UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Computer Science Curriculum

Yuliia Siur

Prosumer Net Consumption Forecasting: The Impact of Behind-the-Meter Self-Consumption and Weather Forecast

Master's Thesis (30 ECTS)

Supervisor(s): Novin Shahroudi, MSc
Jean-Baptiste Scellier, MSc

Tartu 2024

Prosumer Net Consumption Forecasting: The Impact of Behind-the-Meter Self-Consumption and Weather Forecast

Abstract:

In recent years, the adoption of renewable energy sources has significantly increased. Notably, solar photovoltaic (PV) panels are gaining widespread popularity, particularly among private households. Residential rooftop PV systems enable private households to generate and utilize their own electricity. Private households with a dual role in electricity production and consumption, also known as "prosumers", establish direct market relationships with energy companies, facilitating the sale of surplus energy to the grid and purchasing when energy production is insufficient. The boost of prosumer-driven energy generation shifts energy companies' electricity flow management. Traditional Consumption Metering is no longer sufficient, as it fails to capture prosumers' interactions with the grid. Instead, energy companies adopt Net Purchasing Systems. These systems measure both the electricity consumed from the grid and the electricity exported back to the grid by the prosumers. The adoption of new metering systems gives rise to the development of novel methodologies for forecasting Net Consumption. However, the task of forecasting Net Consumption presents challenges arising from the three primary factors: a) the diverse behavioral consumption patterns exhibited by private households, reflecting complexities observed in Consumption Forecasting; b) the inherent variability of solar energy production, which is influenced by fluctuations in weather patterns and solar positioning across various time frames; c) the nature of the Net Purchasing Metering does not consider behind-the-meter values of production and consumption but rather accounts for energy injected to and withdrawn from the grid. The discrepancy between total and monitored values equals the prosumers' self-consumption of generated energy, which remains unmonitored, thereby increasing the complexity of modeling relationships indicated in (a) and (b). In our study, we focus on advancing day-ahead Net Energy Forecasting techniques using Estonian prosumers as a case study. We introduce novel types of Additive and Integrated Models by incorporating distinct input features, aiming to mitigate uncertainty originating from weather forecast variables and unmetered self-consumption. Our approach enables the models to effectively capture complex relationships between input and target variables. Experimental results provided empirical evidence of an enhanced capacity to address uncertainty originating from weather predictions and unmetered self-consumption in our Consumption model developed for the Additive method. In contrast, other models did not exhibit any improvement. These findings establish a foundation for further research focused on understanding how the models capture the relationships between input and target variables.

Keywords: Time Series Forecasting, Net Consumption Forecasting, Self-consumption,

Prosumer, Electricity Consumption, Electricity Production

CERCS: P176 - Artificial Intelligence, T140 - Energy Research.

Prosumerite netotarbimise prognoos: mõõtmata enesetarbimise ja ilmaennustuse mõju

Lühikokkuvõte:

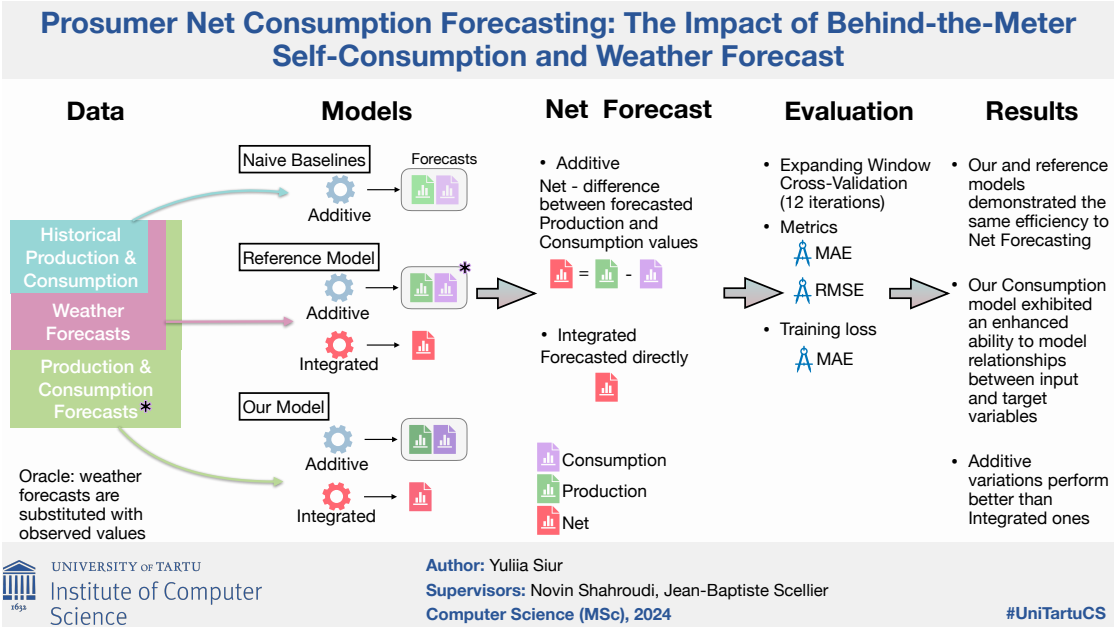
Viimastel aastatel on taastuvate energiaallikate kasutuselevõtt märkimisväärselt suurenenud. Eelkõige on laialdast populaarsust kogumas päikeseenergia fotogalvaanilised (PV) paneelid, eriti eramajade seas. Elamute katusel asuvad fotoelektrilised süsteemid võimaldavad majapidamistel toota energiat ja kasutada oma elektrit. Majapidamised, kes võrguelektrit nii toodavad, kui ka tarbivad, on tuntud kui "prosumerid". Need loovad otsesed turusuhted energiaettevõtetega, hõlbustades energia ülejäägi müüki võrku ja ostes, kui nende energiatootmine ei ole piisav. Prosumerite arvu kasv mõjutab energiaettevõtete elektrivoogude juhtimist. Traditsiooniline tarbimise mõõtmine ei ole enam piisav, sest see ei hõlma prosumerite suhtlemist võrguga. Selle asemel võtavad energiaettevõtted kasutusele kahesuunalised net ostusüsteemid. Need süsteemid mõõdavad nii võrgust tarbitud elektrienergiat kui ka tarbijate poolt võrku tagasi eksporditud elektrienergiat. Uute mõõtesüsteemide kasutuselevõtt annab hoogu uudsete meetodite väljatöötamiseks netotarbimise prognoosimiseks mitmetele prosumerite jaoks. Netotarbimise prognoosimisel esinevad aga väljakutsed, mis tulenevad kolmest asjaolust: a) majapidamiste erinevad tarbimisharjumused, mis kajastavad tarbimise prognoosimisel ilmuvat keerukust; b) päikeseenergia tootmise olemuslik varieeruvus, millele mõjuvad ilmastikutingimuste kõikumised ja päikese positsioneerimine erinevates ajaraamides; c) elektriarvesti olemus ei arvesta tootmise ja tarbimise väärtused, vaid pigem võrku saadetud ja sealt välja võetud energiat. Erinevus reaalse ja arvesti poolt mõõdetud väärtuste vahel võrdub prosumeri enesetarbimisega, mis jääb järelevalveta. See suurendab punktides (a) ja (b) osutatud modelleerimissuhete keerukust. Uurimuses keskendume päev-ette netotarbimise prognoositehnikate arendamisele, kasutades Eesti prosumerite andmeid. Meie tutvustame uudseid aditiivseid ja integreeritud mudeleid kasutades erinevad tunnused, mille eesmärk on leevendada ilmaprognooside muutujatest ja mõõtmata enesetarbimisest tulenevat ebakindlust. Meie lähenemine võimaldab mudelitel tõhusalt jäädvustada keerulisi seoseid sisend- ja sihtmuutujate vahel. Eksperimentaalsed tulemused andsid empiirilisi tõendeid selle kohta, et ilmaennustustest ja mõõtmata omatarbest tulenev ebakindlus on paranenud meie tarbimismudelis, mis töötati välja additive-meetodi jaoks. Seevastu teised mudelid ei näidanud mingit paranemist. Need leiud loovad aluse edasiseks uurimistööks, mis keskendub arusaamisele, kuidas mudelid hõlmavad sisend- ja sihtmuutujate vahelisi seoseid.

Võtmesõnad: Aegridade prognoos, Netotarbimise prognoos, Enesetarbimine, Prosume-

rid, Elektritarbimine, Elektritootmine

CERCS: P176 - Tehisintellekt, T140 - Energeetika.

Visual Abstract:



Contents

1 Introduction

Climate change remains a highly debated topic in modern society in recent years. The European Commission has outlined an objective for achieving carbon neutrality across the European Union by 2050 [SP21]. Developing sustainable, environment-friendly, and competitive energy sources is crucial in attaining this goal. Solar power is one of the most widely adopted forms of renewable energy worldwide. In Europe alone, the capacity of PV electricity generation increased from 1.9 GW to over 150 GW in the past decade [AJW20].

As the adoption of PV panels has risen, their cost has steadily declined. Government support, including subsidy programs, has further facilitated the uptake of PV energy among private households as a leading choice for sustainable energy generation [Pal18]. Besides environmental considerations, private households transition their perspective from a "consumer" to a "prosumer" (*pro*ducer-con*sumer*) due to financial benefits arising from achieving energy self-sufficiency and the potential for energy storage and resale at favorable rates.

The rising trend of prosumers, while advantageous for the environment, presents notable challenges, particularly for energy grid operators. Primarily, the stochastic nature of solar energy contributes to operational complexities in aligning demand with electricity supply generated by photovoltaic (PV) panels. Additionally, installing PV panels behind-the-meter obliges a transition from traditional Consumption Metering to systems like Net Metering, which monitor grid electricity injection and withdrawal [YWX18].

Adopting the new metering systems gives rise to the development of novel methodologies for forecasting Net Consumption. However, the task of forecasting Net Consumption presents challenges arising from the three primary factors: a) the diverse behavioral consumption patterns exhibited by private households, reflecting complexities observed in Consumption Forecasting; b) the inherent variability of solar energy production, which is influenced by fluctuations in weather patterns and solar positioning across various time frames; c) the nature of Net Purchasing Metering does not consider the actual values of production and consumption but rather accounts for energy injected to and withdrawn from the grid. The discrepancy between total and monitored values equals the prosumers' self-consumption of generated energy, which remains unmonitored, thereby increasing the complexity of modeling relationships indicated in (a) and (b).

Energy companies trade electricity through electricity markets, where they submit bids at intra-day and day-ahead auctions. Energy trade, based on demand and supply forecasts, helps ensure a balance between energy supply and demand across participating entities and regions. Notably, the leading power market in Europe, Nord Pool, has traded 1,077.35 TWh of power in 2022, of which 1,077.35 TWh belong to the day-ahead auction [Poo23].

In the past, day-ahead bidding emerged as electricity production relied heavily on flexible power plants that could adjust their generative power by manual regulation. How-

ever, the growing utilization of intermittent renewable energy sources has significantly heightened the unpredictability of day-ahead energy production. Inaccurate forecasts can lead to financial penalties for energy companies that fail to meet contractual obligations, and they may be forced to purchase electricity at higher prices during peak demand hours when supply falls short. Therefore, developing an efficient method of high prediction accuracy for forecasting Net Consumption is crucial for effective day-ahead energy management.

Although the forecasting methods of Solar Energy Production and Electricity Consumption have been extensively researched, the task of Net Consumption Forecasting has received limited attention in the literature. It has been demonstrated that Net Forecasting poses a greater complexity compared to Consumption Forecasting, primarily due to additional uncertainty originating from the PV energy generation [SERM20].

Several distinctive approaches to Net Forecasting have been developed. In [AK16], the authors proposed and compared Additive and Integrated Models, demonstrating the superiority of the Integrated Model on a microgrid scenario where solar energy covers 33% of the annual energy demand. In [YWX18], the authors addressed the problem of the distributed PV energy, invisible to the distribution system operators. They proposed a separation and aggregation strategy to tackle this challenge. This strategy involves decomposing the Net Energy profile into three components: PV energy production, consumption, and residual. Each segment is forecasted individually, and they are subsequently aggregated to obtain a Net Energy forecast. The authors demonstrated that this approach outperformed the traditional regression method of direct Net Consumption Forecasting in both point and probabilistic forecasting, particularly under high energy penetration scenarios.

However, none of the aforementioned studies considered data from northern regions like Estonia. In Estonia, winters are longer than summers, with daytime lasting around 6-7 hours in winter and 18-20 hours in summer, resulting in less solar potential compared to southern regions. In compliance with EU directives, starting in 2021, all new buildings in the EU should achieve nearly zero energy standards. Although Estonia has limited solar energy potential, it is sufficient to meet the energy needs of small households and residential buildings. It makes solar energy an effective solution in Estonia, particularly with the support of the Estonian government [NS22]. The growing adoption of solar energy in northern regions underscores the importance of researching use cases specific to these areas.

Our research focuses on advancing day-ahead Net Consumption Forecasting techniques using Estonian prosumers as a case study. We introduce novel types of Additive and Integrated Models by incorporating distinct input features, aiming to mitigate uncertainty originating from weather forecast variables and unmetered self-consumption. Our approach enables the models to effectively capture complex relationships between input and target variables. Our study is conducted in collaboration with Eesti Energia (EE),

recognized internationally as Enefit, the largest energy producer of renewable energy in the Baltics.

2 Background

We start this section with an introduction to the energy market and its participants, followed by a description of the Time Series Forecasting problem, encompassing its definition, an overview of utilized forecasting models, and subsequent evaluation methodologies. Furthermore, we delve into feature engineering and selection procedures, discuss various covariate types and an oracle model, and conclude by exploring the Net Consumption Forecasting problem.

2.1 Energy market

An electricity market represents a structured system where electricity is traded among diverse participants. It encompasses various entities, including electricity producers, distributors, retailers, consumers, grid operators, and regulatory authorities. Electricity markets function within specific geographical areas or countries, providing participants with various product options, such as purchasing electricity on a day-ahead, intraday, or hour-ahead basis, among others. Figure 1 depicts market participants and their relations.

Electricity producers generate electricity from various sources, such as fossil fuels and renewable energy, operating power plants of different types and capacities. Distributors manage the physical infrastructure required for transmitting and distributing electricity to consumers, ensuring the secure and reliable operation of substations, transformers, and transmission lines. Retailers function as intermediaries, purchasing electricity and distributing it to consumers through various supply agreements. Consumers, including residential, commercial, and industrial entities, are end-users of electricity. Regulatory institutions supervise electricity market operations by establishing rules, managing compliance, and ensuring fair competition and consumer protection.

Transmission system operators (TSO), also called grid operators, manage the transmission grid, which transports electricity from power generation facilities to distribution networks and end-users. TSOs are responsible for maintaining a real-time supply-demand balance. This balance is crucial for sustaining grid stability, ensuring the grid operates within safe operating limits. When supply and demand are balanced, the grid's frequency and voltage remain steady, preventing disruptions of the electricity flow and maintaining the integrity of the grid infrastructure.

