Local Energy Oxfordshire

April 2023 | Version 1

Project LEO Baselining Working Group Summary Report

Scot Wheeler, University of Oxford

UNIVERSITY OF OXFORD
OXFORD BROOKES UNIVERSITY
low carbon hub
OXFORD CITY COUNCIL
OXFORDSHIRE COUNTY COUNCIL
Scottish & Southern Electricity Networks
piclo
equiwatt
NUVVE
Origami POWER OVER ENERGY
INDUSTRIAL STRATEGY
UK Research and Innovation

Report Title:	Project LEO Baselining Working Group – Summary Report
Author(s):	Scot Wheeler (scot.wheeler@eng.ox.ac.uk)
Organisation(s):	University of Oxford

Version:	2.0	Date:	29/03/2023
Workpack*:	4, 5	Deliverable:	N/A
Reviewed by:	Timur Yunusov		
Date:	28/03/2023		
Signed off by:	Project LEO Project Delivery Board		
Date:	17/04/2023		

Can be shared (Y/N):	Internally	Y	Externally	Y

Context

The UK Government has legislated to reduce its carbon emissions to net zero by 2050. Meeting this target will require significant decarbonisation and an increased demand upon the electricity network. Traditionally an increase in demand on the network would require network reinforcement. However, technology and the ability to balance demand on the system at different periods provides opportunities for new markets to be created, and new demand to be accommodated through a smarter, secure and more flexible network.

The future energy market offers the opportunity to create a decentralised energy system, supporting local renewable energy sources, and new markets that everyone can benefit from through providing flexibility services. To accommodate this change, Distribution Network Operators (DNOs) are changing to become Distribution System Operators (DSOs).

Project Local Energy Oxfordshire (LEO) is an important step in understanding how new markets can work and improving customer engagement. Project LEO is part funded via the Industrial Strategy Challenge Fund (ISCF) who set up a fund in 2018 of £102.5m for UK industry and research to develop systems that can support the global move to renewable energy called: Prospering From the Energy Revolution (PFER).

Project LEO is one of the most ambitious, wide-ranging, innovative, and holistic smart grid trials ever conducted in the UK. LEO will improve our understanding of how opportunities can be maximised and unlocked from the transition to a smarter, flexible electricity system and how households, businesses and communities can realise the benefits. The increase in small-scale renewables and low-carbon technologies is creating opportunities for consumers to generate and sell electricity, store electricity using batteries, and even for electric vehicles (EVs) to alleviate demand on the electricity system. To ensure the benefits of this are realised, Distribution Network Operators (DNO) like Scottish and Southern Electricity Networks (SSEN) are becoming Distribution System Operators (DSO).

Project LEO seeks to create the conditions that replicate the electricity system of the future to better understand these relationships and grow an evidence base that can inform how we manage the transition to a smarter electricity system. It will inform how DSOs function in the future, show how markets can be unlocked and supported, create new investment models for community engagement, and support the development of a skilled community positioned to thrive and benefit from a smarter, responsive and flexible electricity network.

Project LEO brings together an exceptional group of stakeholders as Partners to deliver a common goal of creating a sustainable local energy system. This partnership represents the entire energy value chain in a compact and focused consortium and is further enhanced through global leading energy systems research brought by the University of Oxford and Oxford Brookes University consolidating multiple data sources and analysis tools to deliver a model for future local energy system mapping across all energy vectors.

Table of Contents

Table of Figures

1 Executive Summary

The flexibility market trialled as part of Project LEO and TRANSITION calls for relative changes in power usage in response to market instruction. To verify the service delivery, the metered data is compared to the counterfactual or baseline. As this baseline cannot be explicitly measured, it must be retrospectively estimated from historic data or agreed upon.

Throughout Project LEO and TRANSITION trials, two options were available to service providers: a mid-8-in-10 Historic Baseline with Same Day Adjustment whereby the market facilitator (in this case the DNO) uses recent historic data submitted by the service provider to estimate the baseline; or a Nomination Baseline whereby the service provider submits their own counterfactual a day ahead of the service.

This report aims to summarise the discussion and learnings relating to baselining that arose throughout the real-world trials carried out as part of Project LEO and TRANSITION. These trials ran between November 2021 and February 2023. The work includes an analysis of the accuracy of the methods used, alongside variations of Historical Baselining and other separate methods, applying these to a range of DER types that took part in the real-world trials. The work also highlights some of the challenges and shortcomings of the baselining methodology and the impact these have on the wider market process.

Some of the key findings and recommendations include:

- The context specific analysis of baseline errors applied across different DERs provides valuable insights into method performance and suitability. The process should be available to market facilitator and industry actors to provide greater market transparency and could be included within the baselining process itself.
- The accuracy of the baselining and verification process has impacts on the wider operation of the market. The settlement rule put in place to discourage under-delivery is asymmetric with a grace window of only 5%. For the data analysed, baseline error analysis suggests that on the order of 15% of events could be underpaid because of baseline errors. The settlement rule should be modified to reduce this and avoid potential capacity sterilisation due to flexibility providers holding capacity in reserve.
- The simplest method of Meter Before Meter After (MBMA) which just uses an interpolation of the profile from the points immediately before and after the event, consistently saw the best accuracy for the methods and data analysed. It also appears to be a close match to the way small scale DER flexibility is controlled (the response being set relative to usage at the time of the event). However, it is the most prone to manipulation if the service provider has prior knowledge of the event. Further study is needed to see if the risk of manipulation materialises within local flexibility markets that overcomes the benefits of a simple and transparent method.
- The use of Same Day Adjustment (SDA) to correct for daily variations in usage resulting from external factors such as temperature typically provides a more accurate (smaller average error) historic baseline. However, the trials exposed practical limitations to the implementation of SDA if the DER undergoes any pre-conditioning, also exposing the method to manipulation. The use of SDA needs careful consideration if the gain in accuracy

is to be realised over the potential for manipulation or greater inaccuracy. It is most suitable for services with immediate real-time instruction with little to no warning.

- The ability to stack services within both local and national flexibility markets will be important in any DERs business case. Service stacking can impact the baseline if the stacked services occur within historical days contributing to the baseline calculation or within the SDA window. To overcome this, market facilitators need information about a DERs participation in other services; this could either be through notification by the flexibility provider as part of the market rules, or better, managed through a centralised database that all market facilitators (DSOs, ESOs, registered aggregators etc) have access to – this reduces burden on the flexibility provider.
- To baseline at scale with low transaction costs, baselining must be an automated process which is integrated into the wider market processes. To achieve this, data must be of sufficient quality and in the correct format. During the Project LEO trials, quality assurance and data cleaning was a manual (and laborious at times) process. Industry wide data standards and integrated cleaning/formatting tools are needed but must be a balance so as not to make participation too strict that discourages participation.
- In general, the baselining processes were seen by market participants as complicated and not transparent. Alternative flexibility markets that do not rely so heavily on baselining for verification, such as firm capacity markets, should be explored as an alternative.

The insights presented herein are an output of the baselining working group within Project LEO with input from all project partners. Work is ongoing as part of TRANSITION with collaboration with the FUSION project.

2 Introduction

Baselining in the context of Project LEO and TRANSITION refers to the process within the end-to-end procedure whereby the magnitude of service delivery is verified by the DSO relative to an estimation of standard DER (Distributed Energy Resource) behaviour (the Baseline) had the DER not been taking part in the specific LEO/TRANSITION market service. For a flexibility market based on relative changes in power consumption (rather than one of absolute capacity limits), a baseline is critical in calculating an Industry Actor's (IA - a provider of flexibility to the market) response. This is primarily used to settle the market (paying the IA for the response) but can have far reaching consequences throughout the market relating to reliability indices used for dispatch, capacity sterilisation, over procurement requirements, IA offered capacity, or even IA participation in the market in the first place.

This report aims to summarise discussions and work within the Baselining Working Group over the three trial periods of LEO and TRANSITION that ran between November 2021 and February 2023. It details different methodologies for baselining in the context of particular market services and DER types, and highlights some of the issues and anomalies that have arisen related to baselining. As part of TNEI's scope of work for TRANSITION, an analysis of the performance of historic baselining methods was produced and the reader is directed here for further specific insights into these methods [1].

2.1 Baselining Methods

There have been numerous studies undertaken on baselining methods on behalf of energy industry bodies, innovation projects, and as part of academic research [2]–[6]. For reference, this section highlights some known methods that have motivated potential options and discussions within Project LEO.

Baselining methods can be grouped into 6 general categories[7]. These include:

- **Historic Averaging** – the baseline is calculated by taking the average of a set of historic days preceding the event. There are many variations on how the set of days is selected and how the average is taken [8]–[12].
- **Control Groups** – the baseline is calculated from the metered data of a peer control group that did not participate in a service [13], [14].
- **Scheduling** – an ahead of time forecast provided by the industry actor is used as the baseline [15], [16]. Nomination Baselines and Zero Baselines are examples.
- **Interpolation** – Considers the metered data in the periods before and after the event and interpolates values within the period window [17].
- **Regression** – The baseline is calculated from historic data using a regression model where feature variables might include temperature, sunrise/sunset times, irradiance, or temporal offset amongst many others [10], [18], [19].
- **Machine Learning methods** – A large suite of machine learnings methods, such as neural networks, clustering, and support vector machines, are available and have been explored [20]–[23]. These might be incorporated into a hybrid with other methods.

- **Same Day Adjustment** - Not strictly a baseline method, same day adjustment is often an optional second step whereby the baseline is adjusted based on a reference point prior to the event to correct for external effects such as weather [2].

3 Project LEO Baselining Methods

The DNV-GL's 2020 report to the ENA [2] recommended that the Historic Baseline methodology is used for all flexibility products as the default option (**without SDA** for Secure services due to complexities of service stacking in adjustment window, and **with SDA** for real-time, Dynamic and Restore services); it is also recommended that the DNO calculate and share the baseline prior to the utilisation period. An option of Nomination Baseline where it improves accuracy (expected to be better for EVs and primary generators) should also be offered. TNEI, through Open Networks and TRANSITION, have created a baselining tool for the ENA (accessed here) which offer adjustable historic averaging (with or without SDA), nominated baseline and zero baseline.

Project LEO and TRANSITION provided two options for baselining during the trials:

3.1 Mid X-in-Y approach with Same Day Adjustment

Also commonly referred to as Historic Baseline with SDA, the method and possible variations is described in detail in the Open Networks' Flexibility Baselining Tool – Mathematical Specification [24]. For Project LEO, the key parameters selected for Project LEO's market trials were: [1]

- X = 8 for weekdays, 2 for weekends; Y = 10 for weekdays, 4 for weekends.
- In LEO, the middle X days are chosen based on event period mean usage. Alternatives include ranking by peak or average energy throughout the day, or within the service window period.
- Same day adjustment uses the difference in average power between event day and unadjusted baseline over a 2-hour window prior to event delivery, adjusting the baseline by this value.

The baseline was calculated retrospectively following the bulk submission of data by the IA to the DNO covering the full range of required data for the month of delivery (up to 8 weeks of data from the end of the month in which delivery occurred). This reduced the time burden for what was a manual data submission process (versus submitting before and after every event), however, does not allow for baselines to be calculated and shared prior to utilisation (reducing transparency). The baseline tool developed for Open Networks was available to IAs should they wish to calculate the baseline themselves.

3.2 Nominated Baseline

Service providers also have the option to submit their own nominated baseline, including a zero baseline. The baseline must be submitted by 17:00 on the day before event delivery. Service providers are free to implement their own forecasting methodology without reporting this. The FSA (Flexibility Services Agreement) states that the DNO will check the accuracy of the nominated

[1] During Trial Period 1, the NTVV method of 6-in-10 days was used with the nearest 6 days ranked based on total daily energy. SDA used a 4-hour window immediately prior to the event.

baseline against days without utilisation and reserves the right to require an alternative baseline method be used in the event of inaccuracies or manipulation suspicion.

