The Price of Uncertainty: Analyzing the Impact of Weather Forecast Inaccuracy on Energy Grid Imbalance

Korin Aldam-Tajima, Liam Graham, Tyler Headley, Julian Piedrahita

5C Datathon 2025 Report

1 Executive Summary

Estonia is undergoing a rapid energy transition. Once heavily dependent on oil shale, the country has set ambitious climate goals—100% renewable electricity by 2030 and carbon neutrality by 2050. A growing number of Estonian households and businesses have installed solar panels, becoming "prosumers" who both consume and generate electricity. While this shift toward decentralized renewable energy is a crucial step in climate action, it also introduces new challenges for maintaining a stable and efficient electricity grid.

One of the biggest challenges is managing energy imbalances—the difference between the energy that is forecasted and the energy that is actually used or produced. These mismatches can strain the grid, increase costs, and create instability. Our analysis investigates a central question: Do inaccuracies in weather forecasts, especially for solar-related variables, contribute to these energy imbalances—and could better forecasting improve grid reliability?

We examined hourly electricity production and consumption data across Estonia, enriched with weather observations and forecasts. Our results show that Estonian counties with higher solar panel adoption tend to experience larger energy imbalances, particularly during sunny midday hours when solar overproduction is most common. We also found that inaccuracies in weather forecasts—especially cloud cover and temperature—are strongly associated with these imbalances. In short, when the weather doesn't behave as predicted, the grid suffers.

To analyze this relationship more deeply, we developed a time series model called SARIMAX, which allowed us to capture patterns in energy imbalance over time and account for forecast errors in temperature, solar radiation, and cloud cover. The model found that even small inaccuracies in weather forecasts could lead to meaningful shifts in energy imbalance, especially for cloud cover and temperature.

Based on our findings, we recommend a combination of technical and policy interventions. These include investing in community-scale energy storage in high-imbalance regions, improving weather forecasting systems, adopting smart-grid technologies, and creating pricing incentives for consumers to use electricity when solar power is most abundant.

Looking ahead, the importance of accurate weather forecasts will only grow. As climate change increases weather volatility, our energy systems must be prepared to adapt. Forecasting isn't just a planning tool—it's a foundation for making renewable energy reliable and scalable. For Estonia and countries following similar paths, getting the forecast right could be the difference between a stable, low-carbon future and a costly, unpredictable one.

2 Introduction

The global energy system is currently detaching itself from centralized power sources such as fossil fuel plants in favor of decentralized, renewable energy solutions. With the popularized renewable energy solution of the early 2000's—the solar panel—the quantity of *prosumers* (individuals or businesses that both produce and consume electricity) has grown immensely. This shift has introduced new complexities into energy management: balancing the dynamic supply and demand of owner-specific and variable sources, like solar power, becomes more difficult to control. Adaptive and advanced strategies are needed to address issues such as the increasing *energy imbalance*, which occurs when the electricity consumed or produced differs from what was expected or planned. These imbalances can lead to higher operational costs, market penalties, and even threats to grid stability (the ability to adjust to and meet fluctuating energy demands) [4].

A particularly compelling case study for the impact solar panels have on the modern energy market is Estonia, a Baltic nation of 1.3 million people. Historically dependent on oil shale for over 70% of its energy, Estonia has set ambitious targets to reach 100% renewable electricity by 2030 and carbon neutrality by 2050. In response to both environmental goals and rising energy costs—household electricity prices nearly doubled in Europe between 2021 and 2022—many Estonians have adopted photovoltaic (solar) systems. The number of prosumers in Estonia is expected to increase two- to four-fold by 2040, making the need for intelligent, data-driven energy management more urgent than ever [2], [3].

Yet, the integration of renewables introduces new forms of uncertainty. Solar energy production is highly sensitive to environmental conditions, and small deviations in weather forecasts can lead to significant mismatches between expected and actual generation. This generates the question:

> **Do weather forecast inaccuracies drive energy imbalances in increasing populations of prosumers such as Estonia—and can better weather forecasting serve as a means to more efficiently use decentralized forms of solar power and improve grid stability?**

The findings from this study aim to offer valuable insights into the role of weather forecast accuracy in harming the grid's ability to balance the fluctuating supply and demand of energy in increasingly decentralized energy system populations.

3 Data

The dataset covers energy consumption and production across Estonia's counties, including:

Geographic information: counties.

Temporal features: datetime, month, week, date, hour.

Customer segmentation: business vs. non-business customers.

Product types: Combined, Fixed, General service, Spot.

Energy metrics: consumption and production in kWh.

Environmental data: solar radiation, temperature.

3.1 Data Cleaning

Our data preparation process began by loading and preprocessing the training features from the dataset 'train_features.csv'. We created descriptive mappings for categorical variables to improve interpretability, standardized datetime fields to ensure consistency, and carefully separated and then merged consumption and production data to facilitate accurate comparisons. Missing values were addressed using appropriate fill methods tailored to each variable type, and incomplete records were filtered out to maintain data integrity. Additionally, we standardized county names across datasets and GeoJSON files to ensure geographical consistency, and finally, removed outliers identified during solar efficiency calculations to refine the quality of our analysis.

