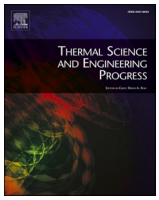
Thermal Science and Engineering Progress 70 (2026) 104498

Contents lists available at ScienceDirect

Thermal Science and Engineering Progress

journal homepage: www.sciencedirect.com/journal/thermal-science-and-engineering-progress



Unlocking thermal flexibility through demand-side response: baseline methodology assessment and heating electrification in the Baltic region☆

Deividas Šikšnys [a,*], Jonas Vaičys [a,b], Saulius Gudžius [a], Roma Račkienė [a], Matas Grigošaitis [b]

[a] Department of Energy, Kaunas University of Technology, Studentu 56, LT-51424 Kaunas, Lithuania
[b] EDIS Lab UAB, V.Kudirkos 4-1, LT-56126 Kaišiadorys, Lithuania

ARTICLE INFO

Keywords:
Power-to-heat (P2H)
Baseline methodology assessment
Demand-side response (DSR)
Heating electrification
Aggregators
Market welfare analysis

ABSTRACT

Demand-side response (DSR) flexibility is gaining increasing attention across power systems undergoing the energy transition, where renewable generation now dominates supply patterns. However, its reliable integration remains constrained by baseline methods that fail to accurately capture the operational characteristics of distributed demand resources, particularly thermally driven loads. This research provides a practical, decision-oriented framework for baseline selection, combining a two-stage process of technical feasibility assessment and multi-criteria performance evaluation. Eight baseline-method families are systematically evaluated, with empirical validation using 39 Latvian consumption sites and a Lithuanian hybrid heat-pump and photovoltaic system demonstration. Results show that static historical baselines are insufficient to capture thermal inertia, cyclic heat-pump operation, and cyclic compressor behaviour, while adaptive, weather- and PV-sensitive methods substantially improve accuracy (MAE reduction from 6.65 to 3.62 kWh), ensuring robust and transparent flexibility quantification. Market welfare simulations using 85 days of Baltic 2024 summer day-ahead market data indicate that even modest volumes of price-responsive DSR (5–50 MW) can reduce scarcity-hour market-clearing prices by up to 33 €/MWh and increase substantial social welfare gains (0.59–4.3 million euros) highlighting the tangible economic benefits of improved baseline accuracy. Overall, the study establishes that accurate, integrity-preserving baselines coupled with digital metering infrastructure unlock significant short-term and intraday flexibility, bridging technical precision with system-level market and welfare outcomes.

1. Introduction

The ongoing transition of power systems towards carbon–neutral operation is driven by the integration of renewable energy sources (RES), which by their nature are variable in generation. Since the balance between generation and load must be maintained at all times, this shift, together with the electrification of consumption, creates a need for system wide flexibility. Today, the flexibility of the European electricity system is still provided largely by dispatchable fossil-fuel units such as gas turbines, which can respond to variable demand and supply but emit significant greenhouse gases and air pollutants. Continued reliance on these resources is increasingly incompatible with the 2030 EU greenhouse gas reduction target and the 2050 climate neutrality objective. To align with these ambitions, the EU's climate and energy goals have been translated into national commitments through Member States' National Energy and Climate Plans (NECPs). For example, Lithuania's NECP foresees that the share of new flexible resources will increase more than sixfold by 2050, with contributions from technologies such as batteries, power-to-gas, power-to-heat and dispatchable low-carbon generation [1,2].

The decarbonisation of the heating sector has become one of the most critical components of this transition. The International Energy Agency (IEA) report The Future of Heat Pumps (2022) identifies heat pumps as one of the fastest and most cost-effective pathways to decarbonise space and water heating while improving energy security. Tripling global heat pump installations by 2030 could reduce annual CO_2 emissions by more than 500 Mt and reduce European natural gas use by nearly 7 %. Moreover, smart operation of heat pumps, particularly when coupled with flexible electricity markets-can provide up to 20 % of total system flexibility in high-renewable scenarios [3]. This impact is reflected in the wholesale market price and rising negative price hours.

However, economic barriers remain a major obstacle in parts of

☆ This article is part of a special issue entitled: 'David Reay?s 80th' published in Thermal Science and Engineering Progress.

* Corresponding author.
E-mail address: deividas.siksnys@ktu.lt (D. Šikšnys).

https://doi.org/10.1016/j.tsep.2026.104498
Received 30 November 2025; Received in revised form 2 January 2026; Accepted 6 January 2026
Available online 7 January 2026

Nomenclature

COP	Coefficient of performance
ERA	Electricity reduction amount
MWh	Megawatt-hour – unit of energy
MW	Megawatt – unit of power
P2H	Power-to-heat
PV	Photovoltaic
PV/T	Photovoltaic-thermal
$P_s(V)$	Hourly supply
$P_d(V)$	Hourly demand
ΔTS	Change in Social-welfare change

Abbreviations

aFFR	Automatic frequency response
ANNs	Artificial Neural Networks
BRP	Balance Responsible Party
CNN	Convolutional Neural Network
DR	Demand Response
DSO	Distribution System Operator
DSR	Demand Side Response
EU	European Union
EE	Estonia
FCR	Frequency Containment Reserve
FSP	Flexible Service Provider
IEA	International Energy Agency
KTU	Kaunas University of Technology
Litgrid	Lithuanian transmission system operator
LSTM	Long Short-Term Memory
LT	Lithuania
LV	Latvia
LGBM	Light Gradient Boosting Machines
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MBMA	Metre Before Meter After
ML	Machine Learning
mFFR	Manual frequency response
MTU	Market Time Unit
NECP	National Energy and Climate Plans
PCA	Principal Component Analysis
SO	System Operator
SDA	Same Day Adjustment
ToU	Time-of-Use
TSO	Transmission System Operator
UK	United Kingdom
V2G	Vehicle to grid

Europe, especially in the Baltic region. The Green Heat for All report (2023) highlights that heat pumps are among the least financially competitive in Lithuania, Latvia, and Estonia, where payback periods often exceed 12–20 years-among the longest in the EU. This is driven by high upfront costs, limited subsidies, and unfavourable electricity-to-gas price ratios. Without stronger financial incentives and tariff reform, the electrification of heating will lag, constraining both the progress of climate and the potential for flexibility. Addressing these barriers is essential to make heat pumps a viable alternative to gas and oil boilers and unlock their role as flexible assets in future electricity markets [4].

The analysis of the European Commission's Joint Research Centre [5] and several recent studies highlight that large-scale heat pump integration can both decarbonise the building sector and provide operational flexibility to the power system. Heat pumps, particularly when combined with thermal storage or photovoltaic-thermal (PV/T) hybrid systems, can act as controllable loads within demand-side response (DSR) programmes, allowing the temporal shift of heating demand to periods of low electricity prices or high renewable generation. This interaction effectively transforms the heating sector into an active component of the power system, capable of balancing short-term fluctuations and reducing overall system costs.

Recent research supports this emerging role for the heating sector. Studies have demonstrated that up to 10 to 56 % of total heating demand could be covered by electric heat pumps depending on their coefficient of performance (COP) and local climatic conditions [6]. Moreover, heat pumps equipped with thermal storage have been shown to provide short-term flexibility and participate in intraday market arbitrage, achieving annual savings between €17 and €692 per household depending on building size and efficiency [7]. District heating networks coupled with power-to-heat (P2H) units can also deliver ancillary services and balancing support, generating operational profit increases of up to 9.7 % [8]. Additionally, integration of heat pumps in local energy markets can lower household costs by around 5 %, highlighting both the economic and environmental benefits of heating electrification [9]. Taking most referable and liquid day-ahead market in the Fig. 1, we can indicate negative price number per MTU increasing year by year which gives positive economical signal to flexible thermal systems payback time.

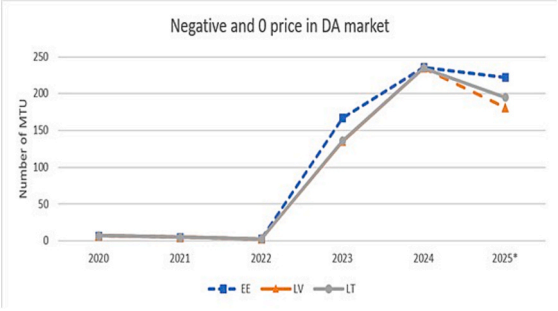


Fig. 1. Historical wholesale DA market negative and 0 price number of MTU.

In combination with electric vehicles, heat pumps form the backbone of future distributed flexibility portfolios, enabling seasonal balancing and reducing the need for stationary storage [3].

Electricity remains a non-storable commodity that must be generated and consumed in real time, yet is traded across several structured market timeframes, ranging from long-term contracts to real time balance [10]. Among these, the day-of-market is the most liquid and serves as a reference point for the generation and consumption scheduling [11]. Transmission System Operators (TSOs) ensure real-time balance by activating reserves and flexibility products whenever deviations occur, supported by structured balancing markets. This market structure is presented in the Fig. 2 gives understandable full picture of the markets. European experience shows that market design is central to unlocking both electrical and thermal flexibility.

In the Baltic region, a joint study conducted in 2017 explored harmonised approaches to DSR aggregation and resulted in pilot projects [12,13]. Some studies were manned to highlight the benefits of inclusion of DSR in the Baltics balancing market with the central settlement model presented for the integration of aggregators [14,15]. Lithuania subsequently introduced independent aggregators as licenced market actors, but in practice only a limited number are active [16,17]. Estonia's TSO, Elering, is now conducting a public consultation on how to extend DSR participation from TSO-regulated balancing services to

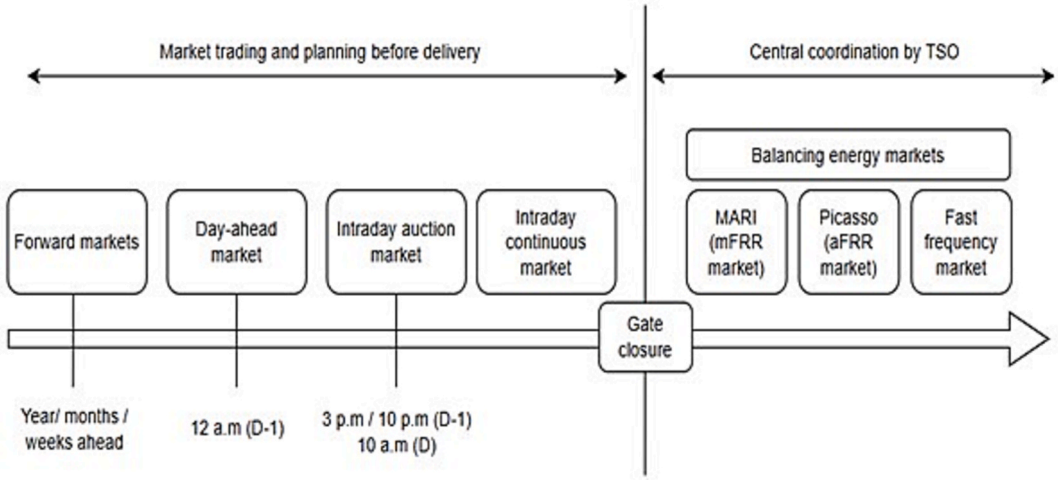


Fig. 2. Overview of different time frames of the wholesale and balancing markets.

wholesale markets, including the day-ahead (DA) and intraday segments [12]. This initiative aligns with EU electricity market design reforms and the forthcoming demand response network code [18,19].

As the heating and power sectors converge, heat pumps emerge as a key resource for explicit participation in DSR, particularly in northern European regions with significant heating demand. In Latvia, the inclusion of DSR in the wholesale market was shown to reduce day-ahead electricity prices by approximately €0.025/MWh per 1 MWh/h of flexible demand, resulting in measurable socio-economic benefits [20]. With average annual electricity prices in Latvia nearly doubling between 2019 (€46.3/MWh) and 2024 (€87.43/MWh), the economic incentive for heat pump-based flexibility is expected to increase further and we present in Section 4 welfare of the flexibility inclusion in the DA Baltic market based on the 2024 summer data.