However, the classification of participants within the energy market is not straightforward. Companies in the energy sector often engage in multiple activities beyond their primary operations. For instance, Estonia's Eesti Energia, primarily an electricity producer, extends its activities to include retail and other energy-related services. Another noteworthy example is prosumers, a distinctive category within the electricity market that challenges conventional boundaries between consumers and generators by consuming and producing electricity. Thus, prosumers concurrently function as consumers and small-scale electricity producers.

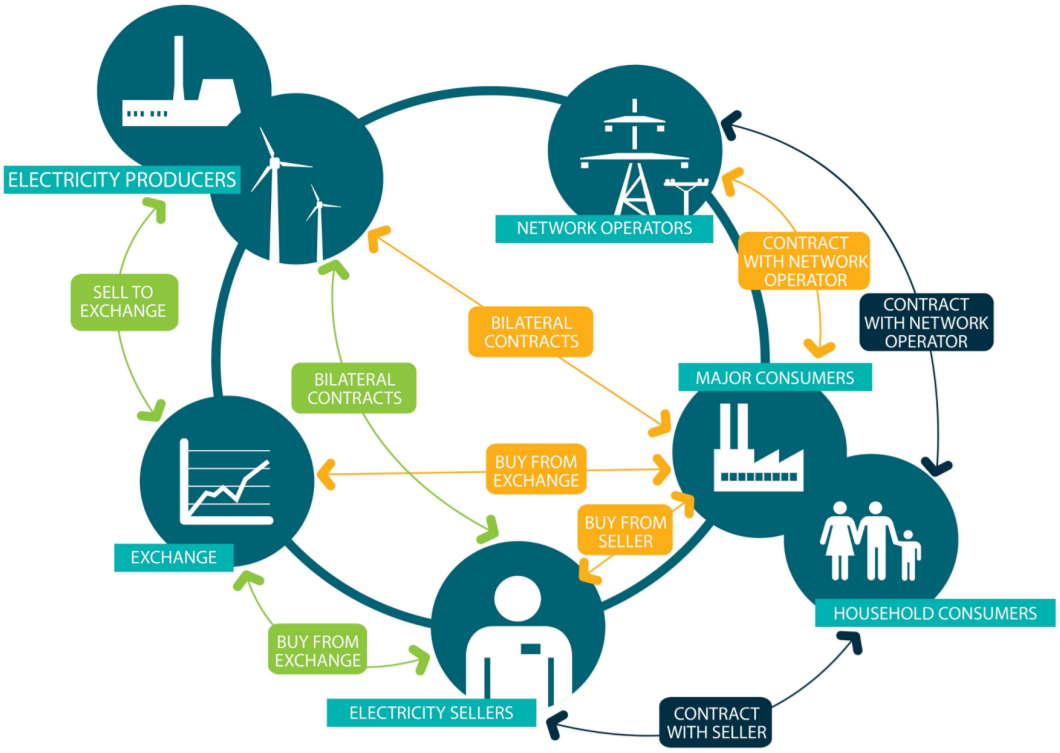


Figure 1. Energy market participants [Ele]. Electricity producers distribute their electricity to energy retailers through electricity markets. Retailers, in turn, supply electricity to various consumer segments, ranging from individual households to large-scale commercial entities, in accordance with supply contracts. Network operators maintain the reliability of the energy grid by managing the supply-demand balance. Major energy corporations often engage in multiple functions: electricity generation, energy import, and supply.

In power markets like Nord Pool, various energy producers participate in auctions to distribute their electricity. At the same time, consumers submit bids to acquire the necessary volume of electricity to satisfy their demand. Energy companies from different countries participate in these markets to trade electricity across national borders, taking advantage of price differences and ensuring the fulfillment of anticipated demand within the constraints of available resources. Such entities have the flexibility to operate as both producers and consumers within the market. Auctions are organized for intra-day trading, enabling prompt electricity transactions, and day-ahead trading, where agreements are made for future electricity delivery.

Market participants present their offers and bids in the bidding process, considering anticipated demand, production expenses, and market dynamics. Nord Pool facilitates

this bidding procedure, aligning supply with demand to determine the market clearing price, thereby optimizing the distribution of electricity resources throughout the grid.

Non-compliance with bid obligations in electricity markets leads to financial penalties for market participants. These penalties are designed to maintain the integrity of electricity markets by ensuring commitment, obedience, and obligation fulfillment among participants. They push producers to accurately evaluate their capacity to deliver electricity and adjust their bids accordingly.

Moreover, another substantial financial challenge for energy companies arises from inaccuracies in forecasting future demand and supply. When consumption exceeds estimations, energy providers must acquire additional energy power from the system operator, potentially experiencing higher expenses, to meet demand. Conversely, if consumption is lower than anticipated, retailers would have surplus electricity that they may be unable to sell, resulting in financial losses or inefficient resource utilization.

2.2 Time Series Forecasting

Time series data consists of chronologically arranged sequential data samples collected over consistent intervals, such as every 15 minutes, 30 minutes, hourly, or daily, enabling monitoring changes over time. Time series data is characterized by stationarity, seasonal and periodic components, trend, and noise, collectively contributing to its temporal structure.

Stationarity represents the constancy of statistical properties over time, indicating that the data's mean, variance, and autocorrelation remain steady across different time intervals. On the contrary, non-stationarity in time series data refers to the absence of stable statistical properties, appearing as trends, shifts, or fluctuations in mean, variance, or autocorrelation.

Seasonal and periodic components represent recurring patterns and fluctuations occurring at fixed intervals, such as daily, weekly, or yearly, caused by seasonal factors, environmental changes, or other external events. Noise, conversely, encompasses random fluctuations, measurement errors, and outliers that disrupt underlying patterns and introduce uncertainty into the data.

The trend in the time series demonstrates its change over an extended time period. It indicates whether the series is increasing or decreasing over time. This trend can be linear, displaying consistent growth or decline, or nonlinear, where the rate of change is not constant.

Effective Time Series Forecasting relies on understanding the temporal relationships within the time series data and its underlying components. Machine Learning (ML) models utilized for Time Series Forecasting must be designed to capture and learn these dependencies, maintaining causal relationships, underlying patterns, and seasonal variations between individual observations to achieve accurate predictions.

A time series forecasting model $g(\cdot)$ estimates a target variable Y based on the information set Ω, and model parameters Θ at a defined forecast origin time t. The forecasted value $\hat{y}_{t+k|t}$ for each time step k in the forecast horizon is computed as the expected value of Y_{t+k}, given model g, the information set Ω_t, available up to the time step t, and estimated parameters $\hat{\Theta}$, denoted by $\mathbb{E}$, as

$$\hat{y}_{t+k|t} = \mathbb{E}[Y_{t+k}|g, \Omega_t, \hat{\Theta}], \tag{1}$$

where $k = \{k_{min}, k_{min} + 1 \cdot \Delta, k_{min} + 2 \cdot \Delta, \ldots, k_{max}\}$.

A time increment Δ represents a fixed time interval, or time resolution, between consecutive time steps. k denotes the interval from the forecast origin to the forecasting time step, with k_{min} indicating the minimum lead time, a time gap between the forecast origin t and the first forecasting time step, and k_{max} representing the maximum lead time. Forecast horizon H specifies the number of forecasting time steps in k, which also can be expressed as $H = k_{max} - k_{min}$ (see Figure 2). The choice of the temporal components for a time series forecasting task differs due to the forecasting objective, data availability, and operational needs.

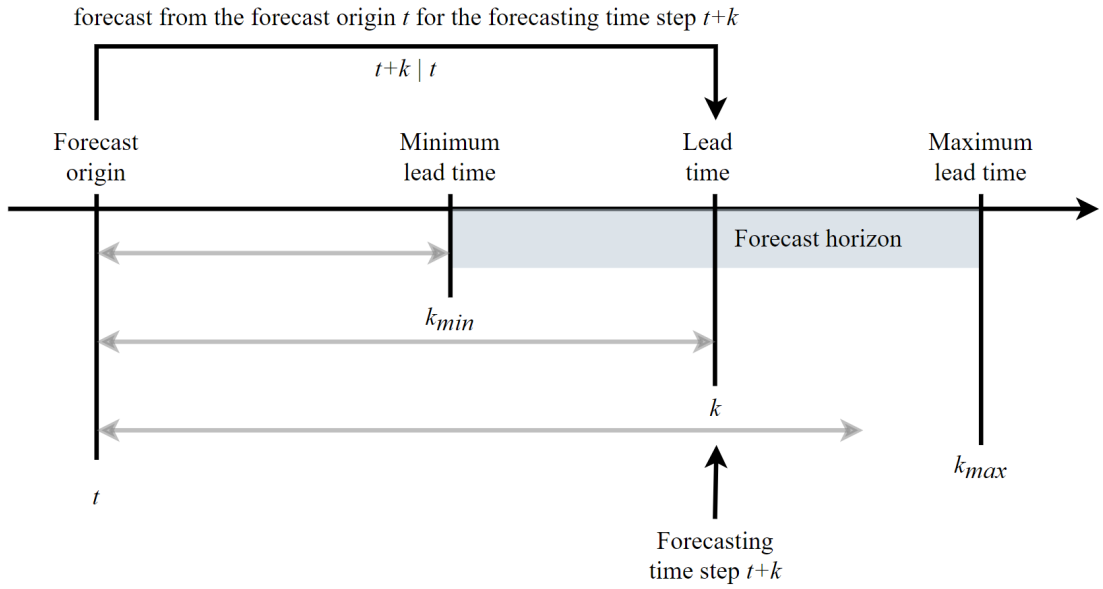


Figure 2. Nomenclature of the temporal components of the time series forecasting task.

2.3 Time Series Forecasting Models

Traditionally, fundamental statistical methods like Autoregressive Integrated Moving Average (ARIMA) or Exponential Smoothing were considered state-of-the-art for time-series modeling [JGDG06]. Despite the simplicity and interpretability of the statistical methods, many traditional parametric models cannot capture intricate non-linear relationships and struggle to model long-term dependencies.

Significant advancements in computational resources have prompted a lifted interest in Machine Learning algorithms. Over the past years, various non-parametric Machine Learning models have been widely exploited owing to their ability to learn intricate patterns within data. ML algorithms demonstrated superior performance in comparison with statistical methods. This supremacy is attributed to their capability to capture non-linear relationships in the data and their ability to utilize exogenous variables [CSB21]. These variables represent external factors that impact the target variable, which is the variable being forecasted, without being directly influenced by it in return. It indicates that variations in the exogenous variable can result in changes in the forecasted variable. However, fluctuations in the forecasted variable do not affect the exogenous variable. For instance, in the context of renewable energy forecasting, exogenous variables include weather features and temporal factors such as time of day and day of the week.

Time Series Forecasting is often a supervised problem, where true values of the target variable are employed to train ML models. Gradient Boosting (GB) methods stand out as most favored supervised learning ML algorithms for regression tasks [CSB21]. GB methods demonstrated exceptional performance relative to other popular models, such as Random Forests and Decision Trees, particularly in short-term wind energy forecasting tasks [SU21].

Regression is one of the types of supervised learning. Many specialized ML models have been developed specifically for regression tasks. Gradient Boosting Decision Trees (GBDT) are a family of gradient boosting predictive algorithms composed of a gradient boosting model and decision trees as base learners that address regression problems. GBDT models are favored by many winners of ML competitions, such as Kaggle, due to their simplicity and exceptional performance [CSB21]. Notable implementations of the GBDT algorithm are XGBoost [CG16], CatBoost [Han20], and LightGBM [GK17].

LightGBM significantly outperforms XGBoost and CatBoost in the matter of computational efficiency and memory utilization while demonstrating a similar measure of accuracy [GK17, E.19].

By default, LightGBM adopts a leaf-wise tree growth strategy, potentially enhancing performance, but poses a risk of overfitting [Shi07]. Overfitting can be mitigated by fine-tuning key parameters such as max_bin, num_leaves, feature_fraction, lambda_l1, lambda_2, max_depth, and others.

2.4 Evaluation of Time Series Models

Assessing the performance of a Machine Learning model is an essential aspect of its development. Performance evaluation is vital for understanding the model's ability to adapt to new, unseen data beyond that it was trained on, a concept known as generalization capability. Given that the training dataset exclusively comprises patterns and features already learned by the model, it does not comprehensively reflect real-world scenarios. Evaluating the model's performance on unseen data is crucial, as it demonstrates its effectiveness in projecting real-world scenarios beyond the training data.

Identifying overfitting and underfitting is another aspect of performance assessment. Overfitting occurs when the model excessively adapts to the training data, capturing noise or irrelevant patterns, potentially causing unsatisfactory performance on unseen data. Conversely, underfitting occurs when the model is too simplistic, failing to capture the underlying patterns in the data, leading to insufficient performance. Lastly, performance evaluation helps determine the optimal model parameters, ensuring more efficient model performance.

A commonly used technique for evaluating ML models is Cross-Validation (CV). K-fold CV is one of the most popular CV methods, where the dataset is divided into subsets, or folds, with one reserved for evaluation while the rest are utilized for training. This iterative process involves each subset serving as the validation set in turn, while the others function as the training data. K-fold CV assumes that observations are independent and identically distributed, allowing for random partitioning of the dataset into folds without a specific order. It ensures unbiased estimation of the model's performance across diverse data subsets.

However, time series data observations are not independent; they inherently maintain a temporal order of samples. Therefore, any method for estimating time series models must consider this characteristic by preserving the chronological relationships among observations. Cross-validation techniques, such as K-fold, are unsuitable for time series data as they disrupt the chronological order between observations.

Specific methods like Rolling Window CV and Expanding Window Cross Validation (EWCV) have been developed for time series data (see Figure 3). These techniques are designed to preserve the temporal order of the data during the model evaluation process, ensuring reliable performance assessments for this type of data. Additionally, they prevent data leakage, where future information unintentionally influences model training.

In Time Series Forecasting, data leakage occurs when future data is accidentally incorporated into the training set, granting the model access to information not available in real-world situations and compromising its ability to generalize to unencountered data. Thus, data leakage arises during K-fold validation of time series data due to data shuffling, introducing future samples into the training set. Conversely, Exponentially Weighted Cross-Validation mitigates this risk.