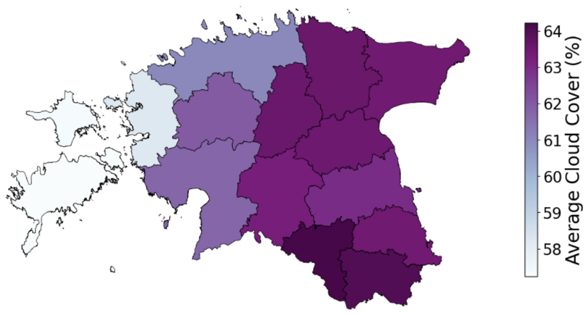

Figure 1: Average Cloud Cover

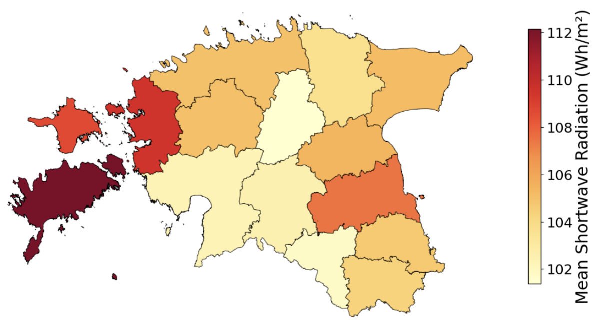

Figure 2: Short Wave Radiation

3.2 Feature Engineering

We engineered several features to enable our analysis:

Energy Metrics. We converted raw consumption and production values from kWh to MWh, calculated the *absolute imbalance* to focus on magnitude regardless of direction, and computed the financial impact by multiplying the imbalance with the electricity price.

Solar Production Metrics. We calculated the *ideal efficiency* based on installed capacity and solar radiation, estimated excess solar production using seasonal sunlight hours, and converted that excess production to MWh for direct comparison with overall imbalances.

Geospatial Integration. We mapped county IDs to standardized names, linked our dataset with GeoJSON files for visualization, and aggregated metrics at the county level to support regional analysis[1].

Production and Consumption Pivot. We reshaped the data to pivot production and consumption into separate columns for easier comparison and computed the hourly net imbalance directly within the pivoted DataFrame.

3.3 Data Exploration

3.3.1 Weather of Estonia

In order to gain a better understanding of how weather forecasting could impact grid stability, we analyzed several weather patterns over the counties of Estonia:

Average Cloud Cover (%)
Figure 1 shows average cloud cover in Estonia. Using all available elevations of cloud cover data, the lowest cloud coverage is around 55% in the western region of the country in counties such as Hiiu, Saare and Lääne. Geographically, these counties lie on the Baltic Sea. Further inland into Estonia and towards Latvia, counties such as Valga, Võru, and

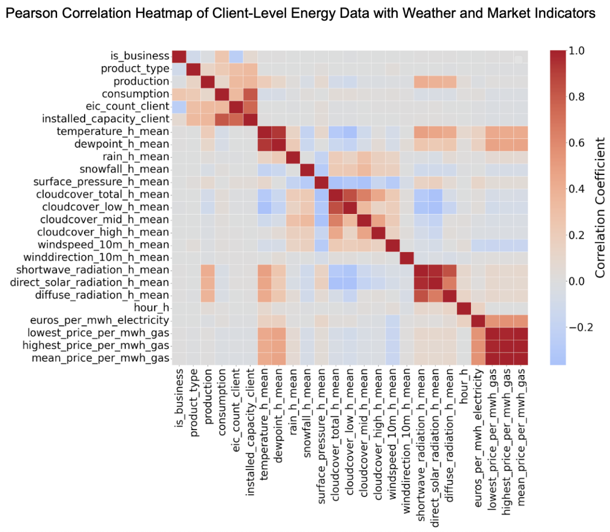

Figure 3: Feature Correlation Matrix

Põlva experience, on average, increased cloud cover near 65%.

Mean Shortwave Radiation
Figure 2 shows mean shortwave radiation received for each county in Estonia. With data of shortwave radiation (wavelengths of light from the sun), we saw that western counties of Estonia experience greater solar radiation (approximately 112 Wh/m^2), specifically the island Saare. Further inland into Estonia we see that shortwave radiation decreases, staying on average between 102 Wh/m^2–106 Wh/m^2, except for Tartu.

3.3.2 Weather Forecast and Prosumer Energy Planning

To see which features correlate with production and consumption, and as an initial exploratory entry point, we generated a correlation matrix (Figure 3) with selected features, displaying the Pearson correlation coefficients. The matrix highlights clear clusters of correlated features in the dataset. Energy-related variables like production and installed capacity show strong positive correlations. Weather variables, notably radiation measures, form their own tightly correlated group. Similarly, gas pricing variables are closely related internally and with historical temperature and dew point but largely independent from other features.