Despite this growing potential, the reliable market integration of DSR, especially from thermally driven and hybrid electric–thermal loads, critically depends on the baseline methods used to quantify flexibility delivery. Baselines define the counterfactual consumption against which flexibility is settled and remunerated. Inappropriate baseline selection can therefore lead to biased settlements, reduced trust in flexibility markets, and inefficient activation signals. These challenges are particularly pronounced for electrified heating systems, whose consumption patterns are shaped by thermal inertia, weather dependence, and user behaviour, features that are not adequately captured by many conventional baseline approaches.

Existing research on DSR baselines has predominantly focused on isolated accuracy comparisons or algorithmic improvements, often without sufficient consideration of practical applicability, data availability, settlement requirements, and regulatory constraints. As a result, a methodological gap remains between baseline performance demonstrated in controlled studies and their suitability for real-world market implementation, particularly for thermal and hybrid flexibility resources.

To address this gap, this paper advances research on DSR baseline methodology by adopting a practical, decision-oriented approach to baseline selection. Rather than proposing a new baseline algorithm, the paper makes three main contributions. First, it introduces a two-stage baseline selection framework that combines a technical feasibility assessment, considering data availability, metering resolution, and operational constraints, with a structured, multi-criteria performance evaluation that explicitly accounts for settlement integrity and regulatory requirements. Second, it presents a graded evaluation matrix assessing accuracy, simplicity, and integrity, supplemented by scalability and technology-neutrality criteria, including the use of a mean absolute error (MAE) metric, enabling consistent comparison across different load capacities and aggregation levels. Third, the framework is empirically validated using real consumer data from the Baltic region and complemented by a quantitative welfare analysis that illustrates how baseline choices influence the economic value of DSR participation in the day-ahead electricity market.

Together, these contributions establish a practical methodological foundation for market-oriented baseline selection and demonstrate that baseline methodology is a critical enabler of real power system flexibility and welfare-efficient market outcomes.

2. Methodological framework

Instead of proposing a new baseline algorithm, this article introduces a decision-oriented framework that combines existing baseline method families into a systematic selection process suitable for real-world deployment. This article examines eight families of baseline methods, analysing their underlying principles, technical requirements, and suitability to assess the flexibility of distributed resources such as DSR and small-scale electrified heating systems. Section 1 presents the proposed methodological framework for baseline selection, structured as a two-stage evaluation process.

Section 2.1 first provides a detailed overview of the baseline method families, outlining their underlying assumptions, principal advantages and limitations, and their typical applicability to different asset types and flexibility services. As the first step, a set of technical requirements is established in section 2.2 to determine whether a given baseline family can be feasibly applied to specific flexibility resources and services. These requirements include data resolution, responsiveness to dynamic consumption patterns, computational simplicity, and compatibility with the available metering infrastructure. For electrified thermal systems, baselines must be able to capture thermal inertia and cyclic operation characteristics of heating devices, ensuring realistic modelling of demand variations caused by temperature dynamics or user behaviour this in more details is presented in the indicated Section 2.2.

As the second stage, Section 2.3 an assessment matrix is developed to evaluate the remaining baseline families. The matrix compares methods using a structured set of main criteria (accuracy, simplicity, integrity) and supplementary criteria (technology neutrality, transparency, scalability, and automation). By combining quantitative accuracy metrics with qualitative evaluation principles presented in Table 4, the framework provides a comprehensive tool to identify the most suitable baseline approaches to enable flexibility from both electrical and thermal loads. Based on the literature analysis focusing on the DSR asset group Table 5 created to demonstrate, how each baseline method is evaluated by main criteria seeking to compare these results with the Baltic region

use case data and baseline methodology suggested in this article. This structure of the methodological framework is presented in Fig. 3.

To validate the proposed approach, in Section 3 the methodology is applied to several real-world use cases in the Baltic region. The first case tests the applicability of selected baseline families on 39 diverse consumption objects within the Latvian DSO network, while a second Lithuanian hybrid demonstration system is being tested to evaluate joint electrical–thermal flexibility. By simulating implicit DSR integration of 5 MW and 50 MW flexibility, the analysis quantifies potential economic benefits from demand reduction, including price decreases and total welfare gains in the day-ahead market.

2.1. Baseline method families

The purpose of this section is to define a finite and practically relevant set of baseline families that can be evaluated within the proposed selection framework. Most often applied is historical baseline methodology, purely data driven baseline (b_t) = average of X highest values among the last Y comparable days (d_t). Rolling variants often add a Same Day Adjustment (SDA) using the hour(s) just before the event. Other variations do not use the same weights over a set of admissible days but put larger weights on more recent consumption. In article [21], several variations as presented. For the metric HighXofY, which is based on historical daily data, a dataset of size Y is considered, and the agreed number of X data points is selected by ordering the values from largest to smallest. Conversely, for the LowXofY metric, the same range of Y data points is used, but the agreed number of X values is selected by ordering the data from smallest to largest, representing the lowest electricity consumption. This method formula:

$$b_t \equiv d_t + SDA \tag{1}$$

The main advantages of this method are that it is transparent, easy to audit, minimal computational burden, works well for stable and repetitive loads, that is why this method is usually the best for DSR assets and small consumer load forecasts. On the other hand, this method has disadvantages which appear in specific groups. If service supplier object profile is volatile or weather-sensitive (prosumers, PV generation in same location, electrical heating etc.) forecasting error increase. In addition, it can be susceptible to gaming if events are anticipated.

The second methodology is comparable day, selects a past non activation day whose weather & calendar conditions best match the event day, and uses that profile as the baseline. As advantages are naturally capturing exogenous factors (weather, holidays) and avoids averaging unrelated outliers, but for disadvantages needs robust similarity metrics and weather data, can be manipulated by pre-event load shifting.

Statistical & Regression use parametric models (linear, polynomial, random forest) predict consumption as a function of temperature, humidity, calendar effects, prior demand, and other described variables. Regression-based, calculated, and control group baselines are less common; these baselines are usually preferred when the common baselines are not suitable. In particular, regression baselines are used for longer usage periods, as they can increase accuracy using external variants. There can be many variations for this type of methods depending on the objectives, as in the article [22] for the aggregated flexibility of the residential thermal load using only behind the meter measurement data 6 different baseline methods were presented and compared in the regression family group identifying that the benefits of decomposition and regression increase the accuracy of the baselines. For pros is high accuracy for weather-sensitive loads and can be adaptable via periodic re-calibration. Cons require correct model specifications; analysis indicated that administrative costs and the associated complexity of regression approaches are significantly higher than those of averaging approaches, and calibration workload grows with portfolio size which decreases simplicity [23,24].

Machine learning (ML) methods represent a class of non-parametric modelling techniques that are increasingly applied for electricity baseline estimation. These algorithms – such as Random Forests, Gradient Boosting Machines, Artificial Neural Networks (ANNs), Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks – are capable of learning complex non-linear relationships from high-resolution smart metre data [21,25,26]. Unlike traditional statistical models, ML approaches can automatically detect patterns and interactions among multiple exogenous variables, including weather, occupancy, time-of-use, and socio-economic factors, without requiring explicit model specification. As a result, they often achieve state-of-the-art predictive accuracy in baseline construction.

Recent studies have confirmed these advantages in renewable energy forecasting, showing that ensemble learning models such as Light Gradient Boosting Machines (LGBM) and Random Forests outperform conventional regression methods in capturing variability in renewable generation and consumption patterns [27].

However, despite their performance advantages, ML methods present several challenges in the context of regulatory or policy-driven applications. First, the “black box” nature of many ML models limits transparency and interpretability, making them difficult to audit or validate by external stakeholders. Second, their application typically requires substantial computational resources, technical expertise, and ongoing maintenance – including periodic retraining to prevent performance degradation due to evolving consumption patterns (i.e., model drift). These limitations may hinder their acceptability in conservative regulatory frameworks, where explainability (simplicity) and reproducibility (accuracy) are essential.

The Meter-Before Meter-After (MBMA) approach defines the baseline by calculating the average consumption or output over a short time window immediately before activation of a response event. The energy delivered or curtailed is then determined by comparing this baseline to the post-activation measurement over a corresponding time. This method relies exclusively on real-time metered data, eliminating the need for prior modelling or historical analysis. MBMA is particularly well-suited for applications involving fast-response reserves such as FCR and aFRR, mFFR, where rapid activation and verification are essential. Its transparency and resistance to manipulation make it highly reliable in operational contexts – especially when a secure, high-resolution metering infrastructure is in place. However, MBMA has notable limitations. The assumption of a flat or stable pre-activation load may not hold for slow-varying or erratic demand profiles, reducing accuracy in such cases. The method is also sensitive to noise or fluctuations in the pre-activation period, which can distort baseline estimation. Additionally, MBMA requires high-resolution metering (e.g., 1-second to 1-minute granularity) and secure communication protocols, which may not be universally available, especially in legacy systems.

The Zero Baseline approach assumes that baseline consumption or export during an activation event is zero. Assets are expected to be idle

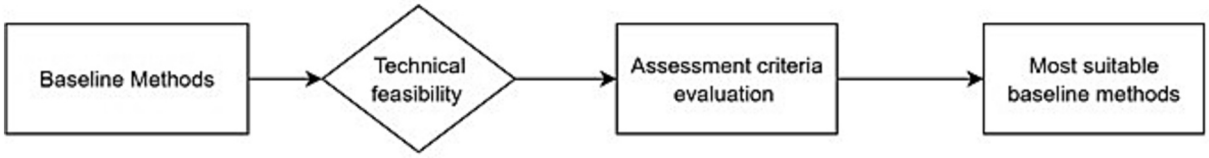


Fig. 3. Framework for the methodology to find the best suitable baseline methods.

in the absence of a demand response signal. Under this assumption, any net export to the grid during activation is fully attributed to the flexible response. This method is primarily applicable to assets that are typically inactive under normal conditions, such as battery energy storage systems or standby diesel generators. The key advantage of the Zero Baseline method lies in its simplicity and robustness: it eliminates the need for counterfactual modelling and avoids estimation errors associated with predicting baseline behaviour. This makes it especially attractive for operational programmes focused on fast, dispatchable capacity. However, the method has several limitations. It may overestimate delivered flexibility when the asset simultaneously meets onsite demand (i.e. self-consumption), unless separate metering points are used to isolate export to the grid. This raises concerns around double counting of energy volumes or emissions reductions. Furthermore, the method is not applicable to demand reduction measures, where consumption is curtailed rather than energy exported. Therefore, while useful for specific asset types, the Zero Baseline approach requires careful metering design and application scope definition.

The Control Group baseline approach estimates counterfactual consumption by using the average load profile of a matched group of non-participating customers who were not subject to activation during the demand response (DR) event or as in the Ref [28] create virtual control group. These control customers are selected to share key characteristics with the treated entity, such as sector, climate zone, size, and historical consumption behaviour, ensuring that they represent a valid reference. This method allows for the isolation of causal impacts, correcting for real-time external shocks such as weather changes, price volatility, or system events. Control group baselines are well suited for programme level impact evaluation and large scale behavioural DR assessments, where individual modelling is impractical. They offer a strong statistical basis for estimating what would have happened in the absence of intervention. However, the method involves several practical challenges. It requires access to large, anonymised datasets and the use of matching algorithms to construct a statistically valid comparison group. There is a risk of selection bias or cohort drift over time, particularly if control group behaviour diverges from the treatment group due to unobserved factors. Moreover, implementation often places an administrative burden on Distribution System Operators (DSOs) or aggregators, who must manage data privacy, participant matching, and ongoing monitoring [29].

The Self-Declared or Nomination baseline method allows the Flexibility Service Provider (FSP) to submit its own forecast of expected consumption or generation shortly before the delivery period. This self-nominated baseline is then compared against actual metered performance, and deviations are assessed ex post by the system operator. To ensure integrity, the mechanism typically includes financial penalties or performance adjustments for significant discrepancies between the declared and actual values. This approach offers several advantages. Changes the modelling responsibility to the aggregator or FSP, enabling the use of asset-specific knowledge and proprietary forecasting models. This approach can give benefits to the SO to more accurate load and energy flow profile in the specific area the FSP is providing compared to the casual SO system forecast measures. In addition, it can be more compatible with data privacy requirements as it avoids the need for operators to access granular historical data directly. This makes it particularly relevant for commercial or behind-the-metre flexibility resources. However, the Self-Declared method requires a robust validation framework and enforceable penalty structures to deter strategic inflation of baseline values. If penalties are weak or are applied inconsistently, there is a risk of intentional overestimation, compromising the credibility of the flexibility delivered. Furthermore, verification becomes resource-intensive in large portfolios, as each nomination must be audited for accuracy and compliance, posing a significant operational burden for system operators or market platforms [30,31].