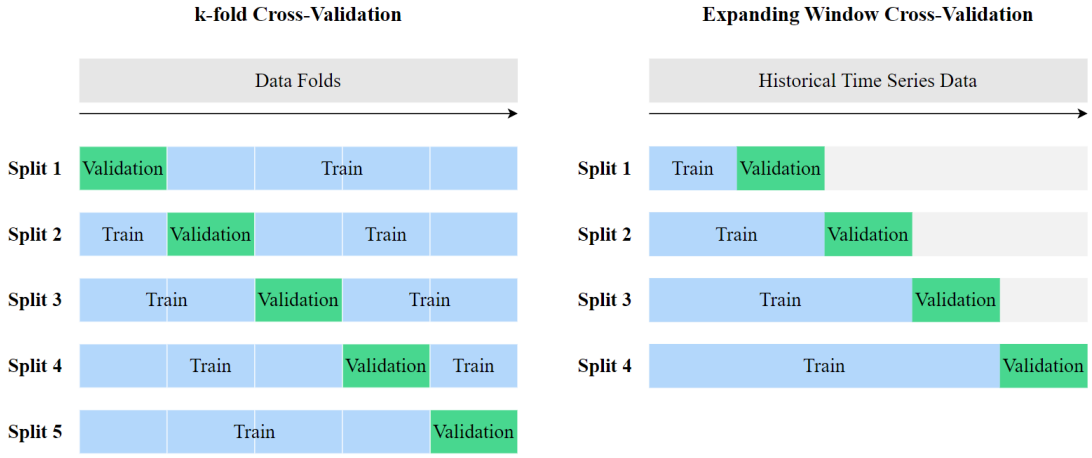


Figure 3. Comparison of K-fold Cross-Validation and Expanding Window Cross-Validation methods. In K-fold Cross-validation, observations are randomly shuffled before splitting, and models are evaluated on each fold iteratively. On the contrary, Expanding Window Cross Validation maintains the temporal order of observations throughout the process, preventing data leakage by estimating model performance on future data with respect to the training set.

In the domain of Time Series Forecasting evaluation, a range of metrics, or estimators, are employed to evaluate the effectiveness of ML models. These metrics provide quantitative assessments of accuracy and predictive performance. Frequently utilized metrics encompass Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and others. The selection of the appropriate metric depends on various factors, including the specific characteristics of the forecasting task, the nature of the data, and the researcher's preferences.

2.4.1 Mean Absolute Error

Mean Absolute Error stands as a foundational metric extensively utilized across various domains of predictive modeling. MAE estimates an average error across all samples. The Mean Absolute Error is defined as

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{2}$$

where n is a number of observations, y_i represents an actual value of the observation i, and $\hat{y}_i$ is the estimated value of the observation i.

MAE provides easily interpretable estimates, as it provides an average magnitude of errors between predicted and actual values directly related to the scale of the problem. Besides, MAE is more robust to outliers than MSE and RMSE, which makes it suitable for datasets with significant variability or with extreme values.

2.4.2 Root Mean Square Error

Root Mean Square Error estimates the square root of the average of the squared differences between actual values and forecasted. RMSE is defined as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{3}$$

where n is a number of observations, y_i represents an actual value of the observation i, and $\hat{y}_i$ is the estimated value of the observation i.

RMSE is particularly advantageous when the dataset contains outliers or when penalizing large errors more significantly, which is essential due to its ability to weigh larger errors more extremely.

2.4.3 Expanding Window Cross Validation

Expanding Window Cross Validation is a resampling technique commonly employed to evaluate time series forecasting models. The main idea of this approach is to expand the training dataset incrementally over time on a specific window size, allowing the model to learn from progressively larger sets of historical data while evaluating the model on a subsequent portion of data, the validation set. The model is retrained each iteration (see Figure 3).

The choice of the window size should consider the specific characteristics of the time series data, such as seasonality or trend, as well as the size of the available data and the objectives of the forecasting task. As the data may vary in time, assessing the model's predictive performance over diverse time periods is imperative.

Each model's performance can be finally estimated using the mean metric across all CV iterations as

$$\text{MeanMetric} = \frac{1}{v}\sum_{i=1}^{v}\text{Metric}_i, \tag{4}$$

where v is the number of validation iterations, and Metric is a chosen estimator, such as MAE.

The iterative process of the EWCV simulates real-world forecasting scenarios, where the model is continuously updated with new observations and assessed for its performance

on unseen data. EWCV provides valuable information about the model's ability to perform across various time frames of the time series data.

2.5 Feature Engineering and Feature Selection

Feature engineering is an important step in preparing data for ML model training. Its objective is to derive new features with relevant information from the data, thereby improving the model's performance and enabling a more effective representation of inherent data patterns. Feature engineering comprises various techniques, such as feature transformations like scaling or normalization, creation of novel features, encoding of categorical variables, and other methods.

However, a high number of features presents certain drawbacks. These include the curse of dimensionality, which can result in overfitting and potential multicollinearity among data features, where two or more data variables correlate with each other. These factors collectively have a negative impact on model performance and increase computational complexity.

To address the complexities associated with multiple features, feature selection methods are utilized. These methods aim to identify the most informative data features while discarding irrelevant or redundant ones within the dataset. By reducing the dimensionality of the feature space, feature selection helps to mitigate the risk of overfitting, enhancing model performance, and optimizing computational efficiency.

One feature selection method is a manual feature selection, a systematic process where a practitioner rigorously chooses features based on their understanding of the problem domain and data characteristics. This approach suggests an iterative evaluation and selection of features considered or demonstrated relevant for a specific task.

2.6 The Role of Covariates

Covariates, alternatively referred to as predictors, are additional variables incorporated in the dataset along with the historical time series data of the target variable. In Time Series Forecasting, covariates play a vital role in attributing the variance of the target variable by incorporating information about external influencing factors. These covariates are typically classified based on their temporal relationship with the primary time series into past, future, and static categories (see Figure 4). Future, past, and static covariates hold significant importance in Time Series Forecasting tasks as they encompass information of a diverse nature.

- *Past covariates* represent information available at the forecast origin time t, aiding in capturing historical events relevant to the future target.
- *Future covariates* encompass data about forthcoming events. They are used to model relationships between target and anticipated circumstances. Future

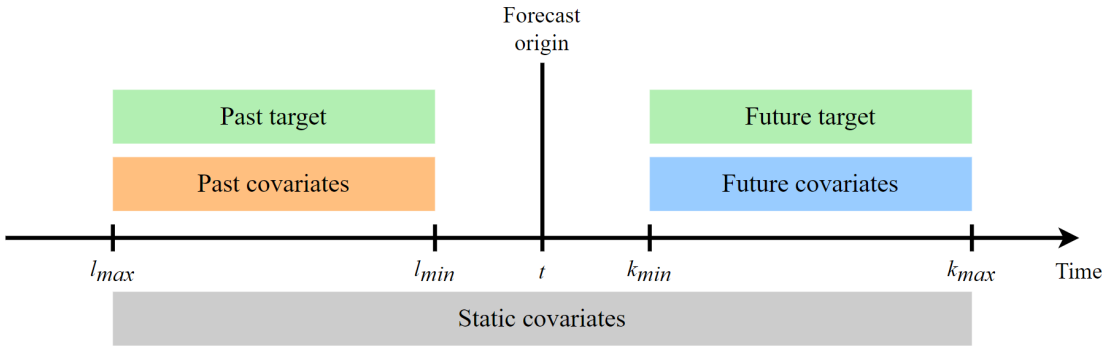


Figure 4. Temporal relationships of the past, future, and static covariates in relation to the forecast origin t, forecasting time steps k, and past time steps l. Past covariates always preceed the time of t, while future covariates exceed it.

covariates may be categorized as observable, such as holidays marked on a calendar, or unobservable and unknown at time t, as is the case with weather conditions. In the latter scenario, forecasts, such as weather predictions, serve as future covariates.

- *Static covariates* are predictors that remain constant over time.

Typically, future covariates provide information regarding forthcoming events occurring during time steps k within the forecast horizon H. Meanwhile, past covariates present information of former occurrences within the look-back window L of time steps $l = \{l_{max}, l_{max} + 1 \cdot \Delta, l_{max} + 2 \cdot \Delta, \ldots, l_{min}\}$, the size of which is selected based on temporal dynamics of the time series data.

2.7 Oracle model

Forecasted future covariates introduce additional inherent uncertainty into the dataset due to the inherent errors in the forecasting models. Models operating in real-world settings encounter challenges when incorporating forecasted future covariates, resulting in reduced prediction accuracy and model performance compared to ideal conditions with error-free data.

The oracle model is employed to assess a model's performance under ideal conditions. In the oracle model, uncertainty originating from the future covariates is eliminated by substituting them with the actual values of the respective predictors. In such a manner, the oracle model has a complete knowledge of the future.

The practical application of the oracle model aims to establish an empirical lower performance boundary for a given forecasting problem. It demonstrates the best achievable performance of a time series forecasting model $g(\cdot)$ under the condition of a perfect

knowledge of the future. Any error reported is solely attributed to $g(\cdot)$ and estimated model parameters $\hat{\Theta}$ alone.

The oracle model requires a distinct input dataset as it utilizes data of a different nature. Instead of using forecasted future covariates, the oracle model assumes complete knowledge of the future and thus utilizes actual observed values instead of forecasted. Although the oracle model's demand for complete knowledge of the future makes it unrealistic, establishing it as a benchmark for ideal conditions provides valuable information about the capabilities and limitations of real-world forecasting tasks.

2.8 Net Consumption

2.8.1 Consumers, Prosumers, and Net Purchasing Systems

Energy companies or electricity retailers employ Consumption Metering Systems to monitor the consumption of electricity. They are installed at individual households, business sites, or other entities that utilize electrical energy to power their operations. These meters record the total amount of consumed energy over a specific period, such as monthly or quarterly.

As end-users of electricity, consumers interact directly with the centralized grid, or energy grid, to satisfy their power requirements, thus operating without their own microgrids. However, prosumers operate within their own microgrid (see Figure 5), a decentralized electrical network that incorporates distributed energy resources like solar panels, batteries, and generators. Additionally, control systems are integrated within the microgrid to manage power generation, storage, and consumption. These microgrids interact with the centralized grid through bidirectional energy exchange, import and export, while maintaining the ability to operate autonomously, particularly during grid outages.

Establishing bidirectional relations between prosumers' microgrids and the centralized grid requires transitioning from traditional unidirectional meters, like the Consumption Metering System, to Net Systems. Two widely used Net Systems for metering are Net Purchasing Systems and Net Metering Systems. The Consumption Metering System measures only the incoming flow of electricity, denoted as C^{ttotal}. In contrast, the Net Purchasing System monitors energy flows at the prosumers' microgrid connection points to the energy grid using two separate unidirectional meters. These meters record energy flows in both directions: from the centralized grid to the prosumer's microgrid (import), denoted as C^{grid}, and from the microgrid to the centralized grid (export), denoted as P^{grid}. The Net Metering System, a bidirectional system, registers only the net difference between imports from and exports to the grid. In the Baltics, however, Eesti Energia employs only Net Purchasing Systems, and thus, our focus is on this type of the Net Systems.

Conventional Consumption Metering and Net Purchasing Systems serve primarily

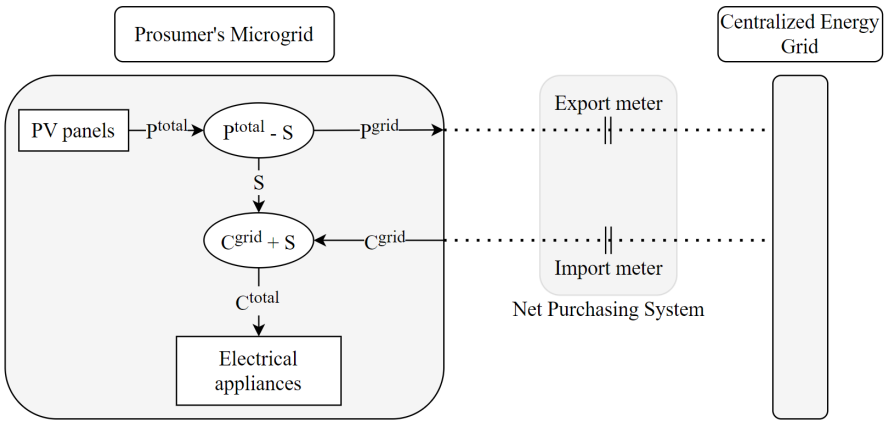


Figure 5. Depiction of the relationships between the total and measured variables within a prosumer's microgrid and their interrelation with the Centralized Energy Grid. Energy generated by PV panels, denoted as Total Electricity Production P^{total}, is partially consumed before its transmission to the Centralized Energy Grid. This partial consumption is called Self-consumption, denoted as S. If the locally generated electricity is insufficient to fulfill the electricity demands, supplementary energy is acquired from the grid, represented as the Grid Energy Consumption C^{grid}. Total Consumption C^{total} is a combination of S and C^{grid}. Electricity exchanged between a prosumer's microgrid and the Centralized Energy Grid is metered by the Net Purchasing System, which records P^{grid} and C^{grid} separately.

for billing purposes. On-site consumption, or Self-consumption S, the Total Energy Production P^{total}, and Total Energy Consumption C^{total} are typically not measured as they are irrelevant for billing. Metering of S, P^{total}, and C^{total} at the sources of production and consumption require additional metering equipment and infrastructure, leading to increased expenses for both prosumers and energy companies.

2.8.2 Net Consumption Definition

Net Consumption Y, also referred to as Net Load or Net Energy, represents the difference in Electricity Production P and Electricity Consumption C within a specific time period t. Y can be derived as

$$Y_t^{total} = P_t^{total} - C_t^{total} \tag{5}$$

when metering is performed at both the production and consumption sources. However, when metering is conducted at the grid connection point, it is expressed as

$$Y_t^{grid} = P_t^{grid} - C_t^{grid}. \tag{6}$$

The difference between total and metered values corresponds to the Self-consumption S, energy consumed from the energy generated by prosumers. Consequently, the relationships between P^{total} and P^{grid} can be expressed as

$$P_t^{total} = P_t^{grid} + S_t, \tag{7}$$

and the relationships between C^{total} and C^{grid} as

$$C_t^{total} = C_t^{grid} + S_t \tag{8}$$

where P_t^{total} and C_t^{total} denote the total values of produced and consumed energy, respectively, measured at the source, and P_t^{grid} and C_t^{grid} represent values measured at the grid connection point.

The consistency of the Net Consumption value remains unchanged regardless of the meter placement. This can be verified using Equations (5), (6), (7), and (8) as illustrated below:

$$Y_t^{total} = P_t^{total} - C_t^{total} = (P_t^{grid} + S_t) - (C_t^{grid} + S_t) = P_t^{grid} - C_t^{grid} = Y_t^{grid} = Y_t. \tag{9}$$

This consistent behavior of Y, regardless of the meter placement, ensures methodological comparability across various studies of Net Consumption.

2.8.3 Forecasting methods

Several approaches to the Net Forecasting Task have been explored in the literature. One of the popular methods, as shown in [SERM20] and [CX24], concerns extension of the Load Forecasting Task to the Net Load Forecasting by including historical weather data W and weather forecast W', typically used in Production Forecasting Models, to the Consumption Forecasting Models to be utilized in a Net Forecasting Model for a direct estimation of the Net Consumption variable.

In [AK16], the researchers introduced and compared two fundamentally different methodologies to the Net Load Forecasting: Additive and Integrated Models (see Figure 6a). While in the Additive method, Net Consumption is obtained by subtracting consumption estimations from production estimations generated by separate models, the Integrated approach suggests developing a Net Forecasting Model for a direct Net Consumption prediction.

