Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis International Econometrics and Operations Research

From Stability to Turbulence: GAS Models for Volatility Forecasting in Electricity Markets

Antoni Borys (613195)

Abstract

This thesis investigates the application of Generalized Autoregressive Score (GAS) models in forecasting one-day-ahead conditional volatilities of daily electricity returns. The primary goals were to evaluate and compare their forecasting performance and examine their robustness across varying market conditions. The models were estimated using relatively low-volatile data to then forecast the variability of returns during periods of increased fluctuations, which is of particular relevance in managing the stability of electricity prices. The main contribution of this paper lies in extending the log-scale score-driven models with Gaussian and Student's t-distribution of errors to a more flexible Generalized-t specification, as well as incorporating asymmetric EGB2 within the observation-driven class of models. Additionally, exogenous covariates are introduced to the scale-updating equation. The results of the paper show that the predictive performance of the models depends on the positivity bias and sensitivity to large errors of the loss function used. Models based on thin-tailed Gaussian and asymmetric EGB2 errors score lowest in QLIKE, with the latter appearing to be the most robust. Fatter-tailed Generalized-t and Student's t-based models outperform the previous two in MSE results. The covariates describing weather conditions do not significantly improve the in-sample fit or forecasting accuracy; however, their choice was made without sufficient domain knowledge. The findings underscore the potential of GAS models as reliable tools for stakeholders in electricity markets, aiding in informed decision-making and risk management.

Supervisor:	prof. Donker van Heel, Simon
Second assessor:	prof. Robin Lumsdaine
Date final version:	2nd July 2024

1 Introduction

The volatility of electricity prices in wholesale markets is a significant concern for market participants, including Regional Transmission Organizations (RTOs), electricity suppliers, and utilities. These stakeholders require accurate volatility predictions to manage financial risks and maintain operational efficiency. Understanding and forecasting the conditional volatility of electricity prices is essential for effective market operations and planning.

This paper contributes to the research on volatility modelling which is a major challenge in electricity markets. Electricity must be produced and consumed simultaneously because it cannot be economically stored in large quantities. This requirement for instantaneous balance between supply and demand in the power grid, coupled with the inflexibility of generation capacities, can lead to significant price volatility. Energy price spikes bring risk and uncertainty to businesses and consumers, while sudden price cuts, often caused by an oversupply of electricity, can be costly for producers and may even lead to infrastructure damage if the excess is not absorbed. Accurate volatility modelling can improve the efficiency of electricity markets and allow businesses to hedge against price spikes, optimize generation schedules, integrate renewable energy sources, and maintain the balance of the power grid by anticipating price fluctuations and peak levels of consumption and supply.

The primary objective of this research is to explore the potential of Generalized Autoregressive Score (GAS) models for forecasting the conditional volatilities of daily electricity returns. The study aims to evaluate the performance of different score-driven models under various distributional assumptions and to determine their practical applicability in real-world scenarios. Additionally, models investigate the significance of regressors in the volatility modelling besides standard autoregressive and innovation terms. This research contributes to the broader field of volatility modeling, which is a major challenge in electricity markets, and has potential applications in financial and other commodity markets. The findings from this research have significant implications for stakeholders in the electricity market such as RTOs that are responsible for managing stability of electricity prices. By providing robust and accurate volatility forecasts, GAS models can help in better risk management, more efficient market operations, and informed decision-making.

The research is conducted using daily returns data from the PJM market, covering a specific period to ensure consistency and relevance. The study employs a fixed-window approach for model estimation and evaluation, ensuring that parameters remain consistent throughout the analysis period.

This thesis seeks to address the following research questions:

1. Can GAS models provide reliable volatility forecasts for turbulent market conditions having been estimated using low-volatility data?

2. What are the comparative performances of different GAS models under various distributional assumptions for error terms and optional covariates?

3. To what extent the one-day ahead electricity forecasts provided by the GAS model are affected by outliers and periods of persisting volatility?

1.1 Theoretical framework

In statistical modelling, it is crucial to evaluate how well the parameters fit the observed data which is commonly described by the total (log-) likelihood that aggregates the probabilistic fit of all observations. Then, the score refers to the (partial) derivative of the natural log of the likelihood function (parameter-wise) and its sign indicates the direction in which a parameter should move to increase the function. The Fisher information, defined as the variance of the score (given the set of parameters), measures the amount of information the observed data provides about the parameter. Higher values indicate that the observed data are more informative about the parameter, leading to more precise parameter estimates.

This paper deals with univariate score-driven models, where the dependent variable is a single-dimensional series of observations. Multivariate score-driven models, which examine interactions between multiple variables over time, are also explored in the literature, for example in [Artemova et al., 2022b].

Suppose y_t represents a univariate time series observation at time t, for $t = 1, ...T$. Let y_t be characterized by the predictive conditional density $p_y(y_t \mid f_t, \mathcal{F}_{t-1}; \theta)$, where f_t is the time-varying parameter and θ be the set of fixed coefficients that describe the predictive density of y_t. Additionally, let $\mathcal{F}_{t-1}$ be the set of all information available at time $t-1$ concerning the observations $y_1, ..., y_{t-1}$.

The observations are assumed to be sequentially available over time. Upon the arrival of observation y_t, the time-varying parameter f_t is updated using the equation $f_t = \omega + \alpha S_t \nabla_t + \beta f_{t-1}$ where $\omega, \alpha, \beta \in \theta$ are the time in-varying coefficients that describe the process. This resembles a simple autoregressive model of order 1 with the difference of the $S_t \nabla_t$ innovation term. The ∇_t denotes the score of the predictive conditional density with respect to f_t and S_t be the scaling function that determines the magnitude that ∇_t impacts the updating at time t, to obtain the successive f_{t+1}. The score causes f_{t+1} to update from f_t in the direction that would have increased the probability of observation y_t (depending on the sign of αS_t).

This yields the below specification of the score-driven model as in [Creal et al., 2013]:

$$
\begin{aligned}
&y_t \sim p_y(y_t \mid f_t, \mathcal{F}_{t-1}; \theta), \quad f_{t+1} = \omega + \alpha s_t + \beta f_t, \\
&s_t = S_t \cdot \nabla_t, \quad \nabla_t = \frac{\partial \log p_y(y_t \mid f_t, \mathcal{F}_{t-1}; \theta)}{\partial f_t}
\end{aligned}
\tag{1}
$$

A similar model formulation was proposed by [Harvey, 2013].

Different choices of S_t result in different significance of the innovation relative to the autoregressive term in the f_t process. Typically, a natural selection of the scaling factor is a function of the variance of ∇_t, $\mathscr{J}_{t|t-1}$ (known as Fisher information), to account for the inherent variability in the score given the nature of the data. At any time point t, under regularity conditions $E_{t-1}[\nabla_t] = 0$ since the estimated parameters under MLE are consistent hence in expectation the f_t should be such that improvement in the log-likelihood is not possible if evaluated in true parameters. Therefore, the asymptotic variance of the score $\mathscr{J}_{t|t-1} = Var_{t-1}[\nabla_t] = E_{t-1}[\nabla_t \nabla_t']$.

Since the S_t scales the score ∇_t, a popular approach which is also used in this paper, is to set it equal to an inverse asymptotic variance $\mathscr{J}_{t|t-1}^{-1}$. This choice of scaling factor accounts for different variability of ∇_t over time and the variance of the scaled scores s_t become $\mathscr{J}_{t|t-1}^{-1}$. As observed by [Creal et al., 2011], then the updating equation of f_t as shown in the Equation 1,

is equivalent to estimating f_t over time using a Gauss-Newton algorithm. Using this approach, several popular autoregressive models can be generalized to a score-driven framework such as GARCH of [Engle and Bollerslev, 1986] or multiplicative error MEM by [Engle and Gallo, 2006]. Other possible scaling functions beyond application in this paper, include the square root of the inverse information matrix or more straightforwardly, a constant such that no additional scaling is applied to the score.

The models discussed in this paper fall in the class of observation-driven specifications for which time-varying parameters are treated as functions of lagged dependent variables and exogenous covariates as explained by [Artemova et al., 2022a]. This ensures that conditional on past and concurrent information the parameters are fully known and don't have their own source of errors. The discussed score-driven models in this paper are of log-scale specification with non-zero intercept as explained later in this paper. The two benchmark models from [Artemova et al., 2022a] include Gaussian and Student's t distribution of errors. Additionally, three more extensions are proposed that use zero-mean, unit-scaled, standardized Generalized-t and EGB2 distributions. Finally, the alternative Gen-t-based model of two exogenous covariates (describing weather conditions) is investigated.

The rest of this paper is organized as follows. First, the datasets used in this research are introduced in Section 2. These include the electricity prices that the dependent variable is derived from (subsection 2.1), and weather records based on which exogenous covariates are extracted (subsection 2.2). Then, Section 3 contains an overview of the models, score-driven specifications and metrics reported along with the results later in the paper. Subsection 3.1 describes the target variables and loss functions used to evaluate the out-of-sample performance of the models. The benchmark and extension models explored in this research are discussed in subsections 3.2 and 3.3 respectively. Subsequently, Section 4 presents the estimated models along with the prediction results. The estimates of the parameters along the information criteria and in-sample fit discussion are featured in subsection 4.1. The estimated models are used to produce the forecasts whose accuracy is evaluated using MSE and QLIKE loss functions as shown in subsection 4.2. The research is concluded in Section 5 along with suggestions for further research.