All reviewed and presented baseline methods with the identified advantages and disadvantages are presented in the Table 1:

2.2. Technical requirements for applying the best suitable baseline method

Technical feasibility assessments are often implicitly assumed or overlooked in baseline comparison studies. This section clearly outlines the procedure for evaluating the technical aspects of baseline methods, which is crucial for creating a decision-oriented framework. The first step to guide the most suitable baseline methods is to identify the availability to meet the technical requirements that each method can require. For measurement data timeframe, MBMA and zero-baseline require Real-time data fulfilment, in some cases ML and regression method families can request as well, but rest of the methods can be applied using historical measurement data, as it typically includes a collection of measurements recorded over a specific period in the past. The control group and self-declaration should be noted as methods that do not request historical or real-time data, as they compare activation moment measurements with similar group data and or data provided from the service provider. This can include data on hourly, daily or seasonal patterns, as well as long-term trends. Real-time data typically include current measurements of electricity demand, weather conditions, market prices, and other relevant variables; for this, we exclude and marked methods which request additional data source and can by challenge applying this group of methods.

Baselines can be calculated before, in real-time or after the event. Ex-ante baseline refers to calculating the baseline before the event or activation of flexibility resources. Real-time baselining involves continuously updating and adjusting the baseline demand during the event. The primary value of real-time baselining is the provision of immediate information about the activations, enabling, for example, real-time monitoring of the grid status. Ex-post baselining refers to calculating the baseline after the event. Measurement data granularity is crucial for specific markets, e.g., balancing market where market time unit is 15 min for activation and delivery, for this purpose usually MBMA is selected methodology (due to identified specific in previous chapter)

Table 1
Identified baseline methodologies with indicated pros and cons.

Baseline methodology	Advantages	Disadvantaged	Most suitable for
Historical	Transparent, minimal resource needed to apply, simple	Suspect of gaming, low accuracy for volatile profiles	DSR, customers with low volatility
Comparable day	implicitly accounted for external factors;	Needs robust similarity metrics and weather data	Weather dependent objects profile
Statistical & Regression	Increase the accuracy for weather-sensitive loads	Correct model specification	Difficult to define or attach to other group profiles
Machine-Learning	Best predictive accuracy handles multiple non-linearities	Low transparency for “black box”, high costs	All type objects
MBMA	Hard to game	High-resolution metering data	Balancing market service
Zero-baseline	The highest simplicity	Inapplicable to DSR, self-consumption separation	Batteries, small generators, V2G
Control Group	Corrects for real-time shocks	Needs large, anonymised datasets & matching algorithms	Large group DSR assessments
Self-declare (nomination)	Can represent behind the meter changes	Requires robust expost validation & penalties	Mature markets

fulfilling this requirement. Based on the detailed technical requirements Table 2 presents, how each baseline method meet them.

2.3. Assessment criteria for baseline methods

One of the main challenges in selecting an appropriate baseline method lies in defining and applying suitable evaluation criteria that can balance precision, practicality, and robustness. While previous studies usually evaluate baseline methods mainly on predictive accuracy, this work expands the assessment to include flexibility, integrity, and substantial compatibility, emphasizing real settlement and regulatory constraints. In this research, we adopt the methodology of three primary criteria—accuracy, simplicity, and integrity (robustness), which have been extensively discussed in previous works [32–35]. Building on this established foundation, we specify how these criteria can be consistently measured in the context of flexibility estimation and extend the assessment framework by introducing several additional criteria, as proposed by other authors [23,27,36,37]:

- Cost & Administrative Effort;
- Transparency & Explainability;
- Scalability & Automation;
- Data Privacy & Availability;
- Technology Neutrality / Inclusivity.

Starting with the accuracy measure, the main goal of it is to correctly forecast the level of consumption or production which will be compared with the actual measured data to identify provided service size. The major impact on accuracy can have the decision and principles of measuring the point of the object. If metering data are used from the connection point accuracy can be lost, but empowering sub-metering principle can bring accuracy to the high level. Identifying the accuracy and baseline methodology can be done by MAE as main measurement criteria if comparing results equal results reached mean absolute percentage error (MAPE) as sub measurement will be used.

The second measurement is simplicity, ease of understanding, implementation, audit, and communication with customers and regulators. Simplicity can be crucial for a not developed market or for a seeking to attract new market participants into the dedicated flexibility service or market. Identified criteria can be exposed as replicability if flexibility market design is created indicating the responsibility for baseline application to the market participants chosen baseline method has to be easily replicable otherwise it can become the market barrier. It can be measured by implementation costs that can appear for the FSP and service procurement side, SO. This criterion can be assessed through replicability, which depends on the data inputs, assumptions, procedural steps, and formulas used by the baseline methodology. [38].

The third main assessment criteria are integrity. Resilience against strategic manipulation, data tampering, or incentive misalignment; supports fair settlement and maintains market confidence [36]. Integrity is evaluated qualitatively through susceptibility analysis and verification feasibility, rather than through direct numerical measurement. This approach reflects regulatory and market practice, where robustness against gaming and auditability is assessed through procedural and behavioural indicators rather than error metrics. Practical example, simulate user behaviour from historical data and compare with the drastic changes as it started participating in the service. The other way is based on the chosen baseline method to reduce impact with the pre-event signals to the baseline calculations (in the historical approach the adjustment factor is used, or in the MBMA method). The second measure is regular verification by pre-qualification and post-qualification checks let service procurement party validate and reduce gamming approach by identifying them and stopping until it can damage the market. Valuable point to address customised methods has verry low integrity compared with the standardised. Indicated assessment criteria with the measurements presented in the Table 3.

To establish a practical assessment matrix to select the most suitable baseline method, we defined stepwise criteria as summarised in Table 4. In particular, the accuracy criterion is evaluated through the indicators MAPE and MAE, whose magnitude depends on both the scale of DSR consumption and the application context (individual buildings versus aggregated groups of objects). The article proposes using MAE as the primary indicator for grading accuracy, as MAPE is not reliable and can be misleading in cases involving a single object or very low consumption levels, so it suggested to be used only as sub-measurement criteria. To determine the appropriate value ranges, evidence from recent

Table 3
Baseline evaluation assessment criteria and suggested measurements.

Assessment criteria	Measurement 1	Measurement 2	Valuable notes to address
Accuracy	MAE	Sub measurement – MAPE	Sub-metering vs metering
Simplicity	Implementation costs for SO and FSP	Replicability	x
Integrity	Susceptibility analysis	Regular verification	Standardised method vs customised
Supplementary criteria			
Technology Neutrality / Inclusivity	Technology coverage can be applied in a wide range of technologies	Can the methodology be used for both large and small users	
Transparency & Explainability	It is necessary to facilitate the required degree of trust that the market parties have in the fair outcome of the settlement process. Transparency ensures market parties know exactly how the baseline is produced and which data are used.		
Scalability & Automation	These criteria can be used in the specific evaluation to better reflect criteria that can be important to fulfil correctly.		
Data Privacy & Availability			
Cost & Administrative Effort			

Table 2
Technical requirements for different groups of baseline methods.

Technical requirements		Baseline methodologies							
		Historical	Comparable day	Statistical & Regression	Machine-Learning	MBMA	Zero-baseline	Control Group	Self-declaration
Additional data source required				X	X			X	
Timeframe of data included	Historical	X	X	X	X				
	Real-time			X	X	X	X		
Timing of the baseline calculation	Ex-ante	X		X	X		X		X
	Real-time				X	X			
	Ex-post	X	X					X	
Measurements data granularity	High resolution					X			

Table 4
Assessment criteria grade value definitions and requirements.

Criteria	Grade A	Grade B	Grade C
Accuracy	Excellent accuracy; suitable for high-stakes regulatory or settlement use MAPE $\leq$ 5 % MAE $\leq$ 0,05 x average hourly load or $\leq$ 2 % of peak	Small, medium, and high variance depending on the time within the service window 5 % $\leq$ MAPE $\leq$ 10 % MAE $\leq$ 0,1 x average hourly load or $\leq$ 5 % of peak	High variance for a larger part of the service window MAPE $>$ 10 % MAE $>$ 0,1 x average hourly load or $>$ 5 % of peak $>$ 5 % of peak
Simplicity	Simple, straightforward calculation that can be made up without additional investments	Slightly complex calculations, additional data near the consumption measurement required. (e.g weather data)	Highly complex calculations will be required, and additional investments will be needed to run and maintain this method. More complex data required and to run the calculations investment to the software may be required.
Integrity	No foreseeable gaming options, simple verification procedures	Some gaming options with little mitigation options. Gaming can appear less than couple times of the month; verification requires resource and knowledge	Obvious gaming options, appearing more than couple of time in the month with little mitigation options
Technology Neutrality / Inclusivity	Full technology and customer class applicability	Part of technology and customer class applicability	Favouring specific technologies or customer classes

forecasting studies was considered. For example, Guoyoa Wu et al. (2025) [39] demonstrated that incorporating spatiotemporal and application-expansion data in short-term building load forecasting achieved MAPE values consistently below 5 %, which can be interpreted as a high-accuracy benchmark. Similarly, Lago et al. (2018) showed that advanced machine learning and regression models for energy consumption prediction frequently deliver MAPE $<$ 5 %, with MAE varying seasonally but typically remaining below 0.05 $\times$ average load, supporting their classification as Grade A under the proposed criteria [40]. In electricity price forecasting, studies have also identified MAPE $<$ 5 % as an indicator of excellent short-term prediction accuracy. Further evidence from large-scale applications, such as aggregated demand forecasting in Queensland, Australia (average load of 6.3 GW), reported MAPE $<$ 5 % and MAE within the threshold of 0.05 $\times$ average hourly load, thereby reinforcing the robustness of our proposed grading thresholds [41]. Importantly, our suggested scaling of MAE by average hourly load introduces an innovative adjustment that accounts for differences in system size and measurement aggregation. As summarized in Table 5, the highest accuracy levels are consistently achieved by machine learning and statistical regression models, which aligns with findings from comparative studies in diverse urban environments, where similar error levels were reported [42]. This convergence of empirical evidence supports the validity of our grading scheme and underscores the role of accurate baselines in enhancing flexibility and reliability across energy systems.

Second, simplicity is easy to measure, as reaching the Grade A baseline methodology can be applied using excel or similar low resource requiring programme. If methodology requires additional data near the consumption measurements, for example, whether forecasts, or relevant objects measurement data, this will decrease simplicity to Grade B. For methodologies requesting more mathematical understanding to apply the calculation or even a dedicated mathematical model, simplicity is the lowest.

Integrity can be the most important measurement for the procurement side to avoid market gaming and paying for simulated market behaviour. The best methodologies do not leave this opportunity for gaming or suggest easy, applicable verification procedures to verify this type of behaviour.

Based on the matrix created, we can identify each baseline method with the Grades according to different assessment criteria. To reflect the replicability of the system more detailed principle is described. In the article the weight of the criteria is split into equal parts, this can be adapted to other cases increasing the weight to more important criteria as splitting it equally can give similar results for different methods. Each Grade stands for different points (e.g A = 3, B = 2, C = 1) giving total score range [3;9]. To be transparent in final score and method Grade are grouped by this principle, Grade C if total score is below 5 points, Grade B if the score is between 6–7 points and Grade A if the score is above 8 points. By performing this evaluation of each baseline method described in this article, additionally considered the results of other authors rating each method according to the main assessments criteria [35,36]. The authors in article [35] split the evolution of the assessment criteria by different types of distributed energy resources (flexible load, non-controllable and controllable distributed generation, energy storage systems). Focusing on the flexible load as DSR in Table 5, each baseline family method is rated in three main assessment criteria giving first view of the compatibility of the method to this group of assets. Table 5 presents a broader rating, which is subsequently compared with the proposed baseline methodology and regional data in Section 3.