Outside of trivial clusters, we see solar radiation correlating with production. This informed our initial exploration on the impact of pre-planning energy usage according to the weather forecast.

3.3.3 Solar Radiation Forecast Error and Energy Imbalance

This analysis (Figure 5) visualizes the relationship between absolute energy imbalance and the average direct solar radiation forecast error, using hexagonal bins to aggregate and count data points. Each hexagon's color corresponds to the number of data points within it, displayed on a logarithmic scale. Darker colors indicate fewer observations, while brighter colors indicate densely populated bins. Outliers of 1% are filtered out. The plot illustrates a noticeable trend: as forecast error in direct solar radiation increases, there's generally a decrease in energy imbalance, supported by a negative correlation coefficient of approximately -0.44. This suggests that higher errors in predicting solar radiation—potentially indicative of unexpectedly higher solar input—may correspond with smaller imbalances, possibly due to unantic-

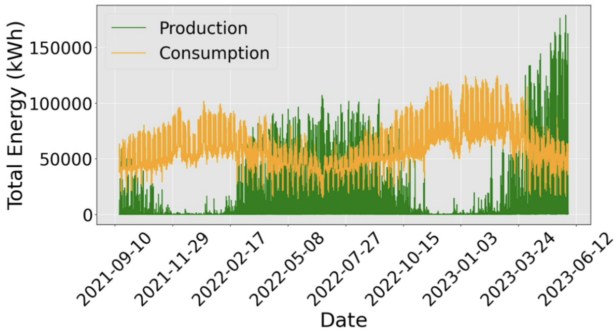

Figure 4: Supply Demand Timeseries

ipated excess energy production. The key takeaway here is the moderate negative relationship, implying that forecasting accuracy in solar radiation significantly influences energy imbalance management.

3.3.4 Weather Forecast Error

In our ongoing analysis of the correlation between weather features and production and consumption, we investigated whether prosumers were utilising weather forecast information to plan their energy consumption. To ascertain this, we developed a metric for the forecast error (difference between forecast and historical data) and measured correlations with energy imbalance. We compared these correlations with the extremity of weather experienced on that day (measured by its deviation from the average using a z-score).

Correlation between historical weather variables and energy imbalance:

- `temperature_h_mean`: -0.5236
- `dewpoint_h_mean`: -0.4546

Correlation between forecast error (forecast − historical) and energy imbalance:

- `temperature_error`: 0.0265
- `dewpoint_error`: 0.0216
- `snowfall_error`: -0.19997
- `cloudcover_total_error`: -0.26745
- `cloudcover_low_error`: -0.31860
- `cloudcover_mid_error`: -0.13638
- `cloudcover_high_error`: -0.08186
- `direct_solar_radiation_error`: -0.34439

Our findings indicate a strong correlation between energy imbalance and actual weather conditions, while there is a lower correlation between energy imbalance and forecast error. This suggests that prosumers in Estonia may have an opportunity to better use forecast data to optimise their energy patterns.

3.3.5 Energy Imbalance Analysis

The supply-demand graph (Figure 4) underscores the dynamic and seasonal nature of Estonia's energy imbalance, defined as the difference between production and consumption. Clear seasonal patterns emerge, with production often exceeding consumption significantly during summer months, reflecting the heightened solar generation from residential installations. Conversely, winter months typically exhibit a deficit, highlighting the critical need for accurate, adaptive forecasting and management solutions. This distinct Estonian seasonality, along with observable daily fluctuations, sets the stage for our application of the SARIMAX time-series model, which leverages temporal dependencies and weather-related exogenous variables to effectively predict and manage these imbalances. Understanding these temporal dynamics informs our subsequent exploration of regional imbalance distributions and associated cost impacts.

4 Methods

4.1 Model Formulation

To explore the question "Do weather forecast inaccuracies drive energy imbalances — and can better forecasting improve grid stability?", we constructed a time series model to characterize and predict hourly energy imbalances across Estonia. Our modeling approach is grounded in the understanding that electricity demand and supply are temporally dependent, influenced both by past imbalances and by external covariates such as meteorological conditions. To that end, we employed the Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors (SARIMAX) framework.

SARIMAX models extend the classical ARIMA family by incorporating external predictors (exogenous variables) and by allowing for seasonal components. Specifically, the SARIMAX model forecasts a univariate response variable (in our case, the energy imbalance, defined as the absolute difference between aggregate production and consumption at each hour) based on its own past values, past forecast errors, and concurrent values of exogenous predictors. The SARIMAX model is typically written in the form:

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \sum_{k=1}^{r} \beta_k x_{k,t} + \epsilon_t$$

with optional seasonal counterparts for the autoregressive and moving average terms. Here, y_t is the target variable at time t, ϕ_i are the autoregressive (AR) coefficients, θ_j are the moving average (MA) coefficients, $x_{k,t}$ are the exogenous variables, and ϵ_t is the residual white noise error. The model accommodates seasonality via terms such as seasonal AR(P), seasonal MA(Q), and seasonal period s, enabling it to capture periodic fluctuations in the series.

