This article is a preprint. Please cite the published version: https://doi.org/10.1016/j.jup.2023.101688

Baseline methods in the context of modern distributed flexibility: an evaluation considering multi-DER types, markets, and product characteristics

Leandro Lind [a,*] , José P. Chaves [a], Orlando Valarezo [a], Anibal Sanjab [b,c], Luis Olmos [a]

[a] Institute for Research in Technology (IIT)-ICAI School of Engineering, Universidad Pontificia Comillas, Santa Cruz de Marcenado 26, 28015, Madrid, Spain
[b] Flemish Institute for Technological Research (VITO), Boeretang 200, 2400 Mol, Belgium
[c] EnergyVille, Thor Park 8310, 3600 Genk, Belgium
[*] Corresponding author. Email address: llind@comillas.edu

ABSTRACT

Emerging flexibility markets are increasingly characterised by a multitude of product characteristics, differing flexibility buyers, and a disparate set of potential flexibility service providers (FSPs), including those connected at the distribution networks. This paper investigates the fitness of various baseline methodologies while considering different possible Distributed Energy Resource types, multi-DER FSPs, as well as diverse products and flexibility markets characteristics. As a result, a novel decision framework is proposed, aiming at providing a guideline for the choice of suitable baseline. The analysis showcases that no one-size-fits-all baseline method exists while proposing alternatives for various FSP types and flexibility market conditions.

Keywords:

Baseline, Distributed Energy Resources, Flexibility Service Provision, Aggregation.

1. INTRODUCTION

The recent call for the decarbonisation of power systems has triggered a range of new strategies, business models and use cases for enabling the efficient and reliable operation of the grid. Among those is the use of operational flexibility provided by Distributed Energy Resources (DERs) (Villar et al., 2018). The use of distributed flexibility can be seen both as a way to contribute to decarbonisation goals, as well as supporting the safe and secure operation of the grid (and its economic efficiency) through the provision of grid services to network operators (Lind et al., 2019; Valarezo et al., 2021).

Distributed flexibility can be categorised into two types, namely, (i) implicit flexibility and (ii) explicit flexibility (SEDC, 2016). The former refers to a voluntary change of consumption (or production) in reaction to an external signal. Reactions to wholesale energy prices or dynamic network tariffs are considered implicit flexibility. Explicit flexibility, on the other hand, refers to a contractual agreement in which the Flexibility Service Provider (FSP) agrees to change their schedule at specific conditions following specific requirements. This agreement can be the outcome of an organised market (e.g. balancing or congestion management markets, local flexibility markets) or another acquisition mechanism (Martín-Utrilla et al., 2022). Therefore, in the case of explicit flexibility, a product is possibly traded (e.g. manual Frequency Restoration Reserves-mFRR) and a service provided (e.g. balancing service to the Transmission System Operator - TSO). In such a case, the verification of product delivery (or service provision) is required to ensure the proper market operation and settlement mechanisms.

Typically, flexibility products are traded in the form of capacity, energy or both. The former can be defined as a band of available flexibility with a fixed duration to be (partially or fully) activated or not by the buyer (e.g. the TSO or Distribution System Operator - DSO) under certain conditions. This means that during the service provision, the FSP is available for its flexibility to be activated in the amount defined in the product, and this availability is usually remunerated in €/MW (which could be complemented by remuneration for activation). An energy product considers a pre-defined activation, meaning that both buyer and seller know when the activation starts and ends and the quantity of flexibility to be delivered by the time the product is traded. This type of product is usually remunerated in €/MWh.

Among the most well-established flexibility markets in Europe are the balancing markets. These markets trade both capacity and energy products. According to (ENTSO-E, 2021), capacity products are procured in a market-only fashion by TSOs for Frequency Contingent Reserves (FCR) in 14 Member States (MSs), automatic Frequency Restoration Reserves (aFRR) in 16 MSs and mFRR in 17 MSs. Energy products, too, are procured for FCR in 3 MSs, aFRR in 10 MSs and mFRR in 19 MSs. While the provision of capacity products may take place in different ways depending on the mode of activation (e.g. activation when needed or through mandatory participation in energy markets), energy products generally follow a more standardised activation process. The verification of the product delivery in balancing markets usually takes place by comparing the flexibility provided (metered data) with the generation or consumption schedule of the market agent. This schedule is determined in the wholesale energy markets. The agent is, therefore, a Balance Responsible Party (BRP) over that schedule. In the absence of such a schedule, alternative verification methods would be required to estimate the counterfactual position (i.e., injection or consumption levels) of an FSP (i.e., individual or collection of flexibility resources) had they not activated their flexibility, which would allow estimating the value of flexibility delivered. In other words, the delivered flexibility would be the difference between the meter outcome at the time of activation and the estimated counterfactual position. This counterfactual position is known as the baseline and is of key importance for the participation of DERs in flexibility provision.

Indeed, the participation of DER in flexibility markets is being proposed, tested and implemented as part of the decarbonisation and decentralisation strategies in power systems. The possibility for the participation of demand-side response in balancing markets is already mandated by the EU regulation and is currently being implemented in many countries (smartEn, 2018). The provision of distributed flexibility to DSOs is also at the centre of the debate. Current EU regulation also calls for the use of local flexibility by DSOs as a means to help manage the grid, possibly deferring or avoiding network reinforcements (*CEP Electricity Directive*, 2019). Therefore, many initiatives have focused on demonstrating and even implementing large-scale local flexibility markets (Ruwaida et al., 2022; Valarezo et al., 2021). However, a key challenge for distributed flexibility provision lies in the verification of energy product delivery. In traditional balancing markets and wholesale energy markets, a schedule exists, which serves as the baseline for delivery verification. For DERs, this schedule does not individually exist. When DERs are active consumers, a schedule exists in an aggregated form, which the retailer is responsible for balancing. However, no individual schedule exists for the specific DERs participating in flexibility markets. Therefore, a key challenge for the successful deployment of distributed flexibility is to unambiguously define the level against which to measure the amount of flexibility (product) delivered, or in other words, a baseline (Schittekatte et al., 2021). With the publication of the proposal for the Electricity Market Reform in Europe, the development of baseline methods become again a necessity, now expressed in the proposal (European Commission, 2023). The proposal mentions that the baseline should reflect “the expected electricity consumption without the activation of the peak shaving product”, in the specific case of the new flexibility product to be procured by the TSO.

Baselining DERs is not a new problem. With the first demand response programs came the need for the definition of a baseline method. Several were developed and implemented, mostly relying on statistical information on past consumption (AIEC, 2009; EnerNOC, 2009). Most of these methods, however, were

tailored-made for demand response programs in which the only type of service provider was the active consumer, and the only product was upward flexibility provisions (e.g. reduction of consumption). Therefore, the suitability of these methods has mostly been analysed from the perspective of demand response (Antunes et al., 2013; Jazaeri et al., 2016a; Mohajeryami et al., 2017a; Wijaya et al., 2014). Adaptations to the consumer's baseline have been proposed and analysed for specific groups of DERs. For example, in (Fonteijn et al., 2021), a variation is proposed for PVs, while in (Arunaun and Pora, 2018), a method is proposed specifically for industrial loads.

