag{1}$$

where η is a learning rate and the binary variable $\text{err}_t \in \{0, 1\}$ indicates whether the true value fell outside the interval. When coverage is too low ($\text{err}_t = 1$ more often than α), the threshold increases, widening intervals. When coverage is too high, the threshold decreases, tightening intervals.

Error-Quantified Conformal Inference (ECI) extends OGD by incorporating the magnitude of prediction errors. The update rule becomes

$$q_{t+1} = q_t + \eta \cdot [(\text{err}_t - \alpha) + (s_t - q_t) \cdot f'(s_t - q_t)] \tag{2}$$

where s_t is the nonconformity score and f is a shaping function, such as a sigmoid or a Gaussian kernel. As such, the additional term $(s_t - q_t) \cdot f'(s_t - q_t)$ adjusts the update based on how far the score was from the threshold. The choice of shaping function comes with considerations: the sigmoid saturates for large deviations, giving extreme misses similar corrections to moderate ones. The Gaussian kernel goes further, actively dampening corrections for extreme deviations, treating them as potential outliers that should have less influence on the threshold.

2.3. Data

We use hourly electricity market data from England spanning February 1, 2022 to February 1, 2023, comprising 8,760 observations. The target variable is the day-ahead spot price (GBP/MWh), while the raw features include wind power forecasts from a wind farm near London (MW), temperature forecasts (°C), historical normal temperatures (°C), and timestamps.

This period coincides with the European energy crisis, characterized by significant price volatility and distribution shifts. These conditions stress-test conformal methods designed for non-stationary environments.

2.3.1. Feature engineering

With only a single year of data, capturing seasonal patterns directly is challenging. To address this, we engineer additional features from the available measurements.

From the timestamp, we derive hour of day (1–24), day of year (1–365), week number (1–52), and month (1–12). To represent annual seasonality without discontinuities, we encode day of year cyclically:

$$\text{day_sin} = \sin\left(\frac{2\pi \cdot \text{day}}{365}\right), \quad \text{day_cos} = \cos\left(\frac{2\pi \cdot \text{day}}{365}\right)$$

We also construct features based on energy market dynamics. Following WHO guidelines that recommend a minimum indoor temperature of 18°C [2], we define heating demand as $\max(0, 18 - \text{temperature forecast})$. We then multiply this by consumption forecasts, based on the hypothesis that cold weather combined with high demand creates price pressure.

Additionally, we compute temperature deviation from historical norms, based on the intuition that unusual weather may drive price volatility.

Finally, we include lagged spot price $(t-1)$ as a feature, reflecting the realistic constraint that only past prices are observable at prediction time.

This yields 14 input features in total. The data is split 80/20 into training and test sets, preserving temporal order to prevent information leakage.

2.4. Forecasting Models

We use two forecasting models to generate point predictions, representing deep learning and classical time series approaches.

2.4.1. Deep Learning Model

Sequential data like this calls for a recurrent architecture. As covered in the lectures, standard feedforward networks treat each input independently, failing to capture positional relationships within a sequence. They also require fixed input dimensions, making them unsuitable for variable-length sequences. Recurrent neural networks (RNNs) address both issues by processing inputs sequentially and maintaining a hidden state that carries information across time steps.

However, traditional RNNs suffer from vanishing or exploding gradients. During backpropagation, gradients in early layers are computed via the chain rule, resulting in a product across all time steps:

$$\frac{\partial \mathcal{L}}{\partial h_0} = \prod_{t=1}^{T} \frac{\partial h_t}{\partial h_{t-1}} \cdot \frac{\partial \mathcal{L}}{\partial h_T}$$

For long sequences, this product either explodes or vanishes. If the intermediate gradients are consistently greater than one, the product blows up. If they are consistently smaller than one, it collapses to zero.

LSTMs address this through a gating mechanism, as we saw in lecture. The cell state acts as a highway that allows gradients to flow across many time steps with minimal transformation. The forget gate controls what information to discard, the input gate controls what new information to store, and the output gate controls what to expose to the next layer. These gates use sigmoid activations to produce values in $[0, 1]$, enabling smooth gradient flow during training.

We considered adding attention mechanisms or moving to a full Transformer architecture, but opted against this for practical reasons. Both require substantially more data to train effectively, and our dataset spans only a single year. They also introduce computational overhead that slows down iteration during experimentation. Given these constraints, a standard LSTM provides a reasonable balance between expressiveness and trainability.

Our implementation uses PyTorch and consists of two stacked LSTM layers with 64 hidden units each, followed by a linear output layer. The network processes sequences of 48 hourly observations (two days), with only the final hidden state passed to the output layer to produce a point prediction for the next hour. The implementation can be found in `models.py` in the linked GitHub repository.