In our case, SARIMAX is particularly well-suited due to three key characteristics of the data. First, the energy imbalance time series exhibits strong temporal autocorrelation, with clear persistence at lag 1. Second, there is evident seasonality on a 24-hour cycle, reflecting consistent daily usage and production patterns. Third, imbalance is affected by exogenous drivers — namely, errors in weather forecasts, which can lead to misestimation of solar energy production and weather-dependent demand (e.g., temperature-sensitive HVAC usage).

To capture these influences, we constructed three exogenous regressors to represent forecast inaccuracy:

1. Temperature error: the difference between forecasted and observed 2-meter air temperature,
2. Cloud cover error: the difference between forecasted and observed total cloud cover percentage, and
3. Solar radiation error: the difference between forecasted direct solar radiation and observed shortwave radiation at the surface.

This formulation allows the model to account for autocorrelation and seasonality in imbalance while simultaneously estimating the associative relationship between weather forecast errors and grid stability outcomes. Importantly, while SARIMAX provides interpretable coefficients on exogenous variables, it does not support causal inference. The relationships identified are correlational and must be interpreted within the limits of observational data and model assumptions.

This approach aligns with recent work by

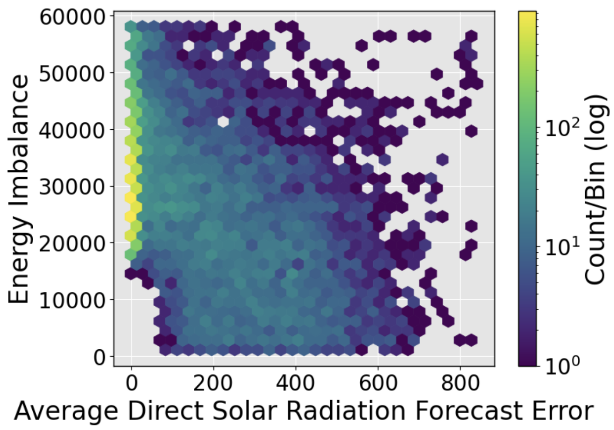

Figure 5: Solar Radiation and Energy Imbalance Correlation

Goodarzi et al. (2019), who examined the German electricity market to evaluate how forecast errors in wind and solar energy affect imbalance volumes and intraday spot prices[5]. Using a combination of regression and autoregressive models with high-frequency data, they identified a strong association between wind forecast inaccuracies and grid-level imbalances. While solar forecast errors had a smaller effect, their analysis demonstrated the broader value of modeling renewable forecast uncertainty for market stability. Following this precedent, we apply a SARIMAX framework to the imbalance forecasting setting, extending the ARX family of models by incorporating both short-term and seasonal temporal structure and using solar-relevant forecast error terms as exogenous inputs.

4.2 Verification of Assumptions

To ensure the reliability of our SARIMAX model estimates, we assessed the primary assumptions associated with time series regression: stationarity, residual independence, linearity, and homoskedasticity. Our final model—SARIMAX$(2, 0, 2) \times (2, 0, 2, 24)$, incorporating temperature, cloud cover, and solar radiation forecast errors as exogenous regressors—shows substantial improvements across diagnostic checks.

4.2.1 Stationarity

The SARIMAX model assumes the underlying time series is weakly stationary. We verified this with an Augmented Dickey-Fuller (ADF) test on the imbalance series, which returned a p-value of 4.45×10^{-9}, providing strong statistical evidence to reject the presence of a unit root. As a result, differencing was not required, and the model was fit in levels.

4.2.2 Residual Independence

A key improvement of the enhanced SARIMAX specification was in the reduction of autocorrelation in residuals. The autocorrelation function (ACF) of the residuals revealed that the largest autocorrelation coefficient—previously around 0.2 at lag 24—was reduced to approximately 0.1. The Ljung-Box test at lag 1 returned a p-value of 0.55, indicating no evidence of significant residual autocorrelation at short lags. The inclusion of additional seasonal AR and MA terms (lags 24 and 48) appears to have successfully accounted for much of the daily and multi-day structure in the data. While some minor residual autocorrelation remains as seen in Figure 6, it is well within acceptable bounds for applied time series modeling and does not materially affect the model's validity.

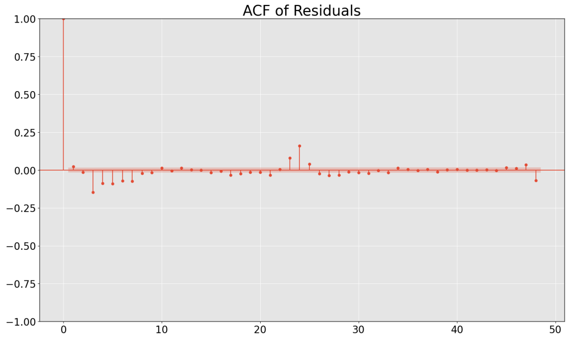

Figure 6: Residuals ACF for the SARIMAX model on time series data.