The participation of distributed flexibility in the present and future flexibility markets is notably more complex. The types of DERs are many: demand-side resources, distributed generation, storage systems, and aggregated DERs in different forms (different resources behind the meter, energy communities, independent aggregators). Markets and products are more complex, and upward and downward flexibility may be traded. Market timing may differ from 30 seconds (e.g. FCR) to weeks ahead of delivery. Such differences pose a fundamental question to the baselining of DERs: are the available methods suitable for the efficient participation of DERs in current and future flexibility markets?

This paper sheds light on this discussion by analysing the existing baseline methods proposed by academia and practitioners through an evaluation framework considering the characteristics of modern distributed flexibility provision. Firstly, baseline methods are characterised in terms of data needs for their application, who is responsible for the baseline calculation, if close-to-real-time adjustments are allowed, and if the baseline is dynamic or static. These features serve as input for the subsequent analysis of baselines. Secondly, the baseline methods are evaluated according to their accuracy, simplicity and integrity in a variety of modern flexibility provision use cases, including different types of DER, aggregation, different directions of activation and market timing. As a result, this paper proposes a novel set of guidelines for the selection of a baseline for DER participation in flexibility markets. This guideline aims at providing a reference for the selection of suitable baseline methods for flexibility trading in future flexibility markets.

The remainder of this paper is organised as follows. Section 2 provides an overview of the baseline methods proposed, used and analysed in the literature and by practitioners. Section 3 introduces the proposed methodology and analyses the different baseline methods according to the proposed criteria. Section 4 concludes the paper.

2. BASELINE METHODS

Different baseline methods have been progressively proposed both in the academic literature and by practitioners. Valentini et al. (2022) show that most methods currently used in international practice and European research projects rely on historical data, i.e., using metered data at the same time as the activation time (i.e., the time at which the flexibility is supposedly delivered) but from previous days that share similar criteria with the activation day (such criteria can be simply the type of day, e.g., weekday vs weekend). A common type of baseline method based on historical data is what is known as the **XofY** method (and its variations) (Arunaun and Pora, 2018; Mohajeryami et al., 2017b, 2017a; Ramos, 2019; Rossetto, 2018; Wang and Tang, 2022). The HighXofY, for instance, considers the average profile of the X days with the highest consumption within the set of Y-eligible previous days (e.g. weekdays if activation takes place on a weekday). Another method involving the use of historical metered data only is the **rolling average** of the previous X days of the same type (e.g. weekdays), potentially increasingly weighting the most recent days in order to capture the most current determinants for consumption, e.g., temperature and weather conditions (Holmberg et al., 2013; Miriam L. Goldberg and G. Kennedy Agnew, 2013; Tufts and Breidenbaugh, 2010). In addition, a **comparable day** to the activation day was proposed as a baseline for demand response. In this method, the flexibility provider chooses ex-post a non-activation day in the past that would reflect the conditions of the event day (EnerNOC, 2011). These methods have in common the fact that only metered

data is required for the calculation of the baseline. Eventually, a Same-Day Adjustment (SDA) can be performed using additional data, such as weather-based adjustments.

Another common method that relies on statistical data is the **regression method.** Regression models (mostly linear or polynomial) use a set of historical data to estimate a function that represents the relationship between the dependent variable (baseline consumption) and the independent variables (e.g. past consumption, season, weather, day-of-the-week) as elaborated in (Arunaun and Pora, 2018; Mohajeryami et al., 2017a; Vagropoulos et al., 2022).

More recently, novel methods based on neural networks and other **machine learning (ML)** techniques emerged, promising a higher accuracy for baseline estimation (Park et al., 2015).

The abovementioned baseline methods (except for the **comparable day** method) involve the ex-ante calculation of a baseline, mostly relying on historical data. Alternatively, other methods exist that do not require an ex-ante estimation. The simplest and most straightforward is the **Meter-Before-Meter-After (MBMA)**, which has been used in several countries, especially for short-time delivery products (DNV-GL, 2020). For specific use cases, a **zero baseline** was proposed, especially for behind-the-meter backup generation units (Miriam L. Goldberg and G. Kennedy Agnew, 2013). This method assumes the baseline is equal to zero. Therefore, any power injection during activation is considered a flexibility provision. A **control group** baseline has also been used in several countries, consisting of using a set of non-activated end-users with similar characteristics to those of the flexibility providers on the event day. The average profile of these end-users would then serve as a baseline for the FSPs (DNV-GL, 2020; Wang et al., 2018).

Finally, (Ziras et al., 2021) argue that the current baselining methods are not suitable for local flexibility markets, proposing that **capacity limitation** products should be used instead. A capacity limitation product limits the maximum power withdrawn or delivered by the end-user (and FSP) during the activation hours, eliminating the need for a baseline. In this method, the normally contracted capacity is temporarily reduced during the activation period. This approach, however, also means that the flexibility product (and therefore clearing mechanism) has to be defined in terms of maximum capacity, requiring the usage of specific market clearing algorithms, considering that the typical pay as-bid, uniform pricing and Vickrey-Clarke-Groves are not appropriate for this type of product trading (Heinrich et al., 2021).

The different methods lead to baselines with different characteristics that need to be considered for their implementation. Based on baseline methods' analysis in the literature above, we identify several dimensions along which different methodologies can be classified. In what follows, a mapping and classification of baseline methods is presented. First, some baseline methods require some form of ex-ante calculation – such as, e.g. baseline methods based on historical data as well as statistical and machine learning-based methods – while others either rely on very close to real-time data (e.g. *MBMA*), compute the baseline ex-post (e.g. *control group, comparable day*), or do not involve a specific individual baseline calculation (e.g. *zero baseline*, *capacity limitation*). Another characteristic of a method is the party that is held responsible for the baseline calculation, which can be the buyer (e.g. TSO, DSO) or the FSP. The calculation could be carried out within a dedicated platform on their behalf. The data required for the baseline calculation is also a key differentiating characteristic of each method, ranging from real-time metered data only, historical metered data, and other historical data (e.g. weather related). The possibility for SDA is also a criterion of differentiation between different methods. While some methods offer that possibility, for others, it is Not Applicable (NA), such as for ex-post methods (e.g. comparable day, control group). Finally, a baseline can be static (the same baseline value is considered during the whole activation period) or dynamic (a baseline profile is considered during the activation period). An *MBMF*, for example, is a static method by design, while a self-reported baseline can have different values for different points in time. Table 1 provides a classification of different methods and their characteristics. While some methods are well defined and characterised by their intrinsic

properties, others may depend on the method employed by the buyers or FSPs for the calculation and reporting of its baseline (e.g. *self-reported* or *capacity limitation*).