4 Systematic Methodology Scan

As described in Section 2.1, there are multiple different methods that can be used for baselining and many different ones are used across different markets. This section presents a systematic methodology scan across some common methods (and parameter variations) and asset types to demonstrate variations in accuracy across methodologies to examine suitability.

4.1 Error metrics and methodology

The choice of error metric is important when comparing different baselining methodologies. Common metrics for accuracy (on average, the degree to which the baseline can calculate the usage) include Mean Absolute Percentage Error (MAPE) and Relative Root Mean Square Error (RRSME). A metric for bias (the degree to which the baseline method consistently under-estimates or over-estimates usage) is Average Relative Error (ARE).
For the empirical analysis presented here, each dataset is randomly sampled 5000 times to obtain a test period of random width between 30 minutes and 4 hours; known flexibility event periods are excluded. For each sample, the baseline method is applied as if this test period was a flexibility event, resulting in a baseline forecast. The error for each sample (e_n) is calculated by taking the difference between this baseline (b_n) and the actual value (a_n) over this test period:

$$e_n = a_n - b_n$$

.
The relative error (r_n) is typically calculated by dividing the error through by the actual value $r_n = \frac{a_n - b_n}{a_n}$. However, when considering flexibility from numerous different assets with different flexibility capacities, that have very different standard behaviour (batteries might have zero as the counterfactual whereas an office or PV will have a larger baseload), the typical relative error may not reflect the error in the context of the size of flexibility being offered. An alternative is to use the DERs flexible capacity (C_{fx}) as the divisor. This is how flexibility response will be judged by the market and can be directly compared with the settlement rules. This analysis calculates the relative error as:

$$r_n^{fx} = \frac{a_n - b_n}{C_{fx}}$$

. It thus follows that ARE is calculated:

$$ARE_{fx} = \frac{1}{n}\sum_{n}^{N} \frac{a_n - b_n}{C_{fx}}$$

and MAPE:

$$MAPE_{fx} = \frac{1}{n}\sum_{n}^{N} \left|\frac{a_n - b_n}{C_{fx}}\right|$$

And finally, RRMSE:

$$RRMSE_{fx} = \frac{\sqrt[2]{\frac{1}{n}\sum_{n}^{N}(a_n - b_n)^2}}{C_{fx}}$$

.[24]

4.2 Variations on Historic Baseline

The following variations on a historic baseline are considered within this analysis. More detailed descriptions can be found in Open Networks' Baselining Tool Mathematical Specification:

Mid 8-in-10 Historic Baseline – the middle 8 of 10 historic weekdays (excluding previous event days and public holidays – mid 4 of 5 for weekends) based on average historic energy during the event timeframe. This is the recommended ENA method.

Mid 8-in-10 Historic Baseline with SDA – as above but with a same day adjustment based on average energy in the 2 hours prior to the event.

Nearest 8-in-10 Historic Baseline – the nearest 8 of 10 historic weekdays (excluding previous event days and public holidays – mid 4 of 5 for weekends) judged based on total daily energy compared to the event day.

Nearest 8-in-10 Historic Baseline with SDA – as above but with a same day adjustment based on average energy in the 4 hours prior to the event.

Nearest 8-in-10 Historic Baseline with SDA – as above but with a same day adjustment based on average energy in the hour before and hour after the event.

4.3 Alternative methods

4.3.1 Meter Before Meter After (MBMA) Linear Interpolation

This method simply assumes a straight line between the time-period before the event and the time-period after the event. The advantage of this method is its simplicity to implement and that it requires no historic data beyond the event day. The issue is that it is entirely dependent on the point in time immediately before and after the event which could be impacted by rebound effects and would be easy to manipulate given prior knowledge of the event – neither of which are accounted for within this analysis.

4.3.2 Same days

Like the historic baseline method, this is also an averaging method that takes the same day of the week for the previous *n* weeks, i.e., if the event day was a Thursday, it would average the previous *n* Thursdays.

4.3.3 Clustering

Clustering is a very broad term. In general, these methods implement an unsupervised clustering algorithm on non-event day data, followed by classification of the event day. There are many different clustering algorithms such as k-means and DBSCAN. A centroid, mean or median cluster profile for the cluster to which the event day is assigned, is used as the baseline. Clustering could be trained on profile data or a combination of summary metrics (e.g. event day max, range etc). There are additional processes such as normalisation or dynamic time warping that might be applied. A post adjustment step such as SDA might also be integrated.

This study utilises the HDBSCAN algorithm with a minimum cluster size of 8, applied to raw energy profiles, excluding the event window; the baseline is the mean profile of cluster members. The same SDA method as with the historic baseline is used. Figure 1 provides an example of calculated clusters for an example office building demand. More extensive work is needed into the different variations on a clustering approach. It should be noted that of the methods tested, this was most disrupted by poor data quality, particularly missing data.

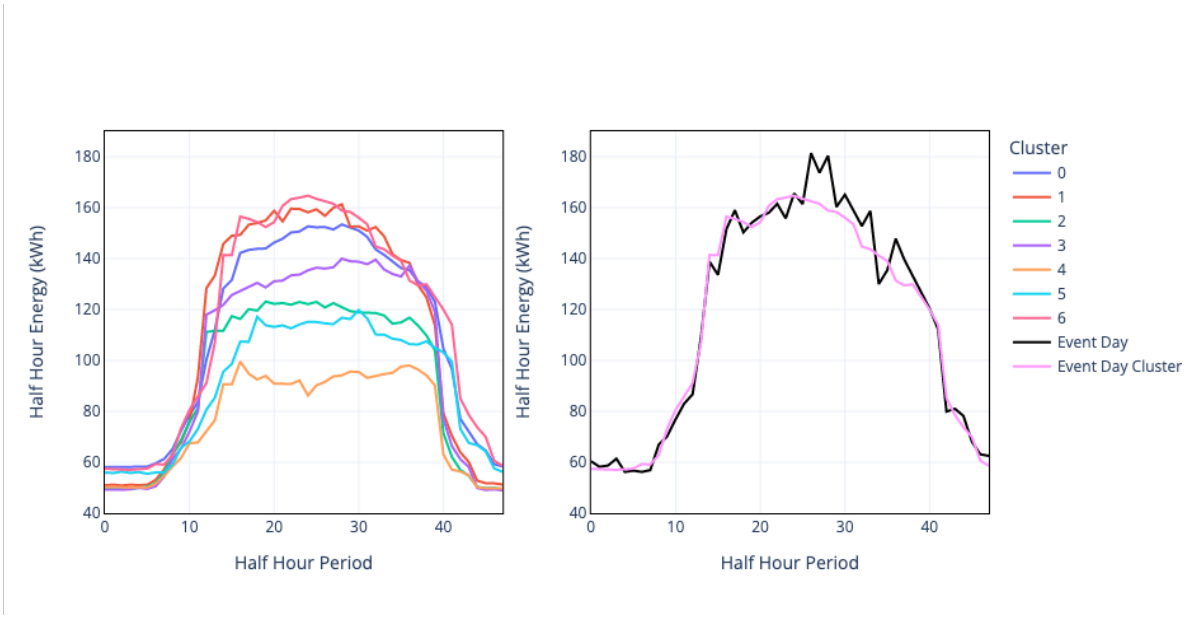

Figure 1: A example of the clustering method applied to an example office building. (a) shows the mean cluster profiles for the generated clusters while (b) shows the event day in comparison to its associated cluster profile.

4.3.4 Regression

Regression methods which train algorithms against feature vectors such as day of the week, temperature, irradiance etc, have been developed and tested for TRANSITION by TNEI. Early results suggest significant improvement can be made for solar PV assets. However, regression methods require access to a greater amount of historic data for a range of variables.

4.4 Methodology scan

The accuracy of the baselining methods described above were tested on multiple different DER types using datasets collected as part of Project LEO. The DER types tested include a battery located on a business site, a ground mount solar PV installation, a commercial rooftop solar PV installation, office building, a domestic property with EV and a LV secondary substation (11kV to 400 V).

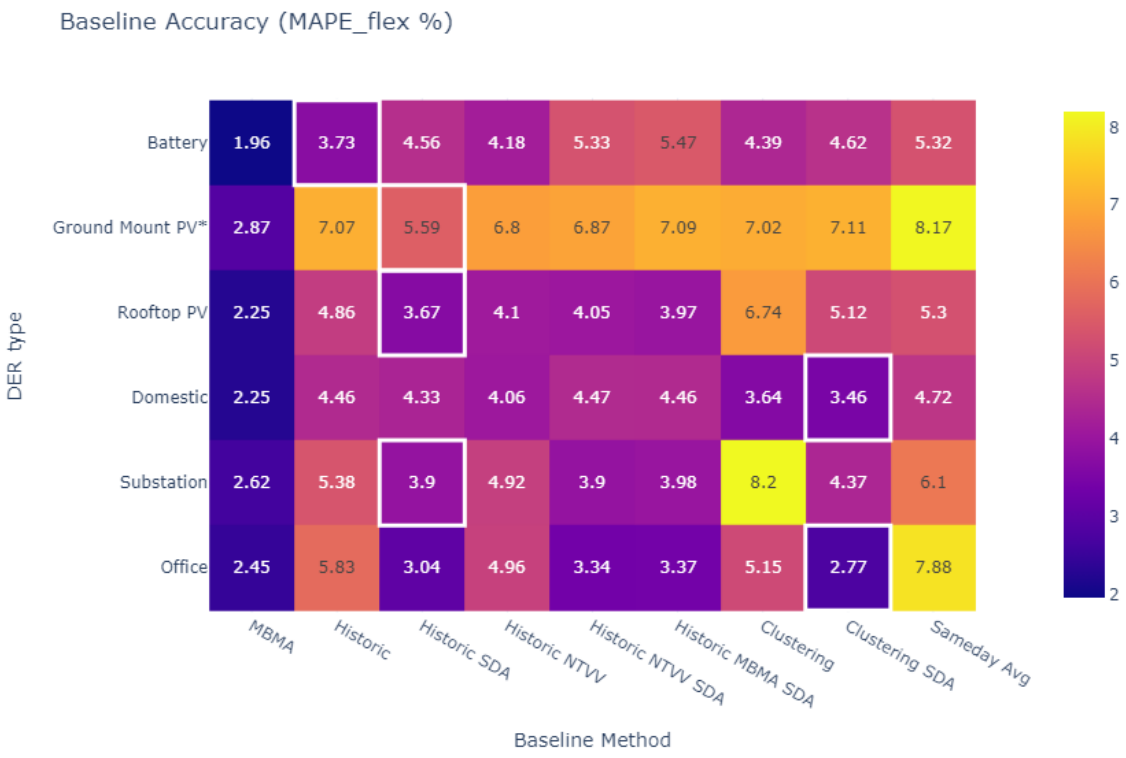

*Figure 2: Baseline method accuracy (MAPE flex) matrix for various baselining methods considered within Project LEO, tested against various DER types available to the Project LEO flexibility market; method with smallest mean error is (excluding MBMA) is highlighted with a white outline. *the ground mount PV is a limited dataset only covering 4 months of data.*

Figure 2 shows the accuracy as measured by the MAPE_flex metric described above for various baselining methods applied to different DERs that took part in the Project LEO trials. All metrics show mean errors below 10% relative to the estimated size of available flexibility from the DER. White outlines indicate the method with the smallest error for each DER type (excluding MBMA).

The **MBMA (interpolation)** method, which is the simplest, is consistently the best performing baselining methodology across DER types. However, as described above, this is the easiest method to manipulate hence why previous studies have ruled it out for services procured with advanced notice (greater than a few minutes-hours). However, experience from Project LEO suggest there might be a case for considering it, particularly for an immature market which is struggling with low liquidity and competition. The simplicity of MBMA is attractive to DER operators and aggregators alike, which is coupled with low data requirements for IAs and the DNO. The benefits of increasing participation may outweigh the negative effects of malicious manipulation. One of Project LEO's partners, Equiwatt, utilised a simple Meter Before method when baselining customers using smart plugs. The potential for manipulation was avoided through real-time instruction without prior warning of the event.