Lastly, in [YWX18], the authors employed a data-driven approach, proposing the decomposition of the Net variable into three distinct components: P^{total}, C^{total}, and residual. Each component is subsequently forecasted with separate models, with the outcomes aggregated to obtain Net Consumption.

2.8.4 Additive and Integrated methods for forecasting Net Consumption

Additive model. The Additive method for the Net Forecasting involves the development of two distinct models. The first model is a Production Forecasting Model g^{prod} that generates forecast $\hat{p}$ of Grid Production P^{grid} or Total Production P^{total} using an information set Ω^{prod} excluding any data from the consumption information set Ω^{cons}. The Consumption Forecasting Model g^{cons} objective is to estimate $\hat{c}$ of Grid Consumption C^{grid} or Total Consumption C^{total} utilizing Ω^{cons} excluding information from Ω^{prod}. Subsequently, Net Consumption Y is derived from $\hat{p}$ and $\hat{c}$ in accordance with Equation (5) or Equation (6), depending on the nature of production and consumption data.

Integrated model. Integrated method for Net Forecasting proposes developing a unified model g^{net} for direct estimation of Y that extends information set Ω^{cons} of the Consumption Forecasting Model, designed for the Additive method, by injecting $\hat{p}$, generated by g^{prod} of the Additive method, forming the information set Ω^{net}.

3 Methodology

The main distinction between our study and the setting in [AK16], further referred to as the original study, lies in the nature of employed production and consumption data. While [AK16] operates with P^{total} and C^{total}, our research operates with P^{grid} and C^{grid} in all experiments. According to the Equations (7) and (8), the unmonitored Self-consumption S equals the difference between total and grid energy values.

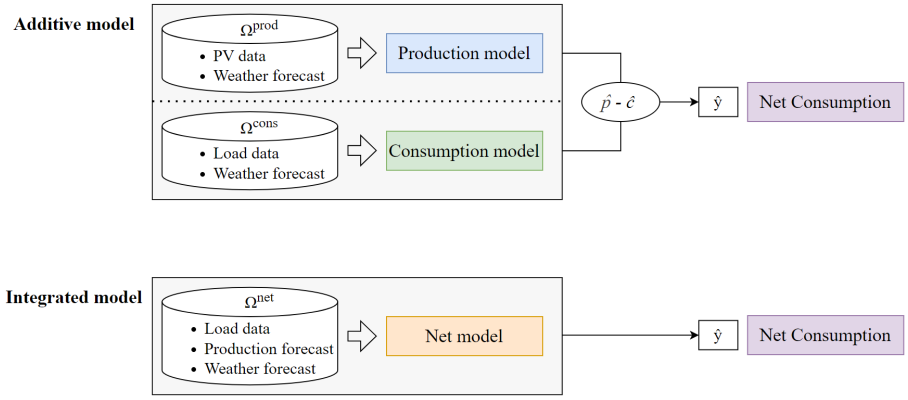
Modeling the relationships between P^{grid}, weather patterns, and solar positioning presents a higher level of complexity compared to P^{total}. When only Net Purchasing Systems are employed, only P^{grid} is accessible, while P^{total} and S remain unknown. The absence of the Self-consumption component S in P^{grid} introduces uncertainty about S into the P^{grid} forecasting task. On the contrary, with additional on-site metering at the sources of production and consumption, both P^{total} and S become available. Since P^{total} comprises both P^{grid} and S (see Equation (7)), there is no uncertainty from S in the P^{total} forecasting task. It makes the task of forecasting P^{total} comparatively simpler than forecasting P^{grid}. Therefore, the absence of S requires its implicit consideration to forecast P^{grid} effectively.

Similarly, modeling the relationships between C^{grid} and various consumption patterns observed in private households presents a higher level of complexity compared to C^{total}. The dependency of C^{grid} on the component S, as defined in Equation (8), introduces additional uncertainty in forecasting C^{grid} when on-site metering is unavailable. Conversely, both C^{total} and S variables are known when on-site metering is in place. As C^{total} encapsulates S, forecasting C^{total} involves no uncertainty related to S, simplifying the forecasting process compared to forecasting C^{grid}. Effective forecasting of C^{grid} needs an implicit consideration of the S variable.

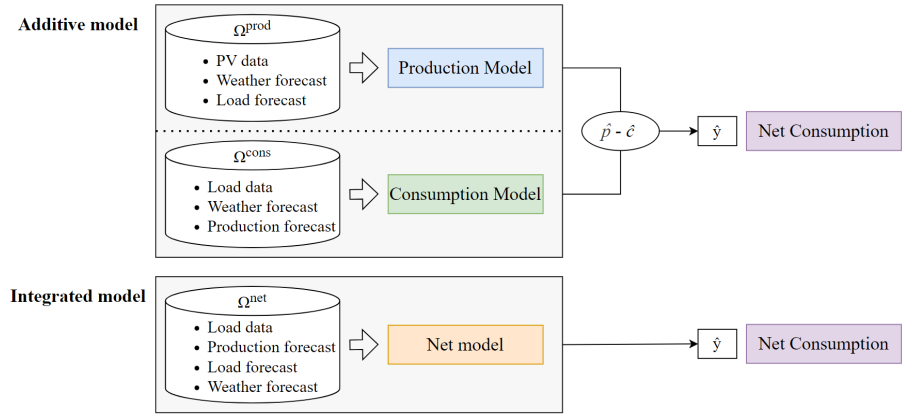
In households without batteries, which represent a majority in the dataset of interest, locally generated electricity is typically consumed immediately (self-consumption) to meet the immediate demand, thereby reducing values of the Grid Electricity Consumption C^{grid}. Conversely, raised C^{grid} may indicate a shortfall in on-site generation P^{total}, leading to lesser values of P^{grid} sent to the grid.

In our study, we hypothesize that the enhanced modeling of the underlying relationships between S and P^{grid} can be achieved by incorporating Consumption forecast as an input variable for forecasting P^{grid}. Similarly, improving the modeling of the underlying relationships between S and C^{grid} can be accomplished by integrating Production forecast as an input variable for forecasting C^{grid}. Our hypothesis aims to mitigate uncertainty associated with Self-consumption in the context of Net Energy Forecasting for prosumers.

Our work proposes enhancements to the Additive and Integrated methods by introducing variations of the Additive and Integrated Models (see Figure 6b). The improvements are designed to increase the accuracy of the Production, Consumption, and Net Consumption Models by integrating forecast incorporation techniques into the existing framework.



(a) In the original study, the Additive method derives Net Consumption as a difference of P and C according to the Equation (5), generated separately, employing different data sets. In the Integrated approach, a unified Net Forecasting Model is designed for direct Net Consumption Forecasting. The Integrated Net Consumption model is trained on the input data of the Consumption Model, extended by a production forecast.



(b) In our Additive method, $\hat{p}$ is estimated by g^{prod} employing Ω^{prod} extended by $\hat{c}$, and $\hat{c}$ is estimated by g^{cons} having Ω^{cons} extended by $\hat{p}$. $\hat{y}$ is subsequently calculated according to Equation (6). In our Integrated method, $\hat{y}$ is generated by a unified Net Model, which incorporates both $\hat{p}$ and $\hat{c}$ into Ω^{net}.

Figure 6. Abstract representation of the Additive and Integrated Net Consumption Forecasting Models of the original research (6a) and our study (6b).

In the following subsection, we elucidate our Additive approach and the Production and Consumption Models it involves. Next, we present our Integrated methodology for Net Consumption Forecasting. We will compare our Additive and Integrated Models

with those proposed in the original study, which serve as reference models.

3.1 Additive method

3.1.1 Production model

In our suggested Additive method, point estimation $\hat{p}$ of the variable P^{grid} is determined as

$$\hat{p}_{t+k|t} = \mathbb{E}[P_{t+k}|g^{prod}, \Omega_t^{prod}, \hat{\Theta}^{prod}], \tag{10}$$

where $\hat{\Theta}^{prod}$ are estimated parameters of the Production Forecasting Model g^{prod}. Point estimation $\hat{c}$ of the variable C^{grid} is incorporated into information set Ω^{prod} so that it comprises:

- $\mathcal{P} = \{p_{t-l_{max}}, p_{t-l_{max}+1}, \ldots, p_{t-l_{min}}\} = \{p_{t-i}\}_{i=0}^{L}$ - a set of historical values of the target variable P^{grid} of the number of time steps observed in the past within look-back window L.
- $\mathcal{W}' = \{\hat{\mathbf{w}}_{t+k_{min}|t}, \hat{\mathbf{w}}_{t+k_{min}+1|t}, \ldots, \hat{\mathbf{w}}_{t+k_{max}|t}\} = \{\hat{\mathbf{w}}_{t+k|t}\}_{k=1}^{H}$ - a set of the future values of the weather variables that is provided by *numerical weather prediction* (NWP) models as forecasts for time steps k within the forecast horizon H.
- $\mathcal{C}' = \{\hat{c}_{t+k_{min}|t}, \hat{c}_{t+k_{min}+1|t}, \ldots, \hat{c}_{t+k_{max}|t}\} = \{\hat{c}_{t+k|t}\}_{k=1}^{H}$ - a set of the future values of Consumption estimated by the Consumption Model g^{cons}, adapted from the original study, for time steps k within the forecast horizon H.

3.1.2 Consumption model

In our proposed Additive method, $\hat{c}$ is estimated as

$$\hat{c}_{t+k|t} = \mathbb{E}[C_{t+k}|g^{cons}, \Omega_t^{cons}, \hat{\Theta}^{cons}], \tag{11}$$

where $\hat{\Theta}^{cons}$ are estimated parameters of the Consumption Forecasting Model g^{cons}. Information set Ω^{cons} is extended with $\hat{p}$, encompassing:

- $\mathcal{C} = \{c_{t-l_{max}}, c_{t-l_{max}+1}, \ldots, c_{t-l_{min}}\} = \{c_{t-i}\}_{i=0}^{L}$ - a set of the historical values of the target variable C^{grid} of the number of time steps observed in the past within look-back window L.
- $\mathcal{W}' = \{\hat{\mathbf{w}}_{t+k_{min}|t}, \hat{\mathbf{w}}_{t+k_{min}+1|t}, \ldots, \hat{\mathbf{w}}_{t+k_{max}|t}\} = \{\hat{\mathbf{w}}_{t+k|t}\}_{k=1}^{H}$ - a set of the future values of the weather variables that is provided by NWP models as forecasts for time steps k within the forecast horizon H.
- $\mathcal{P}' = \{\hat{p}_{t+k_{min}|t}, \hat{p}_{t+k_{min}+1|t}, \ldots, \hat{p}_{t+k_{max}|t}\} = \{\hat{p}_{t+k|t}\}_{k=1}^{H}$ - a set of the future values of P^{grid} that are estimated by g^{prod}, adapted from the original study, for time steps k within the forecast horizon H.

3.2 Integrated method

The Integrated Net Model g^{net}, introduced in the original study, is an extension of the Consumption Model g^{cons} that appends $\hat{p}$ in Ω^{cons}, resulting in Ω^{net}. The Integrated method proposed in our study further appends the predictions, $\hat{p}$ and $\hat{c}$, into the information set Ω_t^{net} of the Net Forecasting Model g^{net}.

Our approach's logic is designed to enable the Net Model to learn the patterns of both forecasts, $\hat{p}$ and $\hat{c}$, including their error patterns. External data, like weather forecasts, is also present in Ω_t^{net}. It provides additional information that enables g^{net} to model relationships between Y and weather variables. This dependency emerges inherently because Y is derived from the variable P, which is strongly influenced by weather conditions, and C. A comparison of both the original and our proposed methods is depicted in Figure 7.

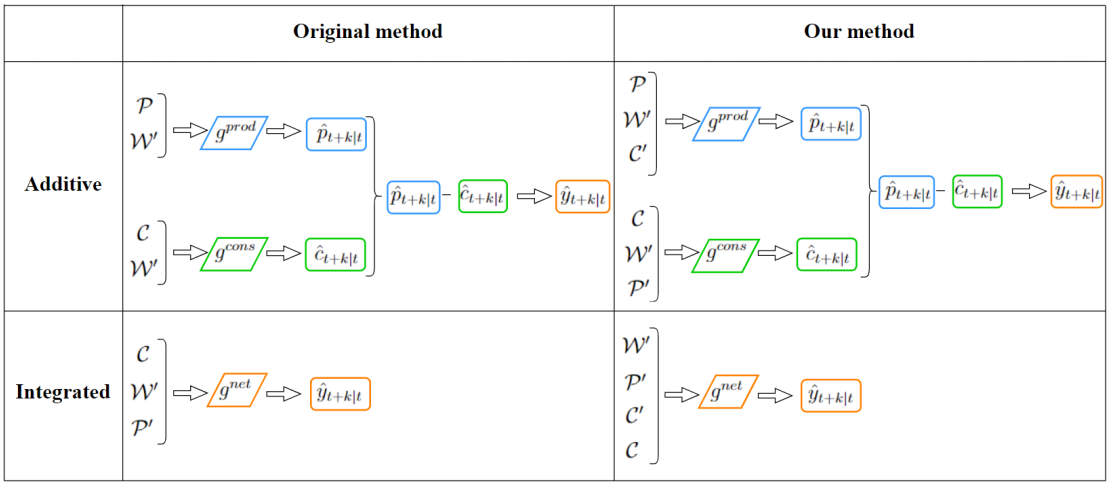


Figure 7. Comparison of the Additive and Integrated methods from the original study and our research. A key distinction is in their respective inputs to the models, also referred to as information sets Ω^{prod}, Ω^{cons}, and Ω^{net}. In our proposed methodology, Ω^{prod} incorporates additional component $\hat{c}$, Ω^{cons} integrates $\hat{p}$, and Ω^{net} combines $\hat{p}$ and $\hat{c}$, in contrast to the inputs of the Additive and Integrated methods introduced in [AK16].

In our suggested Integrated Net Model, estimation $\hat{y}$ of the Net variable Y is determined as

$$\hat{y}_{t+k|t} = \mathbb{E}[Y_{t+k}|g^{net}, \Omega_t^{net}, \hat{\Theta}^{net}], \tag{12}$$

where $\hat{\Theta}^{net}$ are the estimated parameters of the Net Forecasting Model g^{net}.

Production estimations $\hat{p}$ and consumption estimations $\hat{c}$ are incorporated into Ω^{net}, as shown in Figure 7. Resulting Ω^{net} comprises:

- $\mathcal{C} = \{c_{t-l_{max}}, c_{t-l_{max}+1}, \ldots, c_{t-l_{min}}\} = \{c_{t-i}\}_{i=0}^{L}$ - a set of the historical values of C^{grid} within the number of time steps observed in the past within look-back window L.

- $\mathcal{P}' = \{\hat{p}_{t+k_{min}|t}, \hat{p}_{t+k_{min}+1|t}, \ldots, \hat{p}_{t+k_{max}|t}\} = \{\hat{p}_{t+k|t}\}_{k=1}^{H}$ - a set of P^{grid} future values that are estimated by our proposed g^{prod}, explained in Section 3.1.1, for time steps k within the forecast horizon H.