Table 1: Summary of baseline methods assessed

Baseline Technique	*Short description*	*Need for ex-ante calculation*	*Responsibility for baseline provision*	*Data requirement*	*Close to real-time adjustments possible (SDA)*	*Static or dynamic*
XofY Baselines	The average of the last high, middle or low consumption X days in a Y days list of eligible days is considered. Close-to-delivery adjustments are possible (e.g. weather differences).	Yes	Buyer or FSP	Historical metered data from individual FSPs	Yes	Dynamic
Rolling average	A rolling average of the past X days of the same type. Usually considers a higher weight to days close to the activation day.	Yes	Buyer or FSP	Historical metered data from individual FSPs	Yes	Dynamic
Comparable day	The FSP chooses data from the past that they consider similar to the activation day. The baseline choice is made ex-post.	No	FSP (as proposed in the literature. However, variations could be possible - e.g. defined by the buyer)	Historical metered data from individual FSPs	No	Dynamic
Regression methods	Past data is used to build a baseline function. The baseline function can then be used on the activation day, having different input parameters (e.g. past consumption data, temperature, season, etc.).	Yes	Buyer	Historical metered, weather-related data and others. Not necessarily from the individual FSPs	Yes	Dynamic
Machine learning techniques	Machine learning techniques are used to estimate the baseline for the activation day.	Yes	Buyer	Historical metered and weather-related data etc Not necessarily from the individual FSPs	Yes	Dynamic
Meter-Before-Meter-After (MBMA)	A reading of the meter is performed right before activation, which serves as the baseline.	No	Buyer (metering responsible)	Real-time metered data from individual FSPs	NA	Static
Zero baseline	The baseline is equal to zero. This method is mostly used for backup generators.	No	NA	NA	NA	Static
Control group	A group of non-FSP customers sharing similarities with the FSP being baselined is considered. Their average profile during the	No	Buyer	Real-time metered data from end-users other than the FSPs	NA	Dynamic

	activation is used as a baseline.					
Capacity limitation	A temporary limitation is placed on the maximum power the FSP can withdraw from the grid or deliver to it.	Depends on how the DSO computes the limitation.	NA	Depends on how the DSO computes the limitation.	NA	Static
Self-reported	The FSP is requested to report a profile.	Depends on how the FSP computes the limitation.	FSP	Depends on the methods chosen by FSP	Yes	Dynamic or static

3. BASELINE EVALUATION WITH RESPECT TO FLEXIBILITY MARKET CHARACTERISTICS

Typically, baseline methods are evaluated in terms of their accuracy, simplicity and integrity. These criteria were first introduced by (EnerNOC, 2009) and later used by other authors (Arunaun and Pora, 2018; Jazaeri et al., 2016b; Valentini et al., 2022). In order for a method to be **accurate**, the baseline computed should correctly estimate the level of consumption or production if the available flexibility is not activated. In addition, the method should be **simple** enough for stakeholders to understand, implement, and verify its outcome. Simplicity is a desired attribute not only from the implementation point of view but also as a way to ensure that the method is transparent for both flexibility buyers and sellers. Additionally, criterion **integrity** is used to determine to what extent a baseline method does not allow for the seller or buyer to misrepresent the flexibility delivered.

In this paper, the three baseline evaluation criteria are used to evaluate the suitability of the baseline methods for the three different key flexibility market characteristics, which are: (i) baselines for different DERs, (ii) multi-DER baseline (e.g. behind the meter, aggregators), and (iii) the characteristics of flexibility services. First, the DER technology used in the flexibility provision is analysed from the baseline perspective. In this paper, the assessment is done not only for demand-reduction FSPs, but also for controllable and non-controllable distributed generation (DG) and Energy Storage Systems (ESS). Second, baselines for multi-DER FSPs are analysed. These result from aggregating different technologies behind or in front of the meter. The suitability of each baseline method in this case is assessed. Finally, the baseline methods are assessed against the characteristics of the flexibility services, more specifically, their timing (e.g. week-ahead vs close to real-time) and direction of activation (upward vs downward flexibility).

Following this qualitative discussion on the fitness of the different baseline types, a decision framework is built in order to identify the possible baseline types for the different use cases in flexibility markets. Figure 1 depicts a representation of the assessment framework considered in this research work. According to it, three intrinsic baseline criteria are used to evaluate the different baseline methods in the context set by the several flexibility provision characteristics considered.

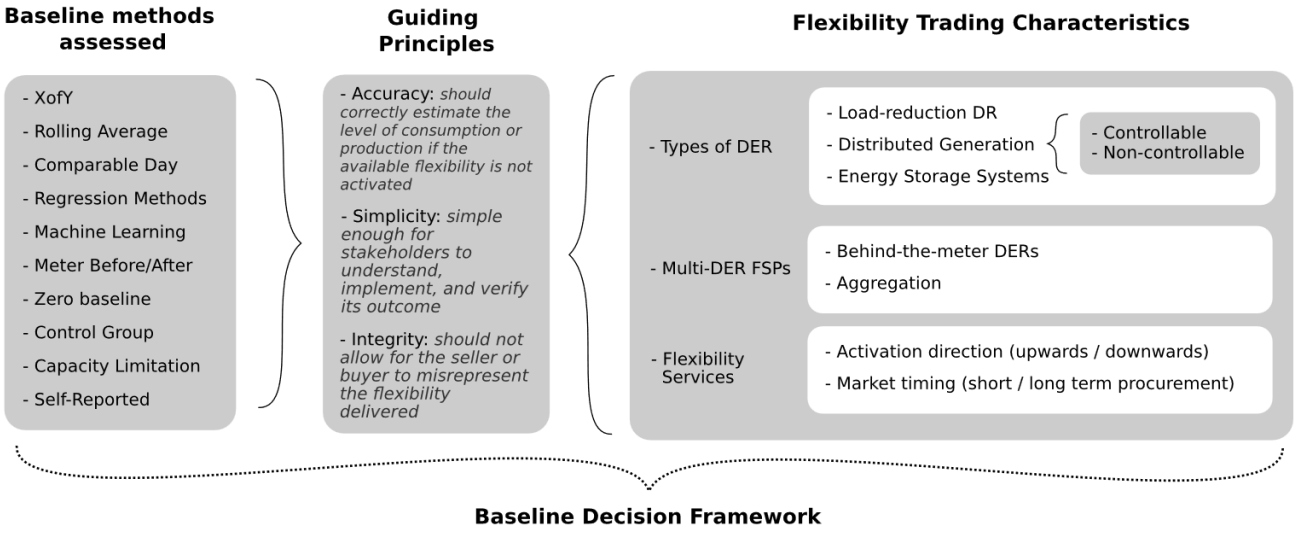

Figure 1: Representation of assessment framework

3.1 DIFFERENT BASELINE METHODS FOR DIFFERENT TYPES OF DER

In order to generalise the analysis of baseline methodology evaluation, four types of DER are considered, namely Load-DR, Controllable DG, Non-controllable DG, and Energy Storage Systems (ESSs). Load-DR is understood as a DER that can mostly reduce consumption. Controllable DG is the one that can modify its output both upwards and downwards following the DG operator commands (e.g. backup generators). Non-controllable DG stands for wind and solar, primarily, as these DERs cannot increase their output. They should always be producing as much power as possible since any primary energy resource not transformed into electricity is lost. Finally, ESS involve both stationary batteries as well as electric vehicles (EVs).