4.2.3 Linearity and Error Structure

The model assumes a linear relationship between the response variable and its predictors. Inspection of the residuals versus fitted values confirms an approximately linear trend with residuals symmetrically distributed around zero. There is a slight wedge-shaped pattern, suggesting that variance may increase with predicted values; however, this pattern is relatively modest. Overall, the residual structure does not show strong signs of misspecification, and the linear model form remains an appropriate choice.

4.2.4 Homoskedasticity and Residual Behavior

We also assessed whether the residual variance remained constant over time. While the residuals-versus-fitted plot shown in Figure 7 shows a slight wedge-shaped pattern—suggesting some mild heteroskedasticity—this pattern appears limited in scope and does not exhibit strong volatility clustering. Formal testing using the ARCH test returned a low p-value ($\approx 2.49 \times 10^{-50}$), indicating some degree of time-varying variance.

However, this result is not uncommon in applied energy data, which often display modest fluctuations in variance due to operational and environmental factors. Importantly, such heteroskedasticity does not affect the consistency or unbiasedness of coefficient estimates, and its influence on standard errors is mitigated by the large sample size of our dataset. Since our primary focus is on understanding the relationship between forecast inaccuracies and energy imbalance—rather than formal hypothesis testing—the observed variance behavior is unlikely to materially distort our conclusions. Moreover, the strong significance of the exogenous predictors and the model's stable residual structure support its robustness for interpretative and predictive purposes.

In future extensions, incorporating a conditional variance model (e.g., GARCH) could offer a more detailed treatment of volatility, but for the current analysis, the SARIMAX model provides a reliable and informative framework.

4.3 Training

The SARIMAX model was implemented using the statsmodels library in Python, which provides robust tools for fitting seasonal time series models with exogenous regressors. Model training involved both exploratory analysis of the time series structure and a targeted process for selecting appropriate hyperparameters.

We began by inspecting the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the energy imbalance time series, as shown in Figures 8 and 9. These diagnostics revealed strong autocorrelation at lag 1 and a pronounced seasonal structure at multiples of 24 hours, corresponding to a daily cycle. This informed our decision to

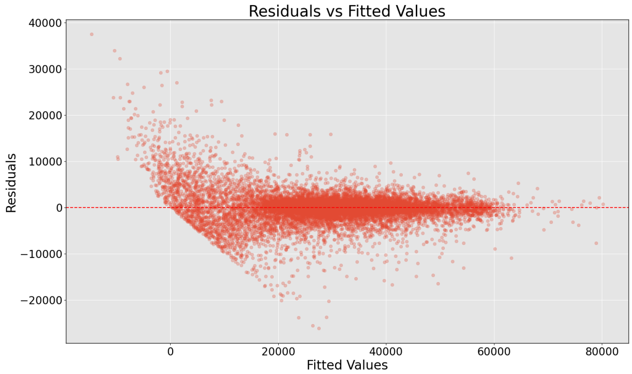

Figure 7: Residuals for the SARIMAX model on time series data. A minor wedge pattern is visible.

include both short-term and seasonal components in the model. Initial experiments used a simple specification with non-seasonal order $(1, 0, 1)$ and seasonal order $(1, 0, 1, 24)$, which significantly reduced residual autocorrelation compared to an unregularized baseline. However, residual diagnostics indicated that additional structure remained, particularly at seasonal lags. We iteratively expanded the model to include additional autoregressive and moving average terms, both seasonally and non-seasonally. Ultimately, our final specification used:

- Non-seasonal order: $(p, d, q) = (2, 0, 2)$,
- Seasonal order: $(P, D, Q, s) = (2, 0, 2, 24)$.

This structure was chosen to balance model complexity with predictive performance and residual whiteness. Each term was justified by examining lag patterns in the ACF/PACF plots and evaluating improvements in model diagnostics. We prioritized reduction in residual autocorrelation and model parsimony over marginal improvements in fit metrics to prevent overfitting. The model was trained on the full hourly imbalance series using maximum likelihood estimation. Forecast error variables—temperature, cloud cover, and solar radiation error—were used as exogenous regressors. These were standardized to improve convergence and ensure numerical stability during optimization. The model was trained on approximately $15,000$ hourly observations, providing ample data to support estimation of the relatively small number of parameters.

The final trained model offered stable convergence, interpretable coefficients, and minimal residual autocorrelation, making it a strong candidate for understanding the dynamics of energy imbalance in relation to forecast error.