3. Use cases of the methodology application

Section 3 validate the proposed selection framework by practically applying baseline-selection methodology to real consumption data and flexibility activation scenarios in the Baltic region. The validation occurs in two steps: (i) to confirm that the integrity- and feasibility-based screening developed in Section 2 effectively narrows down the set of usable baseline families, and (ii) to demonstrate how the accuracy and robustness of different baseline groups vary when applied to diverse load types, including industrial, commercial, and thermally driven consumption. Three groups of use cases were analysed: heterogeneous loads in the Latvian DSO network, targeted industrial and HVAC real case studies, and a hybrid Lithuanian demonstration integrating battery storage, power-to-heat and PV units, but as our main goal is DSR we run testing on the consumption part including heat-pump activity behind the meter. Together, these cases illustrate how to select the most appropriate baseline model for specific flexibility resources and market applications.

3.1. Testing on various consumption objects

A data set comprising 39 consumption objects connected to the

Table 5
All baseline methods are evaluated using an created assessment matrix applied on the DSR.

Baseline family	Accuracy	Simplicity	Integrity
Historical	Grade B	Grade A	Grade B
Comparable day	Grade B	Grade A	Grade C
Statistical & Regression	Grade A	Grade C	Grade B
Machine-Learning	Grade A	Grade C	Grade A
MBMA	Grade C	Grade A	Grade C
Zero-baseline	NA	Grade A	NA
Control Group	Grade B	Grade A	Grade A
Self-declare (nomination)	Grade B	Grade A	Grade C

Latvian DSO network was used as the primary testbed for the baseline evaluation. The measurement period is from 2024 to 01-01 to 2025–05-14 and contained high-quality time-series data with less than 0.1 % missing values. Consumption values ranged from 7 to 300 kWh per interval, representing a mixture of small commercial, industrial, and mixed-use profiles. Auxiliary meteorological data (temperature, precipitation, wind speed, radiation, and cloud cover) were collected from Meteoblue to evaluate weather-dependent baseline methods and presented in the Table 6 [43].

Several objects exhibited characteristics highly relevant for baseline modelling, including daily cycling, quasi-constant load, and daytime net-load reversals indicative of behind-the-meter PV installations. This diversity enabled a systematic comparison of baseline families across heterogeneous load behaviour.

The technical feasibility criteria in Section 2 were first applied to determine which baseline groups could be used with the available data. Methods such as MBMA were excluded due to insufficient temporal resolution from measuriment devices, control-group baselines were only partially feasible due to limited data, and clustering size and zero-baseline and self-declaration approaches were excluded based on integrity considerations. As a result, three baseline families were selected for full testing: historical baselines, regression-based baselines, and matched-control group baselines.

3.2. Historical baseline method testing

Historical baselines are among the most widely used for DSR settlement, but their accuracy depends heavily on parameter choices such as the number of historical days used, the selection logic (high, middle or average values) and the same-day adjustment (SDA) cap. This method takes same hour measurement data from previous days indicating the range of the historical days will be taken and based on the type indicated days measurements are chosen to be used setting the baseline value. Sometimes to increase accuracy holidays and weekends are treated separately, not included in the historical data set, but here these principles were not included as other factors where tested. Five configurations derived from existing DSO practice and the academic literature were initially tested and results is presented in the Table 7.

As the best accuracy results were reached in this step using 20 historical days data and only taking 15 highest consumption points to calculate average point adding SDA in 20 % weight CAP this is presented in Fig. 4 comparing with the actual consumption data.

The results showed that the best performing variant within this group was the High 15-of-20 approach with a 20 % SDA cap, producing an aggregated MAPE of approximately 15.7 %. Based on our assessment criteria it accuracy is defined as Grade C. A systematic refinement process then replaced high-value selection with the middle-X-of-Y approach in which the difference lies in the selection of the X data points. Specifically, from the historical dataset of size Y, the values are first ordered, and the agreed number X is selected from the middle of the ordered sequence, thereby representing the central tendency of the data. This adjustment significantly reduced errors, with several variants achieving aggregated MAPE values between 6 % and 11 % by increasing accuracy grade level to B as MAE fit indicated level for middle grade category. Testing results presented in the Table 8.

As a third approach, the average historical values were tested with different percentages of SDA and presented in the Table 9:

Second historical baseline method approach where applied and best result generated principles is presented graphically in Fig. 5:

Further testing showed that averaging all historical days within a selected window produced the highest accuracy and robustness. The best historical method—using a 10-day average with a 50 % SDA cap—achieved a MAPE of 4.6 % and a MAE of 6.65 kWh, surpassing the accuracy threshold required for Grade B classification. As for simplicity criteria it indicated Grade A classification as well as it was run by Microsoft excel with the basic commands, third measure integrity is Grade B as some gaming options can be applied in this calculation. Historical methods performed particularly well for loads with stable, repetitive daily patterns and for industrial sites with consistent operational profiles. However, they were less accurate for temperature-sensitive loads and prosumers with PV-driven load reversals, where weather-normalized regression or control-group methods produced superior results.

Table 6
Data used for the testing of the baseline methods on 39 LV objects.

Category	Details
Consumption Objects Data	
Number of Objects	39
Time Period	2024–01-01 00:00:00 to 2025–05-14 09:00:00
Average Consumption per Object	Minimum: 7.12 kWhMaximum: 300.47 kWh Mean average: 82.93 kWh
Missing Values	~0.1 % (very few missing data points)
Additional Insights	• Consumption values range from small (~7 kWh) to large (~300 kWh). • Some objects show daily cycles, others more constant.• Two objects flagged for possible solar production (low/negative daytime consumption). Weather Data (Meteoblue)
Temperature (°C)	Range: –23.25 to 31.41Average: ~7.76
Precipitation Total (mm)	Mostly low/zero, peaks up to 9.3 mm
Wind Speed (m/s)	Range: 0 to 42.66Average: ~13.7
Wind Direction (°)	Range: 0–360Average: ~198
Cloud Cover (%)	High, Medium, Low clouds: 0–100 %
Shortwave Radiation (W/m^2)	Direct, Diffuse, Total Nighttime: 0Peaks: >800 (direct radiation)

3.3. Regression baseline method testing

To evaluate more adaptive and data-driven approaches, three regression models were applied using the input data described in Table 6 to the Latvian objects: linear regression, polynomial regression, and random forest regression. Comparing the previous method this requires additional information next to the object consumption data that's why weather data form Metoeblue where collected. Linear and polynomial models showed moderate performance, with aggregated MAPEs between 11 % and 13 %. In contrast, the random forest model captured nonlinearities in consumption behaviour and achieved significantly better results, with an aggregated MAPE of 2.78 % and MAE of 3.62 kWh indicating accuracy Grade as A.

Regression models proved especially valuable for temperature-dependent loads and objects with significant PV contributions—cases where weather variables and irradiance data materially influence baseline accuracy. While regression baselines are computationally more demanding and require periodic re-training, their performance suggests they are highly suitable for aggregators or DSOs managing a portfolio of thermally driven or prosumer loads.

To represent graphically 3 subgroup regression baseline method results in Fig. 6 and compare it same time interval is taken.

The summary of the applied regression baseline methos and results presented in the Table 10:

3.4. Similar-Objects control group baseline testing

A baseline control-group approach was tested by identifying clusters of similar consumption objects using principal component analysis (PCA) and clustering algorithms. The treatment objects were compared with the aggregated behaviour of the matched groups. This approach mimics the logic of causal impact estimation and is widely used in pilot programmes internationally.

Although the control group method achieved a competitive MAE (3.76 kWh) fitting with Grade A accuracy, its aggregated MAPE exceeded 130 %, mainly due to low baseline denominators for certain objects and substantial behavioural variance within small clusters and falling to

Table 7
Historical baseline parameters and their MAE and MAPE results with accuracy grade.

#	High X of Y Historical Values	SDA Applied	SDA Timing	SDA Cap	MAE Score	MAPE (Aggregated)	Accuracy Grade level
1	High 10 of 20	Yes	1 h before delivery	20 %	16.68	22.04 %	Grade C
2	High 10 of 15	Yes	1 h before delivery	20 %	14.08	16.09 %	Grade C
3	High 5 of 15	Yes	1 h before delivery	20 %	19.09	24.24 %	Grade C
4	High 5 of 10	Yes	1 h before delivery	20 %	15.76	18.01 %	Grade C
5	High 15 of 20	Yes	1 h before delivery	20 %	13.04	15.69 %	Grade C

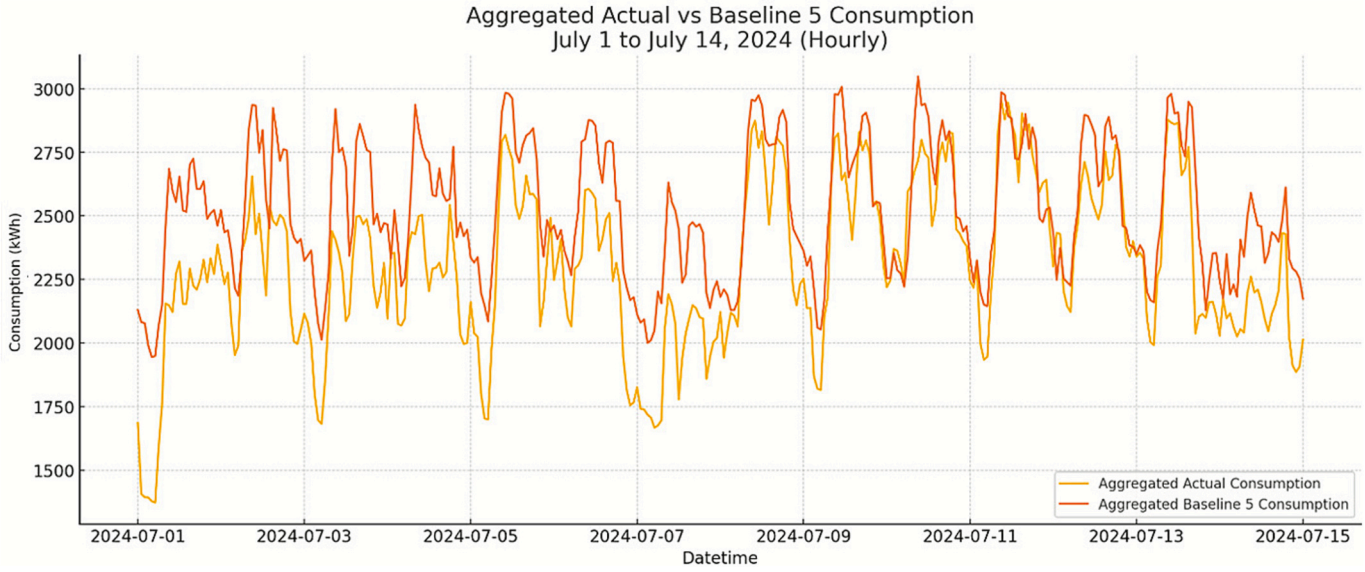


Fig. 4. The best performing set of baseline parameters (High 15 of 20, 1 h before delivery adjustment with 20 % CAP). Yellow curve represents actual consumption and orange curve represents calculated baseline. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 8
Testing of other results of the middle X of Y parameters.

#	Methodology Description	SDA Applied	SDA Timing	SDA Cap	MAE	Aggregated MAPE (%)	Accuracy Grade level
1	Middle 3 of 7 historical values	Yes	1 h before delivery	50 %	9.59	6.26 %	Grade B
2	Middle 10 of 14 historical values	Yes	1 h before delivery	50 %	8.39	7.13 %	Grade B
3	Middle 3 of 5 historical values	Yes	1 h before delivery	40 %	9.23	5.98 %	Grade B
4	Middle 8 of 10 historical values	Yes	1 h before delivery	50 %	8.12	6.06 %	Grade B
5	Middle 12 of 16 historical values	Yes	1 h before delivery	30 %	9.67	8.85 %	Grade B

Table 9
Baseline results when the high and Middle X of the Y parameter was replaced with an Average of selected days period.