Historical approach methods have been largely used to determine the baseline for load-DR. They aim to estimate what would have been the consumption of a load-FSP if flexibility activation had not taken place. Therefore, methods such as *XofY, rolling average,* and *regression* models have been widely used for this type of DER (Valentini et al., 2022). These methods, however, present different levels of accuracy, simplicity and integrity for this type of DER. According to (Jazaeri et al., 2016b), the *XofY* and *rolling average* are low accuracy, high simplicity and medium integrity methods, while *regression* is a medium accuracy and simplicity with high integrity, and machine learning techniques score high for accuracy and integrity, but low for simplicity.

The simpler *capacity limitation* and the *MBMA* methods are suitable for any type of DER, considering their intrinsic simplicity and no need for ex-ante calculation. The only baseline method clearly not applicable to consumers is the *zero baseline*, as this method is designed for the distributed generation behind the meter of a consumer.

With respect to distributed generation, a differentiation has to be made between the controllable and the non-controllable DGs. The non-controllable units are mainly wind generators and photovoltaic (PV) units. In this case, the size of the units is also relevant. For utility-scale PV and wind farms, it is possible that these units are already scheduled in energy markets, not requiring a specific baseline for other services. However, if these units are at the consumer scale, a baseline is needed. However, not all the baseline methods would provide a satisfactory accuracy level in this case. Thus, methods that are only reliant on historical metered data do not manage to accurately estimate the baseline. Averaging methods, such as the *XofY* or the *rolling average,* are only suitable when combined with SDA and other baselining techniques, as the output of these

resources is more dependent on weather data than on what this output was a number of days before (Fonteijn et al., 2021). In this sense, the *regression* method could be used if appropriate data are available, as it can consider weather-related data that can be used to better estimate the DG output. For the same reason, *learning techniques* could also be used, possibly increasing the accuracy of the baseline estimate at the expense of simplicity. Alternatively, the *self-reported* baseline method can be used, but it can face integrity challenges.

In the case of controllable DG, mainly backup generators associated with a consumer can be used to provide flexibility services. Also, Combined Heat and Power (CHP) units could be used. These two types of units are not the same in terms of generation patterns. Given that the backup generator is not expected to generate electricity constantly, since in the idle case, the output of the DG would be 0, the baseline method providing the most accurate estimate would be the *zero baseline*. With regards to the CHP, generation patterns should be more stable as it is based on heat demand (Haesen et al., 2005). Therefore, historical methods should provide accurate estimates of the baseline.

Finally, the last type of DER to be considered is ESS, more specifically, batteries. This type of DER can be deemed a combination of consumption (charging mode) and controllable distributed generation (discharging). If treated separately from the other DER at the delivery point (e.g. load plus storage), historical data approaches alone may not be sufficiently accurate for this type of DER, or the integrity of the baseline estimate could be at risk. Outputs of batteries are more dependent on other factors (e.g., local renewable generation) than on the charging/discharging patterns in previous days. As for non-controllable DG, *MBMA* is, technically, an accurate baseline for batteries. However, considering the capability of this resource to change the power direction, integrity with *MBMA* could be at great risk. Storage owners may have the incentive to momentarily change the state of the battery before activation only to modify the baseline (e.g. change from discharging to charging before an upward flexibility activation). An alternative is to use the *zero baseline* for batteries, thus mitigating the incentives for the resource operator to change its output direction for the sole purpose of gaming the baseline method.

Table 2 provides an evaluation of the different types of baseline methods per individual type of DER based on the three desirable criteria of a baseline method, namely accuracy, simplicity and integrity.

Table 2: Baselines methods for different types of DER

Baseline Technique	Accuracy				Simplicity				Integrity			
	Load-DR	DG-C.	DG-N.C.	ESS	Load-DR	DG-C.	DG-N.C.	ESS	Load-DR	DG-C.	DG-N.C.	ESS
XofY Baselines	Medium	Low	Low	Low	High	High	High	High	Medium	Low	High	Low
Rolling average	Medium	Low	Low	Low	High	High	High	High	Medium	Low	High	Low
Comparable day	Medium	Low	High	Low	High	High	High	High	Low	Low	High	Low
Regression methods	High	Low	Medium	Low	Low	Low	Low	Low	High	Low	High	Low
Machine learning techniques	High	Medium	High	Medium	Very low	Very low	Very low	Very low	High	Medium	High	Medium
Meter-Before-Meter-After (MBMA)	Low	Medium	Medium	Medium	High	High	High	High	Low	Low	High	Very Low
Zero baseline	NA	Medium	Low	Medium	High	High	High	High	NA	Medium	High	Medium
Control group	Medium	Low	High	Low	Medium	Medium	Medium	Medium	Low	Low	High	Low
Self-reported	NA	NA	NA	NA	High	High	High	High	Low	Low	Low	Low
Capacity limitation	NA	NA	NA	NA	High	High	High	High	NA	NA	NA	NA

In addition to the type of DER, the size of the FSP can also be considered when selecting a baseline methodology. Industrial loads or utility-size DG and ESS can typically perform more complex energy management than residential loads. In this context, the relative importance of simplicity may be reduced in favour of accuracy. Integrity can, in this case, also be checked more closely by the parties procuring flexibility and/or regulatory authorities. In this context, the *self-reported* baseline method becomes a more appropriate option for these types of FSP.

3.2 MULTI-DER FSPS AND AGGREGATION

In future flexibility markets, FSPs may comprise not only one type of DER but several. An active consumer may provide flexibility by reducing its load while having a solar panel and a battery connected behind the meter. Alternatively, the FSP can, in fact, be an aggregation of DERs in the same geographical location (e.g. an energy community). Finally, an FSP can be an independent aggregator that offers the flexibility of sparsely located DERs of different types. This situation of multi-DER FSPs poses a challenge to the definition of baselines.