2 Data

This section outlines the datasets used in this research. First and most importantly, subsection 2.1 discusses the dataset concerning the dependent variable behind the research - electricity returns. This data also motivates the research question and goals of this paper as explained below. Secondly, subsection 2.2 provides details on the dataset used to extract regressors as discussed in subsection 3.3.

2.1 Electircity returns

The main dataset considered in this research is a sequence of daily returns from the real-time PJM market. PJM (Pennsylvania-New Jersey-Maryland) is a regional transmission organization (RTO) in the United States coordinating the movement of wholesale electricity in 13 states and the District of Columbia. Importantly, it operates the day-ahead and real-time energy markets where electricity can be competitively traded.

First, the hourly prices of each day from April 6, 2008, to December 31, 2015, were collected, as can be found under "Real-Time Hourly LMPs" on PJM's website. Then, the daily electricity prices were constructed as the average hourly prices every given day, as presented in Figure 1. Third, the corresponding daily spot price returns were calculated as $(p_t - p_{t-1})/p_{t-1}$, where p_t denotes the daily (average) price on day t. This produced the 2826 returns which were finally scaled by a factor of 10 for the sake of numerical stability. Such processed returns are displayed in Figure 2.

Models' coefficients were estimates based on daily returns obtained from daily prices over the period April 6, 2008, to December 31, 2013. This constitutes 2096 observations that are marked with colour black in the below Figures. This dataset was chosen such that it is identical to the one used by [Artemova et al., 2022b] to estimate the score-driven scale models in the illustration of electricity spot price returns. Further, the observations in red denote the forecasting sample covering the range of 2 years from January 1, 2014, to December 31, 2015 (730 observations).

The choice of the testing sample window can be motivated by the following reasons. First, as depicted in Figures 1 and 2 the concerned testing sample exhibits significantly higher price fluctuations. Forecasting returns' behaviour during increased volatility periods is of high relevance because it helps to avoid high financial and efficiency losses during turbulent market conditions. This is essential for optimal electricity supply management and stakeholders such as RTOs. Secondly, increased volatility periods are expected to be more challenging to forecast due to larger uncertainty and unpredictability of market behaviour. It is specifically of interest to observe how well score-driven models trained on low-volatility data can predict price movements in a high-volatility environment. This contributes to the discussion of generalization, adaptation, and robustness of different approaches within the GAS framework as potentially minor differences in specifications and model assumptions will produce magnified performance discrepancies due to challenging data. Beyond 2015, the electricity prices stabilize and no price shocks in the returns can be observed. The reaction of the models to this stabilization is beyond the scope of this research. Finally, the employed division of data into training and testing samples produces 2096 and 730 observations, respectively. This constitutes approximately a 75%-25% split which provides a balanced approach to model training and evaluation and is a common interval practice in the forecasting literature as found for example in [Joseph, 2022].

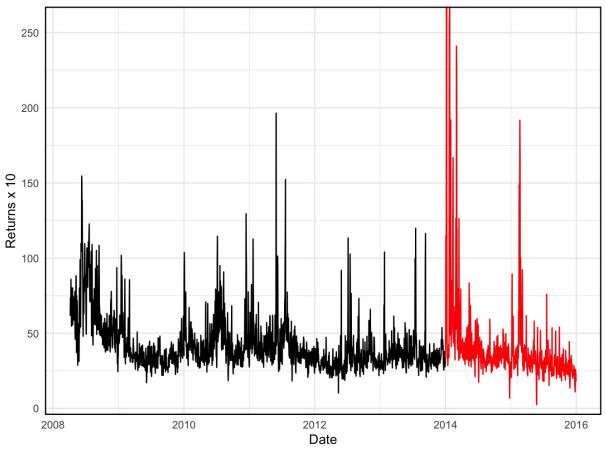

Figure 1: Average daily electricity prices in PJM market over train (black) and test samples (red)

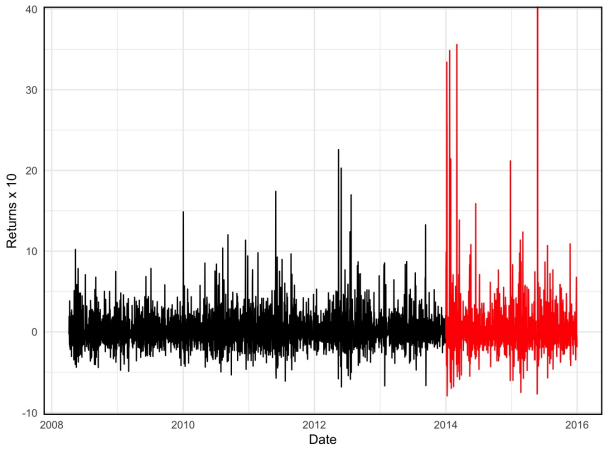

Figure 2: Average daily electricity returns in PJM market, scaled by a factor of 10)

The above figures present several interesting features of electricity markets.

First of all, electricity markets are known for sudden price spikes due to unexpected supply disruptions or extreme weather events. This behaviour finds evidence in the sharp surges that can be seen in Figure 1 that further translate into significant returns' volatilities in the periods of intensified price movements as Figure 2 indicates. Notably, positive price jumps are completely prevailing both in terms of frequency and magnitude over the downward direction.

Additionally, as pointed out earlier the testing sample was chosen such that it comprises a period of increased volatility. This seems to be reflected in the data since beyond 2014, price surges occurred more often and peaked much higher specifically in early 2014. Out of the 15 largest average daily electricity prices recorded, 14 occur in the testing sample and all before mid-March 2014 after which the price falls back to sub-100s. Correspondingly, more extreme return observations intensify post-2013 as can be seen in Figure 2 and only one of the top 8 returns observations occurs within the training sample.

This suggests the models are estimated based on the data from a relatively stable period to then forecast the heightened volatilities of the returns. This supports the goal of the research to investigate the performance of score-driven models going from low to high volatility time and evaluate their robustness and responsiveness in different specifications. Additionally, the data underscores the practical importance of filtering volatility in electricity markets and seems appropriate to achieve the attempted contribution of the research in helping RTOs to efficiently manage electricity prices in the future, particularly in times of heightened market dynamics that are the most challenging to address in practice.

Furthermore, as pointed out earlier, the major source of electricity prices' volatility that can be observed in Figure 1, is introduced by positive upswings followed by quick drops to pre-shock levels, in opposite to down-then-up changes. However, the downtrend recovery is noticeably more lasting than the initial triggering price surge as can be concluded from the prevailing positive returns among extreme observations recorded. As Figure 2 indicates, the distribution of the returns seems positively skewed as judged by greater number and variability among positive spikes in the data. As Table 1 displays below, the returns (scaled by a factor of 10) seem to be centred around 0, with a mean of 0.35. As for the 90% observations closest to 0 (in absolute value), the data seems relatively symmetric and well-balanced, with 1140 positive and 1403 positive records. However, in the top 10%, roughly 76% of the returns are positive (215 out of 283) and the positive returns were found to reach more extreme values than their negative counterparts (mean of positive returns 7.41 vs -5.02 in the top 10%). This implies that the electricity price is more sensitive in the upward direction and suggests to use of skewed distribution that could accommodate this feature. This direction is further investigated in section 3.

Finally, another relevant feature in the data is the discrepancy in the volatility clustering between the training and testing periods. As observed in Figure 2, the returns before 2014 seem to all fall within a similar range of magnitudes, with any shocks being occasional and short-lived. The only exception might be mid-2012, where a slightly larger concentration of relatively extreme returns can be found. This can be considered a heightened variability moment; however, besides that, the returns display one-time 'incidental' outliers that do not necessarily imply a fundamental change in the underlying conditional volatility, and the models should be robust to this. This

	Top 90% obs closest to 0	Top 10% obs least close to 0	All obs
Mean	0.10	4.43	0.35
Num positive (**mean**)	1140 (**1.38**)	215 (**7.41**)	1355 (**2.34**)
Num negative (**mean**)	1403 (**-1.31**)	68 (**-5.02**)	1471 (**-1.48**)

Table 1: Skewness in the returns scaled by a factor of 10 data. Mean of returns in the 90% of observations closest to 0 and furthest away from 0. Number and mean of positive and negative observations in each sample.

motivates using a fatter-tailed distribution that would be less influenced by large observations in updating its score-driven component, which is further explored in section 3.

On the other hand, the beginning of 2014 seems to display more persistent high volatilities. Hence, if too little significance is given to the innovation term in the model, this can hinder its performance. Beyond that, again, one-time incidental outliers can be observed—notably the one exceeding the y-axis limit in 2015. These alternate with similar volatility periods across consecutive observations. Thus, the above may be challenging for the models as an appropriate balance between innovation and lagged behavior should be reached. Combined with the different behavior of returns in training and forecasting samples in terms of magnitude and persistence of volatilities, the above can offer an interesting test for the robustness and flexibility of GAS models, which is another goal of this research.

2.2 Weather variables

The second dataset used in this research comprises past weather data made available by the National Oceanic and Atmospheric Administration (NOAA). The data can be freely accessed through the organization's website (and Climate Data Online). It is used to extract variables that serve as exogenous covariates in one of the models discussed in subsection 3.3.