In general, **the use of SDA** improves the accuracy of all method types across all DERs except for the battery asset. This analysis would therefore support the use of SDA, particularly for DERs where flexibility is on top of a background profile that show daily variations in background consumption – typically driven by weather phenomenon such as temperature of irradiance. However, as per the discussions in section 5.1, there are practical problems that arise with the use of SDA – namely for

DERs that undergo pre-conditioning or adjacent service-stacking, or for the DNO where the market is exposed to manipulation. This error analysis can help market facilitators decide whether the use of SDA to gain greater baselining accuracy (1-2% points in the examples provided) overcomes the risk of manipulation.

Two types of **selection criteria for the X-subset days** have been tested, the NTVV labelled methods uses the nearest X days based on daily energy to the event day, while the unlabelled Historic and Historic with SDA uses the mid-X days based on service window mean power. The results suggest that when applied without SDA, the NTVV method using daily energy has the highest accuracy (smallest error). However, when SDA is included, the mid-X (event period mean) has slight better accuracy. This is not an entirely equal comparison because the SDA for NTVV uses an adjustment window of 4 hours instead of 2 hours.

While a single metric presented in Figure 2 is convenient for comparisons across the many methods and DER types, there needs to be deeper consideration of appropriate methods and their impact on the wider market.

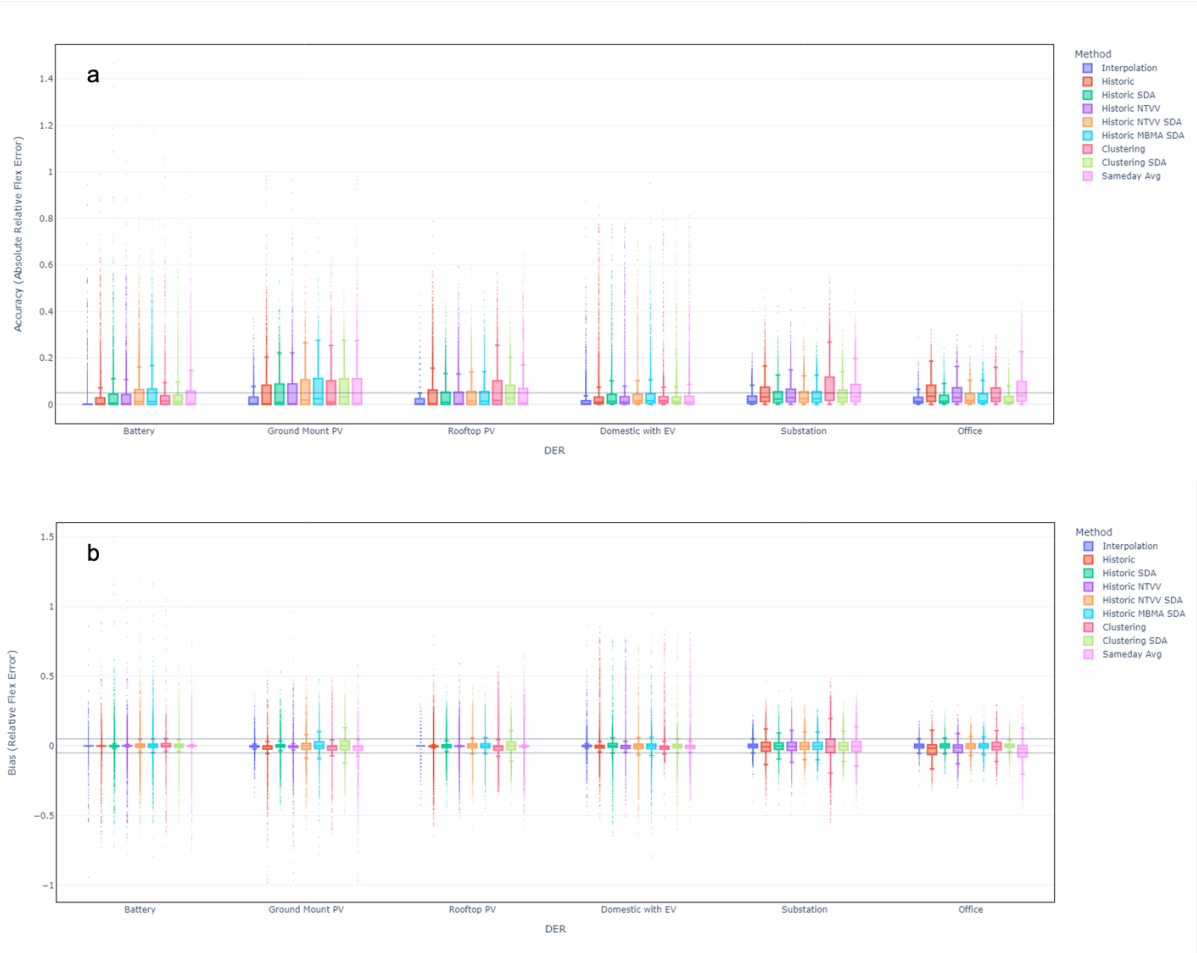

Figure 3: (a) Accuracy and (b) Bias box plots different baselining methods applied to different Project LEO DERs

Figure 3 (a) shows a box plot of the underlying data for the MAPE_flex metric (absolute relative error). While it shows that the interquartile range (defined by the boxes) is within 10% error, the total range of error is quite large, extending beyond 60%, albeit for a small population. Figure 3 (b)

displays relative error which can inform whether there is bias associated with the method. All methods appear to be unbiased with their distributions centred around 0.

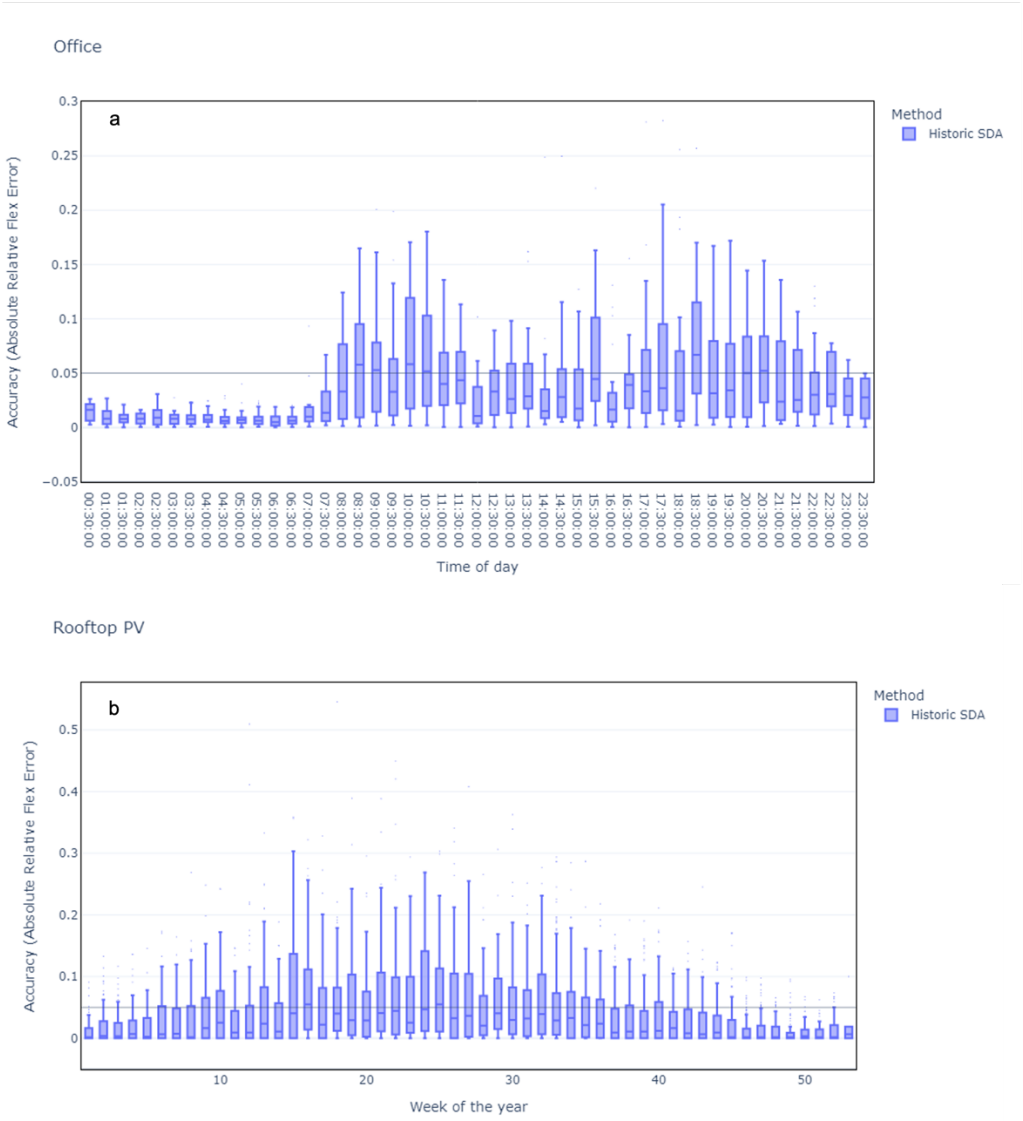

Figure 4: Accuracy for the Historic Baseline with SDA as a function of (a) time of day for an Office building and (b) week of the year for a rooftop PV array.

Figure 4 shows how accuracy of the baseline will change as a function of time of day (a) and time of year (b) depending on DER behaviour; the analysis uses the Historic Baseline with SDA applied to office and rooftop PV data as examples. This highlights how service context should also be a consideration when selecting the appropriate baselining methodology or assessing the impact of the associated baselining error on other aspects of the flexibility procurement such as settlement.

Having this type of error analysis accessible to market facilitators and market participants will help explore the most suitable baselining methods for specific markets, services, and DER types. One topic raised within Project LEO in discussions with project collaborators was that of tailored baselining for DER type, or even each individual DER. The sort of error analysis presented here could be incorporated into the baselining process (or signup process) to establish the most accurate

baseline method for the DER in question (demonstrated by the white boxes in Figure 2). This could be a way of incorporating equity (rather than equality) into the settlement process but comes at the cost of additional computing (and likely personnel) resource.

5 Trial Observations – Deeper Dive

5.1 Variable operation within the SDA window

It was observed that the SDA step can negatively affect baseline accuracy if an irregular event (not captured by historic averaging) or specific service-related action (e.g., preconditioning) occurs in the SDA window. Within the Project LEO trials, this was observed for many battery assets that were set to charge immediately prior to event delivery. It was also identified by DSR assets using HVAC (Heating, Ventilation and Air Conditioning) systems to provide flexibility when utilising a pre-conditioning step to ensure comfort is maintained throughout the event period.

Figure 5 shows the Historic Baseline with SDA applied to a battery asset that was instructed to deliver 30 kWh across a 2 hour sustain peak management service window. On event day, the battery charged immediately before delivery rather than at its regular time in the early hours of the morning. This charging occurred within the 4-hour adjustment window[2] used for SDA. This causes the baseline to be adjusted down by approximately 4 kWh per half hour. The calculated delivery is therefore larger by this amount, wrongfully indicating that 43 kWh has been delivered, a 43% over-delivery.