#	Methodology Description	SDA Applied	SDA Timing	SDA Cap	MAE	Aggregated MAPE (%)	Accuracy Grade level
1	Average of 7 historical values	Yes	1 h before delivery	50 %	6.59	4.62	Grade B
2	Average of 14 historical values	Yes	1 h before delivery	50 %	6.74	4.73	Grade B
3	Average of 5 historical values	Yes	1 h before delivery	40 %	6.85	4.83	Grade B
4	Average of 10 historical values	Yes	1 h before delivery	50 %	6.65	4.60	Grade B
5	Average of 16 historical values	Yes	1 h before delivery	30 %	7.39	5.29	Grade B

Grade C ranges. These findings highlight that control-group baselines require a sufficiently large and homogeneous population to produce reliable results. Therefore, while theoretically strong in integrity, this method is most suitable for large portfolios or community-level pilots and simplicity is in the middle by assessment criteria.

3.5. *Ex-Post baseline evaluation during curtailment events*

To validate the practical settlement accuracy of the preferred baseline families, ex-post evaluation was conducted on two real flexible sites with historical baseline method: an HVAC system (three events) is presented in Fig. 7 and a wooden pole factory (two events) presented in Fig. 8. For each event, the magnitude and timing of curtailment were agreed on beforehand. The baselines were computed for the event hours and compared with the actual consumption.

For HVAC events, the MAPE values ranged from 0.7 % to 50.5 %, with low-error events demonstrating strong alignment between predicted and actual consumption. Variability was associated with rapidly changing cooling demand and weather conditions—factors for which regression baselines may outperform historical approaches. For the wooden pole factory, MAPE values ranged from 10.5 % to 17 %, consistent with expectations for stable industrial loads.

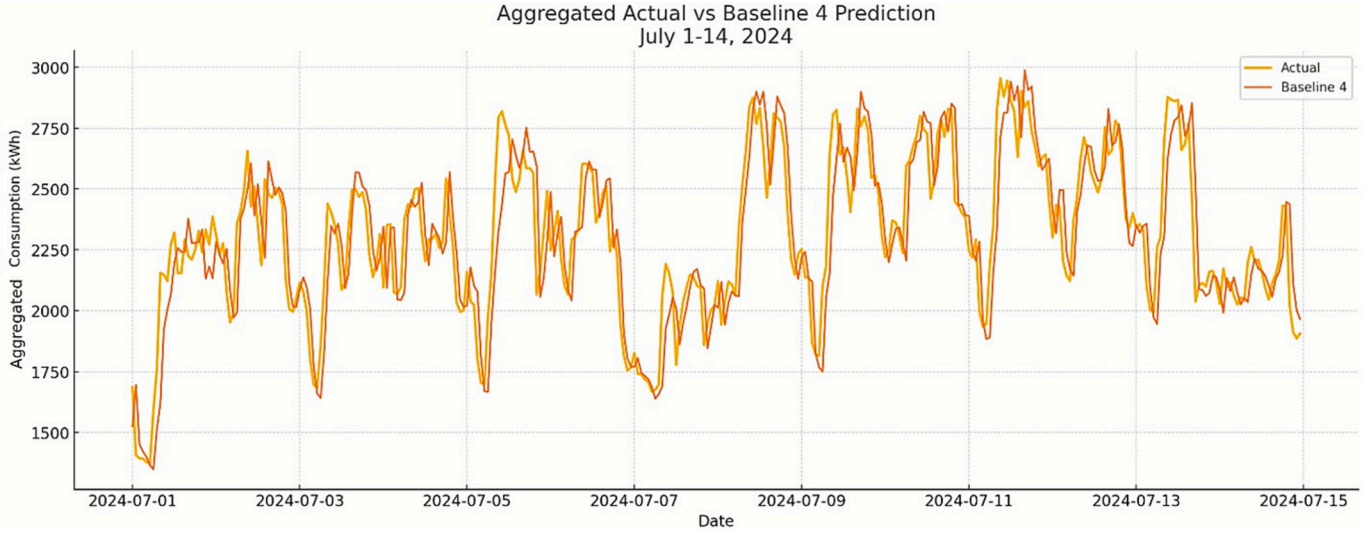


Fig. 5. Historical values and calculated baseline of the best performing set of historical baseline parameters (average of 10 historical days for each period with adjustment on the same day (1 h before delivery) capped at 50 %).

These results confirm that high-quality baselines can reliably reflect true load reductions in real activation scenarios, but they also highlight the importance of aligning baseline selection with the physical characteristics of the flexible resource.

3.6. Lithuanian hybrid energy system use case

Additional validation was performed using the Lithuanian Demand Response Demonstration Project, conducted by Litgrid and Kaunas University of Technology. The project evaluated a hybrid system comprising a 1 MW/1.1 MWh battery energy storage system (BESS), multi-building power-to-heat (P2H) installations with thermal storage, rooftop PV systems, and a cooling centre with cyclic compressors. Based on the principal diagram in Fig. 9, Feeder 1 with the PV generation and heat pump was taken and Feeder 2 as university regular load consumption data.

Collected measurement data used for the baseline method validations for Feeder 1–2 indicated in the Table 11.

We applied previously used Historical baseline with 10 previous days average values for the respective hours of the working day, if it is a weekend average values of previous weekends for the respective hours are used and presented in Fig. 10. The results indicate MAE 2.14 and 4.5 indicating accuracy level as Grade C. In addition, a linear regression baseline was generated using solar radiation, temperature, wind speed and humidity based on the factors presented in the Table 11 giving 5,38 and 5,13 MAE and Polynomial regression with expanded coefficients not increased accuracy significantly indicating MAE 4,91 and 4,48 for this data set. Presenting this both calculation in Figs. 11 and 12.

3.7. Synthesis of use case findings

Across all tested cases, the proposed baseline-selection methodology demonstrated a strong alignment between asset characteristics, technical applicapility and baseline performance.

- Historical baselines performed best for stable industrial loads and repeated patterns.
- Regression baselines produced superior accuracy for weather-sensitive and prosumer loads.
- Control-group methods require large datasets to outperform individual-driven models.

Baseline suitability depends fundamentally on the physical and thermodynamic characteristics of the flexible asset. For thermally driven systems, the baseline must capture heat storage behaviour, compressor cycling, and temperature-dependent consumption dynamics to ensure realistic flexibility quantification. These insights support the broader conclusion of the article: accurate and integrity-preserving baseline selection is essential not only for fair settlement of DSR but also for unlocking flexible thermal and electrical demand in Baltic electricity markets.

4. Impacts of market welfare and the potential for practical flexibility in the Baltic region

Section 3 evaluates the economic value of integrating small-scale flexibility, particularly DSR and thermally driven loads, into the Baltic electricity markets. While Section 2 demonstrated how robust baseline methodologies can measure delivered flexibility, the present analysis focusses on quantifying the system-level welfare impact of activating DSR in the wholesale market. The welfare evaluation complements the baseline assessment by showing why accurate and integrity-preserving baseline methodologies are not just accounting tools but essential for economically efficient flexibility activation. The results provide a numerical foundation for understanding how even relatively small volumes of flexible capacity can contribute to reducing market prices, increasing social welfare, and supporting the integration of renewable energy.

The analysis concentrates on the implicit activation of flexible loads in the DA market. To make this analysis data from active market operator in the Baltic region, Nordpool, aggregate price curves where used. This data set is actual DA market results for the whole region taking in account all complexity behind it with all market participants bids (block bids, linked bids, exclusive bids and etc) flows on the interconnections to find most valuable socioeconomic market results for each MTU.

This approach illustrates the potential system benefits that could arise from the automated, price-responsive behaviour of heat pumps, HVAC loads, and other distributed resources. Given the rapidly growing share of RES in the Baltic region and the increasing electrification of heating, these findings are relevant for both power and thermal sector planners.

4.1. Methodological setup

Hourly supply P_s(V) and demand P_d (V) curves from the Baltic day-ahead market for period of 2024–06-01 to 2024–08-31 were used to

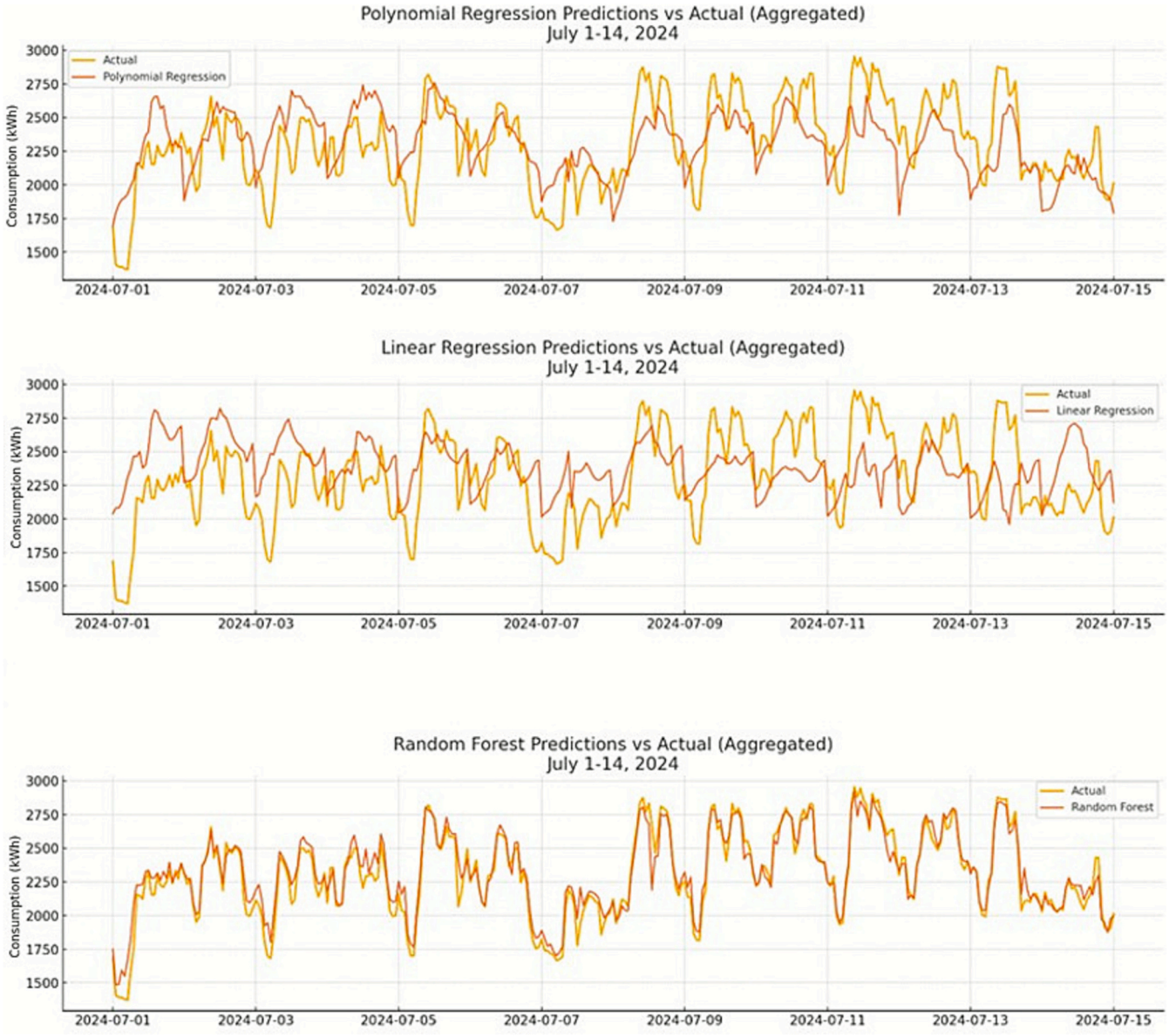


Fig. 6. The actual (yellow) and baseline (orange) curves of the regression model. Linear and polynomial regressions (top two graphs) clearly show higher errors than the random forest model (bottom graph). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 10
The results of the regression method and the assessment criteria.