Since the electricity prices were obtained from the PJM market operating in 13 states and the District of Columbia, the meteorology station where weather data were collected, was chosen to be Philadelphia International Airpot PA US (USW00013739). This region is geographically and population-wise representative of the whole area that PJM is accountable for. Additionally, some stations were collecting limited weather information and those located near airports tended to have the most complete datasets.

The retrieved weather records are the daily summaries from April 6, 2008, to December 31, 2015. The downloaded dataset contained 24 different variables out of which two were used in this analysis: average daily temperature and aggregate daily precipitation. The two data types have been processed to reflect daily nominal changes (since daily percentage changes in electricity prices are the subject of this research) and temperatures were converted from Fahrenheit to Celsius units.

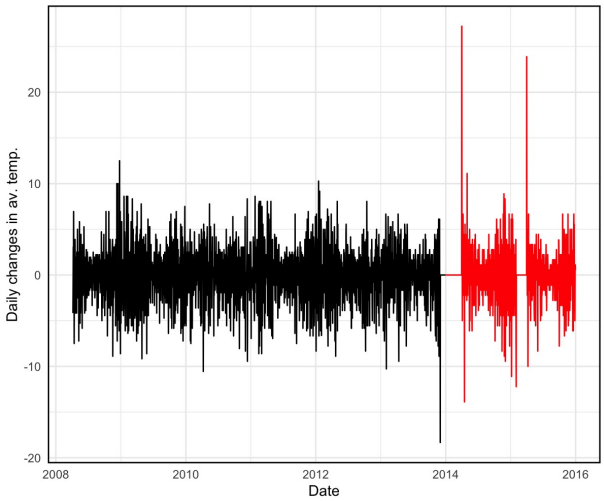

Figure 3: Daily changes in average temperatures from 2008-04-06 to 2015-12-31.

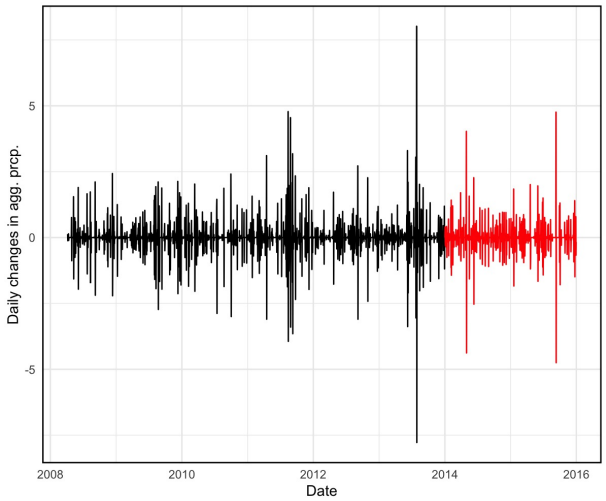

Figure 4: Daily changes in aggregate precipitation. The same periods as Figure 3.

The Figures above display the two variables that will be used as weather covariates, plotted over the duration of training (black) and testing sample (red). As can be seen in Figure 4, zero-valued observations alternate with non-zero daily records which can be interpreted as consecutive days with no rainfall. In general, the black section (2008-2013) shows smaller and more consistent fluctuations in daily precipitation changes, with most values falling between -2 and 2 units. However, the testing sample (2014-2015) exhibits increased variability, with more frequent spikes beyond ± 2 units and seemingly fewer rainless days. Interestingly, a similarly higher-volatility period was observed in Figure 2 concerning electricity returns. Nevertheless, since parameters are estimated solely based on the training sample it is important that the relationship between variables has been consistent also prior to 2014.

On the other hand, Figure 3 shows that the training interval features a relatively stable pattern of daily temperature changes, with fluctuations mostly ranging between -10 and 10 Celsius degrees. To some extent, similar behaviour can be observed post-2014 however the plot there is visibly interrupted by two periods with consecutive days of zero changes in daily temperatures. Additionally, the first observations after temperature stagnations end seem unusually large as well as the last observation preceding the first interval. This is likely indicative of missing observations or data recording issues which can be particularly damaging to the in-sample significance of the temperature regressors, or perhaps more crucially to the accuracy of the model predictions since the majority of the periods seem to belong to the testing sample. This might pose a challenge to predictive accuracy of the log-scale score-driven models discussed in this paper. Hence, it can be interesting to observe to what extent the missing observations lead to biased or inaccurate estimates.

Also, it is worth pointing out that as could be expected, since the original data has been first-differenced, the plotted covariates are symmetric around 0. As discussed in subsection 2.1, the distribution of the returns was positively skewed with the prevailing amount of large positive observations. This suggests that the used covariates don't explain asymmetry among the electricity returns.

3 Methodology

Having discussed some of the features of the data in subsection 2, we proceed to the formulation of the models. All models follow the GAS approach with the scale score-driven specification as explained [Creal et al., 2013] and can be generalized according to Equation 1. The purpose of the models is to filter the conditional volatilities on the daily returns of average electric prices. In particular, the predictive conditional density for observation y_t from Equation 1 is updated to:

$$y_t = \mu + \exp(\frac{1}{2} f_t)\epsilon_t \tag{2}$$

This specification differs from a standard univariate scale model described by [Artemova et al., 2022b]. First, an intercept in the equation modelling the returns hasn't been restricted to 0. This is because the returns data used in this analysis has not been demeaned and as Table 1 implies, the mean of the returns in the investigated dataset was non-zero (approximately 0.35). Hence, the intercept in the model provides flexibility to fit the sample data to a greater extent and allows to reasonably assume zero-mean distribution of shocks ϵ_t. Additionally, our knowledge of electricity, and broader commodity markets, suggests that the prices, and hence the returns, can exhibit persistent trends over time as explored by [Zhang et al., 2022]. While the electricity prices in the time interval discussed in this paper, don't seem to exhibit any trend, if data from another time, region or even commodity market was concerned, the prices could trend upward reflecting a combination of economic, environmental, and regulatory factors. Hence including an intercept μ can make the models more generalizable and robust to the volatility analysis of the returns over different data.

Secondly, as Equation 2 indicates, a scale model of log specification is used to filter the conditional volatility which relaxes the parameter restrictions ensuring that the process is strictly positive as compared to specification $f_t = \sigma_t^2$ in $y_t = \mu + \sigma_t \epsilon_t$ mentioned by [Artemova et al., 2022a]. Hence if ϵ_t are i.i.d. and unit variance, then $\text{Var}(y_t \mid \mathcal{F}_{t-1}) = \exp(f_t)$ such that $f_t = \log(\sigma_t^2)$ is modelled by the score-driven equation.

Finally, for the sake of replication of the results of [Artemova et al., 2022b], the scaling factor S_t is chosen to be the inverse of the conditional (Fisher) information matrix $\mathscr{J}_{t|t-1} = Var_{t-1}[\nabla_t]$. Therefore, while tomorrow's scale f_{t+1} is updated by using the score of today's f_t, its impact on f_{t+1} will be partially standardized according to the variability of the score given specific ϵ_t distribution. If for some models $\mathscr{J}_{t|t-1}$ will be parameter-invariant, its value will be replaced by a convenient constant since its choice will be irrelevant due to parameter α.

Therefore, the univariate log-scale score-driven formulation shared by all models is presented in Equation 3.

$$\begin{aligned} &y_t = \mu + \exp(\frac{1}{2} f_t)\epsilon_t, \quad f_{t+1} = \omega + \alpha(S_t \cdot \nabla_t) + \beta f_t, \\ &S_t = \mathscr{J}_{t|t-1}^{-1}, \quad \nabla_t = \frac{\partial \log p_y(y_t \mid f_t, \mathcal{F}_{t-1}; \theta)}{\partial f_t} \end{aligned} \tag{3}$$

All models will be first estimated using Maximum Likelihood Estimation (MLE) on the training sample of electricity spot returns data discussed in Section 2.1. Akaike (AIC) and Bayesian (BIC) Information Criteria will be used to comparatively assess the in-sample performance of the models. Since the models differ in the number of parameters, the advantage of these measures lies in their ability to balance between the goodness of fit and the complexity of the evaluated models.

3.1 Evaluating model forecasts

As can be seen in subsections 3.2 and 3.3, one of the points of differentiation between the models is the distribution that ϵ_t are assumed to i.i.d. follow. As a consequence, parameters in different model specifications are estimated under MLE to fit different objective functions due to different inherent PDFs. Hence, besides in-sample likelihood fit, another metric is needed such that the models are more easily comparable. Additionally, the practical value for stakeholders such as RTOs lies in the ability to accurately predict volatility in the future. Therefore, the models will be evaluated based on the accuracy of their one-day ahead forecasts. As a proxy for target volatility, daily realized variance is used, computed as a squared demeaned return as can be seen below:

$$\mathrm{TV}_t = (r_t - \hat{r})^2 \tag{4}$$

where $\mathrm{TV}_{t,d}$ denotes the target variance at time t, r_t realized return and $\hat{r}$ mean return over the training and testing sample.

This research uses a fixed-window estimation technique because its focus lies in the GAS nature of the models. If sliding or rolling windows were used, the estimates of parameters like μ, ω, α, or β could be time-varying due to the evolving training sample, however, it would not directly affect the score component ∇_t (or S_t) itself which depends on errors distribution spec. Hence, if all models are evenly treated with a fixed-window specification, and the research aims to compare forecasts among different score-driven models, using a window of another type falls beyond the scope of this research.