Some works in the literature have proposed, for instance, algorithms to identify different DER types and decouple them for baseline purposes. (Li et al., 2019), for instance, proposes a machine learning algorithm to decouple the DG and the load patterns. Although this study advocates for the suitability of this method, it also highlights the challenges in terms of the required data to have a robust model of the distributed generation, including the locational historical information of the output of the PVs. Besides, the machine learning algorithm can be seen as featuring low simplicity.

Historical data approach methods can incur a loss of accuracy in the presence of behind-the-meter generation and/or storage. The *regression* and *ML* type of baseline methods could consider weather-related variables to determine the baseline, but complexity would, then, increase. The *MBMA* method is the one that could be affected the least, not because it accounts for the effects of the DG behind the meter but because of its simplicity. However, this method would probably provide a largely inaccurate estimate of the baseline. The *comparable day* method and the *capacity limitation* could also be considered, provided that enough past data is available to ensure that the conditions of the activation day are considered. Otherwise, efficiency losses could occur for these two methods.

The *zero baseline* is the baseline method that aims to properly account for the controllable DG, considering that an additional meter is installed. This method would remunerate any production from the DG installation (therefore, the baseline is zero), as this energy would have otherwise been withdrawn from the grid.

Multi-DER FSPs can also take the form of an aggregator. In the case of aggregation of the same type of DER, the baseline method can be chosen according to the type of DER being aggregated. Some methods are straightforward, and the aggregated form of the baseline is simply the summation of the individual ones, as for the *MBMA* or the *capacity limitation*. For other methods, such as the *XofY*, it is also possible for the baseline to be calculated directly in an aggregated form (EnerNOC, 2011).

With regard to the aggregation of different types of DERs under the same portfolio, it becomes apparent that one single baseline method can hardly be accurate for the portfolio as a whole. The exceptions to this are the static baseline methods *capacity limitation* and *MBMA*. The *capacity limitation* and the *MBMA* are equally applicable to both individual and aggregated baselines. The latter, given its simplicity, can be applied the same way to compute an aggregated or an individual baseline (assuming that every unit is metered). Finally, with regards to the *zero baseline*, assuming that an additional meter is installed to measure the generation of the behind-the-meter non-controllable DG, the aggregation of all those units would be possible, enabling the separate treatment of backup generation, load, and other DER types.

Another possible baseline method that could overcome the lack of accuracy for aggregation is the *comparable day*, also for multi-DER portfolios. However, this baseline method also has aspects that should be taken into account. It entails the ex-post determination of the baseline by the aggregator, provided there is a large enough set of past data to allow for the selection of a similar day in terms of portfolio behaviour.

If different baseline methods can be considered, a differentiated treatment of the different types of DER would be most accurate. In this context, two approaches may be suggested, namely the grouping of the computation of the baselines by either type of technology or by cluster.

Separate baseline methods can be used per type of technology or type of DER within an aggregated portfolio. In other words, specific baseline methods can be applied to the different DER types according to what fits best each one of them. This approach should not be difficult to implement in the case where the different DER types are all different units, or delivery points, associated with individual meters. However, in the case of multiple DER types per metering point, such as for a consumer with a PV installation and an EV charging station, additional submetering data would be necessary. This would, on the one hand, increase the accuracy of the portfolio's baseline computed but, on the other, decrease its simplicity.

As an alternative to using submetering data to decouple the determination of the baseline for the different technologies, units could be grouped into clusters, and specific baseline methods could be applied to each of those according to the clusters' characteristics (Li et al., 2017; Schwarz et al., 2020; Zhang et al., 2016). A cluster could be defined according to the types of DER combined. For example, residential consumers would form one cluster, and so would consumers with PV, consumers with EV, etc. The advantage of this approach is that no additional data is required apart from the metering data already in place. The disadvantage is that accuracy may be impacted negatively (especially in the absence of appropriate metering). Finally, another alternative is to use the profile of a *control group* of end-users as the baseline for the aggregated FSP (Wang et al., 2018).

3.3 Product Timing and Direction

The flexibility markets of the future are expected to trade different products with different requirements. Firstly, products can be set for upward or downward flexibility. Upward flexibility corresponds to an increase in generation or, similarly, a reduction in consumption, while downward flexibility corresponds to an increase in consumption or a reduction in generation. As of today, balancing markets are trading upward and downward flexibility. Congestion management markets operating in a redispatch fashion, too, require upward activations to be compensated by downward ones.

Secondly, the timing of the products can also have an impact on the choice of the baseline method. As mentioned in Table 1, several baseline methods require the ex-ante calculation of the baseline, which can be challenging if products are procured very close to delivery time.

It is important to note that in this paper, only baselines for active power products are considered. Flexibility for voltage control may also be traded in the future under different product formats (Troncia et al., 2021). Products for voltage control can be set as reactive power or a PQ setpoint. The baselining for such products, however, is outside the scope of this paper.

The following subsections discuss the fitness of the different baseline methods depending on the direction of the flexibility provision (up or downwards) and the market timing.

3.3.1 Flexibility direction

Most baseline methodologies can be considered appropriate for both upward and downward flexibility provision. The main incompatibility concerns the *capacity limitation* baseline/product. This product is designed primarily for DSOs to mitigate congestion in primarily radial networks, as they serve as a means to place a limitation on the power flow in the congestion assets. In this case, the flexibility needed would mostly be upwards.

The other methods would, in principle, be suitable for both upward and downward products without intrinsic limitations to either product. The baseline design, however, may impact the performance of different

products. Some baseline methods are notably biased in a particular direction to incentivise flexibility provision. The *HighXofY* method is known for having an up bias, which is an incentive for upward flexibility provision (Wijaya et al., 2014). For downward flexibility, however, this would act as a disincentive, as the up bias would reduce the amount of flexibility deemed to be provided and remunerated by the buyer. In this case, variations such as the *MidXofY* or the *LowXofY* could be more appropriate.

3.3.2 Market Timing

The trading of flexibility can take place in different timeframes. Frequency control markets can trade products minutes ahead of delivery, while some local flexibility markets can take place years in advance. For the baseline definition, an important aspect from the implementation point of view is the timing of the activation process. Certain baselines are simpler to compute than others. The ability to implement those Baseline methods based on ex-ante computation depends on the time frame of the flexibility activation process. For certain products, activation may take place on very short notice and for short periods of time. This can be the case for capacity products, for instance, in which the system operator is entitled to change consumption or generation when the need arises. In such cases, the need to accommodate the data gathering and computation of the baseline in the activation process has to be considered. If the computation of baselines cannot be done within the activation process timeframe, baseline methods that require no previous computation can be considered, such as *MBMA*, *capacity limitation*, *zero output* and *comparable day*. However, static baseline methods such as the *MBMA* have the drawback of featuring low accuracy in long activations, given that the baseline cannot be changed in them. Conversely, if the activation is for a very short period of time (e.g., a few seconds), *MBMA* may be the most adequate, as historical meter measurements may not provide the right granularity. For this reason, international experience shows that *MBMA* is the most used baseline method for balancing services (DNV-GL, 2020).