Model	MAE	MAPE (aggregated %)	Accuracy Grade	Simplicity Grade	Integrity Grade
Polynomial Regression (deg 2)	15.09	11.74	Grade C	Grade B	Grade A
Linear Regression	17.36	13.40	Grade C	Grade B	Grade A
Random Forest	3.62	2.78	Grade A	Grade B	Grade B

estimate equilibrium quantities (V*, P*) and total social welfare, defined as the integral of the area between demand and supply up to the traded volume. Demand response was modelled as a leftward shift of the demand curve by ΔV, such the demand response was modelled a shift to the left of the demand curve by ΔV, such that:

$$P_{d'}(V) = P_d(V + \Delta V) \tag{2}$$

and the new equilibrium (V_Δ*, P_Δ*,) was determined numerically. The welfare change was computed as:

$$\Delta TS = \int_0^{V_\Delta^*} [P_d(V + \Delta V) - P_s(V)]dV - \int_0^{V^*} [P_d(V) - P_s(V)]dV \tag{3}$$

This isolates the financial benefit (or loss) to society from pure demand reduction, holding supply conditions fixed. Results is presented in the Table 12.

The results confirm that flexible DSR increases social welfare primarily in tight, high-price hours, where the supply curve is steep and the marginal cost of generation rises rapidly. In these hours, small demand reductions remove high-cost MWh from the market and trigger a significant price drop, improving overall welfare.

4.2. Interpretation

For Δ = 5 MW, the market shows welfare gains in about two-thirds of

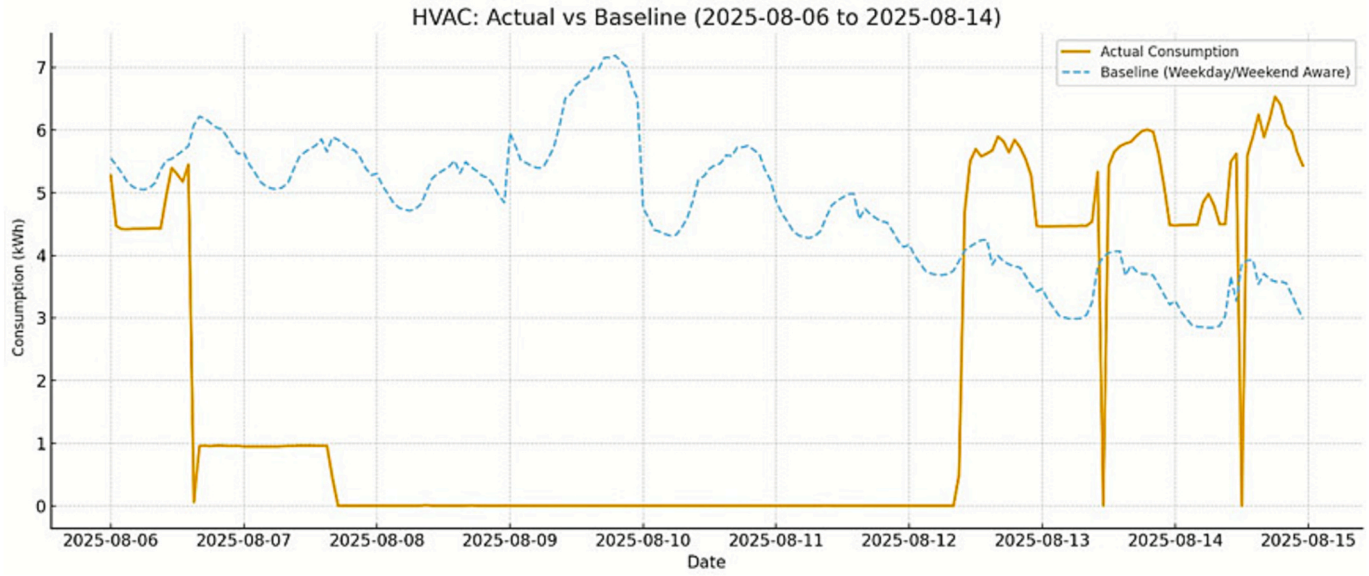


Fig. 7. HVAC actual consumption (yellow) and calculated baseline (blue) curves. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

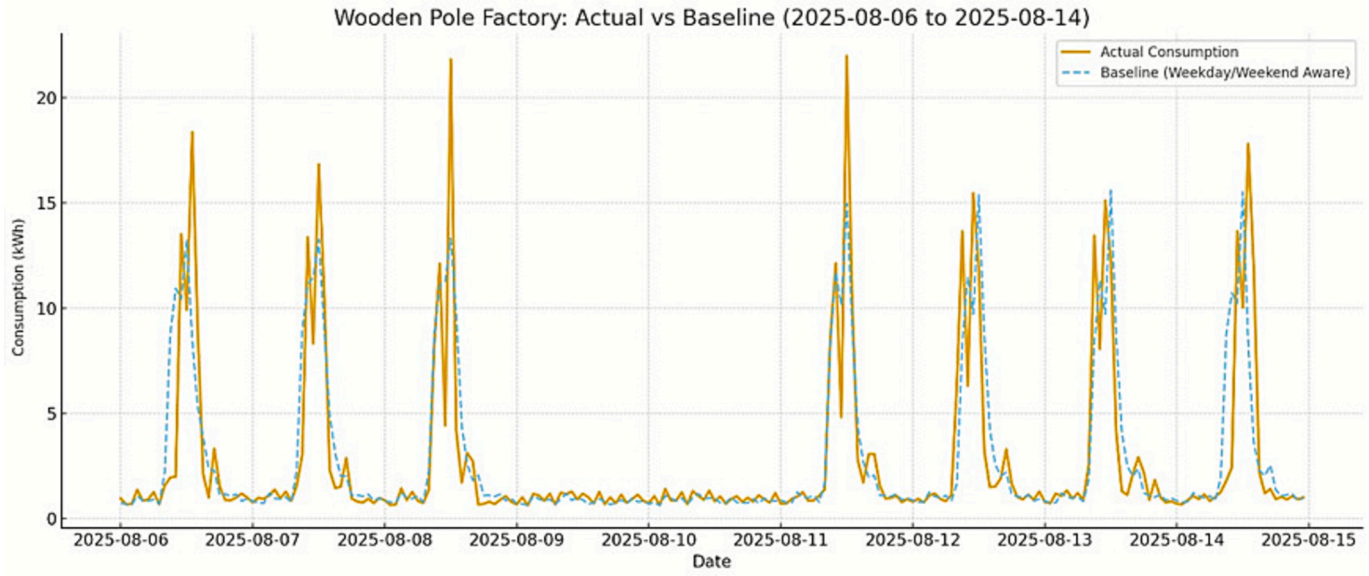


Fig. 8. Actual consumption of the wood pole factory (yellow) and calculated baseline curve (blue). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

hours; prices fall by an average of 6.6 EUR/MWh, producing cumulative welfare (only positive hours) benefits of 0.6 million euros. For $\Delta = 50$ MW, the effect is stronger but nonlinear: welfare gains occur in 56 % of hours, with price drops averaging 33 EUR/MWh. However, when large-scale DSR is applied uniformly across all hours (including low-price ones), infra-marginal welfare losses dominate, resulting in a net welfare decrease of 34 million euros. Visualisation of 50MW mechanism principle presented in the Fig. 13.

Thus, targeting DSR to periods of scarcity or high marginal cost yields strong social and economic benefits, whereas untargeted DSR reduces efficiency.

4.3. Policy and market design implications

The analysis demonstrates that dynamic, price-contingent DSR programmes, where small flexible loads reduce consumption during high-price hours, can produce measurable social welfare gains and lower market-clearing prices without compromising system efficiency. Given the growing penetration of RES in the region, this flexibility becomes increasingly valuable to mitigate volatility and support system balancing. Fig. 14 illustrates that the welfare impact of DSR is critically dependent on its temporal targeting. Static or untargeted load reductions, applied uniformly across all hours, may erode the infra-marginal surplus and even lead to negative welfare effects during periods of low or negative prices. These findings highlight that effective flexibility activation requires dynamic, price-based control, ensuring that demand adjustments occur precisely when the system value of flexibility is highest. Automated, market-responsive DSR mechanisms therefore represent a more reliable pathway to improve efficiency, integrate renewables, and maximise welfare outcomes in Baltic electricity markets characterised by increasing RES variability and tightening capacity margins.

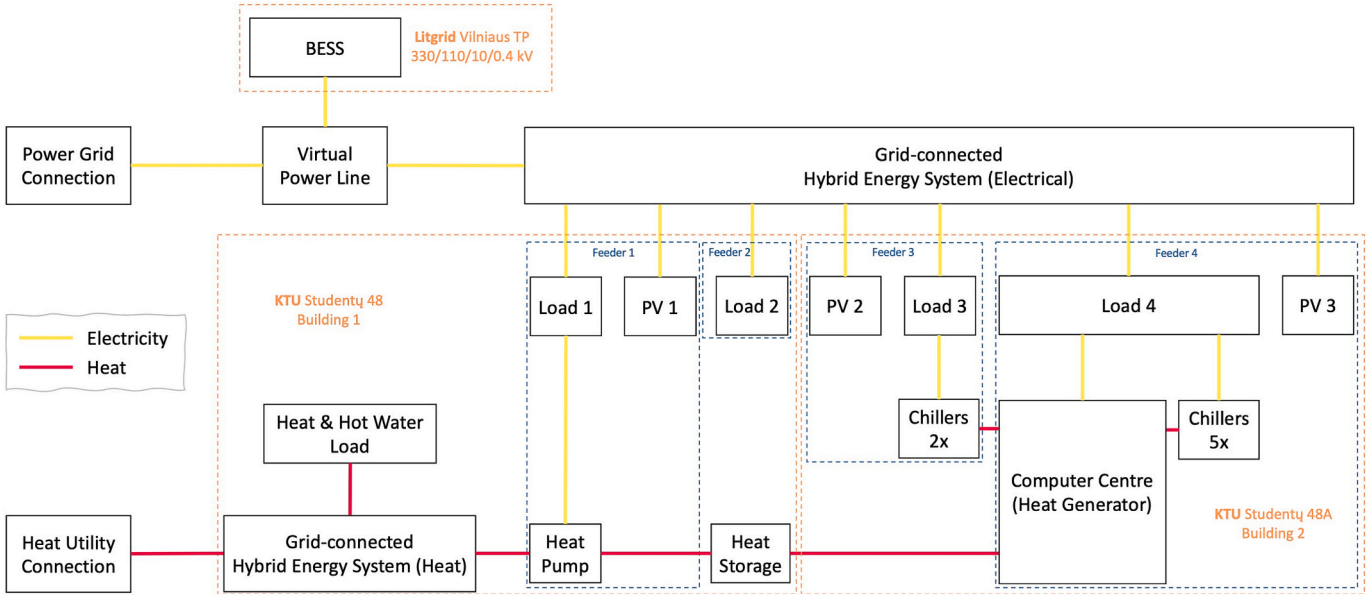


Fig. 9. Principal diagram of KTU-Litgrid testing area.

Table 11
Data used for baseline methods testing on the Lithuania use case.

Category	Details
Consumption Data	
Number of Feeders	2
Time Period	2023-09-01 01:00:00 to 2023-10-31 00:00:00
Average Consumption per Object	KTU1 Feeder Minimum: 3.3 kWhMaximum: 47.27 kWh Mean average: 16.38 kWh KTU2 Feeder Minimum: 0 kWhMaximum: 59.26 kWh Mean average: 30.7 kWh
KTU 1 Feeder	
Intercep	18.49
Temperature coefficient	−0.65
Humidity coefficient	+0.04
Wind Speed coefficient	+0.12
Solar Radiation coefficient	+0.012
KTU 2 Feeder	
Intercep	−10.65
Temperature coefficient	+0.44
Humidity coefficient	+0.45
Wind Speed coefficient	+0.18
Solar Radiation coefficient	−0.023

5. Conclusions

This research develops a comprehensive and reproducible assessment framework for selecting baseline method suitable for DSR and thermal flexibility applications and tested using real consumer data from the Baltic region. By systematically evaluating eight baseline families against defined technical requirements and multicriteria performance metrics, the research identifies approaches that balance accuracy, robustness and overall simplicity for distributed flexibility resources, including electrified heating and hybrid power-to-heat systems.