5 Results

To evaluate the relationship between weather forecast inaccuracies and energy imbalance, we estimated a Seasonal Autoregressive Integrated Moving Average with eXogenous variables (SARIMAX) model using the `statsmodels` library in Python. Our final specification included both short-term and seasonal autoregressive and moving average terms with exogenous regressors for solar radiation error, cloud cover error, and temperature error. The results are summarized in Figure 10.

5.1 Hypothesis testing

To test the central hypothesis of this study—namely, that inaccuracies in weather forecasts significantly contribute to energy imbalances in solar energy systems—we examined the statistical significance of three forecast error variables: solar radiation er-

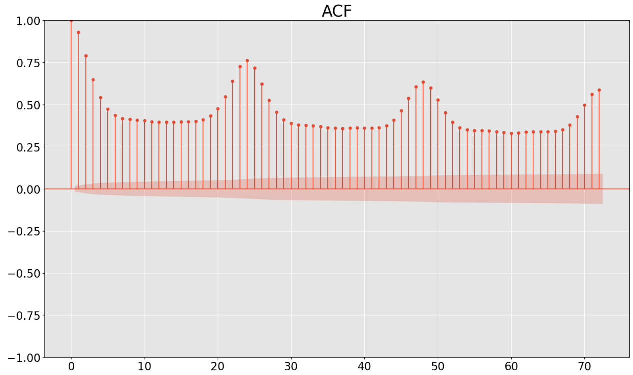

Figure 8: Autocorrelation function for imbalance time series with 72 lags.

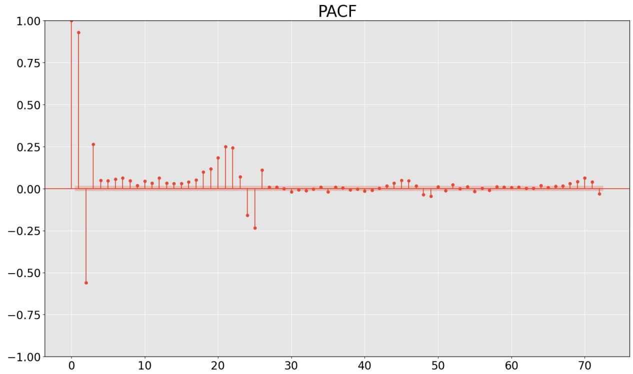

Figure 9: Partial autocorrelation function for imbalance time series with 72 lags.

ror, cloud cover error, and temperature error. These were included as exogenous regressors in the SARIMAX$(2,0,2)\times(2,0,2,24)$ model, and their coefficients are reported in Figure 10.

For each regressor, we conducted a null hypothesis test of the form:

$$H_0 : \beta_i = 0 \quad \text{vs.} \quad H_1 : \beta_i \neq 0$$

where β_i denotes the coefficient on the ith forecast error variable. Statistical significance was evaluated using the z-statistics provided by the SARIMAX model's maximum likelihood estimation procedure.

All three variables were found to be highly statistically significant, with p-values less than 0.001. Specifically:

- Solar radiation error had a coefficient of –50.33 ($z = –61.66$),
- Cloud cover error had a coefficient of –336.37 ($z = –186.39$), and
- Temperature error had a coefficient of 456.09 ($z = 14.01$).

These results provide strong statistical evidence that errors in weather forecasts—particularly those related to cloud cover and temperature—are systematically associated with larger deviations between actual and expected energy flows. The signs of the coefficients are also directionally interpretable: under-predicting temperature, for instance, corresponds to an increase in imbalance, which is plausible given the sensitivity of solar generation to ambient conditions.

Although we do not claim causal identification, the strength and consistency of these relationships support the conclusion that forecast accuracy is a meaningful predictor of energy imbalance, reinforcing the relevance of improved weather modeling for enhancing

Variable	Coefficient	Std. Error	z-value	p-value
Exogenous Variables				
Solar Radiation Error	-50.33	0.82	-61.66	< 0.001
Cloud Cover Error	-336.37	1.81	-186.39	< 0.001
Temperature Error	456.09	32.57	14.01	< 0.001
Non-seasonal ARIMA Terms				
AR(1)	0.041	0.052	0.79	0.431
AR(2)	0.716	0.047	15.27	< 0.001
MA(1)	1.146	0.052	21.88	< 0.001
MA(2)	0.291	0.015	19.19	< 0.001
Seasonal ARIMA Terms (Seasonality = 24 hours)				
Seasonal AR(1) [Lag 24]	1.966	0.007	299.55	< 0.001
Seasonal AR(2) [Lag 48]	-0.968	0.006	-150.79	< 0.001
Seasonal MA(1) [Lag 24]	-1.909	0.006	-296.39	< 0.001
Seasonal MA(2) [Lag 48]	0.916	0.006	152.09	< 0.001

Figure 10: Estimated coefficients from SARIMAX$(2, 0, 2) \times (2, 0, 2, 24)$ model. All reported coefficients are statistically significant at the 0.001 level unless otherwise noted.

grid reliability.