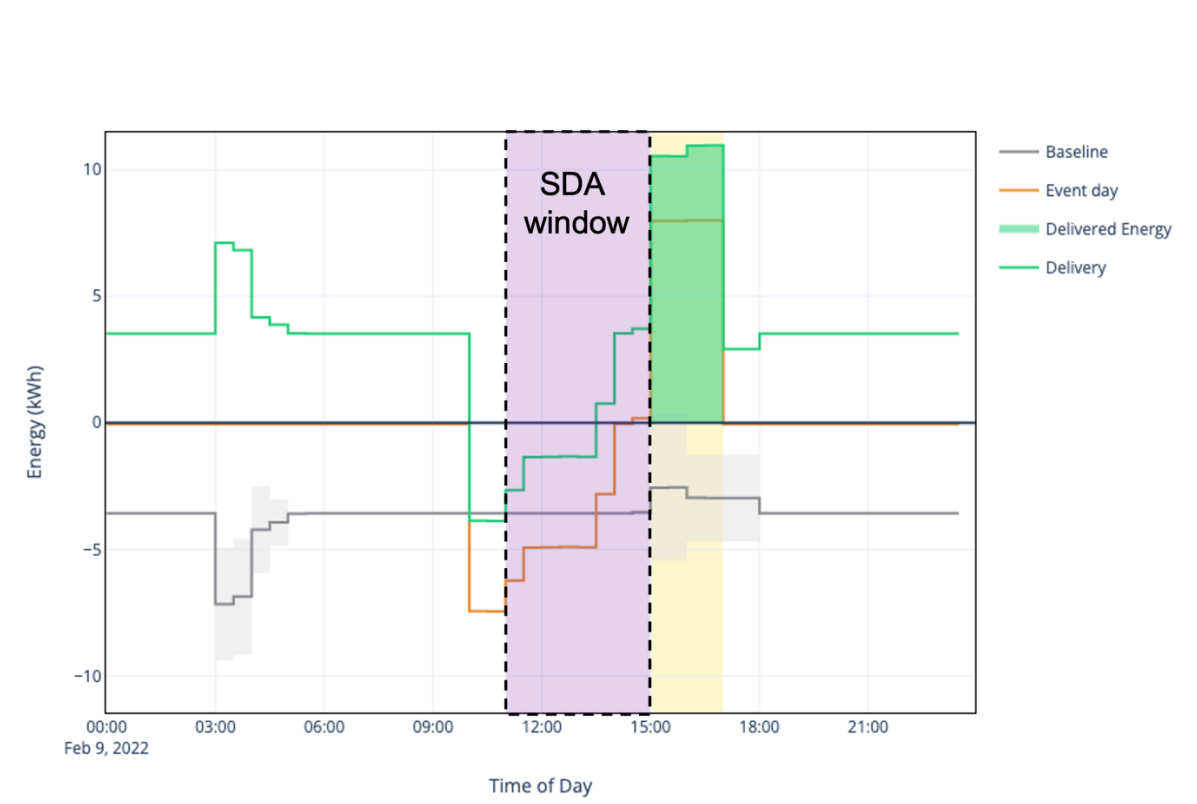

*Figure 5: Historic baseline **with** SDA for a battery asset instructed to deliver 30 kWh across 2 hours in a sustain peak management service. On event day (orange), a pre-charging step occurring immediately prior to the service window and*

[2] 4 hours was used in TP1, this was changed to 2 hours in subsequent trials

within the adjustment window causes the baseline (grey) to be adjusted down to approximately -4 kWh per half hour. The calculated delivery (green) suggests 43 kWh has been delivered, 143% of that actually delivered.

Figure 6 shows the same data but this time baselined using just the Baseline without SDA. In this case, the baseline is not affected by the pre-charging event and remains around 0 as expected.

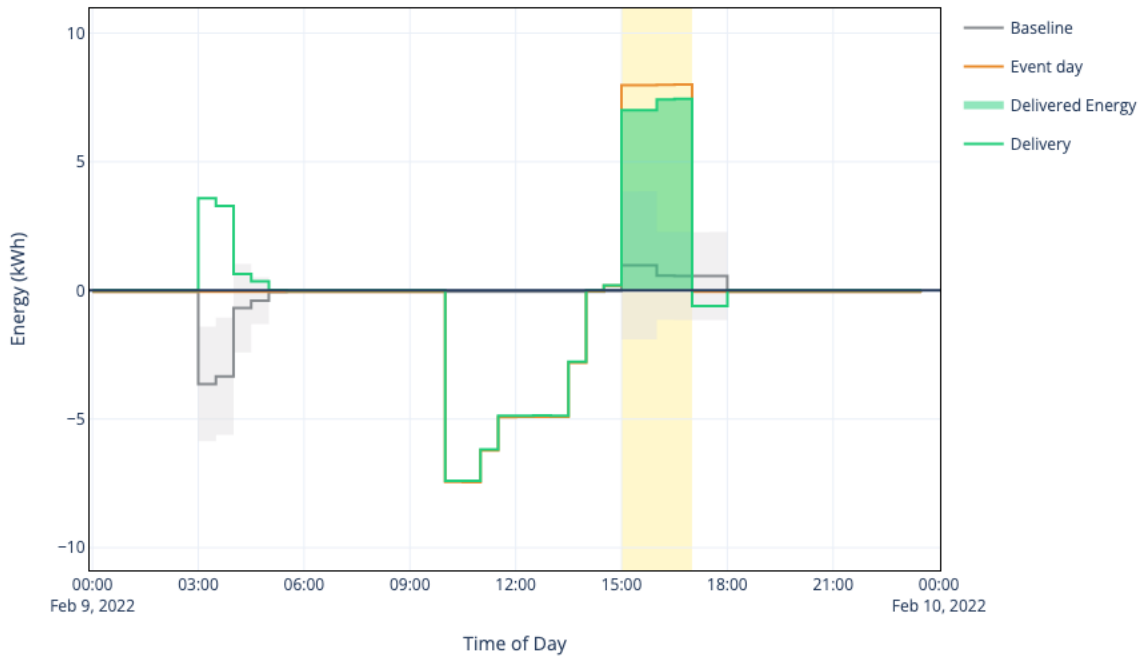

*Figure 6: Historic baseline **without** SDA for the same battery asset service delivery as Figure 5. The baseline (grey) is not affected by the charging event immediately before delivery, instead being centred around 0 as expected. Using this method, the delivery is calculated to be 29kWh.*

In this particular case, the asset will not lose out on revenue because it appeared to be an over-delivery. This is expected for most pre-conditioning steps as they will typically be in the opposite direction to the required delivery. Also, the DNO is not out of pocket either due to a cap on the settlement rule that does not permit payment beyond 100%. However, in the case that this was in the same direction as the flexibility through random DER (or user behaviour) variability, or the DER was taking part in an adjacent service, the DER would be judged to have under-delivered by 43%. In the case of Project LEO, this is compounded by an aggressive settlement rule whereby a 57% delivery would result in only a 17% utilisation payment.

Perhaps more concerning is the potential exposure to gaming. The battery asset could have intentionally delivered only 57% of the contracted amount while still receiving full payment. If widespread, this would either result in a failure to avoid the constraint, or lead to a long-term inefficient market due to unnecessary costly over procurement by the DNO.

Table 1: Domestic battery (3 kW) trial participation showing large over-delivery for SPM and under-delivery for SEPM as a result of pre-conditioning during the SDA adjustment window.

DER	Date	Service	Delivered %
A	15/11/2022	Sustain Peak Management	226
A	18/11/2022	Sustain Peak Management	224
A	22/11/2022	Sustain Export Peak Management	98
B	15/11/2022	Sustain Peak Management	215
B	18/11/2022	Sustain Peak Management	210
B	22/11/2022	Sustain Export Peak Management	80
C	15/11/2022	Sustain Peak Management	169
C	18/11/2022	Sustain Peak Management	168
C	22/11/2022	Sustain Export Peak Management	78

Table 1 contains data from three 3 kW domestic batteries that participated in both Sustain Peak Management (SPM), demand turn-down, and Sustain Export Peak Management (SEPM), demand tun-up services. All three regularly showed over-delivery for the SPM service and under-delivery for the SEPM service. As above, this is due to the batteries pre-conditioning immediately prior to the service during the SDA adjustment window.

> **Recommendation**: The issue is the use of SDA. The DNV-GL's 2020 report to the ENA[2] recommended that SDA is only used for near-real time services such as dynamic and restore because the potential improvement in accuracy with SDA does not overcome the complexities for adjacent service stacking, pre-conditioning, or manipulation. Other options that still include SDA could explore alternative adjustment parameters, for example, using total daily energy or a hybrid regression method utilising secondary data such as temperature.

5.2 Historic Multi-Service Participation

It is likely that IAs, particularly aggregators, will be taking part in both DNO flexibility services and national wholesale flexibility services – like the new NGESO DFS service.[3] There are potentially two problems that were raised during the Project LEO and TRANSITION trials. Firstly, same day delivery – where response to both services coincide on the same day, perhaps during the same (or overlapping) time periods. Secondly, historic service window overlap – where other service delivery occurred during the same time periods on days prior to the DNO service.

For simultaneous delivery with perfect overlap, all baseline methods including the Historic Baseline with (and without) SDA and nomination baseline (if the wholesale service is not included) should be unaffected. There may be settlement conflicts if capacity was split between multiple stacked services within the same market as only the total combined response is measurable. If under-delivery occurred, the contract should make clear how under-delivery is assigned between stacked services (e.g., service importance to DNO, capacity weighted or order of instruction). For same day delivery without perfect overlap of time periods, the issue is the same as discussed in section 5.1 if the additional service occurs within the SDA adjustment window.

[3] Demand Flexibility Service (DFS) - National Grid ESO

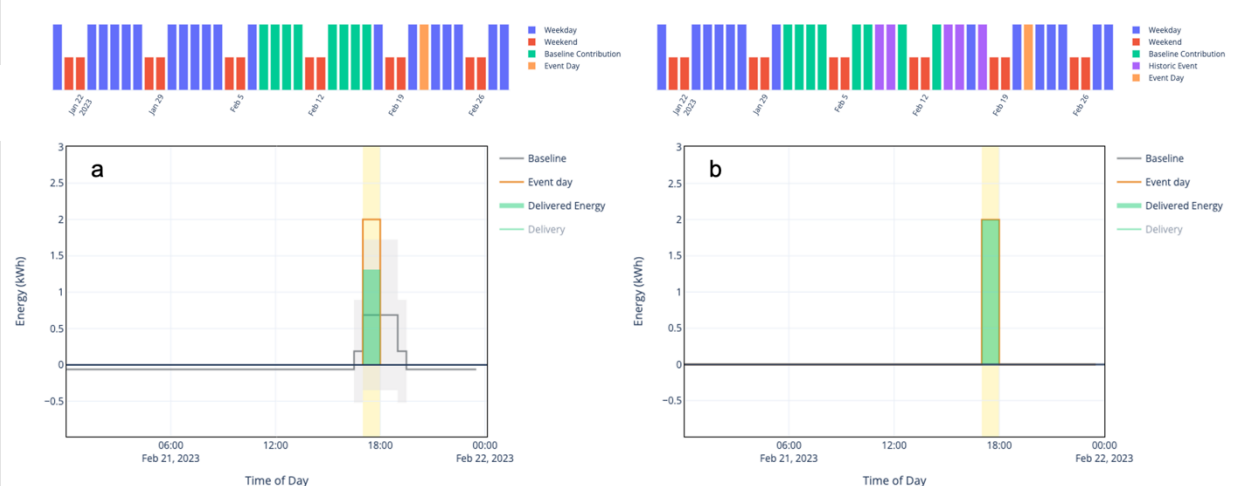

Figure 7: DNO baselining for a DER (synthetic data) providing flexibility to multiple (DNO and ESO) markets during overlapping historic time periods where (a) non-DNO event participation is not shared with the DNO and (b) where non-DNO event participation is shared by the DNO to remove event days prior to baselining. The baseline in (a) is above zero during the event period due to non-DNO flexibility participation being included within historic baselining, which leads to the DER being judged to have under-delivered during the DNO event.