With regards to those energy products cleared some time in advance of the activation, one aspect to be considered is the gaming opportunities the provision of these products may generate for certain baseline methods. If the flexibility market is cleared in the day ahead, for instance, and the baseline calculation includes data collected after the gate-closure time (GCT), the FSP may be able to game the method by trying to inflate the baseline (increasing consumption) during the hours between GCT and the baseline calculation/activation. In this case, an alternative would be to exclude those hours between the GCT and the activation for the baseline methods *XofY* (see Figure 2) and, possibly, for the *rolling average*. However, it is possible that the day of activation, or the day before, is relevant for the accuracy of the baseline calculation. In the *XofY* methods, SDAs are usually calculated based on the operation in the hours before delivery, and the rolling average weights for the samples on the day before activation are larger than those of other samples in the historical series. Therefore, in those cases in which those hours between GCT and activation still have to be taken into consideration, a set of rules should also be put in place to ensure the right of the party procuring the service to verify the need and accuracy of the data close to activation. For instance, the Belgium TSO Elia decided that no adjustments are to be applied in their *HighXofY* baseline method for balancing services. Nevertheless, the FSP could request the calculation of adjustments. In this case, a three-month evaluation is carried out to compare results with and without adjustments. If adjustments are proven to be more efficient, they are introduced for the baseline calculation, but under certain conditions. If the baseline is increased by more than 15% with the adjustment, justifications are required (Elia System Operator, 2019).

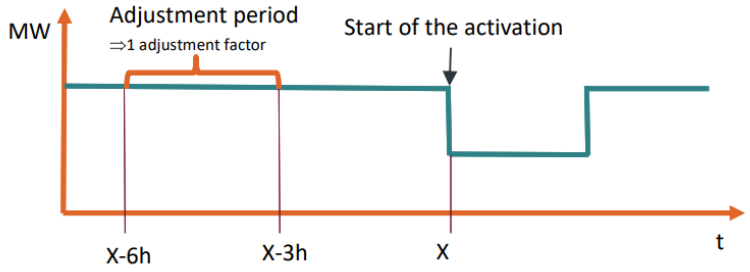

Figure 2: Adjustment period in a HighXofY baseline methodology in Belgium. Hours between the DA GCT and activation are used. Source: (Elia System Operator, 2019)

Another dimension of market timing is the existence of sequential markets that will impact the final output of the FSP. This could be the case even for different buyers, such as for TSOs and DSOs procuring flexibility for their different needs (Lind et al., 2019; Marques et al., 2022). In this context, harmonisation of baselines across interacting markets might be desired in order to avoid distortions created by different methodologies being applied in interrelated processes.

3.4 DECISION FRAMEWORK FOR BASELINE METHOD SELECTION

Considering the assessment above within the proposed methodology, this section summarises this analysis and proposes a tool for the baseline method selection process considering the different characteristics of the product and the participants to which the baseline is applied. This tool adopts the form of a decision tree, as shown in Figure 3.

The starting point is the verification of whether the FSP is already individually scheduled (i.e. the FSP has a demand and generation plan committed for other markets, such as the day-ahead market). In this case, no additional baseline would be required, as the schedule of the FSP would serve as a baseline. The second consideration is whether or not the FSP is aggregated. If the FSP is one individual unit, different baseline methods can be considered for the different types of DER. The specific baseline to be chosen would also depend on the product and unit characteristics. Additionally, the design option to implement should also depend on the market characteristics (e.g. possibility for SDAs).

If the FSP is of a multi-DER type, submetering could be used to calculate baselines per technology. Otherwise, the baseline for clusters could be used, offering greater simplicity but possibly lower accuracy. Alternatively, the *comparable day*, *control group* or *self-reported* baselines are also viable for multi-DER aggregation.

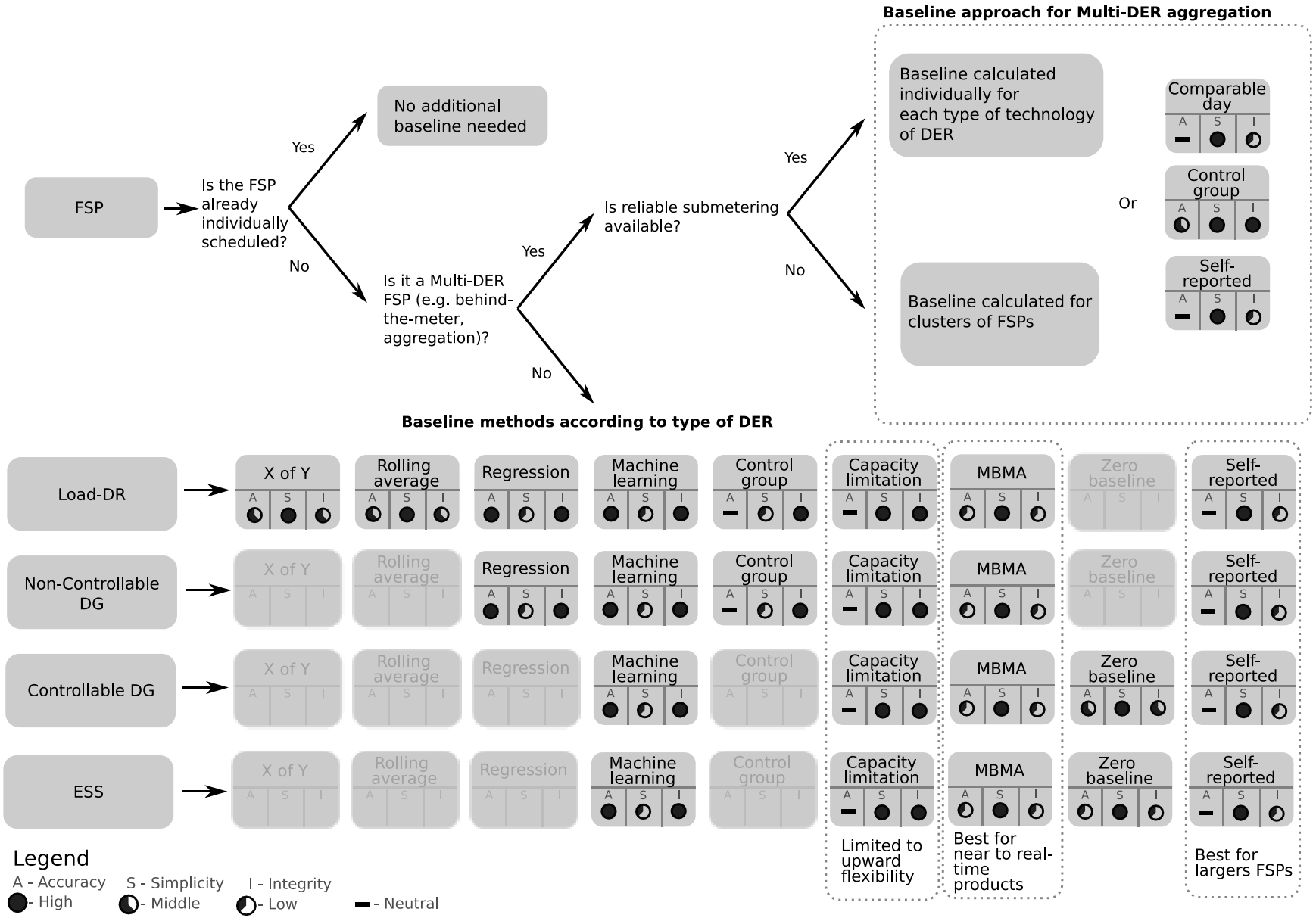

Figure 3: Baseline decision framework according to type of FSP, DER, multi-DER presence, product direction and timing