Moreover, the data from 2008 appears to be equally informative for explaining the volatilities of returns in 2013 as in 2015 as there were no significant structural or organizational changes in the wholesale electricity market that PJM operates. Nevertheless, even if the two datasets were fundamentally different, one of the goals of this research is to examine the robustness of score-driven models by going from a low-, to high-volatility period. Hence, any attempt to incorporate a more volatile sample into parameter estimation could mitigate the performance gaps between models resulting from testing and training observation discrepancies.

Further advantages, of the fixed-window approach, include lower computation time, which could be particularly beneficial for lower-frequency data and more complex models. Additionally, constant estimation sample means time-in varying parameter values hence more consistent and straightforward interpretation, whereas supportive papers can be found in the literature such as [Gómez-González and Cárdenas-Montes, 2021], which discusses the benefits of fixed window sizes for consistent performance for Gaussian Processes in Large Time Series.

Having the target variable, two loss functions are used to evaluate the predictive performance of the models.

First, Mean Squared Error (MSE) is computed, which captures the average squared difference between the estimated and true values. MSE is a standard and widely accepted performance metric in regression tasks, hence it facilitates the comparison of score-driven results from this research against alternative studies. Moreover, MSE penalizes larger errors more significantly than smaller ones due to the squaring term. For the sake of practical application, its sensitivity to large errors could be desired, since in electricity production unanticipated extreme observations can cause disproportionally more damage. Finally, MSE is easily interpretable and when assuming

a Gaussian noise model, its minimization corresponds to the method of MLE.

Below the formula to derive MSE is shown, where n corresponds to the number of observations in the testing sample (730 as stated in Section 2.1).

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5}$$

While sensitivity to outliers of MSE can be beneficial in some applications, in others it might be undesired. A few large errors can disproportionately affect the loss value, potentially undermining promising model performance on low-volatility observations. Additionally, MSE assumes homoscedasticity of errors such that all observations are given equal significance, however by definition of log-scale models, they are constructed to model volatility of heteroskedastic observations as is the electricity dataset discussed in Section 2.1. Therefore, the QLIKE loss function is introduced to address some of these issues, as computed below:

$$\text{QLIKE} = \frac{1}{n}\sum_{i=1}^{n}\left(\log \hat{y}_i + \frac{y_i}{\hat{y}_i}\right) \tag{6}$$

As can be seen in Figure 5, while $\log \hat{y}_i$ is monotonically increasing in $\hat{y}_i$ the ratio term $y_i/\hat{y}_i$ is monotonically decreasing. This holds for the positive side of the graph but since both true daily realized variance and forecasted volatility (by the construction of a log-scale model) are guaranteed to be positive, the negative part is irrelevant to the analysis. The sum of them produces a convex function with a minimum in y_i - the target value that $\hat{y}_i$ was forecasting. Hence, compared to MSE, the QLIKE is much more robust to outliers in the right tail (above the value of the target variable) but on the other hand much more sensitive to (significant) underpredictions of the target value. These can be observed in Figure 5 by the sharp drop visible in the function when approaching y_i from the left, and its flattening when going away from y_i on the right. More generally, QLIKE is not symmetric (opposite to MSE) and it penalizes negative loss more server than positive. This positive bias is taken into account into account in the analysis of the results in Section 4.

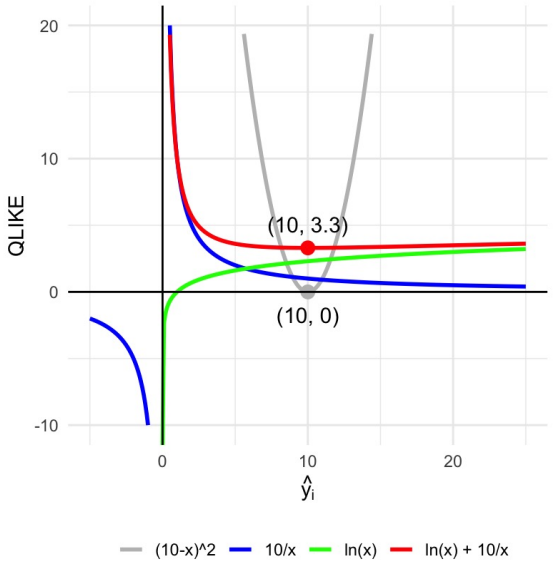

Figure 5: Example of QLIKE (red) and MSE (grey) loss functions for $n = 1$ and $\hat{y}_i = 10$. In red and green the ratio $y_i/\hat{y}_i$ and the logarithmic componented $\log \hat{y}_i$ of QLIKE.

Finally, as shown by [Patton, 2011], amongst nine loss functions commonly used to compare volatility forecasts, only MSE and QLIKE were robust to noise in the volatility proxy. Hence, using a proxy for volatility gives the same model ranking as using the true (unobservable) volatility of an asset. This property can be particularly useful to reliably evaluate the models. Since the two functions differ in the robustness to outliers that are indeed featured in the testing sample, using both will provide a more complete picture.

3.2 Benchmark models

This subsection describes the models used to replicate the results of the electricity spot prices illustrated by [Artemova et al., 2022b]. These include log-scale non-zero intercept models with two different specifications of the innovations' distribution: standard Gaussian $\mathcal{N}(0,1)$ and Student's t $t(\nu)$.

Because of the characteristics of the above distributions, the two models significantly differ in their robustness to outliers and sensitivity of scale parameter f_t. The model with normally distributed errors is the least robust due to its light tails. Consequently, the majority of its density is concentrated in a relatively closer interval than the mean, where the PDF peaks the most. Going away from 0, its PDF will change the steepest which leads to greater gradients as compared to a fatter-tailed distribution. Therefore, the score ∇_t let alone updates the f_t most aggressively since small movement in the scale parameter leads to relatively bigger changes in the gradient. High variability in Gaussian PDF is partially balanced by the score scaling S_t which as explained in Section 1 is the inverse of the conditional information. Coefficient α is also estimated to optimize the score-driven effect on filtering volatility.

The final formulation for $\epsilon_t \sim \mathcal{N}(0,1)$ is presented in Equation 7.

$$\begin{aligned} &y_t = \mu + \exp(\frac{1}{2} f_t)\epsilon_t, \quad f_{t+1} = \omega + \alpha(S_t \cdot \nabla_t) + \beta f_t, \\ &S_t = \mathscr{J}_{t|t-1}^{-1} = \left(\frac{1}{2}\right)^{-1}, \quad \nabla_t = \frac{1}{2}\left(\frac{(y_t-\mu)^2}{\exp f_t} - 1\right) \end{aligned} \tag{7}$$

Since the standard normal is unit variance, the scale parameter directly models the log of conditional variance of the returns i.e. $f_t = \log(\sigma_t^2)$. After evaluating the scaled score $s_t = S_t \cdot \nabla_t$ it can be observed that the left-over expression is the square of standardized return minus 1.

The second model used to replicate the results of [Artemova et al., 2022b] assumes Student's t errors. Importantly, the regular Student's t PDF was used to obtain the results and it hasn't been standardized to unit variance. Hence, if $\epsilon_t \sim t(\nu)$ then $\mathrm{Var}(\epsilon_t) = \frac{v}{v-2}$ for $v > 2$. Thus, $\mathrm{Var}(y_t \mid \mathcal{F}_{t-1}) = \exp(f_t) \cdot \frac{v}{v-2}$ and $f_t = \log(\frac{v-2}{v} \cdot \sigma_t^2)$, such that f_t describes the log of conditional variance incremented by addition log term dependent on estimated v (Therefore is referred to as scale, and not directly volatility).

The final formulation for $\epsilon_t \sim t(\nu)$ is presented in Equation 8.

$$\begin{aligned} &y_t = \mu + \exp(\frac{1}{2} f_t)\epsilon_t, \quad f_{t+1} = \omega + \alpha(S_t \cdot \nabla_t) + \beta f_t, \\ &S_t = \mathscr{J}_{t|t-1}^{-1} = \left(\frac{v}{2(3+v)}\right)^{-1}, \quad \nabla_t = \frac{1}{2}\left(\frac{(v+1)v^{-1}(y_t-\mu)^2\exp(-f_t)}{1+v^{-1}(y_t-\mu)^2\exp(-f_t)} - 1\right) \end{aligned} \tag{8}$$

It should be pointed out that since Student's t is fatter-tailed than standard normal, its sens-

itivity to outliers should be relatively smaller because its PDF is flatter. Therefore, the f_t should be adjusted less aggressively to volatility spikes since the score values ∇_t are expected to be in general lower. Moreover, it can be expected that also MLE parameter estimates can afford to leave out more and larger unexplained outliers since their cost to the overall log-likelihood function is relatively smaller.

Additionally, a model with Student's t innovations, while being more robust is also more general than Gaussian since as $\nu \to \infty$, $t(\nu) \to \mathcal{N}(0, 1)$. Therefore, if a thinner-tailed distribution produced a better fit for the training sample the former can appropriately increase the estimate of degrees of freedom.

The log-likelihood expressions for the above models with Gaussian and Student's t distributions that were maximized with respect to the parameters during the MLE procedure are derived in Appendix sections 5.1.1 and 5.1.2.

3.3 Extension models

The two models discussed in subsection 3.2 used arguably the most common distributions of innovations. This section discusses three models extending on the log-scale score-driven specification presented so far in this paper.