From a technical perspective, the results demonstrate that baseline choice has a measurable impact on flexibility quantification accuracy and robustness, particularly for thermally driven and hybrid electric–thermal loads. Historical averaging methods with same-day adjustment (50 %) achieved moderate accuracy (MAE = 6.65 kWh, Grade B), offering a favorable balance between simplicity and robustness for low-data and settlement-oriented applications. More advanced data-driven approaches, especially random forest regression models, substantially reduced baseline error (MAE = 3.62 kWh, Grade A), confirming their suitability for weather-dependent, prosumer-influenced loads characterized by thermal inertia and cyclic behavior. However, these gains in accuracy come at the cost of increased computational complexity and lower simplicity scores, highlighting a practical trade-off that must be explicitly considered in market implementation. Empirical testing across 39 Latvian DSO consumption sites and complementary industrial and HVAC case studies confirms that adaptive and weather-aware baselines are essential for realistically modelling electrified heating demand.

From a market and system impact perspective, the analysis shows that baseline accuracy is not only a technical concern but a key determinant of market efficiency and welfare outcomes. The welfare analysis based on Baltic day-ahead supply–demand curves further quantified the system-level effects of DSR: implicit demand reductions of 5–50 MW in the 2024 Baltic day-ahead market produced social welfare gains of up to €4.3 million and reduced scarcity-hour clearing prices by 33 €/MWh under 50 MW activation scenarios. In contrast, non-targeted or poorly aligned activation, representative of inadequate baseline selection or settlement distortions, produced social welfare losses of up to €34 million. These results provide concrete evidence that improvements in baseline accuracy directly support more efficient activation decisions, fairer settlement, and higher system-wide welfare.

With respect to regional and broader relevance, the empirical findings are grounded in Baltic market conditions, including high-RES penetration, cold-climate heating demand, and emerging aggregator participation frameworks. While specific welfare magnitudes and activation thresholds are region-dependent, the methodological insights are transferable to other RES-intensive power systems undergoing rapid electrification of heating. In particular, the demonstrated importance of weather sensitivity, thermal inertia representation, and settlement-oriented simplicity is likely to hold across northern and continental European markets with similar flexibility portfolios.

In summary, the article shows that accurate and transparent baseline methodologies are fundamental to unlocking distributed flexibility, not merely a technical refinement. The proposed methodology bridges methodological precision with real-world validation, ensuring reliable DSR measurement and supporting flexibility-driven decarbonisation across the heating and power sectors. Future work will extend the framework to hybrid electro–thermal systems, including the Lithuanian

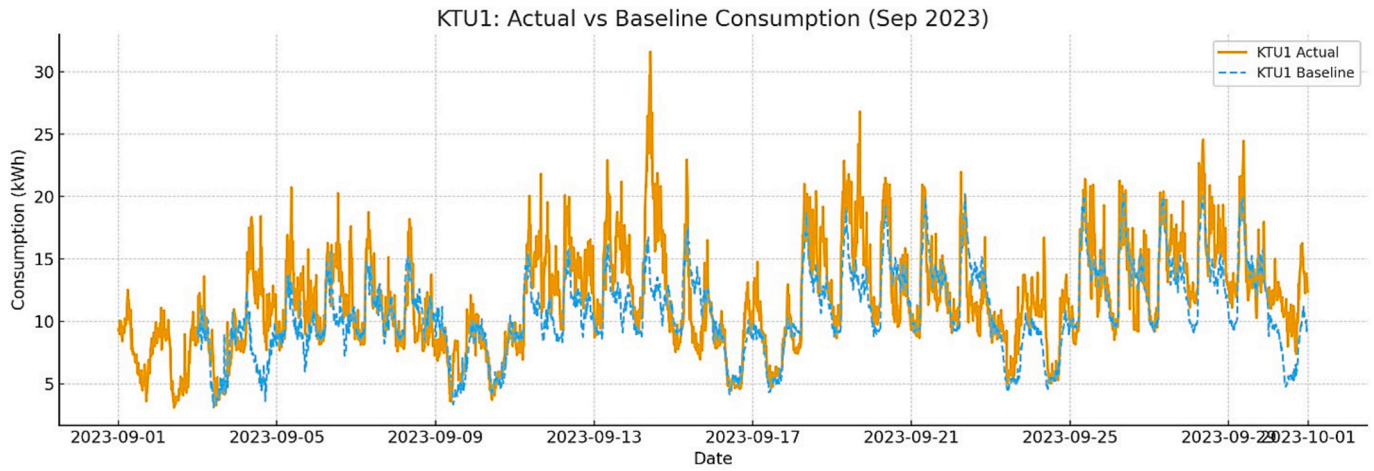


Fig. 10. Comparison of actual consumption data in Feeder 1 with the historical baseline method.

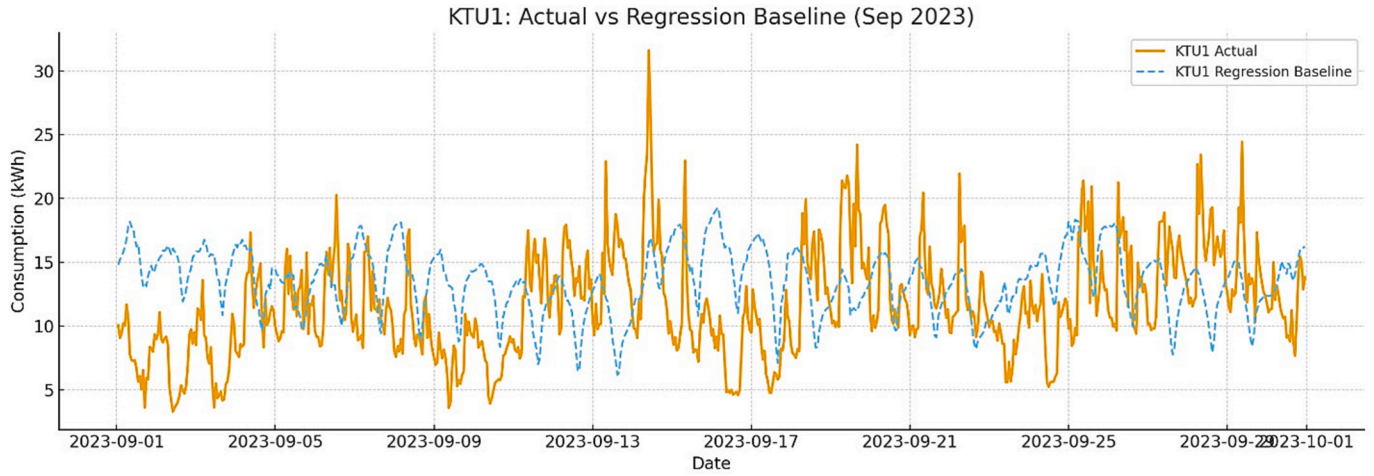


Fig. 11. Comparison of actual consumption data in Feeder 1 with the linear regression baseline method.

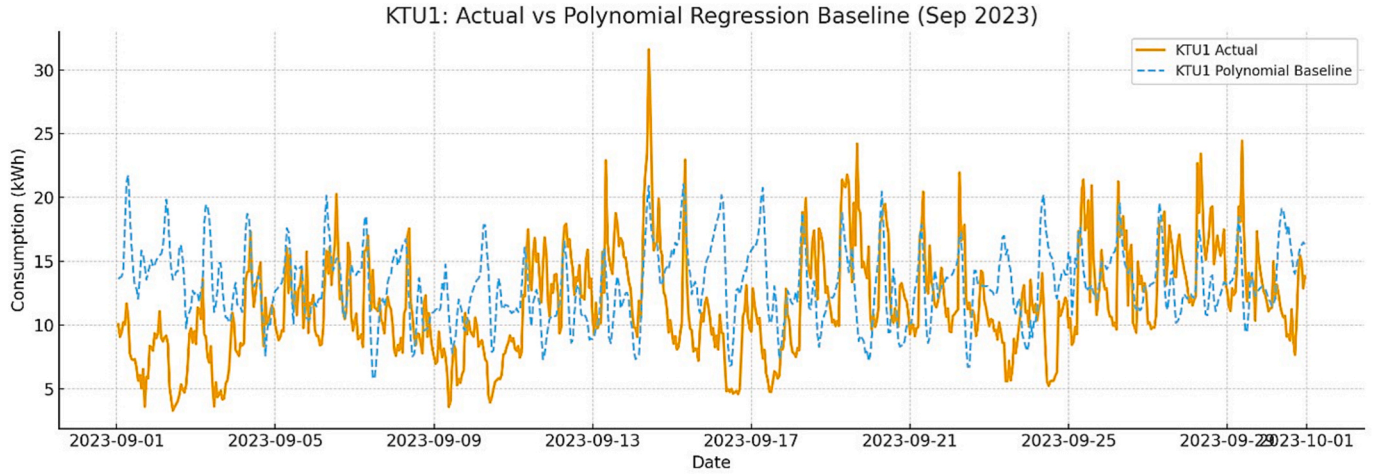


Fig. 12. Comparison of actual consumption data in Feeder 2 with Polynomial regression baseline method.

pilot demonstration to capture cross-sectoral flexibility potential and advance adaptive, AI-assisted baselines for large scale market integration.

7. Funding sources

This research is done based on the EU research project “Development of a new flexibility service-providing solution for efficient management of the network congestion using demand response potential of the

Table 12
Results: 5 MW and 50 MW demand reduction.

Metric	5 MW Reduction	50 MW Reduction
Periods analysed	2,040 h	2,040 h
Hours with positive ΔTS	1,304 (64 %)	1,140 (56 %)
Total welfare gain (positive hours only)	0.59 million euros	4.30 million euros
Average price decrease (positive hours)	−6.6 EUR/MWh	−33.2 EUR/MWh
Typical equilibrium volume	1.5–2.1 GWh/h	1.5–2.1 GWh/h
Welfare change when applied to all hours	−3.0 million euros	−34.0 million euros

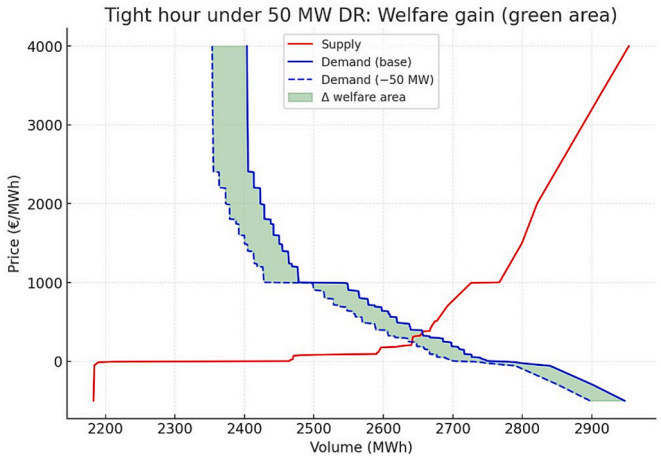


Fig. 13. Tight hours under 50 MW DR — large green shaded area shows the avoided expensive MWh and a sizable price drop, welfare gain. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

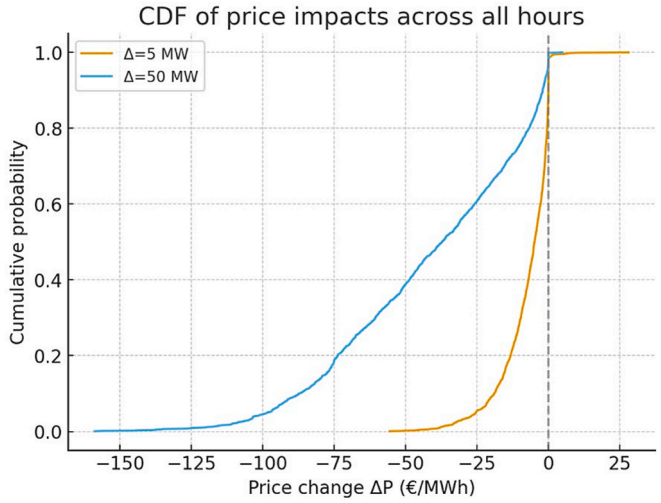


Fig. 14. The curve shows that both 5 and 50 MW demand reductions mostly produce negative ΔP (price drops), but the largest drops occur during the highest-priced and tightest hours. Positive ΔP can occur when the demand bid with negative price is eliminated, which increases the price.

electricity end-users" EU financing nr.: 2.2.1.3.i.0/1/24/A/CFLA/003.