- $\mathcal{C}' = \{\hat{c}_{t+k_{min}|t}, \hat{c}_{t+k_{min}+1|t}, \ldots, \hat{c}_{t+k_{max}|t}\} = \{\hat{c}_{t+k|t}\}_{k=1}^{H}$ - a set of the future values of C^{grid} that are provided by our suggested g^{cons}, introduced in Section 3.1.2, for time steps k within the forecast horizon H.

- $\mathcal{W}' = \{\hat{\mathbf{w}}_{t+k_{min}|t}, \hat{\mathbf{w}}_{t+k_{min}+1|t}, \ldots, \hat{\mathbf{w}}_{t+k_{max}|t}\} = \{\hat{\mathbf{w}}_{t+k|t}\}_{k=1}^{H}$ - a set of the future values of the weather variables that NWP models provide as forecasts for time steps k within the forecast horizon H.

Given that the Integrated Net Model requires estimations of both P^{grid} and C^{grid} as input variables, a prerequisite for its designing involves the development of Production and Consumption Models.

4 Experiments

This section outlines all experiments conducted with the data and models. Initially, we demonstrate the experiments related to data preparation, covering weather data aggregation, feature engineering, and feature selection. Subsequently, we present the naive baseline models, followed by Additive and Integrated models, adapted from [AK16] and proposed in our study.

In the scope of our research, the task of interest is the hourly day-ahead forecast, having $\Delta = 1$ hour and $H = 24$. The forecast origin t is 9:00 UTC, with $k_{min} = 13$ and $k_{max} = 37$ so that $k = \{13, 14, 15, \ldots, 37\}$. This configuration remains constant throughout all experiments, with k, H, and Δ consistently following the specified definitions. During daylight saving time, the forecast origin t is adjusted to 8:00 UTC. The size of the look-back window L varies, and the past time steps l differ across experiments. These parameters are explicitly defined for each specific case.

4.1 Data

4.1.1 Datasets

In our study, we operate with the weather forecast, denoted as W', historical weather W, and Estonian prosumers data. Prosumers data includes variables such as P^{grid} and C^{grid} and details about the prosumers themselves, such as the number of prosumers and the capacity of their installed PV panels. Additionally, we use production and consumption predictions as elaborated in Section 3.

Weather forecast. Weather forecasts $W' = \{\hat{\mathbf{w}}\}_{L \times D}$, where L is the number of weather variables (see Table 1) and D is a number of coordinates, are obtained from the European Centre for Medium-Range Weather Forecasts (ECMWF). The forecasts are issued daily at 00:00 UTC for 48 hours into the future for $D = 2059$ distinct grid coordinates, covering the entire territory of Estonia (see Figure 8). Notably, no anomalies were detected in the acquired dataset.

All weather features, except Direct solar radiation, Surface solar radiation downwards, Snowfall, and Total precipitations, are estimated for the end of the 1-hour period. On the other hand, these specific weather variables are presented as accumulations over the 1-hour period.

Historical weather. Historical weather $W = \{\mathbf{w}\}_{L \times D}$, where L is the number of weather features (see Table 2) and D is a number of coordinates (see Figure 9), is acquired from the Open-Meteo API service [1]. For each hour, historical weather variables are

[1] https://open-meteo.com/en

Table 1. Table of the ECMWF weather forecast variables.

Variable	**Units**	**Description**
Temperature	K	Air temperature at 2 meters above the ground.
Dew point	K	Dew point temperature at 2 meters above the ground.
Cloud cover low	0 to 1	The proportion of the sky covered by clouds in the 0-2 km altitude range.
Cloud cover mid	0 to 1	The proportion of the sky covered by clouds in the 2-6 km altitude range.
Cloud cover high	0 to 1	The proportion of the sky covered by clouds in the 6+ km altitude range.
Cloud cover total	0 to 1	The proportion of the sky covered by clouds in the all altitude range.
u, v wind components	m/s^1	Eastward (u) and northward (v) components of the wind 10 meters above surface.
Direct solar radiation	J/m^2	Amount of direct radiation from the Sun reaching the surface on a plane perpendicular to the direction of the Sun.
Surface solar radiation downwards	J/m^2	Amount of solar radiation, both direct and diffuse, that reaches a horizontal plane at the surface of the Earth.
Snowfall	m	Accumulated snow that falls to the Earth's surface.
Total precipitation	m	Accumulated rain and snow that falls to the Earth's surface.

available for $D = 112$ distinct coordinates, presented in a grid format covering the entire territory of Estonia. No anomalies were detected in the dataset.

All weather features, except Diffuse solar radiation, Shortwave solar radiation, Snowfall, and Rain precipitations, are estimated for the end of the 1-hour period. Snowfall and Rain are provided as accumulations over the 1-hour period, and Diffuse solar radiation and Shortwave solar radiation are averaged across the preceding 1-hour period.

EE prosumers data. The prosumers data contains P^{grid} and C^{grid} data of Eesti Energia prosumers. Available data variables are listed in Table 3. The prosumers are aggregated by a location on a county level, an origin (business building and private residences), and a contract type (Combined, Fixed, General service, Spot). Entities with a number of prosumers less than five are not included in the data due to the low quality of such data. The resulting dataset does not contain artifacts, such as missing values, outliers, or duplicates.

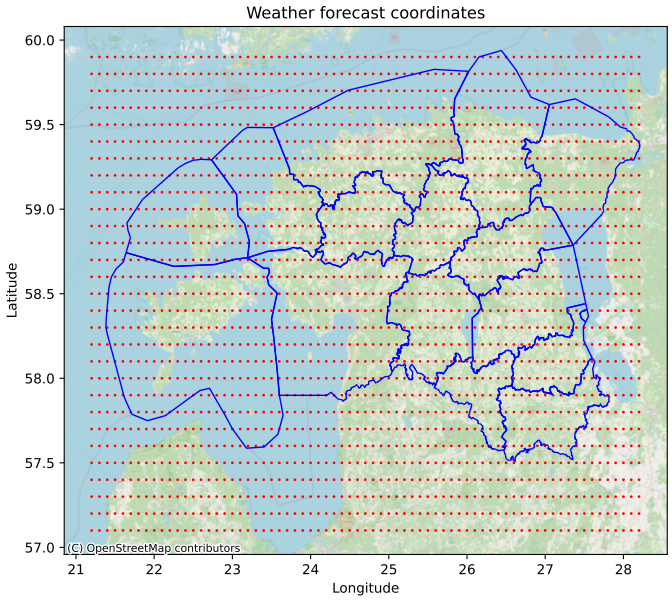


Figure 8. Coordinates (red) represent the locations of the ECMWF weather forecast, illustrated for an individual time step. The blue lines delineate the boundaries of Estonian counties.

4.1.2 Weather data aggregation

The acquired weather forecast dataset contains forecasts of each weather variable for 2059 distinct coordinates per time step, while the historical weather data comprises historical values of each weather observation for 112 distinct coordinates per time step. We aggregated the weather forecast and the historical weather observations to obtain a matching dimensionality of the datasets, resulting in a single value of each weather variable for each of the 15 counties per time step. Also, by this aggregation, we decreased the dataset size, reducing the number of input features for forecasting models.

The aggregation steps of the historical weather dataset include:

- Selection of coordinates within the terrestrial area of Estonia.
- Computation of a mean value for each weather variable within each county boundary.

For the weather forecast dataset, the steps involve:

Table 2. Table of the historical weather variables.

Variable	**Units**	**Description**
Temperature	C	Air temperature at 2 meters above the ground.
Dew point	C	Dew point temperature at 2 meters above the ground.
Surface pressure	hPa	Air pressure at surface.
Cloud cover low	%	The percentage of the sky covered by clouds in the 0-3 km altitude range.
Cloud cover mid	%	The percentage of the sky covered by clouds in the 3-8 km altitude range.
Cloud cover high	%	The percentage of the sky covered by clouds in the 8+ km altitude range.
Cloud cover total	%	The percentage of the sky covered by clouds in the all altitude range.
Wind speed	m/s	Wind speed at 10 meters above the ground.
Wind direction	°	Wind direction at 10 meters above the ground.
Diffuse solar radiation	W/m^2	Average amount of diffuse radiation.
Shortwave solar radiation	W/m^2	Average amount of radiation, both direct and diffuse, that reaches a horizontal plane at the surface of the Earth.
Snowfall	cm	Accumulated snow that falls to the Earth's surface.
Rain	mm	Accumulated rain that falls to the Earth's surface.

- Reduction of the grid size to maintain consistency with the historical weather dataset, selecting 112 coordinates identical to those from the historical weather dataset.
- Selection of coordinates within the terrestrial area of Estonia.
- Computation of a mean value for each weather forecast variable within each county boundary.

The resulting representation of the averaged weather variables corresponds to the averaged coordinates, depicted in Figure 10.

Dimensionality of the weather forecast $W' = \{\hat{\mathbf{w}}\}_{L \times D}$ is reduced from $L \times D = 12 \times 2059$ to $L \times D = 12 \times 15$, and dimensionality of the historical weather $W = \{\mathbf{w}\}_{L \times D}$ changed from $L \times D = 13 \times 112$ to $L \times D = 13 \times 15$, where L is a number of weather variables of a corresponding dataset, and $D = 15$, constant both both datasets, correspond to the number of Estonian counties.

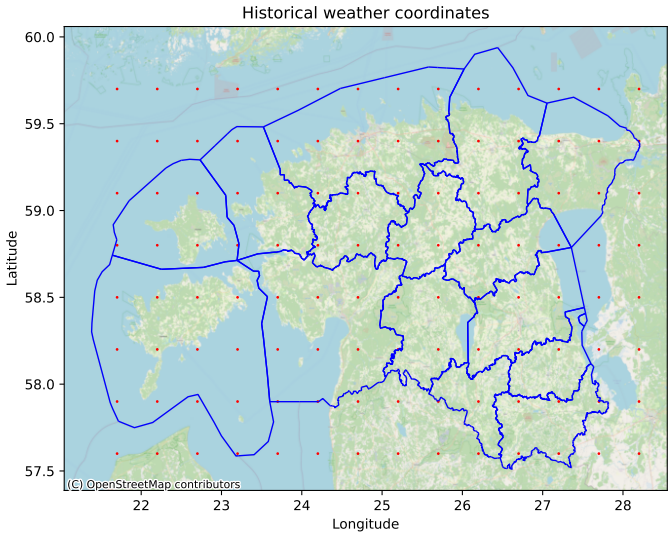


Figure 9. Coordinates (red) represent the locations of the historical weather dataset, illustrated for an individual time step. The blue lines delineate the boundaries of Estonian counties.

4.1.3 Feature engineering and Feature selection

In our work, features are manually engineered, relying on domain knowledge. Weather features are engineered similarly for both datasets. The full list of engineered features is available in Table 4). The main aspects of the experimental feature engineering process are:

- Derivation of new variables serving as targets, such as Relative Production for g^{prod} and Net Consumption for g^{net}, as they are not present in the original dataset.
- One-hot encoding of categorical variables, considering the independence of the categorical variables in the dataset.
- Standardization of features to ensure uniform units convention, such as Celsius (C°) for temperature and dew point, millimeters (mm) for rain and snow, and Watts per square meter (W/m^2) for solar radiation.

Table 3. Data of Estonian prosumers, provided by Eesti Energia.

Variable	Unit	Description
County	-	Categorical. 15 categories.
Origin	-	Categorical. 2 categories
Contract	-	Categorical. 4 categories.
Grid Production	kWh	The production amount for the relevant segment (county-origin-contract) accumulated over the 1-hour period.
Grid Consumption	kWh	The consumption amount for the relevant segment accumulated over the 1-hour period.
Installed capacity	kW	Installed photovoltaic solar panel capacity of the relevant segment.
Estimated yearly Grid Consumption	kW	Amount of electricity consumed preceding year of the relevant segment.
Prosumers number	-	The aggregated number of prosumers in the relevant segment.

- Derivation of new features based on historical targets, weather variables, and temporal information, listed in Table 4.

The feature selection process is conducted manually and subsequently validated with EWCV. Features for Ω^{prod} and Ω^{cons} are chosen independently. Since Ω^{net} is an extension of Ω^{cons}, its features are not explicitly selected. The selected data features for Ω^{prod}, Ω^{cons}, and Ω^{net} can be observed in Table 5. "+" indicates the variable's presence in the information set, while "-" indicates its absence.

4.2 Experiments with Models

4.2.1 Naive baselines

Baseline models establish a fundamental benchmark that subsequently developed models aim to surpass. An interpretable and simple baseline is crucial for setting a reasonable upper threshold for error. This section focuses on experimentation with baseline models, presenting distinct approaches to forecasting Production, Consumption, and Net.

Naive models do not have equivalent oracle predictions. To obtain an oracle prediction, it is required to have additional information beyond the historical time series of target variables. As in our study naive models do not utilize any additional information beyond the historical time series of target variables, there is no additional information

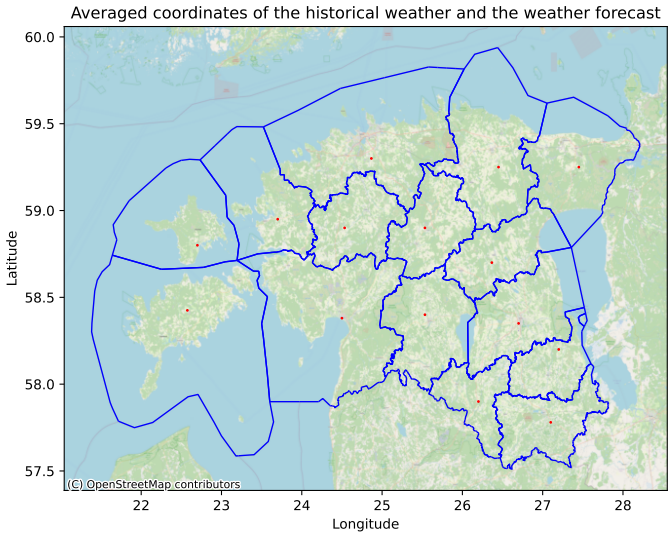


Figure 10. Averaged coordinates (red) representing coordinates of both datasets, historical weather and weather forecast, for an individual time step. Given that the weather forecast utilizes the same 112 coordinates as the historical dataset for aggregation, the averaged coordinates remain consistent across both datasets.

to be learned beyond them. Therefore, the oracle model would replicate the same predictions as the naive model, providing no improvement.

Naive Production. As solar energy implies energy derived from the sun's irradiation, the solar energy generated by the PV panels is influenced by various factors, including weather conditions and the position of the sun. Utilizing recent observations as predictions for the future provides a sensible baseline foundation for P^{grid}, as it considers the time series aspect without overly complicating the approach.