We tune sequence length and number of training epochs via grid search on DTU's HPC cluster. For sequence length, we tested windows ranging from 1 day (24 hours) to 14 days (336 hours), with 48 hours (two days) performing best. The sequence length search is implemented in `seq_len_search.py`, and the epoch search in `LSTM_model_cross_val_train.ipynb`.

Standard k-fold cross-validation is inappropriate for time series, as it allows future information to leak into training folds. Instead, we use chronological cross-validation with an expanding window: train on data up to time t, validate on a fixed horizon ahead, then advance t and repeat.

2.4.2. *Classical time series model*

For comparison with the deep learning approach, we implement a classical time series model using ARIMAX (AutoRegressive Integrated Moving Average with eXogenous variables). This class of models has been the standard approach for time series forecasting for decades and provides a strong baseline for our conformal prediction experiments.

ARIMAX models decompose the forecast into three components, specified by the order (p, d, q).

The AutoRegressive (AR) component captures how the target depends linearly on its own past values. With order p, the model uses the previous p observations:

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t$$

where ϕ_i are the AR coefficients and c is a constant.

The Integrated (I) component with order d applies differencing d times to make the series stationary. For $d = 1$, we model $\Delta y_t = y_t - y_{t-1}$ instead of y_t directly. This removes trends that would otherwise violate stationarity assumptions.

The Moving Average (MA) component models the target as a function of past forecast errors. With order q, it incorporates the previous q error terms:

$$y_t = \mu + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t$$

where θ_j are the MA coefficients, μ is the mean, and ϵ_t represents white noise.

Combining all three components with the exogenous variables X (wind forecasts, temperature, consumption, etc.), the full ARIMAX model becomes:

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \beta^T X_t + \epsilon_t$$

where β are the coefficients for the exogenous variables. This formula applies after taking d differences of the original series.

Selecting the order (p, d, q) manually requires examining autocorrelation plots and testing multiple configurations. To automate this process, we use the `auto_arima` function from the `pmdarima` package, which performs a systematic search over candidate orders. We constrain the search space to $p, q \in [0, 5]$ and use the Bayesian Information Criterion (BIC). The search procedure tests combinations and selects the order that minimizes BIC.

Once the optimal order is identified, we fit the final model using `statsmodels.SARIMAX`. We set avoid seasonality, since seasonal ARIMA requires multiple complete seasonal cycles to reliably estimate seasonal parameters. With only one year of data (Feb 2022 - Feb 2023), we cannot observe repeated annual patterns, making seasonal parameter estimation unreliable.

The implementation can be found in `models.py` and `Time_series_models.ipynb` in the linked repository.

3. EXPERIMENTS & RESULTS

We conduct experiments to evaluate how well online conformal methods construct prediction intervals for energy price forecasting. Our focus is on comparing OGD and ECI across different forecasting models and understanding their performance.

3.1. Experimental Setup

We compare OGD and ECI on both ARIMAX and LSTM models. ECI is evaluated in three variants: basic, cutoff and integral each with sigmoid or Gaussian shaping functions. Performance is measured using coverage rate (the proportion of true values falling within prediction intervals), average interval width, avg median. The target coverage is 90%, corresponding to miscoverage rate $\alpha = 0.1$.

Hyperparameters are selected through random search over 10,000 configurations. For OGD, we sample learning rate $\eta \in [0.01, 10]$ and initial threshold $q_0 \in [5, 100]$. For ECI, we additionally sample sigmoid scaling parameter $c \in [0.1, 10]$ and select between sigmoid and Gaussian shaping functions. The cutoff variant of ECI uses a rolling window for adaptive thresholding, with window length sampled from $\{20, \ldots, 200\}$ and Gaussian bandwidth $h \in [0.3, 2.0]$. From the evaluated configurations, we select the one that minimizes average interval width while achieving coverage between 88% and 92%.

The dataset is split 80/20 along the temporal axis. Training data spans February 1 to November 20, 2022 (7,026 observations), and test data spans November 20, 2022 to February 1, 2023 (1,757 observations). This split ensures the test period includes the volatile winter months of the energy crisis, providing a test of how well the conformal methods handle distribution shifts.

3.2. Results

We present results comparing OGD and ECI variants across both forecasting models, examining both visual behavior and quantitative performance metrics.

Fig. 1. SARIMAX results with Confidence Intervals for both OGD and ECI

Figure 1 presents the SARIMAX model results with similar layout. Compared to LSTM, both conformal methods produce noticeably wider intervals throughout the test period. The predicted prices (orange dashed line) show larger deviations from actual prices, and the intervals remain persistently wide even during stable periods, reflecting the classical model's higher prediction uncertainty.