So far, first ϵ_t were assumed to i.i.d. follow standard Normal distribution. Since the electricity data contains several extreme observations, a less sensitive filtering approach with Student's t errors. This added robustness to the model but unchangeably the coefficients remained being estimated such that the resulting errors fit a single-parameter, bell-shaped distribution.

Hence, the first model discussed in this section goes one step further and lets ϵ_t have a symmetric Gen-t distribution of [McDonald and Newey, 1988] which introduces two parameters that control the kurtosis. Popular distributions such as General Errors Distribution (GED), Laplace, or Student's t are special cases of Gen-t. The distribution has been used in a variety of modelling settings such as famous [Bollerslev et al., 1994] that introduced it for volatility modelling with a family of ARCH models, or notably [Harvey and Lange, 2017] who derived its information matrix and considered a score-driven model applied to stock return data.

A log-scale model based on Generliazed-t provides a more flexible alternative to the models proposed in subsection 3.2 because it generalizes several other distributions thereby reducing the risk of too restrictive model parametrization that can lead to misspecification. It unifies all types of tail decay and allows extra flexibility in the kurtosis of for example t-distribution. Additionally, it allows to capture a variety of shapes in the tails as well as the peak of the distribution. Hence, the new model is capable of extracting and accommodating more nuances from the training data by having more freedom in tailoring the shape of errors' distribution to maximize the likelihood. For the same reason, it is more likely to be suitable with more variety of returns data from different regions, time intervals or even markets (beyond electricity considered in this research) and hence can be of greater relevance across more applications. Additionally, the model based on Generazlied-t can be more robust than its alternatives discussed earlier due to more flexibility in shape and tails' heaviness of the distribution of errors as argued by [McDonald and Newey, 1988].

The distribution assumed for ϵ_t in the first specification presented in this section is the standardized Gen-t distribution, symmetric around zero with a unit scale (to facilitate scale parameter interpretation). The PDF used to derive log-likelihood and the scaled score is presented below,

where $B(...)$ denotes the best function:

$$f(\epsilon_t) = K(\nu, h)\left(1 + \frac{|\epsilon_t|^h}{\nu}\right)^{\frac{-(\nu+1)}{h}}, \text{where} \quad K(\nu, h) = \frac{h}{2\nu^{\frac{1}{h}}} \frac{1}{B(\frac{1}{h}, \frac{\nu}{h})} \tag{9}$$

The parameters ν and h should be positive, but in practice, $\nu < 1$ is unlikely as pointed out by [Harvey and Lange, 2017]. Depending on the estimates of the parameters, the distribution can be interpreted in a number of ways. For example, when $h = 2$, the Gen-t distribution reduces to Student's t-distribution with ν degrees of freedom. When $\nu \to \infty$, a GED(h) is obtained, where GED(2) is equivalent to the Gaussian distribution and GED(1) to the Laplace distribution.

If $\epsilon_t \sim$ Gen-t(ν, h), then $S_t = \mathscr{J}_{t|t-1}^{-1} = \frac{\nu h}{4(1+h+\nu)}$. However, since the scale updating equation includes α coefficient in front of the scaled score $S_t \dot{\nabla}_t$, no matter what ν and h are estimated to be, α can control the scaling. Hence without loss of generality, $S_t = 2$ such that the formulation simplifies. Then the specification of the first model discussed in the subsection based on standardized symmetric Gen-t distribution of errors is as presented in the Equation below:

$$\begin{aligned} &y_t = \mu + \exp(\frac{1}{2} f_t)\epsilon_t, \quad f_{t+1} = \omega + \alpha(S_t \cdot \nabla_t) + \beta f_t, \\ &S_t = 2, \quad \nabla_t = \frac{1}{2}\left((\nu+1)\frac{(|y_t - \mu| \exp(-\frac{1}{2} f_t))^h/\nu}{1 + (|y_t - \mu| \exp(-\frac{1}{2} f_t))^h/\nu} - 1\right) \end{aligned} \tag{10}$$

Since the distribution has been a standardized scale parameter, f_t directly filters log conditional volatility of returns $\log(\sigma^2)$.

The second model discussed in this subsection extends upon the idea of Gen-t distribution to maintain the advantage of a robust and flexible distribution of errors. However, additionally, it modifies the scale-updating expression for f_{t+1} from Equation 10 to include exogenous covariates besides the score-driven and autoregressive terms. Hence, the log-volatility $f_t = log(\sigma_t^2)$ now evolves not only based on past information but also incorporates the effect of the additional regressors that are judged as potentially impactful using domain knowledge. The covariates are added to the f_{t+1} explaining equation and not for example y_t directly because the goal of the research is to evaluate the forecasting performance of conditional volatility using scale score-driven models, log of which f_{t+1} represents. The regressors were chosen such that they improve explaining the volatility of returns.

The discussion of optimal covariates that would be the most significant in explaining the volatility of electricity returns is beyond the scope of this research. The potential candidates include fuel prices, demand-side factors, regulatory and policy variables, etc. Nevertheless, the intuitive choice of regressors that will be used in this research includes weather variables - specifically temperature and precipitation as discussed in Section 2.2, since the two are argued to have a direct impact on electricity consumption patterns and production such as for heating and cooling purposes or hydroelectric power generation. Additionally, papers such as [Sgarlato and Ziel, 2022] showed that meteorological forecasts can be used directly to improve price forecasts multiple days in advance. Since, this research concerns the volatility of daily average returns, the two covariates extending the log-scale model with Gen-t innovations from Equation 10 are: daily average change in temperature ($temp_{t+1}$) and daily average change in precipitation ($prec_{t+1}$). The choice, suitability and the number of regressors used are left for further research. The final formulation of the second

model discussed in this subsection is presented below:

$$
\begin{aligned}
&y_t = \mu + \exp(\frac{1}{2}f_t)\epsilon_t, \quad f_{t+1} = \omega + \alpha(S_t \cdot \nabla_t) + \beta f_t + \gamma_1 temp_{t+1} + \gamma_2 prec_{t+1}, \\
&S_t = 2, \quad \nabla_t = \frac{1}{2}\left((\nu+1)\frac{(|y_t - \mu|\exp(-\frac{1}{2}f_t))^h/\nu}{1 + (|y_t - \mu|\exp(-\frac{1}{2}f_t))^h/\nu} - 1\right)
\end{aligned}
\tag{11}
$$

As can be seen in Equation 10 only the score updating equation changes, since S_t and ∇_t depend on daily returns y_t modelling equation which as motivated earlier remained unchanged. A similar trick as for the model with Gen-t errors without covariates is used, to replace $\mathscr{I}_{t|t-1}^{-1}$ with a constant 2 that would simplify the scaled score s_t expression.

Furthermore, it should be pointed out that while the research question of this paper concerns volatility forecasting, the model uses regressors that are only known at the time of the forecast, for example, $t+1$. The intention of the paper was to first use a historical one-day ahead forecast of daily temperatures and precipitation however these were hardly accessible if going back more than several years ago as needed for the period 2008-2015 when the electricity prices were collected. Nevertheless, for practical application of the model in contemporary volatility prediction, the average daily temperature and precipitation changes can be indeed computed using one-day ahead forecasts of respective variables since these are commonly available. Moreover, the conclusions concerning model performance based on realized and forecasted observations should be almost identical since the modern one-day ahead (short-term) weather forecasts tend to be extremely accurate.

The motivation for extending the previously discussed model with additional covariates relates to the assumption concerning the scale updating equation which underlies the class of observation-driven models discussed in this paper. In opposite to parameter-drive models, time- varying parameters are treated as functions of lagged dependent variables as well as exogenous variables. In such model specifications, when conditioning on past and concurrent observations, the time-varying parameters are perfectly predictable as explained by [Artemova et al., 2022a]. This ensures that the likelihood evaluation is relatively straightforward as compared to dynamic processes that would describe parameters. However, if not all significant variables are included in explaining f_{t+1}, then the model risks misspecification which can result in discrepancies between estimated and true variance of y_t that can't be attributed to f_{t+1} own source of errors (in opposite to parameter-driven models). Hence, this model attempts to emphasize the importance of including all significant information beyond score-driven and autoregressive terms, concerning f_{t+1} particularly under an observation-driven class of models as the log-scale score-driven model is part of. By including the two exogenous covariates, the model is hoped to be (better) specified and can provide more accurate and reliable volatility estimates. Additionally, it can better isolate and quantify the score-driven, autoregressive, and external conditions' effects on the volatility process, providing more accurate and timely forecasts.

Nevertheless, both of the above models are based on a symmetric Gen-t distribution of errors. Therefore, both potentially fail to account for one of the most important features present in the electricity data which was a noticeable positive skewness of the returns as discussed in Section 2.1. The returns were found visibly more volatile if they were tied to a price increase rather than a price decrease. This could have been seen both in greater frequency and magnitude of

positive spikes in Figure 2, prevailing over returns' setbacks. Therefore, the third model proposed in this research aimed to use an asymmetric distribution of ϵ_t. To simultaneously maintain the flexibility and generalizability of the model that can adjust errors' distribution to capture a variety of shapes at its peak and tails an Exponential Generalized Beta Distribution (EGB2) was chosen. An application of the distribution to observation-driven modelling has been already explored by for example [Caivano and Harvey, 2014] or [Ping et al., 2019] and this model attempts to extend upon them using log-scale specification for electricity volatility filtering.