Production estimation $\hat{p}_{t+k|t}$ is assigned with the average historical value p_l of the three most recent known days, forming look-back window L, preceding the forecast origin t at each time step k within the forecast horizon H:

$$\hat{p}_{t+k|t} = \mathbb{E}[P_{t+k}|g^{prod}, \Omega_t^{prod}, \Theta^{prod}], \tag{13}$$

Table 4. Table of the constructed data features.

Variable	**Unit**	**Description**
Relative Production	0 to 1	Relation of produced energy to the capacity of installed PV panels.
Net Consumption	-	A difference between electricity production and consumption (see Equation (6)).
Dew point depression	C°	A difference between dew point and temperature. Indicates the potential for cloud formation, fog, and precipitation.
Temperature previous hour	C°	Temperature at the start of the hour.
Dew point previous hour	C°	Dew point at the start of the hour.
Wind u, v components	-	Derived from wind speed and wind direction of W to complement u, v wind wind components in W'.
Rain	mm	Derived from Snowfall and Total precipitation of W' to complement Rain in W.
Precipitation binary	0 or 1	Indicates whether precipitation occurred within a preceding hour.
Hours since precipitation	-	Indicates number of hours passed from the last occurrence of precipitation.
Time	-	Hour, week, month, day in a year, day of the week. Temporal information.
Sun azimuth and Sun elevation	-	Features indicating position of the sun, calculated from coordinates and time, at the middle and at the end of the hour.
Consumption t-168	kWh	Historical consumption, 168h (7*24h) before the forecasting time step $t + k$.
Production t-96, t-168	kWh	Historical production, 168h (7*24h) and 96h (4*24h) before the forecasting timestamp $t + k$. $t - 168$ represents the most recent historical data available.

where g^{prod} is the Naive Production Model, the information set is defined as $\Omega_t^{prod} = \{\mathcal{P}\}$, $\mathcal{P} = \{p_{l_{max}}, p_{l_{max}+1}, \ldots, p_{l_{min}}\} = \{p_{t-i}\}_{i=0}^{L}$, look-back window $L = 3 * 24$, $l_{max} = 131$ and $l_{min} = 83$, and Θ^{prod} comprises a constant parameter $\frac{1}{|L|}$. Alternatively, the estimator

Table 5. Information sets Ω^{prod}, Ω^{cons}, and Ω^{net}.

	Ω^{prod}	Ω^{cons}	Ω^{net}
County, origin, contract	+	+	+
Temperature	+	+	+
Temperature previous hour	+	+	+
Dew point	+	-	-
Dew point previous hour	+	-	-
Dew point depression	+	+	+
Cloud cover (low, mid, high)	+	-	-
Shortwave solar radiation	+	-	-
Direct solar radiation	+	+	+
Binary precipitation	-	+	+
Hours since last precipitation	+	-	-
Wind u, v components	+	-	-
Prosumers number	-	+	+
Estimated yearly Grid Consumption	-	+	+
Time components (hour, day in a year, day of the week)	-	+	+
Sun azimuth and Sun elevation mid hour	+	-	-
Historical Production $t+k-4*24$	+	-	-
Historical Consumption $t+k-7*24$	-	+	+

of P can be represented as

$$\hat{p}_{t+k|t} = \frac{1}{|L|} \sum_{i=l_{min}}^{l_{max}} P_{t-l}. \tag{14}$$

Naive Consumption. The choice of a baseline approach for C^{grid} is established on the understanding that the prosumer consumption behavior demonstrates predictable patterns influenced by temporal factors like the hour of the day and day of the week. These patterns emerge from consistent human activities, including work, sleep, television viewing, and other daily activities. Utilizing data from the corresponding hours of

the seventh day preceding the forecasting timestamp, this method effectively captures temporal fluctuations in consumption.

At the forecast origin t, consumption estimation $\hat{c}_{t+k|t}$ of time steps k within the forecast horizon H is assigned with historical consumption values c_f of the corresponding day of the preceding week, forming look-back window L, as

$$\hat{c}_{t+k|t} = \mathbb{E}[C_{t+k}|g^{cons}, \Omega_t^{cons}], \tag{15}$$

where g^{cons} is the Naive Production Model, the information set is defined as $\Omega_t^{cons} = \{\mathcal{C}\}$, $\mathcal{C} = \{c_{l_{max}}, c_{l_{max}+1}, \ldots, c_{l_{min}}\} = \{c_{t-i}\}_{i=0}^{L}$, look-back window $L = 24$, $l_{max} = 155$ and $l_{min} = 131$. The Naive Model estimator of C^{grid} is defined as

$$\hat{c}_{t+k|t} = C_{t-l}. \tag{16}$$

Naive Net Consumption. Extracting consistent temporal patterns from the Net Energy data poses a challenge as it incorporates both variables P^{grid} and C^{grid}. Consequently, to establish a baseline for Y, an additive approach is adopted: Net Consumption estimation $\hat{y}_{t+k|t}$ is derived from $\hat{p}_{t+k|t}$ and $\hat{c}_{t+k|t}$, leveraging the Equation (6).

4.2.2 Oracle

As discussed in Section 2.7, the oracle model requires a distinct set of data, where future covariates, represented by W', are substituted with actual historical values W. Therefore, one dataset preserves real-life conditions by aligning target variables P_k, C_k, and Y_k with weather forecast variables W'_k at the corresponding time steps k. In contrast, another dataset, created for training oracle models, contains historical weather data W_k instead of W'_k.

One exception had to be considered. The Direct solar radiation variable, present in W', is absent in W. Therefore, it cannot be replaced, and it is used in the dataset for oracle models. Another aspect to consider is the distinction in the nature of cloud cover data. Although cloud cover variables in W' and W are provided for different altitudes, this factor is neglected.

4.2.3 Baselines for Additive and Integrated Models

In the [AK16], two methodologies for Net Consumption Forecasting are introduced: Additive and Integrated. These methodologies are tested on a microgrid of the University of California campus with high solar penetration, where solar energy covers 33% of the annual energy demand. Results indicated that the integrated model outperformed the additive model by 10.69% in terms of RMSE for the Support Vector Machine model.

In our experiment, the framework of the original study is adapted to the aggregated data of Estonian prosumers. The objective is to asses how well the proposed framework

can be adjusted to our study, considering: a) the use of P^{grid} and C^{grid} instead of P^{total} and C^{total} as target variables in g^{prod} and g^{cons}, respectively; b) aggregated data of numerous prosumers; c) Estonian study case considering the fact that weather conditions in Estonia differ from those in the original study. Therefore, all models developed in this experiment will serve as reference models for those in our methodology.

Within this investigation, the look-back window $L = 24$, and past time steps $l_{max} = 83$ and $l_{min} = 59$.

Additive method. Estimations of variables P^{grid} and C^{grid} are determined as

$$\hat{p}_{t+k|t} = \mathbb{E}[P_{t+k}|g^{prod}, \Omega_t^{prod}, \hat{\Theta}^{prod}] \tag{17}$$

and

$$\hat{c}_{t+k|t} = \mathbb{E}[C_{t+k}|g^{cons}, \Omega_t^{cons}, \hat{\Theta}^{cons}]. \tag{18}$$

These equations closely resemble our Additive methodology proposed earlier (see Equations (10) and (11)), with a primary difference lying in their respective information sets Ω_t^{prod} and Ω_t^{cons}. For Production Forecasting, historical production time series and weather forecasts serve as inputs, represented by $\Omega_t^{prod} = \{\mathcal{P}, \mathcal{W}'\}$. Similarly, historical consumption time series and weather forecasts, $\Omega_t^{cons} = \{\mathcal{C}, \mathcal{W}'\}$, are utilized in Consumption Forecasting.

Historical values of P and C are denoted by $\mathcal{P} = \{p_{t-l_{max}}, p_{t-l_{max}+1}, \ldots, p_{t-l_{min}}\}$ and $\mathcal{C} = \{c_{t-l_{max}}, c_{t-l_{max}+1}, \ldots, c_{t-l_{min}}\}$, respectively, with L as previously noted. Weather forecast variables $\mathcal{W}' = \{\hat{\mathbf{w}}_{t+k_{min}|t}, \hat{\mathbf{w}}_{t+k_{min}+1|t}, \ldots, \hat{\mathbf{w}}_{t+k_{max}|t}\}$ are also included.

In addition, oracle variations of g^{prod} and g^{cons} are developed. $\hat{p}_{t+k|t}$ and $\hat{c}_{t+k|t}$ are estimated as demonstrated in Equations (17) and (18), where weather forecast data W' is substituted with the corresponding observable historical weather values $\mathcal{W} = \{\mathbf{w}_{t+k_{min}|t}, \mathbf{w}_{t+k_{min}+1|t}, \ldots, \mathbf{w}_{t+k_{max}|t}\}$. This results in $\Omega_t^{prod} = \{\mathcal{P}, \mathcal{W}\}$ and $\Omega_t^{cons} = \{\mathcal{C}, \mathcal{W}\}$. Particular weather variables are specified in Table 5.

Additive approach involves deriving the Net Consumption estimations $\hat{y}$ from estimations $\hat{p}$ and $\hat{c}$. Thus, $\hat{y}$ is determined as

$$\hat{y}_{t+k|t} = \hat{p}_{t+k|t} - \hat{c}_{t+k|t}. \tag{19}$$

The oracle estimation of $\hat{y}$ follows the same procedure, utilizing oracle estimations of $\hat{p}_{t+k|t}$ and $\hat{c}_{t+k|t}$.

Integrated method. As suggested in [AK16], the Integrated Net Forecasting Model incorporates solar energy forecasts into Ω_t^{cons}. We adopt the same approach to investigate its applicability to aggregated data of Estonian prosumers. Thus, Ω_t^{net} is an extension of Ω_t^{cons}, incorporating solar energy forecasts denoted as P'.

In the Integrated model, a unified g^{net} is adopted to estimate $\hat{y}$, expressed as

$$\hat{y}_{t+k|t} = \mathbb{E}[Y_{t+k}|g^{net}, \Omega_t^{net}, \hat{\Theta}^{net}], \tag{20}$$

where the information set $\Omega_t^{net} = \{\mathcal{C}, \mathcal{P}', \mathcal{W}'\}$ includes historical consumption time series $\mathcal{C} = \{c_{t-l_{max}}, c_{t-l_{max}+1}, \ldots, c_{t-l_{min}}\}$, future estimations of the Production variable $\mathcal{P}' = \{\hat{p}_{t+k_{min}|t}, \hat{p}_{t+k_{min}+1|t}, \ldots, \hat{p}_{t+k_{max}|t}\}$, and additional weather forecast data $\mathcal{W}'$, specified as $\mathcal{W}' = \{\hat{\mathbf{w}}_{t+k_{min}|t}, \hat{\mathbf{w}}_{t+k_{min}+1|t}, \ldots, \hat{\mathbf{w}}_{t+k_{max}|t}\}$.

In the oracle version, the forecasted future covariates $\mathcal{W}'$ are replaced with their actual values, $\mathcal{W}$. Similarly, the production forecast $\mathcal{P}'$ generated by g^{prod} is replaced with $\mathcal{P}'$ issued by the oracle g^{prod}. This substitution eliminates the inherent uncertainty associated with weather forecasts. Consequently, Ω_t^{net} is defined as $\Omega_t^{net} = \{\mathcal{C}, \mathcal{P}', \mathcal{W}\}$, where historical weather data $\mathcal{W} = \{\mathbf{w}_{t+k_{min}|t}, \mathbf{w}_{t+k_{min}+1|t}, \ldots, \mathbf{w}_{t+k_{max}|t}\}$.

4.2.4 Enhanced Additive and Integrated Models for Net Forecasting

In this section, we detail the experiments conducted with the Additive and Integrated models incorporating proposed advancements.

Within this experiment, the look-back window $L = 24$, and past time steps $l_{max} = 83$ and $l_{min} = 59$.

Additive method: Production model. The Production Forecasting Model g^{prod} is employed to estimate $\hat{p}$ using input features from Ω^{prod} as demonstrated in Equation (10). Ω^{prod} comprises selected input features (see Table 5), in addition to the future values of C^{grid}. Thus, $\Omega_t^{prod} = \{\mathcal{P}, \mathcal{C}', \mathcal{W}'\}$, where $\mathcal{C}' = \{\hat{c}_{t+k_{min}}, \hat{c}_{t+k_{min}+1}, \ldots, \hat{c}_{t+k_{max}}\}$. C' is obtained from the Consumption Model g^{cons} created for the Additive method, adapted from the original study, described in Section 4.2.3.

The oracle estimates of $\hat{p}$ are generated using g^{prod} with an oracle dataset, where $\mathcal{W}$ replaces $\mathcal{W}'$. Future estimations of $\mathcal{C}'$ are replaced by those generated by the oracle g^{cons}, thereby eliminating the inherent uncertainty associated with weather forecasts. Consequently, the information set for the oracle model is defined as $\Omega_t^{prod} = \{\mathcal{P}, \mathcal{C}', \mathcal{W}\}$.

Additive method: Consumption model. $\hat{c}$ is predicted using the Consumption Forecasting Model g^{cons} with an expanded information set Ω^{cons}. In addition to the input features selected for the Consumption Model, as listed in Table 5, $\hat{p}$ is included as an additional input variable. The estimates $\hat{p}$ are generated using g^{prod} of the original study. Consequently, $\Omega_t^{cons} = \{\mathcal{C}, \mathcal{P}', \mathcal{W}'\}$, where $\mathcal{P}' = \{\hat{p}_{t+k_{min}}, \hat{p}_{t+k_{min}+1}, \ldots, \hat{p}_{t+k_{max}}\}$.

The oracle estimations $\hat{c}$ are generated by the model g^{cons} using the oracle dataset, where $\mathcal{W}$ replaces $\mathcal{W}'$. Additionally, P' contains future estimations $\hat{p}$, generated by the oracle g^{prod}, adapted from the original study. Consequently, the information set for the oracle g^{cons} is $\Omega_t^{cons} = \{\mathcal{C}, \mathcal{P}', \mathcal{W}\}$.

Additive method: Net Consumption. The Additive approach involves deriving the Net estimation $\hat{y}$ from $\hat{p}$ and $\hat{c}$, as illustrated in Equation (19). This procedure is repeated to obtain an oracle estimation of $\hat{y}$, employing oracle forecasts of $\hat{p}$ and $\hat{c}$.

Integrated method: Net Consumption Model. In the Integrated method, the Net Consumption variable Y is directly forecasted by the Net Forecasting Model g^{net}, outlined in Equation (12). This experiment upgrades the Integrated Model from the experiment detailed in Section 4.2.3 by extending Ω^{net} with an additional input variable, $\hat{c}$. Here, $\Omega_t^{net} = \{\mathcal{C}, \mathcal{P}', \mathcal{C}', \mathcal{W}'\}$, where future Consumption values $\mathcal{C}' = \{\hat{c}_{t+k_{min}}, \hat{c}_{t+k_{min}+1}, \ldots, \hat{c}_{t+k_{max}}\}$ are generated using g^{cons} developed according to our Additive method, and future Production values $\mathcal{P}' = \{\hat{p}_{t+k_{min}}, \hat{p}_{t+k_{min}+1}, \ldots, \hat{p}_{t+k_{max}}\}$ are generated by our g^{prod}.