A second issue that arose within trials was that of historic service window overlap. One trial participant that was taking part in the newly introduced NGESO DFS service was judged to have under delivered in a DNO service due to other market participation occurring on days used within the subset of historic days. Figure 7a shows an example of how this issue can impact the measured delivery if the previous event days are not known. Partial mitigation is possible through the selection criteria of X days in Y. It will depend on whether these criteria (median days based on event window mean) successfully excludes these other event days, or if the set of Y days becomes exhausted such that the event days are included. The solution used within Project LEO and TRANSITION was for the IA to notify the DNO of other non-DNO event days which would then be excluded from the set of Y historic days (as is the process for known DNO event days). Figure 7b shows how accounting for these historic event days can negate the issue, with the inset showing days contributing to the baseline being taken from further back in history. This works if the IA is happy to share the supplementary event information (which could be considered commercially sensitive), it does not result in too many days being excluded reducing accuracy (see below), or that it is not misused by the IA to exclude convenient days that provide an advantage. The NGESO DFS event requires MPANs to be recorded by aggregators to ensure properties do not participate through two market participants (e.g., an aggregator and their energy supplier).

> **Recommendation**: To avoid this, the DNO will need a log of all recent historic events from both its internal market and other DNO recognised or allowed external markets that the DER participated in. This could be achieved through a central register of all distribution, transmission, and wholesale events alongside MPANs that participated, providing a way for network companies to automatically exclude other event days without any additional burden on the IA.

5.3 Regular Service Instruction

A potential issue flagged regarded regular service instruction. As previous event days are removed from the historic days used for baselining, high utilisation frequency (many event days within a short period) potentially reduces the number of eligible days and/or increases the time difference between eligible days and event days. The latter is expected to have greater negative impact on DERs that show seasonality such as solar or HVAC loads (partly mitigated with SDA).

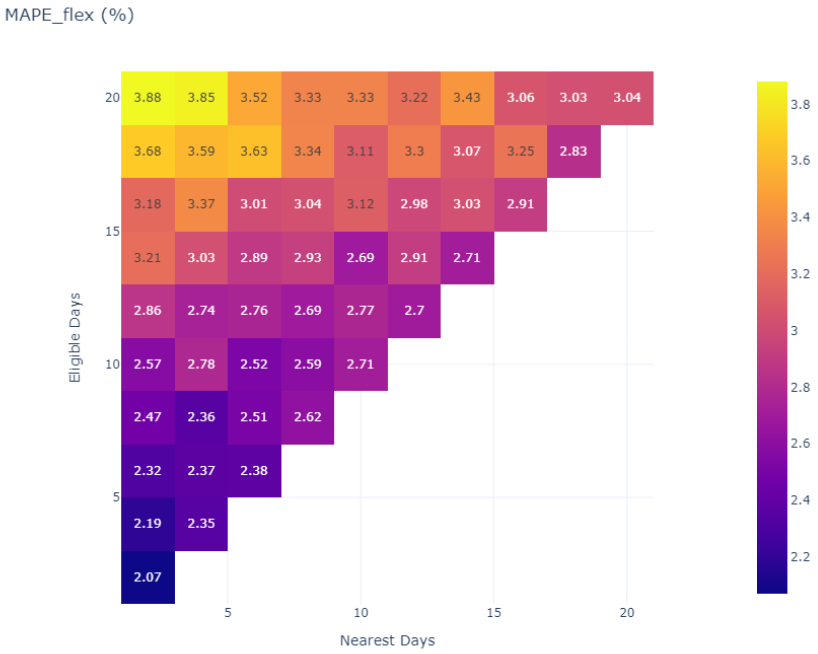

Figure 8: Accuracy measure for an X (Nearest Days) and Y (Eligible Historic Days) hyperparameter scan for the Historic Baseline method for an office building. The MAPE data shows larger errors are observed when a greater number of historic days are eligible i.e. the baseline has contribution from days further away from the event day.

Figure 8 provides an insight into the impact of using historic days further into the past on baseline accuracy. As the number of eligible days is increased, the MAPE error increases for the historic baseline (using the median 'nearest days' subset criteria) without SDA method for an office building with some heating and cooling load. This reduction in baselining accuracy may have to be accounted for if the DER is being called upon with high frequency. The extent of this problem will not be known until DNO flexibility markets become more common places, and the instruction frequency is better understood.

> **Recommendation**: The SDA method should help alleviate this issue as it helps correct for event day variations, however, this will have to be balanced against the issues with SDA highlighted above. Alternative methods that do not rely on historic data close to the event day or even the asset itself, such as clustering, baselining, or control groups, should be explored.

5.4 Behind-the-Meter Optimisation

Local flexibility markets are likely to only account for a relatively small proportion of a DER's operation and business case. Particularly for domestic assets, flexibility will be first utilised for

behind-the-meter optimisation or domestic energy arbitrage. This could be maximising self-consumption of behind-the-meter solar PV or responding to time-of-use tariffs (ToU) – the most well-known example being that of Octopus' Agile Tariff which changes every half hour. A similar effect is observed when business customers optimise their energy use around network charges (e.g. DUoS).

If a DER is regularly taking part in these actions, the response will be present in the majority, if not all, historic days used for baselining DNO (and ESO) flexibility services and likely to be in the same direction. The result will be the same as historic events in Figure 7a, the measured response will be less than the capacity of the asset. The opinion of asset owners is that this is unfair as it reduces the capacity available (up to 100% if historic behaviour has perfect overlap with the service window) to be opportunistic and trade within these less frequent DNO and ESO markets, and does not reflect the true value to the DNO of that asset not operating to its counterfactual (opposite to the DNO's cause). However, market operators such as the DNO would argue that this is long-term behaviour of the asset reflecting a new normal for the network, and likewise would already be accounted for in the DNO's modelling prior to identifying the need for a service.

> **Recommendation**: Firstly, the energy industry (networks, suppliers, or regulators) need to clearly define all the benefits of various schemes that encourage flexibility, ensure that the asset is getting fairly rewarded for this, and clearly explain how this breakdown works. For example, a ToU tariff implemented by an energy supplier initially helps the supplier balance their own energy supply and demand deficits on the wholesale market, however, it encourages behaviour that could provide long-term benefits to the DNO in reducing the need to procure sustain type services. Network companies should explore ways in which this additional value can be paid to the asset, for example, this could be through a dynamic top-up (requiring locational pricing) to the ToU tariff paid through the supplier.

5.5 Influence on Settlement

The Project LEO and TRANSITION market includes a settlement rule which determines payment in the event of under-delivery. The rule is a piecewise linear function with intervals that get more aggressive for higher under delivery and is shown in Figure 9.

$$\phi(\delta) = \begin{cases} 1.0 & \text{if } \delta \geq 0.95 \\ 1.0 - 1.5 \times (0.95 - \delta) & \text{if } 0.85 \leq \delta < 0.95 \\ 0.85 - 2.42 \times (0.85 - \delta) & \text{if } 0.5 \leq \delta < 0.85 \\ 0 & \text{if } \delta < 0.5 \end{cases}$$

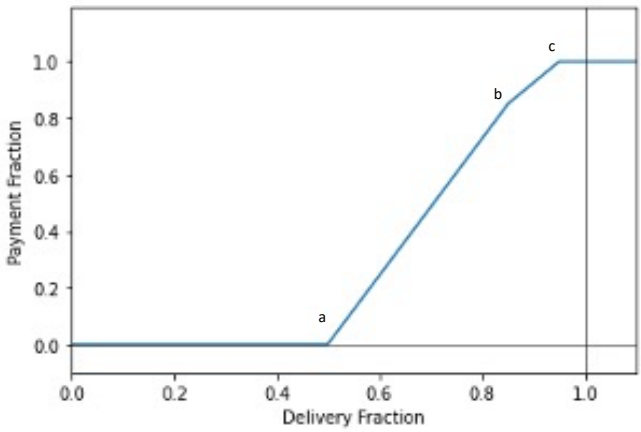

Figure 9: Utilisation Settlement Rule used in Project LEO and TRANSITION trials.

The settlement rule has 5% grace window, anything down to 95% of instructed delivery is paid the full payment. Payment is reduced to 0 for any delivery which is less than 50%. The motivation behind the rule is to encourage high reliability and truthful bidding without the use of penalties.

However, when considering the impact of errors associated with the baseline, the asymmetry and piecewise nature of the settlement rule introduces a bias towards reduced payment, with the more aggressive decline multiplying the error. Figure 10 shows the distribution of relative error for the Historic Baseline with SDA in comparison to the Settlement Rule (black); the black dotted line shows the cut-off for full payment. It is clear some of the relative error is below the 0.95 cut-off meaning payment will be reduced in these cases. Due to the flat cap at 100%, the equivalent over estimation does not result in increased payment to balance this out.

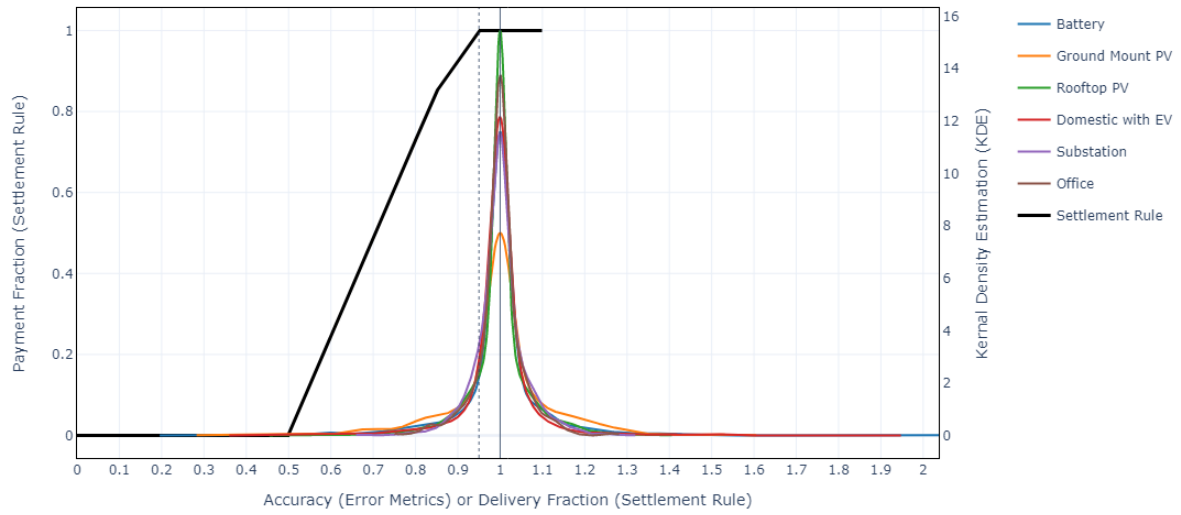

Figure 10: Distribution of relative error for the Historic Baseline with SDA for different DER types (KDE right axis), compared to the Project LEO Settlement Rule (black, left axis). The black dotted line shows the 0.95 cut-off for full payment.

Figure 11 shows the cumulative distribution of payment fraction for different methods across all DER types (a) and for different DER types baselined with Project LEO's Historic Baseline with SDA (b). Cumulative distribution is a measure of the fraction of instances with that payment fraction or less. The point at which the curve dramatically increases for a payment fraction of 1 represents the

probability of some amount of under-payment because of baselining errors. For the different DER types analysed, this is between 10% and 20%.