4. Conclusions and Policy Implications

In this paper, baseline methodologies are analysed from the perspective of future flexibility markets in which FSPs can (i) be of different types of DER, (ii) be composed of multiple types of DER (e.g. behind-the-meter, aggregated), and (iii) may participate in markets procuring flexibility in different directions (upwards or downwards) and of different timings (from years ahead to close to real-time).

In the context of a multitude of flexibility services, markets, and FSPs types, this paper shows that there is no one-size-fits-all baseline method while providing a guideline and framework for choosing the most suitable methods depending on the flexibility trading characteristics. These recommendations are also aligned with the current efforts taking place at the European and international scale for the adequate definition of baselines due to their essential impact in enabling the participation of DERs (of different types) in electricity and flexibility markets. The Framework Guideline on Demand Response published by the EU Agency for the Cooperation of Energy Regulators (ACER) highlights that baseline definition should aim at easy, transparent and accurate methods while preventing gaming opportunities. In addition, ACER's framework guideline states that baseline methodologies could be different for different products and timeframes. In this sense, this paper analyses the different methods and identifies the most appropriate baseline definition and calculation methodology for each of the different flexibility provision contexts. The ACER framework also advocates for technology-neutral baselines. However, this can lead to a decrease in the adopted baseline method's accuracy or integrity as compared to choosing the most adequate baseline method for each different DER technology and aggregation of technologies, the availability of metering/sub-metering, and the flexibility market's characteristics.

The choice of a baseline should be based on the particularities of the flexibility markets, products and FSPs involved. Harmonisation of baselines across interacting markets, however, might be desired in order to avoid distortions created by different methodologies and to allow additional certainty for the FSP regarding its expected remuneration. This can be the case for TSO-DSO sequential markets, for example.

Therefore, this paper proposes a guideline for baseline choice to be used by practitioners and market actors in defining the most appropriate baseline given the flexibility trading characteristics. Nevertheless, future work is required to further investigate these interactions from a quantitative perspective, including the specific conditions found in flexibility markets (e.g. balancing and congestion management redispatch) as well as FSP characteristics.

5. Acknowledgment

This work is supported by the European Union Horizon 2020 research and innovation program under grant agreements No. 824414 – CoordiNet Project; and No. 101075438 – BeFlex Project.

6. REFERENCES

AIEC, 2009. Demand Response Measurement and Verification.

Antunes, P., Faria, P., Vale, Z., 2013. Consumers performance evaluation of the participation in demand response programs using baseline methods, in: 2013 IEEE Grenoble Conference. Presented at the 2013 IEEE Grenoble Conference, pp. 1–6. https://doi.org/10.1109/PTC.2013.6652420

Arunaun, A., Pora, W., 2018. Baseline Calculation of Industrial Factories for Demand Response Application, in: 2018 IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia). Presented at the 2018 IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia), pp. 206–212. https://doi.org/10.1109/ICCE-ASIA.2018.8552114

Directive (EU) 2019/944 of the European Parliament and of the Council of 5 June 2019 on common rules for the internal market for electricity, 2019. , Official Journal of the European Union.

DNV-GL, 2020. Baseline Methodology Assessment.

Elia System Operator, 2019. Transfer of Energy in DA and ID markets.

EnerNOC, 2011. The Demand Response Baseline.

EnerNOC, 2009. The Demand Response Baseline.

ENTSO-E, 2021. Survey on Ancillary Services Procurement, Balancing market Design 2020.

European Commission, 2023. REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL amending Regulations (EU) 2019/943 and (EU) 2019/942 as well as Directives (EU) 2018/2001 and (EU) 2019/944 to improve the Union's electricity market design.

Fonteijn, R., Nguyen, P.H., Morren, J., Slootweg, J.G. (Han), 2021. Baselining Flexibility from PV on the DSO-Aggregator Interface. Appl. Sci. 11, 2191. https://doi.org/10.3390/app11052191

Haesen, E., Espinoza, M., Pluymers, B., Goethals, I., Thong, V.V., Driesen, J., Belmans, R., Moor, B.D., 2005. Optimal Placement and Sizing of Distributed Generator Units Using Genetic Optimization Algorithms. Electr. Power Qual. Util.

Heinrich, C., Ziras, C., Jensen, T.V., Bindner, H.W., Kazempour, J., 2021. A local flexibility market mechanism with capacity limitation services. Energy Policy 156, 112335. https://doi.org/10.1016/j.enpol.2021.112335

Holmberg, D., Hardin, D., Koch, E., 2013. Towards Demand Response Measurement and Verification Standards.

Jazaeri, J., Alpcan, T., Gordon, R., Brandao, M., Hoban, T., Seeling, C., 2016a. Baseline methodologies for small scale residential demand response, in: 2016 IEEE Innovative Smart Grid Technologies - Asia (ISGT-Asia). Presented at the 2016 IEEE Innovative Smart Grid Technologies - Asia (ISGT-Asia), IEEE, Melbourne, Australia, pp. 747–752. https://doi.org/10.1109/ISGT-Asia.2016.7796478

Jazaeri, J., Alpcan, T., Gordon, R., Brandao, M., Hoban, T., Seeling, C., 2016b. Baseline methodologies for small scale residential demand response, in: 2016 IEEE Innovative Smart Grid Technologies - Asia (ISGT-Asia). Presented at the 2016 IEEE Innovative Smart Grid Technologies - Asia (ISGT-Asia), pp. 747–752. https://doi.org/10.1109/ISGT-Asia.2016.7796478

Li, K., Wang, B., Wang, Z., Wang, F., Mi, Z., Zhen, Z., 2017. A Baseline Load Estimation Approach for Residential Customer based on Load Pattern Clustering. Energy Procedia, Proceedings of the 9th International Conference on Applied Energy 142, 2042–2049. https://doi.org/10.1016/j.egypro.2017.12.408