In the oracle model adaptation, Ω^{net} incorporates data from the dataset prepared for the oracle models. $\hat{y}$ is estimated using $\Omega_t^{net} = \{\mathcal{C}, \mathcal{P}', \mathcal{C}', \mathcal{W}\}$, where forecasted future covariates $\mathcal{P}'$, $\mathcal{C}'$ are generated with our oracle g^{prod} and g^{cons}, proposed in our methodology.

4.3 Implementation details

Given the focus on a supervised regression problem, the LightGBM model is used in the experiments, excluding naive baselines, owing to its efficiency in processing large datasets. Since LightGBM models are limited to single-output forecasting, individual time steps within the horizon H are forecasted using distinct LightGBM models.

The Expanding Window Cross-Validation method with the MAE and RMSE metrics is implemented to assess the predictive models' performance. The EWCV process involves the evaluation of 12 validation and 12 test subsets with 1-month length each, covering each month in a year. Validation sets start from December 2022 to November 2023, while test sets start from January 2023 to December 2023. The initial training set contains data from September 2021 to November 2022, extending it by one additional month for each iteration. Both metrics, MAE and RMSE, are calculated for each data subset. This validation approach ensures the assessment of model performance across different weather conditions and seasonal variations inherent in energy production and consumption patterns.

All trainable Production Models g^{prod} utilized Relative Production (see Table 4) as the target variable, calculated by dividing Grid Production by the installed PV panel capacity, offering a standardized measure for Production Forecasting Models. This approach, particularly suitable for aggregated data of numerous prosumers, ensures more consistent and accurate predictions. After forecasting, resulting predictions are then rescaled back to the Grid Production magnitude by multiplication with the corresponding installed PV panel capacity. For evaluation purposes, predictions of all models, g^{prod}, g^{cons}, and g^{net}, are aggregated across all counties, origins, and contracts (see Table 3),

resulting in aggregated variables P, C, and Y, along with their respective estimations $\hat{p}$, $\hat{c}$, and $\hat{y}$, used for MAE and RMSE metrics calculations.

The experiments were conducted using a combination of the Databricks and the local environments. In the Databricks environment, experiments focusing on weather data aggregation were conducted due to the large size of the historical weather and weather forecast datasets. Data aggregation experiments employed the Python programming language (version 3.10.12) and various libraries for data manipulation, machine learning, and visualization. All other experiments, encompassing feature engineering and ML model development, were conducted in the local environment with Python version 3.11.4. Table 6 lists the libraries used in the experiments across both environments, including their respective versions.

Table 6. Versions of Python libraries of the local and Databricks environments.

Library	**Version, local**	**Version, Databricks**
pandas	2.0.3	1.5.3
numpy	1.24.3	1.23.5
sklearn	1.3.0	-
plotly	5.18.0	-
lightgbm	4.0.0	-
matplotlib	3.7.2	3.7.0
seaborn	0.12.2	0.12.2
plotly	5.18.0	-
requests	-	2.28.1
geopandas	-	0.14.4
shapely	-	2.0.4
contextily	-	1.6.0
pyspark	-	3.5.0
contextily	-	1.6.0

In all experiments, the random seed was configured to the value 42. The specific library versions and the random seed value ensure reproducibility and compatibility with the experimental setup.

5 Results and discussion

In this section, we present the results of all developed models, including Naive baselines. First, we report the results of the Production Models. Then, we discuss the performance of the Consumption Models. Lastly, we review Net Consumption Models, paying most attention to Additive and Integrated methods of Net Consumption Forecasting. Each model, including Naive baselines, is evaluated with the Expanding Window Cross-Validation, having MAE and RMSE calculated for train, validation, and test sets at each CV iteration.

5.1 Production Models

In our study, we devised and assessed five distinct models for forecasting Electricity Production. These models are categorized into three groups: Naive Production Model, reference Production Model, adapted from [AK16], and Production Model utilizing the methodology proposed in our study. The latter two models are trainable LightGBM models, while the Naive Production Model functions as an estimator based on average values. Additionally, we developed oracle models for each trainable Production Model.

The results depicting the performance of all Production Models across both the validation and test subsets are provided in Table 7, with the best-performing metrics highlighted in bold. As discussed in Section 4.2.1, Naive Models lack oracle implementations, thereby excluding performance metrics for them. Both Production Forecasting Models, adapted from the original study and introduced in our research, exhibit enhancements over the foundational benchmark established by the Naive Model. This empirical evidence confirms that both models can capture intricate patterns within P^{grid} and accurately model the relationships between P^{grid} and other variables, surpassing the capabilities of the Naive model.

Table 7. The performance of Production Models on validation and test sets, averaged across 12 CV iterations.

Model	**MAE validation/test**	**MAE oracle validation/test**	**RMSE validation/test**	**RMSE oracle validation/test**
Naive	11011.89 / 11020.72	-	13505.94 / 13515.53	-
Reference	**1692.89 / 1702.46**	**1430.19 / 1454.60**	**3358.39 / 3382.87**	**2843.82 / 2894.30**
Our	1765.55 / 1817.99	1470.38 / 1520.55	3417.15 / 3488.84	2853.25 / 2938.75

Across both metrics, MAE and RMSE, it is evident that oracle models outperform other models on both the validation and test data subsets. Eliminating uncertainty originating from weather forecasts reduced models' errors by 14.5% (reference model) and 16.4% (our model) in terms of RMSE and by 14.4% (reference model) and 15.8% (our model) in terms of MAE on the test set. This observation emphasizes the potential for maximum improvement if weather forecasts were error-free. Furthermore, this finding validates the implementation approach of oracle models as performed in our study.

The performance evaluation based on Mean Absolute Error indicated no improvement of our Production Model over the reference Production Model. There is a 4.3% increase in errors on the validation set and a 6.8% increase on the test set. This lack of improvement suggests that including Consumption forecasts in our method did not enhance the Production Model's efficiency; instead, it introduced additional complexity. This complexity may be due to multicollinearity of the forecasted Consumption $\hat{c}$ with other input variables, as the information set Ω^{cons} of g^{cons}, generated $\hat{c}$ for our Production Model, shares several weather input variables with Ω^{prod} of our g^{prod} (see Table 5). Other potential reasons include overfitting and the low quality of Consumption forecasts, which fail to represent inherent Self-consumption patterns accurately. Future research should analyze the correlation between the new variable and other input variables, identifying and excluding redundant features. Additionally, feature importance techniques could assess its contribution to the model's forecast.

We observed a similar outcome in the oracle implementations of Production models. MAE analysis showed that the oracle reference model surpassed our oracle model by 2.7% on the validation set and 4.3% on the test set. Although the discrepancy between model performances decreased in the absence of weather forecast uncertainty, it persisted. This persistence suggests that the primary source of performance variance may arise from the complexity introduced by the additional input variable and potential multicollinearity. However, it is possible that the observed discrepancy could be attributed to noise within the model and the data.

The noise associated with the LightGBM model originates from the stochastic nature of the gradient boosting algorithm, including random feature and data point sampling during tree construction and the random initialization of tree parameters. Additionally, noise in the data encompasses various sources, such as errors or inaccuracies in forecast variables arising from forecasting methods or the underlying data. Moreover, data noise may include fluctuations or anomalies within historical data, along with external factors not captured but impacting the observed data.

Likewise, the reference model exhibited superior performance compared to our model by 1.7% on the validation set and 3.0% on the test set in terms of RMSE. Moreover, the disparity in performance between the oracle reference model and our oracle model did not surpass 0.3% on the validation set and 1.5% on the test set. This minimal difference further supports the hypothesis that the observed variation is attributed to noise inherent

in both the model and the data, certifying that the incorporated Consumption forecast did not provide any additional information to the Production model.

Furthermore, we observe consistent MAE and RMSE performances across validation and test sets for trainable models, suggesting efficient generalization abilities. This consistency extends to the Naive Model, certifying the preservation of the temporal relationship between past and target variables alongside the effectiveness of the chosen validation and test window sizes.

Another direction for assessing the impact of the Consumption variable on g^{prod} involves the creation of an oracle model complementary to the one developed in this study. Besides removing uncertainty stemming from weather forecasts, this supplementary oracle model could also mitigate uncertainty arising from Consumption forecasts. The Consumption forecast utilized in our oracle Production model inherently lacked weather-related uncertainty, as it was generated by the oracle Consumption model. Thus, the only uncertainty in the input data of our oracle Production model originated from the Consumption Forecasting Model. Substituting this forecast with historical observed Consumption values could further advance the investigation of relationships between P^{grid}, C^{grid}, and S under conditions of perfect knowledge of future Consumption values.

5.2 Consumption Models

In analogy to Production Forecasting Models, five distinct models were developed and evaluated for Consumption Forecasting: Naive baseline and trainable models, adapted from [AK16] and proposed in our study. Likewise, each trainable model has an oracle implementation.

The performance evaluation of the Consumption Models is illustrated in Table 8, where the bold values represent the top performance metrics across all models. Notably, the Naive Model lacks an oracle implementation, as detailed in Section 4.2.1. Thus, the results do not include performance metrics for oracle Naive baselines. Both trainable Consumption Models, developed for the original and our proposed methodologies, have empirically exhibited higher efficiency compared to the Naive Baseline. This observation suggests that both trainable models possess a superior capacity for modeling the intricate relationships among the target variable C^{grid}, historical C^{grid} values, and external variables.

All oracle Consumption Models demonstrated decreased errors compared to their real-case scenario implementations. In terms of MAE, the test error decreased by 2.6% (reference model) and 4.3% (our model), while in RMSE, the reduction was 2.7% (reference model) and 4.9% (our model) in the oracle implementations compared to the non-oracle ones. This expected outcome empirically validates the reliability of the developed oracle models, representing the best achievable performance under perfect knowledge of future weather conditions. The more noticeable improvement in the oracle models developed within our framework can be attributed to eliminating inherent

Table 8. The performance of Consumption Models on validation and test sets, averaged across CV iterations.

Model	**MAE validation/test**	**MAE oracle validation/test**	**RMSE validation/test**	**RMSE oracle validation/test**
Naive	5940.15 / 6108.87	-	7311.88 / 7528.14	-
Reference	2306.15 / 2432.17	2253.83 / 2369.83	3274.72 / 3398.46	3197.70 / 3306.69
Our	**2238.15 / 2350.40**	**2121.50 / 2250.17**	**3143.27 / 3255.20**	**2977.18 / 3094.28**

weather-related uncertainty from the Production forecast, a factor not present in the reference models.

The MAE metric revealed that our Consumption model exhibited superior performance compared to the reference model, showing an improvement of 2.9% on the validation set and 3.7% on the test set, averaged across 12 CV iterations. However, given the relatively small percentage difference in the models' performances, it may be considered negligible. It suggests that including the Production forecast in the input of our Consumption model resulted in minimal or insignificant additional information for the model. Alternatively, the slight variance in model accuracy may be caused by noise, originating from both the model itself and the data, as discussed in Section 5.1.

Remarkably, the oracle implementation of our Consumption model demonstrated the superiority of our approach. Our model exhibited a performance advantage of 5.9% on the validation set and 5.0% on the test set compared to the oracle reference model. In the oracle models, the uncertainty associated with weather variables is eliminated. Under these conditions, our Consumption Model may have had an enhanced capability to model the relationships among C^{grid}, weather variables, and Self-consumption. However, further exploration is necessary to understand better how the model addressed each variable. This investigative process may encompass feature importance analysis alongside more sophisticated analytical methodologies.

The RMSE evaluation exhibited similar observations to those made with MAE. Our Consumption model demonstrated a reduction in error of 4.0% on the validation set and 4.2% on the test set, suggesting no definitive superiority of our method. However, the RMSE metrics from the oracle implementations revealed a more substantial error reduction of 6.9% on the validation set and 6.4% on the test set. This trend suggests that our method may enhance the efficiency of the Consumption model by potentially capturing the intricate relationships among C^{grid}, weather variables, and Self-consumption.

However, further investigation is needed to confirm this assumption.

The performance of both models, ours and the one from the original study, exhibited only a slight reduction from the validation to test subsets across both MAE and RMSE metrics. It suggests a satisfactory generalization ability, indicating an efficient window size for the data subsets. The oracle implementation exhibited a favorable generalization ability, showing a performance reduction from the validation set to the test set of 5.15% for the oracle compared to 5.5% for non-oracle implementations in the original study model, and 6.07% for the oracle compared to 5.01% for non-oracle in our model. Similar trends were observed for Naive models, suggesting stability in the relationship between past consumption patterns and target consumption values.

5.3 Net Consumption Models

In our study, we developed various types of Net Forecasting Models. Specifically, we designed 2 Integrated Net Models, each corresponding to the original and our suggested methods, and their oracle implementations. Additionally, we derived Net Consumption using the Additive approach five times, starting with the Naive Baseline, followed by the original and our methods and their oracle implementations. In total, nine Net Consumption Models are created across all experiments.

Table 9 illustrates the evaluation metrics of all Net Consumption Models, averaged across 12 EWCV iterations. In the table, the bold values indicate the best performance metrics across all models, while the underlined values highlight the best-performing models within their respective categories (Additive and Integrated). These metrics represent the models with the lowest MAE and RMSE on validation and test sets, including oracle scenarios. Notably, metrics for the oracle Naive Model are absent from this table, as we did not create oracle implementations for Naive models, as elaborated in Section 4.2.1. Trainable models, devised for the original and our proposed methodologies, demonstrated superior performance compared to the Naive Baseline. This superiority suggests their enhanced ability to capture the underlying patterns and complexities inherent in the dataset.

All oracle Integrated Net Models demonstrated superior performance compared to their non-oracle implementations, confirming the validity of the developed oracle models. Specifically, oracle Additive Net Models demonstrated inherent superiority by leveraging oracle Production and Consumption Models, as discussed in Sections 5.1 and 5.2, which exhibited superior performance. Oracle Integrated Net Models demonstrated a reduction in MAE on the test set by 9.1% for both the reference and our models, as well as a decrease in RMSE by 10.7% (reference model) and 10.9% (our model). These oracle models effectively mitigated uncertainty from weather forecasts and the inherent uncertainty associated with weather in $\hat{p}$ and $\hat{c}$, utilizing $\hat{p}$ and $\hat{c}$ generated by oracle g^{prod} and g^{cons}.

Table 9. The performance of Net Consumption Models on validation and test sets, averaged across CV iterations.