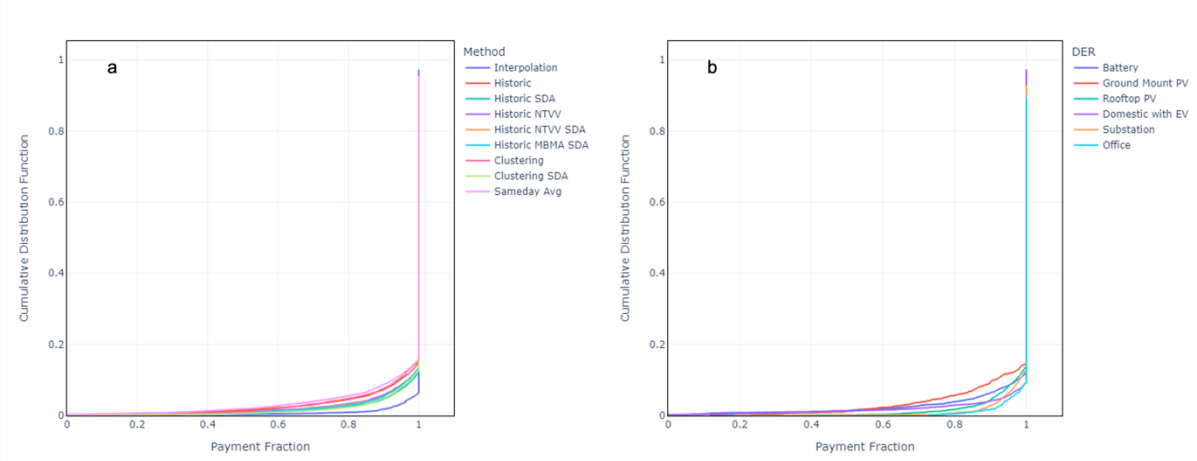

Figure 11: Cumulative distribution of payment fraction for all DERs across different methods (a) and for Historic Baseline with SDA across different DERs (b). The point at which the curves increase dramatically at 1 on the x-axis represents the probability of under-payment because of baselining errors. This is between 10% and 20% for the analysis.

The imbalance in payment due to baseline errors and the settlement rule can be negated if the width of baseline errors is entirely contained within a single linear interval of the settlement rule. If it is flat, nothing is lost or gained to underestimates or overestimates caused by the baseline error. If it is entirely within a non-flat interval, if its linear, any underpayment due to an underestimate will be balanced by an overpayment when overestimated – assuming an unbiased baseline method over many events. The issue is when the baseline error distribution spans multiple intervals.

The settlement rule can be adjusted to reduce the impact of baselining errors on payment. Figure 12 shows how the probability of under-payment due to baseline errors with the Historic Baseline with SDA across the DER type data available varies with the width of the 100% payment cap (or grace interval) in the settlement rule. With the current Project LEO settlement rule ($\beta = 0.95$), the analysis suggests 86% of instances will **not** lead to under-payment due to errors associated with the baseline method. The purple and green dashed lines in Figure 12 represent the 95th and 99th percentile of instances unaffected by baseline errors and can be achieved at $\beta = 0.87$ and $\beta = 0.71$ respectively. This is assuming perfect delivery. If there is genuine under delivery, errors in the baseline could still lead to underpayment if near a transition point in the settlement rule.

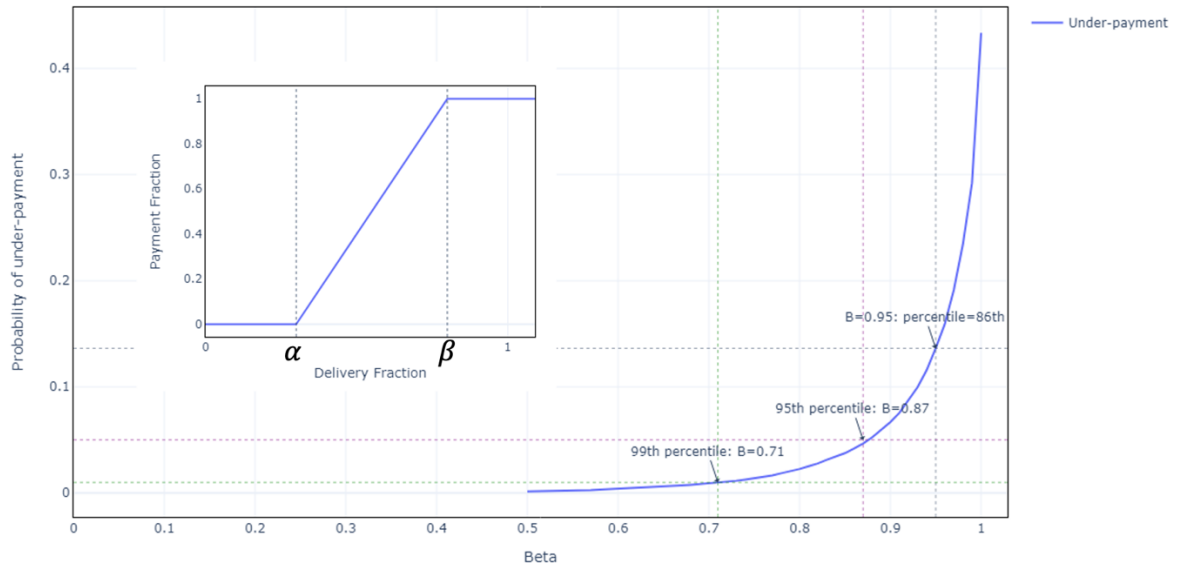

Figure 12: Probability of under-payment due to baselining errors for the Historic Baseline with SDA as a function of (which defines the 100% payment cap or grace interval). Inset: general form of a simplified piecewise settlement rule with 3 intervals defined by

Recommendation: If no improvement in the error associated with baselining can be achieved, the settlement rule should be modified to account for this. The 100% payment cap width, as defined by β in the inset of Figure 12, should be extended. The analysis presented here suggests a cap at 85% delivery should accommodate over 95% of baseline errors. Ultimately, the extent to this will be a balance between accounting for the majority of baseline errors versus the cost of potentially having to over-procure by the same percentage if DERs operate at lower (e.g. 80%) reliability (while still receiving full payment). During market infancy, being more lenient should help increase market participation and liquidity.

5.6 Data Quality and Transaction Costs

For Project LEO and TRANSITION trials, baselining utilised the ENA's Baselining Tool developed as part of TRANSITION by TNEI. To be used, the data must be of sufficient quality and in the correct format. The IA was expected to carry out a manual data quality check and subsequent cleaning if required, ahead of submission. During the trials, SSEN regularly received data of insufficient quality or incorrectly formatted for further processing. Meanwhile, the manual data cleaning step added significant burden of participation for the IA. This problem is largely caused by the trials' openness to different metering requirements as an innovation project and to lower the barrier to participation encouraging greater liquidity in the market.

Another data quality related issue is that of timestamps and dealing with the clock change. Project LEO and TRANSITION required the timestamp to be in 'local time' – as it would appear on a clock. This may require action from the IA to change format pre-submission but means that discontinuities exist in the historic data. Likewise, consistency is required for whether the timestamp is the start or

the end of a period. For some methods that involve relatively short historic data (Historic with SDA, nomination, interpolation), potential issues are minimal, limited only to a few days after the clock change. For methods that require larger historic ranges (clustering, regression, etc) it may have greater influence. There will also be a dependency on the DER type and the degree to which humans influence usage. For instance, the behaviour of a solar PV asset is not affected by a clock change and would be better recorded against a set timezone (e.g. UTC) the year round, whereas demand response is more likely to be strongly affected by people's response to the clock change, therefore local time is the more appropriate index. This gets more complex for combined assets behind a single meter, for example rooftop solar on a building with HVAC providing DSR. It is generally seen as good practice that all timestamps have timezone information included. Conversion is then undertaken as part of the baselining process as appropriate for the baseline methodology and DER type.

> **Recommendation**: To baseline at scale with low transaction costs, baselining must be an automated process which is integrated into wider market processes. To achieve this, data must be of sufficient quality and in the correct format. Industry wide data standards and integrated cleaning/formatting tools are required. These could be developed centrally and used across markets and IAs via open APIs to reduce the many different requirements across different markets – this standardisation would assist aggregators and other IAs taking part in multiple markets. Such tools can also improve transparency of the baselining process.

5.7 Over-delivery from Solar PV

It was reported by Solar PV operators that the Historic Baseline with SDA regularly over-estimated the baseline which manifested in PV assets appearing to over deliver for a SEPM service. While this was initially attributed to the method being inappropriate for PV assets and thus judged inaccurate, this might not be strictly true. The error analysis presented above does not show any strong bias for PV assets. The issue may be a manifestation of the PV operator's control method as demonstrated in the diagram in Figure 13. Flexibility from PV is achieved by restricting the output of the PV to an amount of flexibility capacity below that of the output at the start of the event, rather than dynamically tracking the counterfactual. In this case, the PV technically did over deliver with the baseline reflecting that correctly. However, it highlights that when choosing a baseline (or for DNO forecasting), consideration might have to be made to how the DER is controlled, otherwise the market is at risk of consistently and unintentionally over-procuring flexibility. While this may not put the network at risk, it might make for a costly and inefficient market.

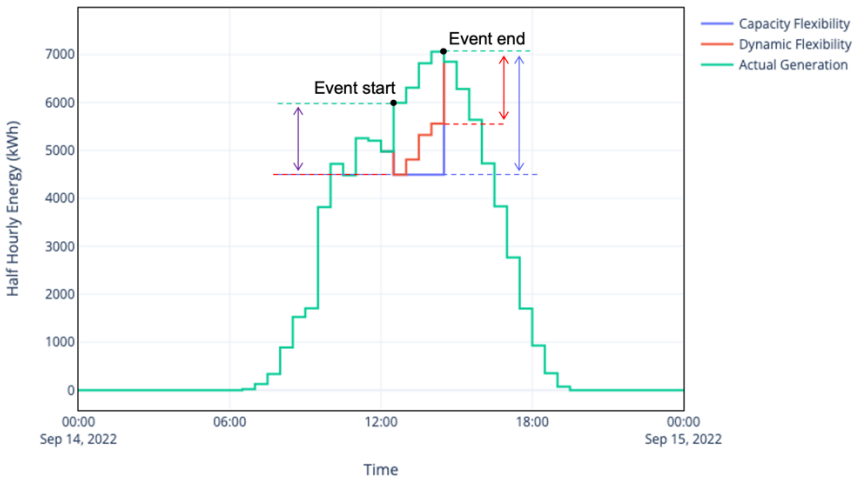

Figure 13: If PV flexibility is controlled by capping output capacity for the whole event window based on generation at the start of the event (blue line), it may lead to genuine over-delivery (or under-delivery) compared to dynamic control (red) which tracks the actual generation potential (green). This is not an issue with the baseline, although a Meter Before baseline method would better align with this control strategy.

5.8 Nomination Baseline

The nominated baseline was not popular with IAs during Project LEO and TRANSITION trials. Throughout all three trial periods, the nominated baseline option was only used on one occasion with IAs preferring the Historic Baseline with SDA. Feedback from IAs suggests this is due to the perception that this method is more difficult, time consuming and costly for the IA. This is due to the required capability on the IA to forecast asset behaviour which, particularly for non-traditional IAs, may require more time, knowledge, and appropriate software tools, and submit it ahead of every event in addition to the standard submission of post event data. This barrier might not be as large as perceived depending on the DER and how accurate or involved forecasting needs to be. For example, if an asset is metered directly and not participating in any other service, a simple baseline of zeros might be appropriate.

The FUSION project,[4] led by SP Energy Networks as part of the ENA funded Open Networks Project alongside TRANSITION, has much higher adoption of the nominated baseline method within the project's flexibility trails. At time of writing, knowledge sharing between TRANSITION and FUSION on baselining learnings is ongoing with the aim of producing shared recommendations.

6 Summary and Recommendations

This report has presented an empirical analysis of the baselining methodologies associated with distribution flexibility procurement trialled within Projects LEO and TRANSITION, specifically how to assess the accuracy of methods and the resulting impact throughout the market. It has also

[4] FUSION project: https://www.spenergynetworks.co.uk/pages/fusion.aspx

highlighted some specific challenges and shortcomings of the Historic Baseline with SDA and wider market operation that were raised as part of real-world trials.