Fig. 2. LSTM results with Confidence Intervals for both OGD and ECI

Figure 2 shows the LSTM model predictions with conformal intervals over the test period. The top panel displays OGD intervals (purple shaded region), while the bottom panel shows ECI intervals (green shaded region). Both methods track the actual prices (blue line) well, with intervals widening during volatile periods (notably around timesteps 400-700) and tightening during stable periods. Visually, the ECI intervals appear narrower, particularly during stable periods between timesteps 800-1200, suggesting more adaptive behavior.

Table 1. Performance comparison of conformal methods. Coverage (%), average width, and median width.

Method / Model	ARIMAX			LSTM		
	Cov. (%)	Avg. W.	Med. W.	Cov. (%)	Avg. W.	Med. W.
OGD	90.0	222	198	90.2	107	106
ECI	88.7	**219**	**197**	88.6	**101**	**101**
ECI-cut.	88.7	220	197	88.6	101	104
ECI-int.	86.0	220	196	82.2	92	97

Table 1 quantifies the performance differences observed in the figures. For the LSTM model, OGD achieves 90.2% coverage with average interval width of 107 DKK/MWh, closely matching the target 90% coverage rate. The basic ECI variant produces narrower intervals (101 DKK/MWh average and median), but with slightly lower coverage at 88.6%. ECI-cutoff shows similar performance (88.6% coverage, 101 DKK/MWh average width), while ECI-integral achieves the narrowest intervals (92 DKK/MWh average) at the cost of significantly undercovering at 82.2%.

For the ARIMAX model, all methods produce substantially wider intervals. OGD achieves exactly 90.0% coverage with 222 DKK/MWh average width, while the ECI variants show marginal width reductions (219-220 DKK/MWh) but fall below target coverage (86.0-88.7%). The difference in interval widths between LSTM and ARIMAX (roughly 2× wider for ARIMAX) confirms the visual observations and reflects the classical model's higher prediction uncertainty.

3.3. Analysis

All methods achieve coverage within a reasonable range of the target 90%. The key differentiator is interval efficiency. For LSTM, basic ECI achieves an average interval width of 101 DKK/MWh compared to OGD's 107 DKK/MWh—a reduction of 6 DKK/MWh, or approximately 5.6%. This improvement is achieved while maintaining 88.6% coverage, only 1.6 percentage points below target. By incorporating error magnitude rather than binary feedback, ECI adaptively tightens intervals during stable periods (visible in Figure 2 around timesteps 800-1200), making it the superior choice when base model quality is high.

The ARIMAX results reveal that interval width is primarily driven by base model quality. ARIMAX ECI shows roughly double the average width (219 DKK/MWh) compared to LSTM ECI (101 DKK/MWh) due to the classical model's larger prediction errors. When base predictions are poor, error-quantification provides minimal advantage over binary feedback—all ARIMAX methods differ by only 3 DKK/MWh. This underscores that conformal methods amplify base model quality rather than compensate for it.

It is worth noting that as the test period progresses, the conformal methods accumulate more observations, allowing the threshold q_t to adapt more effectively. This suggests that

with longer test sequences, average interval widths would continue to improve as the methods learn from additional coverage feedback.

4. CONCLUSION

5. REFERENCES

[1] Junxi Wu, Dongjian Hu, Yajie Bao, Shu-Tao Xia, and Changliang Zou, "Error-quantified conformal inference for time series," 2025.

[2] World Health Organization, *WHO Housing and Health Guidelines*, World Health Organization, Geneva, 2018.

DECLARATION OF USE OF GENERATIVE AI

This declaration **must** be filled out and included as the **final page** of the document. The questions apply to all parts of the work, including research, project writing, and coding.

- I/we have used generative AI tools: yes

If you answered *yes*, please complete the following sections. List the generative AI tools you have used:

- Claude by Anthropic AI
- Claude Code by Anthropic AI

Describe how the tools were used:

What did you use the tool(s) for?

We used the named AI tools for brainstorming- and evaluating ideas, proofreading and correcting parts of the text in the report, bugfixing and generating non-essential, utility code, such as plotting data with matplotlib.

At what stage(s) of the process did you use the tool(s)?

We used the named AI tools during the research-, implementation and proofreading stages of the process.

How did you use or incorporate the generated output?

Some generated output we simply read and responded to, in order to guide our thinking, in a similar way to what we would have done to a TA or professor. Some coding output, such as matplotlib plotting code, we ran directly. And grammatical corrections to sections in our report, we included in the report.