5.2 Effect sizes

In addition to establishing statistical significance, the estimated coefficients from the SARIMAX model provide insight into the practical magnitude of the relationship between weather forecast errors and energy imbalance. Each coefficient represents the expected change in hourly energy imbalance, measured in kilowatt-hours (kWh), associated with a one-unit increase in the respective forecast error.

The most substantial effect was observed for cloud cover forecast error, which had a coefficient of –336.37 Wh. This corresponds to approximately –0.336 kWh per percentage point error in total cloud cover. That is, if cloud cover is underestimated by 10 percentage points, the model predicts an average increase of 3.36 kWh in imbalance, in the opposite direction of the error. This reflects the significant impact that cloud cover has on solar irradiance and photovoltaic generation.

Forecast error in temperature also had a sizable effect, with a coefficient of 456.09 Wh, or approximately 0.456 kWh per °C. A 2°C misestimation in ambient temperature would be associated with nearly 1 kWh shift in imbalance, a meaningful deviation when aggregated across producers and consumers. This is likely driven by both demand-side effects (e.g., heating/cooling needs) and temperature-dependent performance of solar systems.

The solar radiation forecast error had a more modest but still notable effect, with a coefficient of -50.33 Wh, or -0.050 kWh per unit of forecast error (measured in Wh/m^2). Although smaller in magnitude, this result still suggests that inaccuracies in solar radiation forecasting translate into measurable shifts in energy balance, especially over large time horizons or at grid scale.

Together, these results demonstrate that forecast errors can yield imbalance shifts on the order of 0.05-0.45 kWh per unit error, with cloud and temperature errors exerting the strongest influence. While the interpretation is correlational and conditional on the time series model structure, these effect sizes point to a material sensitivity of the energy system to weather forecast accuracy, reinforcing the potential value of improved forecasting for grid stability.

5.3 Feature weights

While absolute effect sizes provide meaningful real-world interpretation, it is also instructive to consider the relative contribution of each forecast variable to the model's predictions of energy imbalance. In the context of the SARIMAX model, the magnitude of each coefficient—along with its statistical significance—offers a proxy for feature importance.

Among the three exogenous predictors, cloud cover forecast error stands out as the most influential variable. Its coefficient of –0.336 kWh per percentage point error was not only the largest in absolute magnitude, but also exhibited the highest z-statistic (–186.39), indicating both strength and stability of its effect. This suggests that cloud cover error contributes most substantially to explaining variation in the target variable.

The second most influential feature was temperature forecast error, with a coefficient of 0.456 kWh per °C. While slightly larger in absolute effect size than cloud error, its standard error was also higher, resulting in a somewhat lower z-value (14.01). Nonetheless, its strong and statistically significant contribution highlights the dual impact of temperature on both generation (via panel efficiency) and consumption (via demand-side behaviors). Solar radiation error played a more modest role, with a coefficient of –0.050 kWh per unit error. Although smaller in magnitude, it remained highly statistically significant (z = –61.66), suggesting that while the signal is subtler, it is consistent and nonnegligible. Its weaker influence may be partially attributable to collinearity with cloud cover forecasts, which are closely tied to surface radiation outcomes.

Overall, this analysis reveals that cloud cover and temperature forecasts are the most critical drivers of imbalance variability in the model. While all three features contribute meaningfully, their weights suggest that grid operators and forecasters may benefit most from improving cloud and temperature prediction accuracy when aiming to reduce forecast-driven imbalance.

6 Discussion

Our analysis provided robust evidence linking weather forecast inaccuracies to energy imbalance issues within Estonia's decentralized energy system. The negative correlation between direct solar radiation forecast errors and energy imbalances ($r \approx -0.44$) demonstrated that increased forecast errors—particularly when solar radiation was underestimated—often corresponded to periods of unexpectedly high energy production. This relationship emphasizes the critical importance of accurate solar radiation forecasts, highlighting the vulnerability of Estonia's power grid to even modest inaccuracies in weather prediction. Delving deeper with a SARIMAX time-series model, we quantified the specific impact of forecast errors on energy imbalances. Cloud cover forecast inaccuracies emerged as the most influential variable, carrying the largest absolute effect size (–0.336 kWh per percentage point of cloud cover forecast error). Temperature forecast error also showed substantial impact (0.456 kWh per °C). Interestingly, direct solar radiation error, while statistically significant, had a comparatively smaller influence (–0.050 kWh per Wh/m^2), potentially due to its correlation with cloud cover error. These findings underscore the nuanced roles different meteorological variables play, guiding strategic prioritization for grid stability interventions.

Regionally, the imbalance analysis revealed substantial differences, with northwestern counties experiencing significantly higher imbalance levels. The alignment of higher imbalances with greater residential solar penetration strongly suggests that midday overproduction from residential solar installations contributes substantially to grid management challenges. Given Estonia's ambitious renewable energy targets, addressing these localized imbalances is critical.