Key observations and associated recommendations are summarised below:

1. **The analysis of baseline errors provides valuable insight for the design of flexibility market facilitation**. The baselining error analysis presented herein demonstrates that: (a) different baselining types might be more of less suitable for both different DER types and services. The analysis presented herein suggests the most simple MBMA interpolation method typically has higher accuracy than the more complicated methods when tested in isolation, however qualitative considerations such as openness to manipulation must also be considered; and (b) it can be used to inform the design and understand (and mitigate negative) impact on other market components such as settlement – the market needs to be considered as a whole integrated system, and it is important to understand how baselining affects other aspects of the market from capacity reserve to settlement. ***Baseline error analysis should be available (through a centralised tool) to market facilitators and industry actors to improve market transparency. It could be integrated as part of the baselining process itself in order in provide a more tailored, equitable, approach to baseline methodology assignment.***

2. **Same Day Adjustment based on usage during narrow window prior to the event is vulnerable to manipulation and not suitable for assets using pre-conditioning.** In general, SDA improves the accuracy of all baselining methods where applied. However, SDA requires a reference point or window prior to the event. If an asset is variable with high uncertainty, or requires a pre-conditioning step prior to utilisation, which occurs in the reference window, the SDA method can dramatically misrepresent the baseline leading to apparent over or under delivery. This feature exposes the method to easy manipulation should an IA decide to game the market. Error analysis using only non-event data will not detect such an attribute. ***The ease or payback of doing this may be reduced by removing SDA, changing, or expanding the reference window, or using a regression type adjustment trained on external features related to the underlying phenomenon that is the focus of the correction. SDA is most suitable for services where instruction has little to no warning, coming close to the real time need for the service. Removing prior notification of the service removes the IAs ability to game the method.***

3. **Service stacking likely to impact the baseline.** Flexibility procured by distribution networks likely only represent a small fraction of available value that can be achieved from flexibility. DERs will regularly be taking part in other flexibility services be that local, national energy services (such as NGESO's new DFS service) or behind-the-meter energy arbitrage activity. For many of the historic methods, this will impact the baseline in a way that negatively reflects on the true capacity of flexibility being delivered by the IA or DER. This either occurs from inclusion of previous flexibility events within the baseline calculation, or a less accurate baseline due to historic days being excluded entirely or taken from further back in time. For some long-term regular activities such as response to time of use tariffs, this activity should be included within the baseline because as this behaviour will already be expected and included within DNO forecasts used to procure flexibility in the first place. ***For***

other more infrequent flexibility activity, it is important that DNOs (and other market facilitators) have knowledge of a DERs previous participation in other markets. This could be in the form of a centralised national database of participation maintained by the market actors to remove any additional burden on flexibility providers.

4. **The nominated baseline was not popular with IAs**. Throughout all three trial periods, the nominated baseline option was only used on one occasion due to the perception that this method is more difficult, time consuming and costly for the IA. Depending on the DER and IAs capability, this might not be the barrier it is perceived to be and might lead to more accurate baselining. ***Market facilitators should help identify cases where a nomination baseline approach improves accuracy without adding additional burden to the IA.***

5. **Poor data quality has negative impact, adding unnecessary transaction cost**. Poor data quality which includes missing data, formatting errors and timestamp inconsistency was a common issue during the Project LEO and TRANSITION trials. Poor quality can lead to less accurate baselining or in the more extreme case, stop any baseline algorithms running. Manual data quality check and subsequent cleaning if required ahead of submission can be very time consuming (up to hours) which adds unnecessary transaction cost to providing flexibility which may make participation infeasible. ***Data quality and metering requires defined standards across markets with cleaning and transformation methods incorporated within centralised market tools or platforms.***

6. **Asymmetric settlement rule.** Project LEO and TRANSITIONS settlement rule which is used to determine utilisation payment as a function of delivery fraction is an asymmetric piecewise function with a cap at 100% payment for any amount of over delivery. If errors associated with the baseline method are larger than the piecewise intervals, baselining errors may cause incorrect underpayment. Compounding this is the asymmetric nature of the settlement rule with a cap that means under-payment is not balanced by over-payment over many events. ***The shape of the settlement rule should be tailored to minimise the impact that baselining errors have following baselining analysis. This can be achieved by increasing the width of the linear regimes, most obviously the grace window where 100% payment is received.***

7. **Non baseline alternatives.** Finally, alternative market mechanisms that do not require baselining should be explored. This will remove errors associated with baselining, which in a minority of cases can be very large. This could be in the form of a dynamic capacity-based flexibility market that only requires absolute measurement of usage from metered data, or integration with other coordination mechanisms such as time-of-use tariffs.

7 Appendices

Appendix A – Error metrics

When considering flexibility from numerous different assets with different flexibility capacities, that have very different standard behaviour (batteries might have zero as the counterfactual whereas an office or PV will have a larger baseload), an alternative to the typical relative error that uses the actual value as the divisor, is to use the DERs flexible capacity (C_{fx}) as the divisor. This is how flexibility response will be judged by the market and can be directly compared with the settlement rules.

For example, a domestic property, for much of the time, the demand is likely to be small in comparison to the size of flexibility offered to the market (a 7kW EV charger or 3kW battery). If the actual demand is close to zero, the relative baseline error will be large (dividing by a small number). However, the absolute error in comparison to the size of a flexibility event might still be very small.

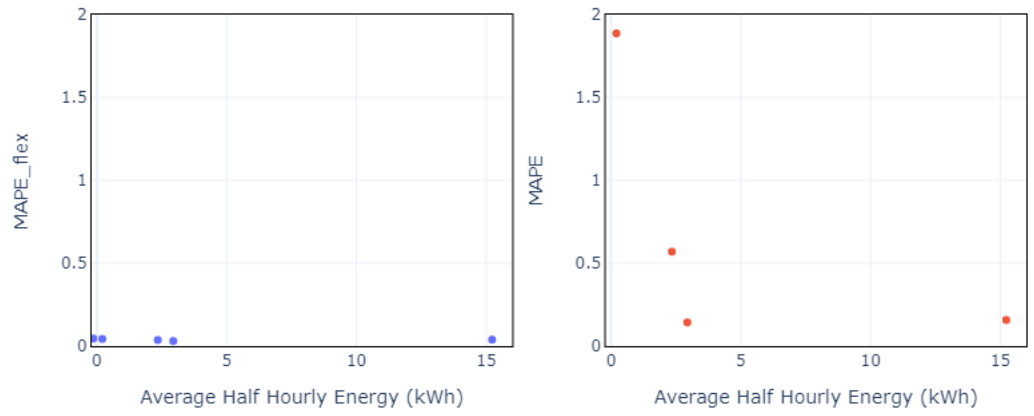

Figure 14: MAPE_flex (left) and the MAPE (right). Using the capacity of flexibility as the denominator for relative errors rather than the average baseload consumption allows for better comparison of accuracy when considering flexibility services for DERs of very different capacities. DERs with near 0 baseload will have an artificially inflated error..

The figures above demonstrate how using a base of average demand during the event interval (right) to calculate relative or percentage errors makes the error metric for assets with near zero baseline appear artificially high compared to using a base which represents the flexibility capacity (left).

Appendix B:

Table 2: Categories of baselining methods commonly used within industry and their advantages and disadvantages.

Category	Description	Advantages	Disadvantages
Historic Averaging	Baseline is calculated by taking the average of X days from a set of Y admissible days preceding the event. Also known as X-of-Y, it is the most common method used in existing flexibility markets. There are many variations on this method and might include weighted averaging (giving higher weighting to more recent days), adjustments to factor in weather-sensitivities for instance and exclusion of atypical days (e.g. holidays, previous event days).	• Low data requirements (small date range, only individual metered data) • Low frequency of data submissions (batch process historic events) • Easy to calculate. • Many variations (possibility to tailor to asset or service) • Good for small perturbation relative to baseline with low variance.	• Evidence suggests not good for uncertainty or small DERs. • Many variations (large parameter space) - can lead to lack of transparency. • Susceptible to large changes in energy profile when used with short historical period duration.
Control Groups	The baseline is calculated from the metered data of a control group that did not participate in a service.	• Hard to manipulate. • Easy to calculate	• Sterilises available capacity to the market. • Requires high market participation (liquidity) or large portfolio. • Control group must be representative of the group delivering flexibility (and experience similar weather conditions). • medium data requirements (meter data from users not participating) • requires statistically significant group size to generate representative baseline.
Scheduling	Using ahead of time forecasts provided by the IA (at aggregator, customer, or asset level). Nominated Baseline is an example of this.	• Can be simple for the IA (depending on asset and ease of forecast) • Minimal DSO data processing	• Can be challenging for the IA (requiring in house forecast). • Higher frequency of data submissions (event-by-event) • Easy to manipulate. • Requires auditing or disincentive to manipulate. • Probability of activation needs to be low to incentivise truthful behaviour – high-capacity sterilisation. • Disadvantageous for non-traditional IAs who may not have sufficient resources to produce suitable accurate baselines compared to a professional IA.
Interpolation	Considers the metered data in the periods before and after the event and interpolates values within	• Simple to calculate. • Minimal data requirements	• Easy to manipulate.

	the period window. This can be a simple linear interpolation or more complex polynomial or hybrid regression interpolation method.	• Low frequency of data submissions (batch process historic events)	• Will not account for variations within (shorter timeframe) width of the service window. • Does not account for devices that would have been used during the service window but avoided.
Regression	The baseline is calculated from historic data using a regression model where feature variables might include temperature, sunrise/sunset times, irradiance or temporal offset.	• Accounts for variability due to external factors •	• High data requirements to get best performance (historic data on order of years) • Secondary data requirement and cost (e.g. weather) • Model training required. • Complex implementation.
Machine Learning and hybrid methods	Large suite of deep-learning methods such as neural networks.	• Reported to be highest accuracy	• High data requirements to get best performance (historic data on order of years) • Secondary data requirement and cost (e.g. weather) • Model training required. • Complex implementation. • Cannot be audited/traced on the outcome process to determine performance on edge cases.
Same Day Adjustment	Not strictly a baseline method in it's own right, same day adjustment is often a second step across other methods whereby the baseline is adjusted based on a reference point prior to the event to correct for background effects e.g. weather.	• Simple correction • Removes the penalty from environmental changes impacting demand response from weather-driven energy use.	• Easily manipulated depending on reference point (and if known)

Appendix C: Impact on Settlement Rule

Below are some additional figures part of section 5.5 exploring the baseline error interaction with the settlement rule. Figure 15 shows how the distribution of accuracy for different baselining methods compares to the settlement rule. Figure 16 show how the majority of settlement is paid at 100% due to the majority of the baseline error being within the settlement grace window (which pays full payment for delivery over 95%).