Li, K., Wang, F., Mi, Z., Fotuhi-Firuzabad, M., Duić, N., Wang, T., 2019. Capacity and output power estimation approach of individual behind-the-meter distributed photovoltaic

system for demand response baseline estimation. Appl. Energy 253, 113595. https://doi.org/10.1016/j.apenergy.2019.113595

Lind, L., Cossent, R., Chaves-Ávila, J.P., Gómez San Román, T., 2019. Transmission and distribution coordination in power systems with high shares of distributed energy resources providing balancing and congestion management services. Wiley Interdiscip. Rev. Energy Environ. 8, e357. https://doi.org/10.1002/wene.357

Marques, L., Sanjab, A., Mou, Y., Cadre, H.L., Kessels, K., 2022. Grid Impact Aware TSO-DSO Market Models for Flexibility Procurement: Coordination, Pricing Efficiency, and Information Sharing. IEEE Trans. Power Syst. 1–14. https://doi.org/10.1109/TPWRS.2022.3185460

Martín-Utrilla, F.-D., Pablo Chaves-Ávila, J., Cossent, R., 2022. Decision Framework for Selecting Flexibility Mechanisms in Distribution Grids. Econ. Energy Environ. Policy 11. https://doi.org/10.5547/2160-5890.11.2.fmar

Miriam L. Goldberg, G. Kennedy Agnew, 2013. Measurement and Verification for Demand Response.

Mohajeryami, S., Doostan, M., Asadinejad, A., Schwarz, P., 2017a. Error Analysis of Customer Baseline Load (CBL) Calculation Methods for Residential Customers. IEEE Trans. Ind. Appl. 53, 5–14. https://doi.org/10.1109/TIA.2016.2613985

Mohajeryami, S., Karandeh, R., Cecchi, V., 2017b. Correlation between predictability index and error performance in Customer Baseline Load (CBL) calculation, in: 2017 North American Power Symposium (NAPS). Presented at the 2017 North American Power Symposium (NAPS), pp. 1–6. https://doi.org/10.1109/NAPS.2017.8107200

Park, S., Ryu, S., Choi, Y., Kim, J., Kim, H., 2015. Data-Driven Baseline Estimation of Residential Buildings for Demand Response. Energies 8, 10239–10259. https://doi.org/10.3390/en80910239

Ramos, A., 2019. Consumer Access to Electricity Markets: the Demand Response Baseline, in: 2019 16th International Conference on the European Energy Market (EEM). Presented at the 2019 16th International Conference on the European Energy Market (EEM), IEEE, Ljubljana, Slovenia, pp. 1–6. https://doi.org/10.1109/EEM.2019.8916212

Rossetto, N., 2018. Measuring the Intangible: An Overview of the Methodologies for Calculating Customer Baseline Load in PJM.

Ruwaida, Y., Chaves-Avila, J.P., Etherden, N., Gomez-Arriola, I., Gürses-Tran, G., Kessels, K., Madina, C., Sanjab, A., Santos-Mugica, M., Trakas, D.N., Troncia, M., 2022. TSO-DSO-Customer coordination for purchasing flexibility system services: Challenges and lessons learned from a demonstration in Sweden. IEEE Trans. Power Syst. 1–13. https://doi.org/10.1109/TPWRS.2022.3188261

Schittekatte, T., Reif, V., Meeus, L., 2021. Welcoming new entrants into European electricity markets (preprint). SOCIAL SCIENCES. https://doi.org/10.20944/preprints202105.0109.v1

Schwarz, P., Mohajeryami, S., Cecchi, V., 2020. Building a Better Baseline for Residential Demand Response Programs: Mitigating the Effects of Customer Heterogeneity and Random Variations. Electronics 9, 570. https://doi.org/10.3390/electronics9040570

SEDC, 2016. Explicit and Implicit Demand-Side Flexibility.

smartEn, 2018. The smartEn Map - European Balancing Markets Edition.

Troncia, M., Ávila, J.P.C., Pilo, F., Román, T.G.S., 2021. Remuneration mechanisms for investment in reactive power flexibility. Sustain. Energy Grids Netw. 27, 100507. https://doi.org/10.1016/j.segan.2021.100507

Tufts, B., Breidenbaugh, A., 2010. ENERNOC - Analysis of Baseline Methodologies and “Best Practice” Recommendations.

Vagropoulos, S.I., Biskas, P.N., Bakirtzis, A.G., 2022. Market-based TSO-DSO coordination for enhanced flexibility services provision. Electr. Power Syst. Res. 208, 107883. https://doi.org/10.1016/j.epsr.2022.107883

Valarezo, O., Gómez, T., Chaves-Avila, J.P., Lind, L., Correa, M., Ulrich Ziegler, D.U., Escobar, R., 2021. Analysis of New Flexibility Market Models in Europe. Energies 14, 3521. https://doi.org/10.3390/en14123521

Valentini, O., Andreadou, N., Bertoldi, P., Lucas, A., Saviuc, I., Kotsakis, E., 2022. Demand Response Impact Evaluation: A Review of Methods for Estimating the Customer Baseline Load. Energies 15, 5259. https://doi.org/10.3390/en15145259

Villar, J., Bessa, R., Matos, M., 2018. Flexibility products and markets: Literature review. Electr. Power Syst. Res. 154, 329–340. https://doi.org/10.1016/j.epsr.2017.09.005

Wang, X., Li, K., Gao, X., Wang, F., Mi, Z., 2018. Customer Baseline Load Bias Estimation Method of Incentive-Based Demand Response Based on CONTROL Group Matching, in: 2018 2nd IEEE Conference on Energy Internet and Energy System Integration (EI2). Presented at the 2018 2nd IEEE Conference on Energy Internet and Energy System Integration (EI2), pp. 1–6. https://doi.org/10.1109/EI2.2018.8582122

Wang, X., Tang, W., 2022. Modeling and Analysis of Baseline Manipulation in Demand Response Programs. IEEE Trans. Smart Grid 13, 1178–1186. https://doi.org/10.1109/TSG.2021.3137098

Wijaya, T.K., Vasirani, M., Aberer, K., 2014. When Bias Matters: An Economic Assessment of Demand Response Baselines for Residential Customers. IEEE Trans. Smart Grid 5, 1755–1763. https://doi.org/10.1109/TSG.2014.2309053

Zhang, Y., Chen, W., Xu, R., Black, J., 2016. A Cluster-Based Method for Calculating Baselines for Residential Loads. IEEE Trans. Smart Grid 7, 2368–2377. https://doi.org/10.1109/TSG.2015.2463755

Ziras, C., Heinrich, C., Bindner, H.W., 2021. Why baselines are not suited for local flexibility markets. Renew. Sustain. Energy Rev. 135, 110357. https://doi.org/10.1016/j.rser.2020.110357