If $y_t = \mu + \sigma_t \epsilon_t$ where $\sigma_t = \exp(\frac{1}{2} f_t)$, and ϵ_t is assumed to follow a standardized EGB2 distribution with unit variance and 0 mean, then:

$$p_y(y_t \mid f_t, \mathcal{F}_{t-1}; \theta) = \frac{h \exp\left(\xi \left(h \frac{(y_t - \mu)}{\sigma_t} + \Delta\right)\right)}{\sigma_t B(\xi, \zeta) \left(1 + \exp\left(h \frac{(y_t - \mu)}{\sigma_t} + \Delta\right)\right)^{\xi + \zeta}}, \tag{12}$$

where ξ and ζ are nonnegative shape parameters, $B()$ is the beta function (as mentioned earlier), $\Delta = \psi(\xi) - \psi(\zeta)$, $h = \sqrt{\psi'(\xi) + \psi'(\zeta)}$, whereas $\psi()$, and $\psi'()$ are digmma and trigamma functions respectively. The shape parameters ψ and ζ determine the magnitude and direction of asymmetry, namely if $\xi > \zeta$ the distribution is positively skewed, if $\xi = \zeta$ symmetric, and negatively skewed otherwise. Since the sample of the returns discussed in 2.1 seemed to have positive extreme observations, it is anticipated that the model is estimated such that $\xi > \zeta$.

Besides asymmetry, one of the advantages of EGB2 distribution is its flexibility to accommodate various shapes of data distributions by adjusting its parameters. For example, the distribution contains the normal distribution for $\xi = \zeta \to \infty$ and the Laplace distribution when $\xi = \zeta = 0$. Additionally, EGB2 features exponential tails, which decay faster than the heavy tails of distributions like the Student's. On the other hand, the distribution allows for leptokurtosis (positive excess kurtosis over three) and heavier tails than normal distribution hence less influence by one-time spikes or drops caused by unexpected events. This can provide a good balance between robustness to outliers and adaptations to periods of persisting volatility as both were present in the data as discussed in Section 2.1.

The final formulation of the model with standardized EGB2 errors is presented in Equation 13. Since $\mathscr{J}_{t|t-1}^{-1}$ is proportional to 1, the score scaling function can be chosen to an arbitrary constant - for example, $S_t = 2$.

$$\begin{aligned} &y_t = \mu + \exp(\frac{1}{2} f_t)\epsilon_t, \quad f_{t+1} = \omega + \alpha(S_t \cdot \nabla_t) + \beta f_t, \quad S_t = 2, \\ &\nabla_t = \frac{1}{2}\left(\left[(\xi + \varsigma) b_t - \xi\right] \frac{h y_t}{\exp\left(\frac{1}{2} f_t\right)} - 1\right), \quad \text{where} \quad b_t = \frac{\exp\left(\frac{h y_t}{\exp\left(\frac{1}{2} f_t\right)} + \Delta\right)}{1 + \exp\left(\frac{h y_t}{\exp\left(\frac{1}{2} f_t\right)} + \Delta\right)} \end{aligned} \tag{13}$$

It is worth mentioning that the model based on EGB2 distributed errors was attempted to be extended to include exogenous covariates similar to the model based on Gen-t however due to computational infeasibilities this was unsuccessful. Nevertheless, the general effect of including covariates on improving the forecasting performance of log-scale score-driven models can be extrapolated from the results comparison of the models based Gen-t errors with and without additional regressors.

4 Results

This section discusses the estimation results obtained using MLE, including in-sample and out-of-sample performance of the five log-scale score-driven models discussed in Section 3. All code used to run the models was implemented in Matlab R20024a. The code used to estimate the two benchmark models with Gaussian and Student's t distributed innovation of subsection 3.2 as well as the three extensions with Gen-t, Gen-t with covariates, and EGB2 of 3.3, was self-written from the beginning without the use of any external packages. Similarly, the program used to generate, process and plot the forecasts of all models discussed in this research was also self-coded and is attached together with this paper. Additionally, the code to retrieve standard errors around parameter estimates was attempted to be self-written, however, due to technical issues, the results couldn't be obtained. Hence, the GASupgraded package which can be found at www.gasmodel.com, was used to compute the standard errors. However, it was only available for the log-scale specifications with errors Gaussian and Student's distributed, therefore, in subsection 4.1 standard errors are only provided for the benchmark models' estimates.

Importantly, the AIC and BIC values presented in the results below are the average scores per observation in the estimation dataset. Hence, they were obtained by dividing the general scores from the Akaike Information Criterion and Bayesian Information Criterion by 2096 (length of the training sample). This is to make the presentation of the results consistent with [Artemova et al., 2022a] which electricity illustration is replicated.

For the ease of replication of the results obtained in this paper, it should be mentioned that initial values for the parameters that the optimizer starts with can be found in the attached code. Additionally, it should be pointed out that the starting value of the filtered sequence f_1 is unobserved hence, it should be replaced by some arbitrary value (or estimated but typically this is not preferred as noted by [Artemova et al., 2022a]). As shown in section 3, $f_t = \log(\sigma_t^2)$ for all models besides Student's t which was the only distribution in which PDF was not standardized to unit scale. Hence, f_1 was assumed to be $\log(Var(y_{\text{train}}))$ for the four models with unit scale error specification, where $Var(y_{\text{train}})$ denotes the sample variance of all returns in the training sample. For Student's t $f_1 = \frac{\nu_0 - 2}{\nu_0} \log(Var(y_{\text{train}}))$ where ν_0 was the initial choice of degrees of freedom. Despite this, f_1 refers to the scale of y_1 only, the sample variance over all returns in the estimation sample was used, for the sake of the stability of the initialization if the training time interval was modified.

4.1 Estimation results

First, the parameter estimates of the benchmark models are presented in Table 2. The results for the total log-likelihood as well as AIC and BIC criteria (as discussed in section 3) are provided to obtain better information concerning the in-sample fit of the model. The estimated parameters shared between the log-scale models with Gaussian and Student's t errors' specification include μ, ω, α and β. The latter model additionally includes ν which controls the amount of probability mass in the tails of the distribution. The total estimation time was 0.68 and 0.93 seconds for models using Gaussian and Student's t respectively. The standard errors were computed using the empirical Hessian as explained in the GASupgraded readme.

As Table 2 indicates, the Student's t model offers lower scores on both information criteria

	μ	ω	α	β	ν	LL	AIC	BIC
Gaussian	0.316	0.123	0.073	0.928		-4771.4	4.56	4.57
	(0.047)	(0.020)	(0.010)	(0.011)				
Student's t	0.029	0.091	0.073	0.916	4.321	-4615.2	4.41	4.42
	(0.045)	(0.093)	(0.025)	(0.055)	(0.413)			

Table 2: The MLE parameter estimates from the score-driven volatility models with Gaussian and Student's t distribution of errors, for the PJM electricity training data (Standard Errors in Parentheses)

implying a better balance between model fit and complexity. As compared to the Gaussian model, Student's t has an extra parameter to be estimated corresponding to the degrees of freedom. While AIC and BIC penalize a number of parameters, the final criteria scores still end up lower, implying a substantially better fit that the Student's t distribution allows to achieve, compensating for increased complexity. The estimated degrees of freedom were approximately 4.3 which suggests that a fatter-tailed distribution performs superior given the electricity returns. This might be explained by numerous short-lived shocks observable in the electricity prices as noticed in section 2.1. A fatter-tailed distribution of innovation with flatter density will produce lower score values, hence the model will be less sensitive in updating the scale parameter and more robust to the incidental outliers and therefore better fit the noisy data of electricity returns.

Additionally, the Student's t is a generalization of the Gaussian model since $t(\nu)$ for $\xrightarrow{\nu\to\infty}$ becomes $N(0,1)$ and already when for example $\nu = 40$ the distributions are hardly distinguishable. Hence Student's t is more flexible and could be anticipated to fit the data better although the information criteria penalize the model's extra parameter.

Set aside the information criteria, as Table **??** presents, the two models feature very similar dynamics in capturing the volatility. The estimated values of α are almost identical whereas the baseline level of volatility ω coefficient persistence term β is very comparable taking into account the corresponding standard errors.

Interestingly, the β parameter has been estimated relatively high, particularly when compared to α estimates. This suggests a strong persistence in volatility in the electricity returns despite the apparent presence of outliers as shown in section 3 which were hypothesized to be short-lived and incidental. Nevertheless, the innovation coefficient seems significantly different from 0 which suggests the advantage of the scored-driven term on top of the sole autoregressive approach.

Noticeably, the estimated mean return μ for the Gaussian model is positive and significantly higher than Student's t which is much closer to 0. The difference in the estimates could be explained by the positively skewed distribution of the returns. After a sharp jump, the price tends to take longer to normalize hence the positive spikes in returns are more frequent. While both distributions are symmetric, Student's t has heavier tails and hence is more robust to outliers. On the other hand, Gaussian might overestimate the mean as it is influenced by large observations to a greater extent and hence is more sensitive to fitting the empirical distribution as if it were symmetric. These findings suggest that an asymmetric alternative might be more suitable to model the data.

Table 3, now presents the parameter estimates for the extension methods discussed in subsection 3.3.