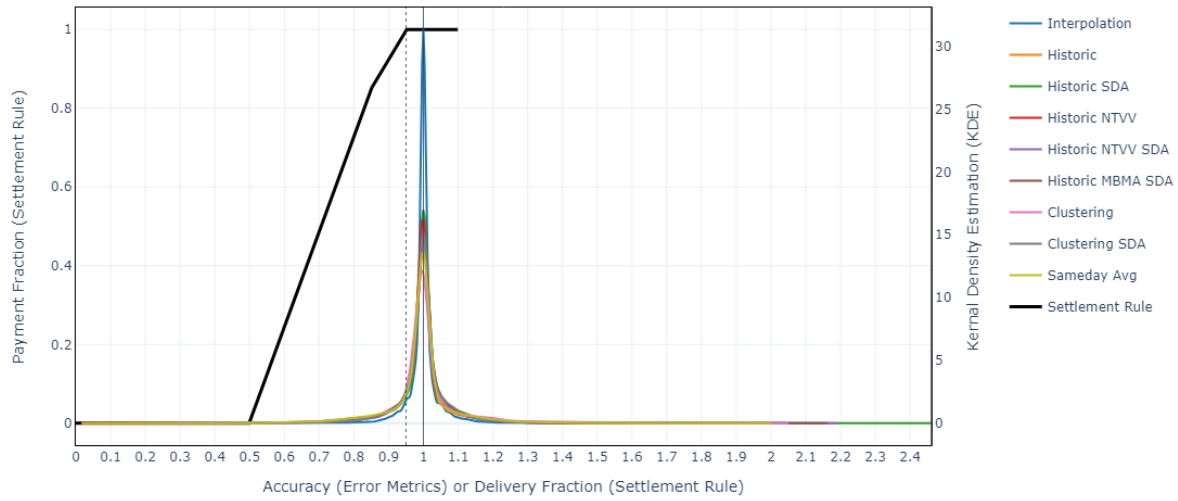

Figure 15: Distribution of relative error for different baseline methods (KDE right axis), compared to the Project LEO Settlement Rule (black, left axis). The black dotted line shows the 0.95 cut-off for full payment.

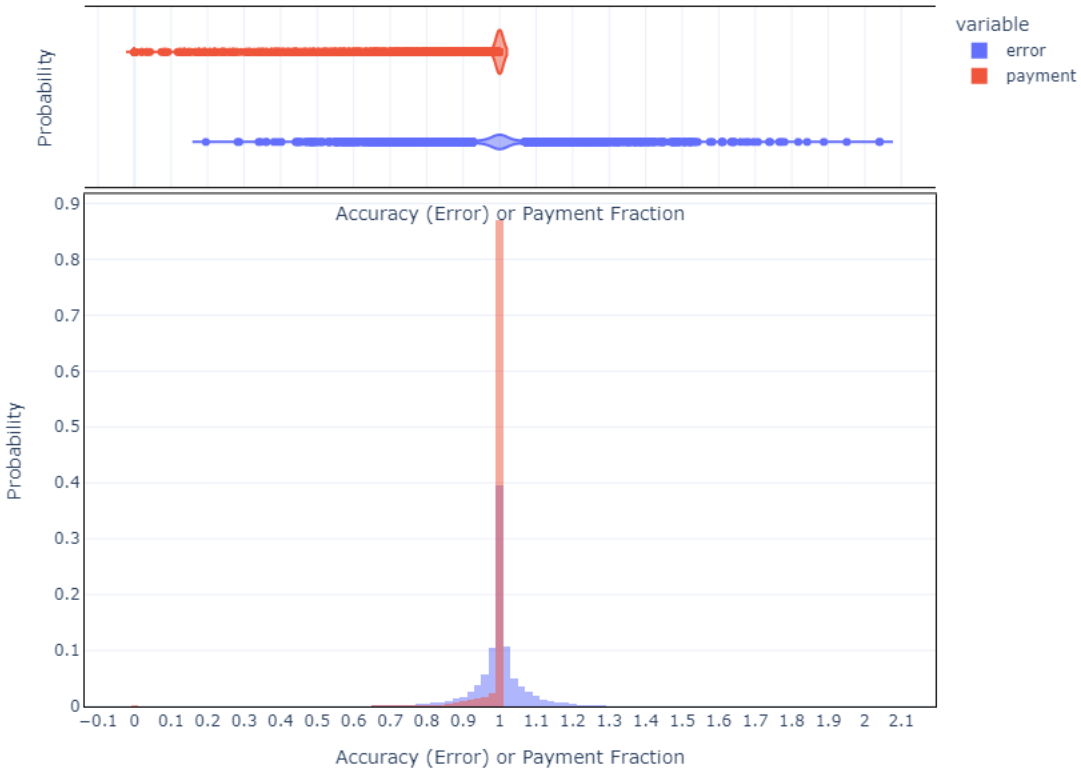

Figure 16: The distribution of settlement payments (red) when the settlement rule is applied to the distribution of baseline accuracy (blue). The majority of settlement is paid at 100% due to the majority of the baseline error being within the settlement grace window which pays full payment for delivery over 95%.

Appendix D – Data

The data used for this version of analysis includes datasets from 6 different DERs that were investigated as part of Project LEO. This includes a ground mount solar farm, a commercial battery storage installation, a commercial rooftop solar installation, an office building, load from a secondary (11 kV – 400 V) substation and a domestic property with electric vehicle.

The baselining methods code and error analysis code will be made available on GitHub.

8 Acronyms

Term	Description
DER	Distributed Energy Resource. Used interchangeably with Asset to refer to equipment capable of providing a flexibility service to the market.
DFS	Demand Flexibility Service
DNO	Distribution Network Operator
DSO	Distribution System Operator
DSR	Demand Side Response
ENA	Energy Networks Association
ESO	Energy System Operator
HVAC	Heating, Ventilation and Air Conditioning
IA	Industry Actor
LEO	Local Energy Oxfordshire
MBMA	Meter Before Meter After
NG ESO	National Grid Electricity System Operator
SDA	Same Day Adjustment
SEPM	Sustain Export Peak management
SPM	Sustain Peak Management

9 Bibliography

[1] SSEN and TNEI, 'Historic Baselining Methods - Performance Assessment', 2022. Accessed: Mar. 28, 2023. [Online]. Available: https://ssen-transition.com/wp-content/uploads/2022/08/14349-007-R0-Historic-Baselining-Methods-Performance-Assessment-1.pdf

[2] DNV-GL, 'Baseline Methodology Assessment: Energy Networks Association', 2020.

[3] KEMA, 'PJM Empirical Analysis of Demand Response Baseline Methods', 2011. Accessed: Mar. 28, 2023. [Online]. Available: https://www2.pjm.com/-/media/markets-ops/demand-response/pjm-analysis-of-dr-baseline-methods-full-report.ashx

[4] DNV-GL, 'Evaluation of 2017 Demand Response Demonstrations: C&I Connected Solutions', 2018. Accessed: Mar. 28, 2023. [Online]. Available: https://ma-eeac.org/wp-content/uploads/National-Grid-Connected-Solutions-Final-Report.pdf

[5] Elia Group, 'Baseline Methodology Assessment', 2021. Accessed: Mar. 28, 2023. [Online]. Available: https://www.elia.be/-/media/project/elia/elia-site/public-consultations/2021/20210927_study_baseline_methodologies_draft_clean_en.pdf

[6] ARENA and Oakley Greenwood, 'Baselining the ARENA-AEMO Demand Response RERT Trial', 2019. Accessed: Mar. 28, 2023. [Online]. Available: https://arena.gov.au/assets/2019/09/baselining-arena-aemo-demand-response-rert-trial.pdf

[7] C. Ziras, C. Heinrich, and H. W. Bindner, 'Why baselines are not suited for local flexibility markets', *Renewable and Sustainable Energy Reviews*, vol. 135, p. 110357, Jan. 2021, doi: 10.1016/J.RSER.2020.110357.

[8] K. Coughlin, M. A. Piette, C. Goldman, and S. Kiliccote, 'Statistical analysis of baseline load models for non-residential buildings', *Energy Build*, vol. 41, no. 4, pp. 374–381, Apr. 2009, doi: 10.1016/J.ENBUILD.2008.11.002.

[9] Y. M. Wi, J. H. Kim, S. K. Joo, J. B. Park, and J. C. Oh, 'Customer baseline load (CBL) Calculation using exponential smoothing model with weather adjustment', *Transmission and Distribution Conference and Exposition: Asia and Pacific, T and D Asia 2009*, Dec. 2009, doi: 10.1109/TD-ASIA.2009.5356984.

[10] K. Coughlin, M. A. Piette, C. Goldman, and S. Kiliccote, 'Estimating Demand Response Load Impacts: Evaluation of BaselineLoad Models for Non-Residential Buildings in California', Jan. 2008, doi: 10.2172/928452.

[11] J. Jazaeri, T. Alpcan, R. Gordon, M. Brandao, T. Hoban, and C. Seeling, 'Baseline methodologies for small scale residential demand response', *IEEE PES Innovative Smart Grid Technologies Conference Europe*, pp. 747–752, Dec. 2016, doi: 10.1109/ISGT-ASIA.2016.7796478.

[12] T. K. Wijaya, M. Vasirani, and K. Aberer, 'When bias matters: An economic assessment of demand response baselines for residential customers', *IEEE Trans Smart Grid*, vol. 5, no. 4, pp. 1755–1763, 2014, doi: 10.1109/TSG.2014.2309053.

[13] F. L. Müller and B. Jansen, 'Large-scale demonstration of precise demand response provided by residential heat pumps', *Appl Energy*, vol. 239, pp. 836–845, Apr. 2019, doi: 10.1016/J.APENERGY.2019.01.202.

[14] E. Lee, K. Lee, H. Lee, E. Kim, and W. Rhee, 'Defining virtual control group to improve customer baseline load calculation of residential demand response', *Appl Energy*, vol. 250, pp. 946–958, Sep. 2019, doi: 10.1016/J.APENERGY.2019.05.019.

[15] J. Vuelvas, F. Ruiz, and G. Gruosso, 'Limiting gaming opportunities on incentive-based demand response programs', *Appl Energy*, vol. 225, pp. 668–681, Sep. 2018, doi: 10.1016/J.APENERGY.2018.05.050.

[16] D. Muthirayan, D. Kalathil, K. Poolla, and P. Varaiya, 'Mechanism design for demand response programs', *IEEE Trans Smart Grid*, vol. 11, no. 1, pp. 61–73, Jan. 2020, doi: 10.1109/TSG.2019.2917396.

[17] Y. Lin, J. L. Mathieu, J. X. Johnson, I. A. Hiskens, and S. Backhaus, 'Explaining inefficiencies in commercial buildings providing power system ancillary services', *Energy Build*, vol. 152, pp. 216–226, Oct. 2017, doi: 10.1016/J.ENBUILD.2017.07.042.

[18] J. L. Mathieu, P. N. Price, S. Kiliccote, and M. A. Piette, 'Quantifying changes in building electricity use, with application to demand response', *IEEE Trans Smart Grid*, vol. 2, no. 3, pp. 507–518, Sep. 2011, doi: 10.1109/TSG.2011.2145010.

[19] E. Larsen, K. Rosenørn, and A. Jónasdóttir, 'Baselines for evaluating demand response in the EcoGrid 2.0 project', 2019. doi: http://dx.doi.org/10.34890/293.

[20] Y. Zhang, W. Chen, R. Xu, and J. Black, 'A Cluster-Based Method for Calculating Baselines for Residential Loads', *IEEE Trans Smart Grid*, vol. 7, no. 5, pp. 2368–2377, Sep. 2016, doi: 10.1109/TSG.2015.2463755.

[21] K. Li, B. Wang, Z. Wang, F. Wang, Z. Mi, and Z. Zhen, 'A Baseline Load Estimation Approach for Residential Customer based on Load Pattern Clustering', *Energy Procedia*, vol. 142, pp. 2042–2049, Dec. 2017, doi: 10.1016/J.EGYPRO.2017.12.408.

[22] Y. Chen *et al.*, 'Short-term electrical load forecasting using the Support Vector Regression (SVR) model to calculate the demand response baseline for office buildings', *Appl Energy*, vol. 195, pp. 659–670, Jun. 2017, doi: 10.1016/J.APENERGY.2017.03.034.

[23] X. Wang, Y. Wang, J. Wang, and D. Shi, 'Residential Customer Baseline Load Estimation Using Stacked Autoencoder with Pseudo-Load Selection', *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 1, pp. 61–70, Jan. 2020, doi: 10.1109/JSAC.2019.2951932.

[24] Open Networks, 'Flexibility Baselining Tool - Mathematical Specification', 2022.

Visit us at

www.project-leo.co.uk/

Stay Connected for news, events and much more…

www.project-leo.co.uk/stay-connected/