The economic implications of these imbalances are noteworthy. The regional disparity in costs associated with managing energy imbalances points towards targeted financial interventions. Our findings suggest that community-scale energy storage and smart-grid technologies in high-imbalance regions like northwest Estonia could mitigate not only technical but also economic impacts. Additionally, dynamic pricing and policy adjustments encouraging installations equipped with energy storage could further incentivize behaviors that align consumption with production peaks. Based on our comprehensive results from both data analysis and the SARIMAX model, we recommend:

1. Targeted Storage Solutions: Deploy community-scale energy storage systems in northwestern counties to absorb excess midday solar production, thus addressing the significant regional imbalance.

2. Enhanced Weather Forecasting: Prioritize investments in improving cloud cover and temperature forecasting accuracy, given their substantial impact on energy imbalance, to enhance grid stability.

3. Dynamic Pricing Schemes: Implement time-of-use pricing structures that incentivize consumers to shift energy-intensive activities to coincide with peak solar production hours, thereby reducing midday overproduction stresses on the grid.

4. Grid Modernization: Invest in advanced smart-grid technologies specifically designed to better integrate fluctuating decentralized energy sources, allowing for real-time balancing and improved resilience.

5. Policy Adjustments: Revise existing solar incentives to encourage the integration of solar installations with battery storage, aligning residential energy production more closely with actual consumption patterns.

Despite the insights provided, our analysis faced limitations. The SARIMAX model, although robust, remains correlational rather than causal; further research employing experimental or quasi-experimental designs could strengthen causal claims. Moreover, the residual variance indicated mild heteroskedasticity, suggesting future models could benefit from conditional variance modeling techniques, such as GARCH, to better capture volatility.

Overall, our study effectively demonstrates that enhancing forecast accuracy, particularly for cloud cover and temperature, could substantially improve energy management in prosumer-rich, renewable energy environments. Implementing the suggested measures could significantly stabilize Estonia's energy grid, reduce costs, and advance the country's progress toward its renewable energy goals.

6.1 Climate insight

The insights we made from this data have broader implications under global climate change. As global warming continues to threaten global climate, it is likely to cause more volatile and unpredictable local weather patterns, such as sudden cloud formation, unseasonal temperature swings, or irregular precipitation. These dynamics make accurate forecasting more difficult, especially in temperate regions like Estonia where variability is already high. Consequently, forecast error may grow as climate volatility increases, worsening the mismatch between solar energy generation and consumption, thus attributing to grid instability and a less efficient future.

This establishes a dangerous feedback loop: more renewables results in more reliance on forecasts; more climate volatility results in less forecast accuracy; less accuracy results in more grid instability. Without significant investments in forecasting, storage, and grid intelligence, this could undermine the very climate goals that Estonia has for its transition into decentralized solar energy systems and for a sustainable future. Thus, our insight highlights a strategic climate vulnerability: accurate forecasting isn't just a tool for efficiency—it's a keystone for sustainable energy transitions in a warming world.

7 Conclusion

Estonian counties face significant energy imbalances, primarily due to midday overproduction from residential solar installations. These regions incur the highest associated costs, emphasizing the importance of targeted energy management strategies. Analysis revealed a strong, statistically significant relationship between inaccuracies in weather forecasts—especially cloud cover and temperature errors—and energy imbalance. Improved forecasting accuracy, particularly for cloud cover and temperature, would substantially enhance grid stability and reduce costs. Using the SARIMAX model, cloud cover forecast error emerged as the most influential variable (effect size: –0.336 kWh per percentage point error), followed by temperature error (0.456 kWh per °C). These findings suggest substantial practical benefits from improved forecast accuracy. We propose targeted, feasible interventions, such as deploying localized energy storage solutions in high-imbalance areas, implementing dynamic pricing schemes, and investing in smart grid technologies. These recommendations provide clear and impactful ways to mitigate the identified problems.

8 Appendix

Code used to generate all figures and models can be found at our github `https://github.com/tylerheadley/datathon-sustainability`.

- `enrgs&dviz.ipynb` (correlation analysis and hexbin plots)
- `moreDataVis.ipynb` (correlation matrix, mosaic plot, supply demand time series)
- `imbalance_patterns.ipynb`
- `model.ipynb`
- `solar_efficiency.ipynb`
- `Weather.ipynb`

References

[1] Abdelrahman Sousa. Estonia administrative boundaries, 2023. Accessed: 2025-04-14.

[2] Colorado Clean Energy Fund. Grid stability, n.d. Accessed: 2025-04-14.

[3] Colorado Clean Energy Fund. Grid stability (google redirect), n.d. Accessed: 2025-04-14.

[4] Global Energy Network Institute (GENI). Energy overview of estonia, n.d. Accessed: 2025-04-14.

[5] Shadi Goodarzi, H Niles Perera, and Derek Bunn. The impact of renewable energy forecast errors on imbalance volumes and electricity spot prices. *Energy Policy*, 134:110827, 2019.