Model or method	**MAE validation/test**	**MAE oracle validation/test**	**RMSE validation/test**	**RMSE oracle validation/test**
Naive	15476.10 / 15640.73	-	18844.77 / 19056.68	-
Reference Additive	**3637.29** / **3771.14**	3262.76 / **3404.90**	**5759.14** / **5856.99**	**5061.90** / **5169.35**
Our Additive	3691.71 / 3849.97	**3254.84** / 3429.21	5838.22 / 5982.38	5080.47 / 5246.28
Reference Integrated	3886.10 / 4343.88	3450.68 / 3948.92	6085.97 / 6641.32	5327.43 / 5929.71
Our Integrated	3839.22 / 4337.28	3433.75 / 3942.97	6071.45 / 6653.62	5317.51 / 5930.40

The Integrated Net Model proposed in our methodology exhibited minimal improvement over the reference Integrated Net Model in terms of MAE. Our model outperformed the reference by 1.2% on the validation set and 0.2% on the test set. The discrepancy in errors between our model and the reference Integrated Net Model may be attributed to noise inherent in the data or arising during the training process of the model. It suggests that the predictions of $\hat{p}$ and $\hat{c}$ appended to the input of our Integrated Model did not contribute significantly meaningful information to the Integrated Net Forecasting Model.

The identical observation applies to the oracle implementations of the Integrated Net Models. Achieving a reduction in error of 0.5% on the validation set and 0.2% on the test set can be solely attributed to noise. It suggests that the model failed to learn the relationships among $\hat{p}$, $\hat{c}$, and their associated errors effectively and that the uncertainty stemming from weather forecast did not influence the models' ability to do it. This limitation may originate from the intricate and dynamic nature of production and consumption forecasts and the uncertainties in these predictions. Moreover, the model may inadequately capture the complex relationships between $\hat{p}$ and $\hat{c}$, resulting in unsatisfactory performance even with eliminated uncertainty associated with weather. Further investigation is required to understand the influence of input data on the model's output and the relationships between them.

The RMSE metric supported the findings observed with the MAE metric for Integrated Net Models. It revealed that our Integrated Net Model surpassed the reference model by 0.2% on the validation set, whereas the reference model outperformed ours

by 0.2% on the test set. A similar pattern emerged in the oracle models, with our oracle Integrated Net Model outperforming the oracle reference model by 0.2% on the validation set. In contrast, the oracle reference model outperformed our oracle model by less than 0.1% on the test set. The negligible nature of these differences further confirms the conclusion that the Production and Consumption forecasts did not provide any additional information to the Integrated Net Model and that the Integrated Net Model failed to model the relationships between the forecasts.

It is worth noting that our Integrated Model employed $\hat{p}$ and $\hat{c}$ generated by our g^{prod} and g^{cons}. While our g^{cons} demonstrated slightly better accuracy, our g^{prod} performed worse than the reference Production Model. Using $\hat{p}$ based on its accuracy, specifically employing $\hat{p}$ from the g^{prod} model adapted from the original study, could have influenced the performance of our Integrated Net Model.

Additive models, whether adopted from the original study or proposed in our research, involve the derivation of Net Consumption estimates $\hat{y}$ from predictions $\hat{p}$ and $\hat{c}$. In our Additive Net Model, $\hat{y}$ is derived from $\hat{p}$ and $\hat{c}$ generated by the Production and Consumption Models developed for our method. Conversely, in the reference Additive Net Model, predictions $\hat{p}$ and $\hat{c}$ are produced using Production and Consumption models adapted from the original study. Our Additive Net Model exhibited slightly higher MAE errors on both the validation and test subsets than the original, with a test set error increase of 2.1% in both MAE and RMSE. However, the differences in performance can be attributed to noise within the models and the data. The MAE and RMSE performance differences of the reference and our oracle Additive Net Models did not surpass 1.5%, indicating no improvement under conditions of eliminated uncertainty associated with weather.

Integrated Net Models exhibited a shift in performance from the validation to the test set. The reference model demonstrated a decrease of 10.5% in MAE and 10.2% in RMSE, while our model experienced a decline of 11.5% in MAE and 10.3% in RMSE. It suggests that the length of the validation and test subsets may not be optimal for capturing changes in the underlying historical data distribution and short-term data fluctuations. Another plausible explanation could be overfitting in the developed Integrated Net Models (see Figure 11). Although, the Additive Models inherited the patterns observed in their corresponding Production and Consumption Models, demonstrating sufficient generalization abilities.

In the original study, the Integrated Net Model outperformed the Additive Net Model by 10.69% in terms of RMSE. In our adaptation of the original study, the Additive Net Model outperformed the Integrated Net Model by 12.81%. At the same time, Additive's oracle estimation maintained the superiority over the Integrated Model by 12.82%, both in terms of RMSE metrics on the test set. In our proposed methodology, we observed a similar trend: the Additive Model surpassed the Integrated Model by 10.09%, while the oracle Integrated Model outperformed the oracle Additive Model by 11.54%. The

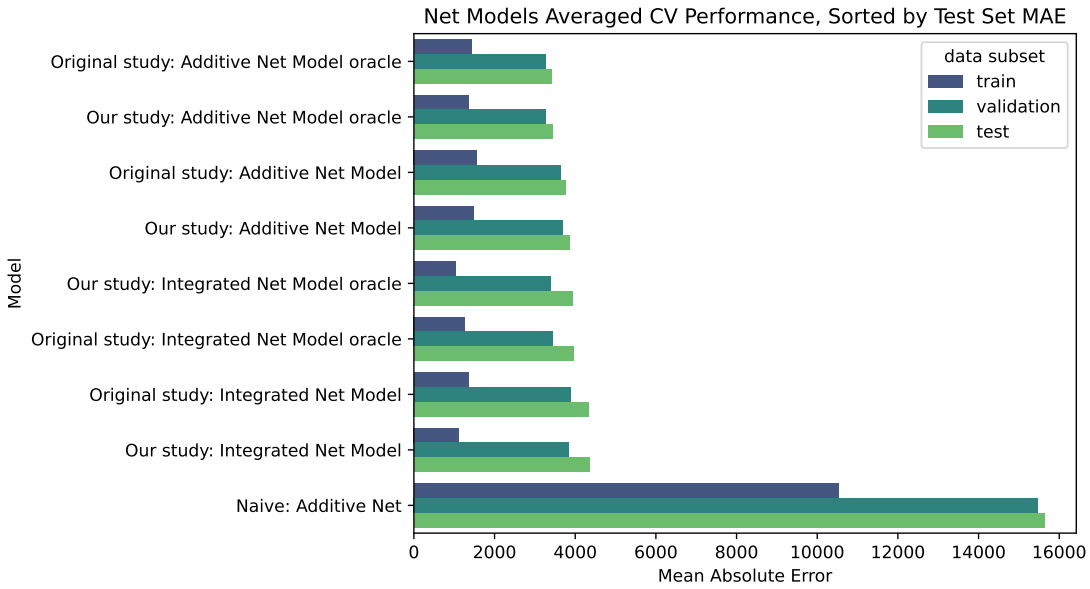


Figure 11. Comparison of train, validation, and test set Mean Absolute Error metrics of all Net Consumption Models, averaged across all twelve Cross-Validation iterations, sorted by the test set MAE in an ascending order.

supremacy of the Additive Model, in contrast to the results of [AK16], can be attributed to our research setup, which operates with the aggregated data of prosumers located in the northern region.

As depicted in Figure 11, all Net Forecasting Models displayed a noticeable discrepancy between their performance on the training set and the validation and test sets, suggesting a potential issue with overfitting. It indicates that the models may have memorized the training data rather than focused on learning significant patterns. Regularization parameters of the model can be adjusted to mitigate this concern.

6 Conclusion

Our study focused on advancing day-ahead Net Energy Forecasting techniques with the Estonian prosumers case study. Specifically, we introduced innovative types of Additive and Integrated Net Consumption Forecasting Models by incorporating distinct input features, aiming to mitigate uncertainty originating from weather forecast variables and unmetered self-consumption.

The primary concept involves the incorporation of Production forecasts into the input of the Consumption Forecasting Model and Consumption forecasts into the Production Forecasting Model for the Additive method. Additionally, Production and Consumption forecasts are used as input into the Integrated Net Energy Forecasting Model. By incorporating the forecasts, we hypothesized that models could effectively capture complex relationships between unmonitored Self-consumption and electricity exchanged between grids, representing the target variables in our study setup. This assumption originated from the understanding that electricity import and export depend on the level of Self-consumption.

The experimental results revealed a slightly superior performance of the proposed methodology for the Consumption model. In contrast, our Production model did not show any improvement and exhibited a slight decrease in performance. Furthermore, our Net Models did not significantly differ from those adapted from the original research. Both methodologies displayed an acceptable decrease in performance from the validation to the test set, indicating satisfactory generalization capabilities, except for the Integrated Net Models. The performance discrepancy between the validation and test sets for the Integrated Models can be attributed to the length of the validation subset and overfitting. Specifically, the one-month interval between the training and test data subsets limited the models' capacity to capture short-term fluctuations in data and to generalize effectively due to shifts in the underlying distribution of the data. All Net Models, including the Integrated Models, exhibited a significant discrepancy between the performance on the training set and the validation and test sets. It indicated that the models memorized irrelevant patterns or noise in the training data, leading to overfitting.

The original study noted that the Integrated Net Consumption Model surpassed the Additive Net Model by 10.69%. However, our investigation revealed a different outcome, showing that the Additive method outperformed the Integrated method in the adaptation of the original study. Specifically, the Additive Net Consumption Forecasting approach outperformed the Integrated method by 12.81% in terms of RMSE on the test set. This discrepancy may arise from variations in the experimental setups between the studies.

The results obtained in our study primarily indicated that the developed models did not adequately address the Self-consumption component, contrary to our expectations. Given that Self-consumption influences both P^{grid} and C^{grid} values, we anticipated that knowledge of future Consumption values would enable g^{prod} to account for S. The empirical evidence for this would be represented as a noticeable difference in

the performance of our models compared to the reference models, indicating that the incorporation of forecasts enhanced the predictive ability of our models. It would suggest that the models could learn dependencies between target values and S. Although our Consumption model showed a slight improvement over the reference model, there is no direct evidence that future Production values enabled g^{cons} to address S.

There are several potential explanations for the observed outcome. Firstly, the chosen modeling approach or the model itself may have had a weak capacity to accurately address self-consumption. It could also stem from overfitting to the training data. Secondly, inaccuracies in the forecasts used to represent Self-consumption and its relationships with target variables may have introduced noise into the data, disrupting the relationships with the underlying Self-consumption component. Further research is necessary to better understand the role of Self-consumption in both Production Forecasting and Consumption Forecasting. Future research should focus on the contribution of uncertainty associated with Self-consumption to the accuracy of the forecasts.

The experiments utilized the LightGBM model, a Gradient Boosting Decision Tree model, due to its computational efficiency. With an improved computational setup, more sophisticated model families, such as Neural Network models, can be explored for our methodology. For example, the Long Short-Term Memory (LSTM), a Recurrent Neural Network (RNN)-based model, is widely employed in Time Series Forecasting tasks due to its capability to capture sequential dependencies. Additionally, Transformer models are renowned for their ability to capture long-range dependencies and interactions, making them particularly attractive for time series modeling. Finally, the study could benefit from testing in different data contexts, such as an individual prosumer household in Estonia or alternative geographical regions.

In summary, our study demonstrated that incorporating the Grid Electricity Production variable into the Consumption Forecasting Model improved the precision of Grid Consumption forecasts. However, incorporating Grid Electricity Consumption into the Production Forecasting Model and including both Grid Consumption and Production variables in the Net Consumption Forecasting Model did not result in more accurate forecasts of Net Consumption. While the models within our framework did not exhibit significantly superior performance results, adopting our Consumption Model can benefit energy companies if the Additive approach is applied to Net Consumption Forecasting. Even a modest reduction in error can result in substantial financial savings for energy companies.

References

[AJW20] Nigel Taylor Christian Thiel Arnulf Jäger-Waldau, Ioannis Kougias. How photovoltaics can contribute to ghg emission reductions of 55 *Renewable and Sustainable Energy Reviews*, 126, 2020.

[AK16] Carlos F.M. Coimbra Amanpreet Kaur, Lukas Nonnenmacher. Net load forecasting for high remewable energy penetration grids. *Energy*, Volume 114, 2016.

[CG16] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *Advances in Neural Information Processing Systems 30*, 2016.

[CSB21] Jens Peder Meldgaard Casper Solheim Bojer. Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting*, 37, 2021.

[CX24] Guo Chen Chongchong Xu. Interpretable transformer-based model for probabilistic short-term forecasting of residential net load. *International Journal of Electrical Power Energy Systems*, 155, 2024.

[E.19] Al Daoud E. Comparison between xgboost, lightgbm and catboost using a home credit dataset. *International Journal of Computer and Information Engineering*, 13, 2019.

[Ele] Elering. Electricity market. Retrieved May 06, 2024, from https://elering.ee/en/electricity-markettab1.

[GK17] Thomas Finley Taifeng Wang Wei Chen Weidong Ma Qiwei Ye Tie-Yan Liu Guolin Ke, Qi Meng. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems 30*, 30, 2017.

[Han20] Khoshgoftaar T.M. Hancock, J.T. Catboost for big data: an interdisciplinary review. *Big Data*, 7, 2020.

[JGDG06] Rob J. Hyndman Jan G. De Gooijer. 25 years of time series forecasting. *International Journal of Forecasting*, 22, 2006.

[NS22] Hadi A. Raja Muhammad Jawad Alo Allik Oleksandr Husev Noman Shabbir, Lauri Kütt. Techno-economic analysis and energy forecasting study of domestic and commercial photovoltaic system installations in estonia. *Energy*, 253, 2022.

[Pal18] Jenny Palm. Household installation of solar panels – motives and barriers in a 10-year perspective. *Energy Policy*, 133, 2018.

[Poo23] Nord Pool. 2022 annual review: Stability through change. Technical report, 2023.

[SERM20] G. Ledwich G. Nourbakhsh D. B. Smith S. E. Razavi, A. Arefi and M. Minakshi. From load to net energy forecasting: Short-term residential forecasting for the blend of load and pv behind the meter. *Access*, 8, 2020.

[Shi07] Haijian Shi. *Best-first decision tree learning*. PhD thesis, The University of Waikato, 2007.

[SP21] Mariano Martin Zdravko Kravanja Sanja Potrč, Lidija Čuček. Sustainable renewable energy supply networks optimization – the gradual transition to a renewable energy system within the european union by 2050. *Renewable and Sustainable Energy Reviews*, 146, 2021.

[SU21] Alaraj M Alsaidan I. Singh U, Rizwan M. A machine learning-based gradient boosting regression approach for wind power production forecasting: A step towards smart grid environments. *Energies*, 14, 2021.

[YWX18] Q. Chen D. S. Kirschen P. Li Y. Wang, N. Zhang and Q. Xia. Data-driven probabilistic net load forecasting with high penetration of behind-the-meter pv. *Transactions on Power Systems*, 33, 2018.

Appendix

I. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Yuliia Siur**,
*(*author's name)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Prosumer Net Consumption Forecasting: The Impact of Behind-the-Meter Self-Consumption and Weather Forecast,
   *(*title of thesis)

   supervised by Novin Shahroudi and Jean-Baptiste Scellier.
   *(*supervisor's name)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Yuliia Siur
12/05/2024