	μ	ω	α	β	ν h	γ_1 γ_2	ξ ζ	LL	AIC	BIC
Gen-t	0.061	0.079	0.106	0.929	3.293 2.730			-4610.4	4.40	4.42
Gen-t covariates	0.060	0.065	0.061	0.942	3.246 2.775	0.008 0.049		-4610.0	4.41	4.43
EGB2	0.106	1.545	0.116	0.888			7.643 0.002	8144463.3	-3204.0	-3203.9

Table 3: The MLE parameter estimates from the score-driven volatility models with Generalized-t (Gen-t), Gen-t with covariates, and EGB2 distribution of errors, for the PJM electricity training data (2008-04-06 to 2013-12-31)

The estimated parameters shared between all the models that results are presented in Table 3 include μ, ω, α and β. As can be seen, the intercept parameters are estimated to be the greatest for the model with EGB2 specification of errors. This might be because the distribution, in opposite to Gen-t, allows asymmetry of ϵ_t hence the models can more freely choose the intercept parameter and fit the distribution shape parameters appropriately. The model implies a strongly left-tailed distribution of errors as shown by a significantly greater ξ estimate than ζ. The two models based on Gen-t distribution of errors have seen all the parameters estimates very close to each other. Combined with seemingly insignificant results for γ_1 and γ_2, this suggests that the chosen covariates don't improve model performance to a great extent, at least in the training sample. It is worth mentioning that similarly to models based on Gaussian and Studnet's t errors, the autoregressive β is estimated to be much greater than the score-driven α. This suggests significant persistence of volatility in the training sample that can be found also if a more flexible and general distribution of errors is allowed.

Moreover, results of AIC and BIC support the observation that the covariates extending the model with Gen-t distributed errors add little to no improvement to the in-sample fit. At the expense of the extra two parameters the log-likelihood barely increased which resulted in slightly lower scores on both information criteria. Interestingly, the AIC and BIC were found smaller for Gen-t-based extensions as compared to the Gaussian benchmark model which parameter estimates can be found in Table 2. Additionally, the two distributions have different PDFs particularly due to the fat-tails of Gen-t distribution as can be seen by estimates of ν and h. Hence, the conclusions drawn by metrics based on log-likelihood results should be limited. This also explains large differences in log-likelihoods of Gen-t and EGB2-based models which can be hardly interpreted.

4.2 Forecasting results

Having discussed the in-sample performance of the benchmark and extension models, this section presents the forecasting results of the models for 2014-01-01 to 2015-12-31 data. As explained in subsection 3.1, two loss functions: MSE and QLIKE are derived to evaluate the predictive accuracy of the models. The obtained scores for the five models are displayed in Table 4.

As can be seen above, the models based on Student's and Gen-t distributions obtain the lowest MSE scores by a significant margin. Nevertheless, the value of approximately 160000 can best still argued to be high since the majority of the returns fell in the sub-20 region as was shown in

	Gaussian	Stundet's t	Gen-t	Gen-t covariates	EGB2
MSE	173616696.0	162292.2	162338.9	162339.8	715595.3
QLIKE	6.6	11.0	14.5	14.6	6.7

Table 4: The average MSE and QLIKE scores of the five score-driven volatility models with: Gaussia, Studnet's t, Generalized-t (Gen-t), Gen-t with covariates, and EGB2 distribution of errors, for the PJM electricity testing data

Figure 2. On the other hand, the average error made in predictions by the Gaussian-based log scale model is even more extreme. This has been probably driven by large errors that have been magnified even more due to the squaring term. Simultaneously, the conclusions that can be drawn from QLIKE results are completely opposite. In fact, the model based on Gaussian errors seems to perform the best in one-day ahead prediction making, closely followed by EGB2. Good results of the latter might suggest that relaxing the assumption of symmetry in the ϵ_t distribution might be beneficial and should be further researched. Interestingly, the Gen-t-based models that had the lowest MSE scores, perform the worst according to the QLIKE loss function. This points back to the discussion of loss functions from subsection 3.1 that emphasized the positive error bias of QLIKE and high sensitivity to large errors of MSE. This might explain so extreme differences in the models and encourage users to select the optimal models according to their individual preferences. Additionally, to obtain a better understanding of results from Table 4, the process of how QLIKE and MSE loss functions were changing as more observations were added to the testing sample is shown below.

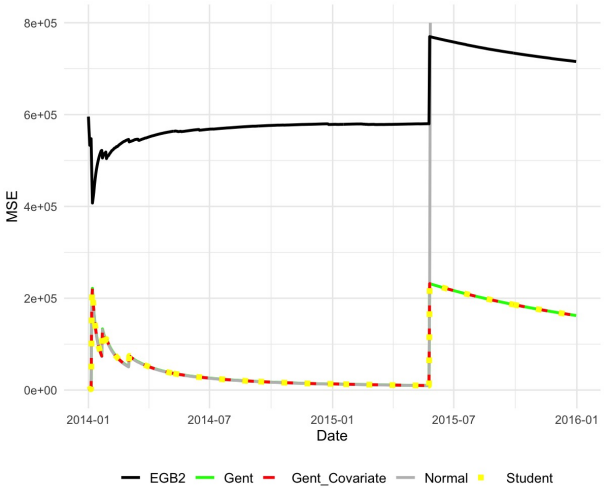

Figure 6: Mean Squared Error as of each date in the testing sample of daily electricity prices in the PJM market (2014-01-01 to 2015-12-31).

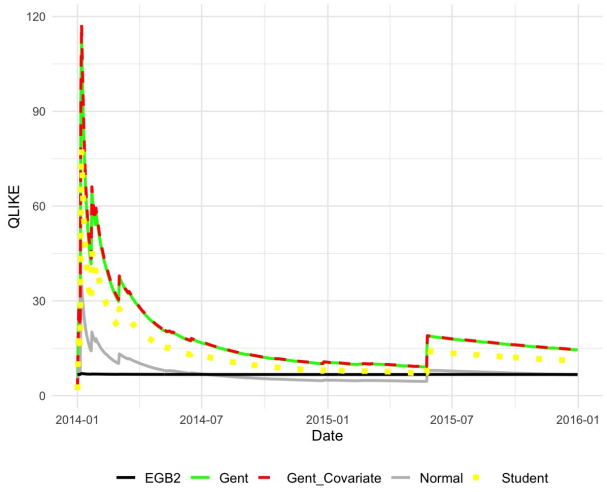

Figure 7: QLIKE as of each date in the testing sample of daily electricity prices in the PJM market (2014-01-01 to 2015-12-31)

As seen in Figure 6, the MSE function followed almost identical shapes for models based on Student's t and Gen-t distributions throughout the whole testing sample. Since the target variable of squared returns was the same, the forecasts made by the three specifications have been very close to each other regardless of which and when volatility was predicted and what was the preceding returns' behaviour. Figure 7 supports this observation since the three models have closely followed

each other also in terms of QLIKE score in all tested observations. Particularly, the quality of forecasts of Gent-t-based models with and without covariates, have been similar, if not the same as implied by identical shapes of the two functions. The above aligns with the implications of Table 3, which suggested the low significance of added covariates due to very similar parameter estimates. This might suggest that the performance of the model with Student's t errors is unlikely to be significantly improved by relaxing the distribution to the more flexible Gen-t, at least for the analyzed datasets.

Interestingly, as indicated by the spike in the curves of Figure 6, all models misforecasted the volatility on 2015-05-27 to a great extent. Particularly, the largest discrepancy between the realized and model-implied volatility has been recorded for the Gaussian specification of errors. This forecasting mistake is likely driven by the outlier in the electricity returns dataset visible in Figure 2 that occurred on that day. It was the largest observation recorded in the sample and was roughly 2.5 times greater than the second-largest return. The fact that a much smaller increase around that date can be seen in Figure 7 suggests the volatility in the successive returns was overestimated due to significant autoregressive parameters found in all model estimates. While this was still costly for MSE, due to the positive bias of QLIKE, the large error due to overestimation was downplayed.

Moreover, as can be deduced from Figure 7, if the testing sample had ended in mid-2015, the log-scale model with Gaussian errors would have the lowest QLIKE score of only 4.9. Nevertheless, excluding EGB2 from this analysis, its advantage in QLIKE-measured accuracy over the other score-driven models can be mostly attributed to better performance in the early stages of the testing sample. As new observations were forecasted, the differences in accuracy functions implied by all the models (besides EGB2) remained relatively steady. Interestingly, this period (early 2014) seemed to exhibit some characteristics of persisting volatility, as hypothesized in subsection 2.1. This should indeed favour a thinner-tailed distribution, which is more sensitive in updating the scale parameter and would faster pick up the volatility movements through the score-driven term. On the other hand, the Gen-t distribution is more flexible, hence if a more sensitive approach was optimal also in the training sample, the distribution's shape parameters would be rearranged to reduce the tail heaviness. This might confirm the idea from subsection 2.1 that the estimation and forecasting datasets exhibit different volatility characteristics in some respects. Noticeably, the model with EGB2 errors, according to Figure 7 could be claimed to be the most robust due to its consistent accuracy and little variable in the function. However, the shape of its MSE accuracy over time is very similar to rest of the score-driven model.

5 Conclusion

The research presented in this thesis explores the application and efficacy of Generalized Autoregressive Score (GAS) models in forecasting the conditional volatilities of daily electricity returns. The models used for this purpose followed the observation-driven, log-scale, score-driven specification with different distributions assumed for the errors and optional exogenous covariates in the scale updating equation. The two benchmark models included Gaussian and Student's t distributions, as presented by [Artemova et al., 2022a]. The three extensions comprised log-scale models with EGB2 distribution of errors, as well as Generalized-t (Gen-t) with and without covariates.