CRediT authorship contribution statement

Deividas Šikšnys: Writing – original draft, Methodology, Investigation, Conceptualization. **Jonas Vaičys:** Resources, Methodology, Formal analysis, Data curation. **Saulius Gudžius:** Writing – review & editing. **Roma Račkienė:** Writing – review & editing. **Matas Grigošaitis:** Writing – review & editing, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was carried out with substantial support from the Technology and Physics Science Excellence Center (TiFEC) under project No. S-A-UEI-23-1, funded by the Research Council of Lithuania. As well authors acknowledge the support and collaboration and support of AS "Sadales tīkls" in the development of this work.

Data availability

Data will be made available on request.

References

[1] "Group Strategy group strategy 2035." Accessed: Nov. 05, 2025. [Online]. Available: https://strategija.epsog.lt/documents/EPSOG_Group_strategy.pdf.
[2] Lithuanian republic energy ministry, National Energy and Climate Action Plan, Available: Nacionalinis energetikos ir klimato srities veiksmų planas - Lietuvos Respublikos energetikos ministerija (accessed Oct. 10, 2025).
[3] IEA. The Future of Heat Pumps – Analysis. IEA, Nov. 2022. https://www.iea.org/reports/The Future of Heat-pumps.
[4] Analysis of the ban of fossil heating technologies on NECPs and national energy dependency 1 GREEN HEAT FOR ALL 2 A review of the necessity and feasibility of a just and green heat transition. Accessed: Nov. 05, 2025. [Online]. Available: https://www.coolproducts.eu/wp-content/uploads/2023/10/Green-Heat-for-All-2-Final.pdf.
[5] D. Koolen, F.M. De, S. Busch, Flexibility requirements and the role of storage in future European power systems, JRC Publ. Reposit. (2023), https://doi.org/10.2760/384443.
[6] Z. Marijanovic, P. Theile, B.H. Czock, Value of short-term heating system flexibility – a case study for residential heat pumps on the German intraday market, Energy 249 (2022) 123664, https://doi.org/10.1016/j.energy.2022.123664.
[7] N. Javanshir, S. Sanna Syri, Tervo, A. Rosin, Operation of district heat network in electricity and balancing markets with the power-to-heat sector coupling, Energy 266 (2023) 126423, https://doi.org/10.1016/j.energy.2022.126423.
[8] Z. You, S.D. Lumpp, M. Doepfert, P. Tzscheutschler, C. Goebel, Leveraging flexibility of residential heat pumps through local energy markets, Appl. Energy 355 (2024) 122269, https://doi.org/10.1016/j.apenergy.2023.122269.
[9] Z. Wang, E. Trutnevyte, Demand-side flexibility of electric vehicles and heat pumps in the swiss electricity system with high shares of renewable generation, Energy 338 (2025) 138903, https://doi.org/10.1016/j.energy.2025.138903.
[10] Market Reports. Entsoe.eu, 2025. https://www.entsoe.eu/publications/market-reports/ (accessed Nov. 05, 2025).
[11] "Single Day-ahead Coupling (SDAC). www.entsoe.eu. https://www.entsoe.eu/network_codes/cacm/implementation/sdac/.
[12] "Elektriturul tarbimiskajas osalemise kontseptsiooni avalik konsultatsioon | Elering. Elering.ee, 2025. https://elering.ee/node/3942 (accessed Nov. 05, 2025).
[13] L. Sadovica, V. Lavrinovics, A.-S. Sauhats, G. Junghans, K.M. Lehtmets, Estimating energy reduction amount in the event of demand response activation: baseline model comparison for the baltic states, in: 2018 15th International Conference on the European Energy Market (EEM), 2018, pp. 1–5, https://doi.org/10.1109/eem.2018.8469796.
[14] P.A.V. Gade, T. Skjøtskift, H.W. Bindner, J. Kazempour, Ecosystem for demand-side flexibility revisited: the Danish solution, Electr. J. 35 (9) (2022) 107206, https://doi.org/10.1016/j.tej.2022.107206.
[15] L. Kurevska, A. Sauhats, G. Junghans, V. Lavrinovcs, Measuring the impact of demand response services on electricity prices in Latvian electricity market, in: 2020 IEEE 61th International Scientific Conference on Power and Electrical Engineering of Riga Technical University (RTUCON), 2020, pp. 1–4, https://doi.org/10.1109/rtucon51174.2020.9316485.
[16] Lithuanian transmission system operator, Litgrid.eu, Baseline methodology for independent aggregator. https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fwww.litgrid.eu%2Fuploads%2Ffiles%2Fdir717%2Fdir35%2Fdir1%2F18_0.php&wdOrigin=BROWSELINK (accessed Nov. 05, 2025).
[17] Lithuania National Energy Regulatory Council, List of independent electricity suppliers. Accessed: Sep. 05, 2025. [Online]. Available: VKEKK List of independent electricity suppliers.
[18] EU DSO Entity and ENTSO-E Proposal for a Network Code on Demand Response V1 2024 May 8th. EU DSO Entity and ENTSO-E Proposal for a Network Code on Demand Response (accessed Sep. 14, 2025).

[19] Regulation - EU - 2024/1747 - EN - EUR-Lex. *Europa.eu*, 2024. https://eur-lex.europa.eu/eli/reg/2024/1747/oj/eng (accessed Jun. 14, 2025).
[20] Liga Sadovica, K. Marcina, Valentins Lavrinovics, and G. Junghans. Facilitating energy system flexibility by demand response in the baltics — Choice of the market model. pp. 1–6. 2017. Doi: 10.1109/rtucon.2017.8124834.
[21] J.Y. Weng, R. Rajagopal, Probabilistic baseline estimation based on load patterns for better residential customer rewards, Int. J. Electr. Power Energy Syst. 100 (2018) 508–516, https://doi.org/10.1016/j.ijepes.2018.02.049.
[22] C. Ziras, C. Heinrich, M. Pertl, H.W. Bindner, Experimental flexibility identification of aggregated residential thermal loads using behind-the-meter data, Appl. Energy 242 (2019) 1407–1421, https://doi.org/10.1016/j.apenergy.2019.03.156.
[23] C. Ziras, C. Heinrich, H.W. Bindner, Why baselines are not suited for local flexibility markets, Renew. Sustain. Energy Rev. 135 (2021) 110357, https://doi.org/10.1016/j.rser.2020.110357.
[24] E. Larsen, K. Rosenørn, A. Jónasdóttir, Baselines for evaluating demand response in the Ecogrid 2.0 project GUIDELINES FOR THE SUBMISSION OF THE FINAL PAPER.
[25] M. Sun, Y. Wang, F. Teng, Y. Ye, G. Strbac, C. Kang, Clustering-based residential baseline estimation: a probabilistic perspective, IEEE Trans. Smart Grid 10 (6) (2019) 6014–6028, https://doi.org/10.1109/tsg.2019.2895333.
[26] S. Park, S. Ryu, Y. Choi, J. Kim, H. Kim, Data-driven baseline estimation of residential buildings for demand response, Energies 8 (9) (2015) 10239–10259, https://doi.org/10.3390/en80910239.
[27] Y. Jo, J. Hur, An improved ramp events forecasting of wind generating resources using ensemble learning of numerical weather prediction: the case of Jeju Island's wind farms, Therm. Sci. Eng. Prog. 66 (2025) 103936, https://doi.org/10.1016/j.tsep.2025.103936.
[28] E. Lee, K.-E. Lee, H. Lee, E. Kim, W. Rhee, Defining virtual control group to improve customer baseline load calculation of residential demand response, Appl. Energy 250 (2019) 946–958, https://doi.org/10.1016/j.apenergy.2019.05.019.
[29] K. Li, B. Wang, Z. Wang, F. Wang, Z. Mi, Z. Zhen, A baseline load estimation approach for residential customer based on load pattern clustering, Energy Procedia 142 (2017) 2042–2049, https://doi.org/10.1016/j.egypro.2017.12.408.
[30] J. Vuelvas, F. Ruiz, G. Gruosso, Limiting gaming opportunities on incentive-based demand response programs, Appl. Energy 225 (2018) 668–681, https://doi.org/10.1016/j.apenergy.2018.05.050.
[31] D. Muthirayan, D. Kalathil, K. Poolla, P. Varaiya, Mechanism Design for Demand Response Programs, IEEE Trans. Smart Grid 11 (1) (2020) 61–73, https://doi.org/10.1109/tsg.2019.2917396.
[32] (a) Anut Arunaun, W. Pora, Anut Arunaun, and W. Pora. Baseline Calculation of Industrial Factories for Demand Response Application. pp. 206–212. 2018. Doi: 10.1109/icce-asia.2018.8552114; (b) J. Jazaeri, S. Tiwari, H. Jethani, V. S. Chauhan. Baseline methodologies for small scale residential demand response. 2016 IEEE Innovative Smart Grid Technologies – Asia (ISGT-Asia), 2016; pp. 747-752. https://doi.org/10.1109/ISGT-Asia.2016.7796478.
[33] Javad Jazaeri, Tansu Alpcan, R. Gordon, M. Brandao, T. Hoban, and C. Seeling. Baseline methodologies for small scale residential demand response. Nov. 2016. Doi: 10.1109/isgt-asia.2016.7796478.
[34] O. Valentini, P. Nikoleta Andreadou, A.L. Bertoldi, I. Saviuc, E. Kotsakis, Demand response impact evaluation: a review of methods for estimating the customer baseline load, Energies 15 (14) (2022) 5259, https://doi.org/10.3390/en15145259.
[35] L. Lind, J.P. Chaves-Ávila, O. Valarezo, A. Sanjab, L. Olmos, Baseline methods for distributed flexibility in power systems considering resource, market, and product characteristics, Util. Policy 86 (2023) 101688, https://doi.org/10.1016/j.jup.2023.101688.
[36] M. Troncia, S. Bindu, J. Pablo, C. Ávila, G. Willeghems, H. Gerard. Techno-economic assessment of proposed market schemes for standardized products D11.2 Authors. Accessed: Nov. 05, 2025, p. 108. [Online]. Available: https://www.onenet-project.eu/wp-content/uploads/2024/01/OneNet_D11.2_V1.0.pdf.
[37] S. Chondrogiannis, J. Vasiljevska, A. Marinopoulos, I. Papaioannou, G. Flego, Local electricity flexibility markets in Europe, JRC Publ. Repository (2022), https://doi.org/10.2760/9977.
[38] "Baseline Methodology Assessment Energy Networks Association." Accessed: Nov. 05, 2025. [Online]. Available: https://www.energynetworks.org/assets/images/ON20-WS1A-P7%20Baselining%20Assessment-PUBLISHED.23.12.20.pdf.
[39] G. Wu, Z. Lan, K. Zhou, S. Huang, A novel model averaging forecasting method for electricity consumption using electricity business expansion data, Energy Rep. 13 (2025) 3898–3914, https://doi.org/10.1016/j.egyr.2025.03.024.
[40] J. Lago, F. De Ridder, B. De Schutter, Forecasting spot electricity prices: deep learning approaches and empirical comparison of traditional algorithms, Appl. Energy 221 (2018) 386–405, https://doi.org/10.1016/j.apenergy.2018.02.069.
[41] M.S. Al-Musaylh, R.C. Deo, J.F. Adamowski, Y. Li, Short-term electricity demand forecasting with MARS, SVR and ARIMA models using aggregated demand data in Queensland, Australia, Adv. Eng. Inf. 35 (2018) 1–16, https://doi.org/10.1016/j.aei.2017.11.002.
[42] J.E. Pesantez, B. Li, C. Lee, Z. Zhao, M. Butala, A.S. Stillwell, A comparison study of predictive models for electricity demand in a diverse urban environment, Energy 283 (2023) 129142, https://doi.org/10.1016/j.energy.2023.129142.
[43] meteoblue AG. Weather – meteoblue. *meteoblue*, Oct. 24, 2025. https://www.meteoblue.com/en/weather/week/vilnius_lithuania_593116 (accessed Oct. 24, 2025).