The primary goal of this research was to explore the potential of GAS models for predicting future volatilities. This was achieved by using daily returns data from the PJM market, split into training and testing samples. The results demonstrate that GAS models can indeed forecast one-day-ahead volatilities with some degree of accuracy, and the score-driven specifications contribute to the enhanced model performance, as judged by significant coefficients of the score-driven term across all five models. This suggests that the GAS modeling framework discussed in this research may hold practical value for stakeholders, such as Regional Transmission Organizations (RTOs) and utilities, who require reliable volatility predictions to manage financial risks and operational efficiency during volatile market conditions.

Additionally, the research assessed the robustness of GAS models across different volatility periods, from stable to turbulent times. The models were challenged to forecast high-volatility periods despite being trained on less volatile data. The models showed some degree of robustness, and alternative distributional assumptions could be investigated to further explore GAS models in their adaptability and reliability for forecasting in dynamic and unpredictable market environments.

The forecasting process used a fixed-window approach to ensure consistency in the estimates of model parameters. Nevertheless, different approaches, such as rolling or sliding windows, are encouraged to be examined in further research. Incorporating realized returns into parameter estimation can extend the data and update the estimates as new information becomes available. Additionally, the significance of observations in the estimation process could be time-weighted, so that the most recent observations influence the model results to a larger extent. This could allow for modeling more dynamic data with more volatile relationships between variables. Furthermore, the forecasting horizon could be modified such that volatility can be predicted at a frequency lower than one-day ahead or, conversely, more long-term if actions taken to address volatility changes require more time. The optimal approach should be selected corresponding to the problem's objectives and characteristics.

Similarly, alternative target variables as proxies for true volatility, beyond the square of demeaned daily returns used in this research, could be investigated. For example, daily returns were constructed by averaging hourly prices over each day. Perhaps target variance computed from intra-day hourly returns or even higher-frequency observations might be considered as well.

Furthermore, the research aimed to compare different score-driven models to identify the most effective approach. The forecasting results provide an ambiguous answer to this question since the two loss functions considered in this research suggest quite opposite model rankings. For example, the Gaussian-based log scale model had the lowest QLIKE score but the worst accuracy according to MSE. The exact opposite holds for specifications with Gen-t shocks. However, an interesting alternative could be an asymmetric distribution of errors, as shown by the EGB2-based model, which displayed the most stable QLIKE results, indicating some degree of forecasting robustness. Additionally, the specification with exogenous covariates in the scale equation should be further researched. Although it didn't improve the performance of the standard Gen-t-based model, the selected regressors described only weather and lacked domain knowledge of electricity markets. Other ideas might include variables describing fuel prices, regulatory changes, or, for example, wind if an economy is rich in wind-sourced energy. Different numbers and choices of covariates could offer more promising results.

References

[Artemova et al., 2022a] Artemova, M., Blasques, F., van Brummelen, J., and Koopman, S. J. (2022a). Score-driven models: Methodology and theory. In *Oxford Research Encyclopedia of Economics and Finance*.

[Artemova et al., 2022b] Artemova, M., Blasques, F., van Brummelen, J., and Koopman, S. J. (2022b). Score-driven models: Methods and applications. In *Oxford Research Encyclopedia of Economics and Finance*.

[Bollerslev et al., 1994] Bollerslev, T., Engle, R. F., and Nelson, D. B. (1994). Arch models. *Handbook of econometrics*, 4:2959–3038.

[Caivano and Harvey, 2014] Caivano, M. and Harvey, A. (2014). Time-series models with an egb2 conditional distribution. *Journal of Time Series Analysis*, 35(6):558–571.

[Creal et al., 2011] Creal, D., Koopman, S. J., and Lucas, A. (2011). A dynamic multivariate heavy-tailed model for time-varying volatilities and correlations. *Journal of Business & Economic Statistics*, 29(4):552–563.

[Creal et al., 2013] Creal, D., Koopman, S. J., and Lucas, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, 28(5):777–795.

[Engle and Bollerslev, 1986] Engle, R. F. and Bollerslev, T. (1986). Modelling the persistence of conditional variances. *Econometric reviews*, 5(1):1–50.

[Engle and Gallo, 2006] Engle, R. F. and Gallo, G. M. (2006). A multiple indicators model for volatility using intra-daily data. *Journal of econometrics*, 131(1-2):3–27.

[Gómez-González and Cárdenas-Montes, 2021] Gómez-González, J. L. and Cárdenas-Montes, M. (2021). Window size optimization for gaussian processes in large time series forecasting. In *Hybrid Artificial Intelligent Systems: 16th International Conference, HAIS 2021, Bilbao, Spain, September 22–24, 2021, Proceedings 16*, pages 137–148. Springer.

[Harvey and Lange, 2017] Harvey, A. and Lange, R.-J. (2017). Volatility modeling with a generalized t distribution. *Journal of Time Series Analysis*, 38(2):175–190.

[Harvey, 2013] Harvey, A. C. (2013). *Dynamic models for volatility and heavy tails: with applications to financial and economic time series*, volume 52. Cambridge University Press.

[Joseph, 2022] Joseph, V. R. (2022). Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 15(4):531–538.

[McDonald and Newey, 1988] McDonald, J. B. and Newey, W. K. (1988). Partially adaptive estimation of regression models via the generalized t distribution. *Econometric theory*, 4(3):428–457.

[Patton, 2011] Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1):246–256.

[Ping et al., 2019] Ping, Y., Jie, W., Ai-jun, Y., and Xiao-xing, L. (2019). Gas-egarch model with egb2 distribution and var forecasting. *Operations Research and Management Science*, 28(11):125.

[Sgarlato and Ziel, 2022] Sgarlato, R. and Ziel, F. (2022). The role of weather predictions in electricity price forecasting beyond the day-ahead horizon. *IEEE Transactions on Power Systems*, 38(3):2500–2511.

[Zhang et al., 2022] Zhang, Q., Hu, Y., Jiao, J., and Wang, S. (2022). Exploring the trend of commodity prices: a review and bibliometric analysis. *Sustainability*, 14(15):9536.

Appendix

5.1 Log likelihood Log scale models

Below are the log-likelihood expressions and their derivations for the log scale models with Standard Normal and Student's t distribution of errors. The functions are maximized in the parameters to obtain Maximum Likelihood Estimates (MLE) of the parameters.

$$\begin{aligned} &y_t = \mu + \exp\left(\frac{1}{2}f_t\right)\epsilon_t, \quad f_t = \omega + \alpha s_{t-1} + \beta f_{t-1} \\ &\text{Let} \quad h_t = \exp\left(\frac{1}{2}f_t\right) \quad \text{such that} \quad y_t = \mu + h_t\epsilon_t \end{aligned} \tag{14}$$

5.1.1 Log likelihood for Log scale with Gaussian distribution of errors

Let $p()$ correspond to the probability density function (PDF) and $\theta = \{\mu, \omega, \alpha, \beta\}$ be the set of estimated parameters. Since $\epsilon_t \overset{\text{iid}}{\sim} N(0,1)$, then $y_t \sim N(\mu, h_t^2)$.

$$\begin{aligned} &p(y_t \mid \mu, h_t) = p(y_t \mid \mu, \omega, \alpha, \beta) = \frac{1}{h_t\sqrt{2\pi}} \exp\left(-\frac{(y_t-\mu)^2}{2h_t^2}\right) \\ &l(\Theta) = \log L(\Theta) = \log\left(\prod_{t=1}^{T} p(y_t \mid \theta)\right) = \sum_{t=1}^{T} \log p(y_t \mid \theta) \\ &= \sum_{t=1}^{T} -\log(h_t) - \frac{1}{2}\log(2\pi) - \frac{1}{2}\epsilon_t^2 = -\frac{1}{2}\left(\sum_{t=1}^{T} f_t + \log(2\pi) + \epsilon_t^2\right) \end{aligned} \tag{15}$$

5.1.2 Log likelihood for Log scale with Student's t distribution of errors

Let $\theta = \{\mu, \omega, \alpha, \beta, \nu\}$ be the set of estimated parameters. Now the shocks are Student's t distributed with degrees of freedom ν. Hence $\epsilon_t \overset{\text{iid}}{\sim} t(\nu)$. Additionally, the PDF is formulated such that the variance is standardized to 1, $V[\epsilon_t] = 1$ variance which introduces minor changes compared to the standard PDF of Student's t. Let $\Gamma()$ be the Gamma function.

$$\begin{aligned} &p(\epsilon_t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{(\nu-2)\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{\epsilon_t^2}{\nu-2}\right)^{-\frac{\nu+1}{2}}, \quad p(y_t \mid \theta) = \frac{1}{h_t}p(\epsilon_t) \\ &l(\Theta) = \log L(\Theta) = \log\left(\prod_{t=1}^{T} p(y_t \mid \theta)\right) = \sum_{t=1}^{T} \log p(y_t \mid \theta) \\ &= \sum_{t=1}^{T} -\log(h_t) + \log\left(\Gamma\left(\frac{\nu+1}{2}\right)\right) - \frac{1}{2}\log\left(\pi(\nu-2)\right) - \log\left(\Gamma\left(\frac{\nu}{2}\right)\right) - \frac{v+1}{2}\log\left(1 + \frac{\epsilon_t^2}{v-2}\right) \end{aligned} \tag{16}$$