CATAPULT
Energy Systems

WP2 – Methods Development

May 2023

Document control

ESC programme name	**SIF ENW Metered Energy Savings**
ESC project number	ESC00858
Version*	1.0
Status	Approved: Contains reviewed and approved content.
Restrictions*	Open
Release date	08/06/2023
External release ID	WP2

Review and approval

	Name	Position
Author	Will Rowe Jessica Steinemann	Data Scientist Senior Data Scientist
Reviewer(s)	Sam Young	Practice Manager, Data Science & AI
Approver	Eveline Sleiman	Senior Commercial Manager - Homes

Revision history

Date	Version	Comments
08.06.2023	V1.0	Initial release

Contents

Glossary

ACRONYM	DEFINITION
AEU	Avoided Energy Use
ASHRAE	American Society of Heating Refrigerating and Air Conditiining Engineers
CTB	Comfort Take-Back
CV(RMSE)	Coefficient of Variation of the Root Mean Square Error
DECC	Department of Energy and Climate Change
EVO	Efficiency Valuation Organisation
FSU	Fractional Savings Uncertainty
HTC	Heat Transfer Coefficient
MAPE	Mean Absolute Percentage Error
MES	Metered Energy Savings
M&V	Measurement and Validation
NEED	National Energy Efficiency Data-Framework
NMBE	Normalised Mean Bias Error
P4P	Pay for Performance
ToU	Time of Use

1. Executive Summary

This report focuses on the technical methods that will be developed to measure energy savings as part of RetroMeter. Developing a methodology that is universally applicable would be extremely challenging due to the broad scope of both retrofits and the potential end goals for which a metered energy savings methodology could be used. The aim is to therefore to develop a broadly applicable methodology that would be effective in the most common use cases, and could potentially be extended in future. To do this, we look at methodologies already in use and potential ways in which these might be combined.

Since the majority of UK properties are currently gas heated, development of the methodology in Alpha and Beta will focus on modelling gas rather than electricity. This allows the evaluation of both fabric-only retrofits on gas heated properties and gas boiler to heat pump retrofits (as long as the energy consumed by the heat pump is directly measured). It has the significant advantage of eliminating the need to develop a half-hourly model, as gas prices are not dependent on time of use.

In addition, requiring internal temperature for a year pre-retrofit poses too significant a barrier to widespread adoption, so methods that require that have been ruled out.

Around half of homes have a smart meter (and therefore should have 13 months of data pre-retrofit), so the core methodology can utilise this – but there needs to be an option for when that data is not available.

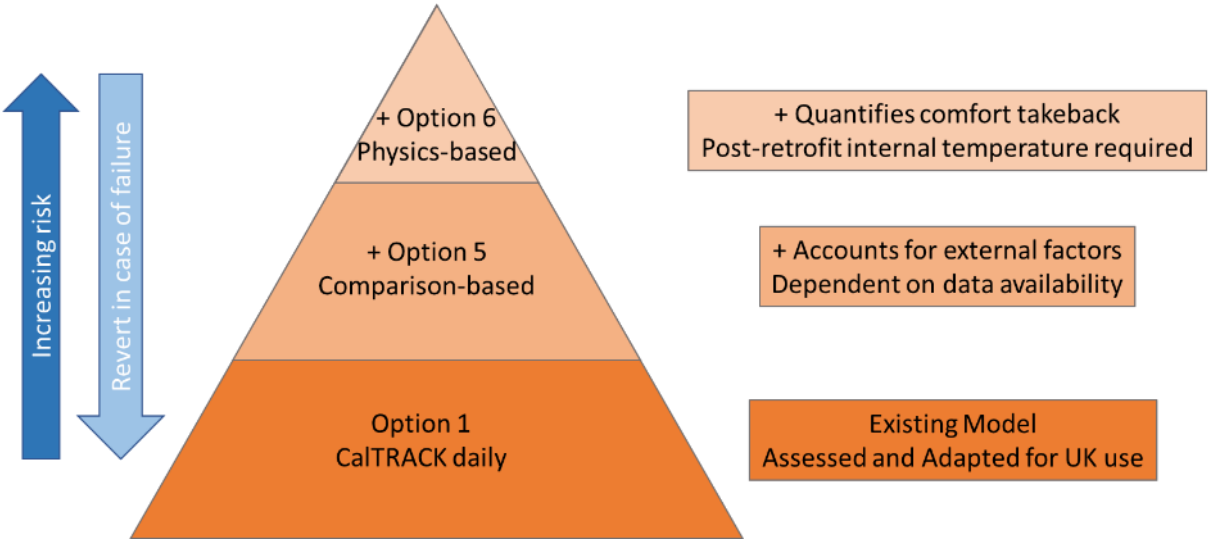
The existing open-source methodology CalTRACK is broadly effective, but because it is dependent on historical energy usage from the same home it fails to account for external changes (e.g. energy price changes and Covid-19) which have been shown to be very important.

Comparison-based methodologies (e.g. GRIDMeter) account for external changes by comparing energy usage to similar homes. However, this is dependent on the availability of lots of smart meter data from homes across the UK to match to, which is not currently available (but may be in future).

Neither of the above approaches account for comfort take-back, which could be a significant factor in overall energy savings, and therefore business models. Direct quantification of comfort take-back could be achieved by combining the above approaches with a physics-based approach. This would require measuring the pre-retrofit heat loss of the property (Heat Transfer Coefficient, HTC) using either smart meter data (for which an algorithm would need to be developed) or a commercial HTC measurement solution (which would eliminate the need for smart meter data). The post-retrofit internal and external temperature measurements would be combined with the pre-retrofit HTC to calculate the counterfactual energy usage given post-retrofit comfort levels. This could be compared to the CalTRACK or comparison-based savings values (which don't account for comfort take-back) to quantify comfort take-back directly.

Method development in Alpha will therefore consist of a layered approach consisting of three components:

1. CalTRACK daily
2. A comparison-based, difference-in-differences approach
3. A physics-based approach that takes post-retrofit internal temperatures and models the pre-retrofit energy usage using an estimated heat transfer coefficient.



This approach mitigates the risks around development and deployment of the comparison-based and physics-based methodologies, whilst hopefully delivering a flexible, scalable approach to metered energy savings that provides the option to account for both external factors and comfort take-back.

Accuracy of these approaches will be assessed using the industry standard metrics (CVRMSE and NBME), with a focus on daily accuracy for individual properties. Alpha phase will also quantify how different sizes of portfolio affect accuracy and uncertainty.

2. Use case summary

Here we give a brief summary of the five use cases outlined in WP1. These cover retrofits of homes with different heating systems and whether they pay based on a Time of Use (ToU) tariff.

Use case 1 – A gas heated home undergoing fabric retrofit only. Gas billing is not ToU sensitive therefore model considers savings on a daily or even monthly basis. A counterfactual is required for gas usage (heating and non-heating) but not electricity.

Use case 2 – Electrically heated home undergoing fabric retrofit only without ToU tariff. Again, savings can be modelled daily or monthly. Counterfactual for electricity heating and non-heating is required, but not for gas use (if any).

Use case 3 – Electrically heated home with ToU tariff, fabric retrofit only. Savings must be modelled on sub-daily basis to account for ToU. Counterfactual for electricity heating and non-heating required.

Use case 4 – Gas heated home, fabric upgrades plus fuel switch to electricity without ToU tariff. Daily or month model is sufficient. Gas heating and non-heating counterfactual required. Electric non-heating counterfactual or post-retrofit sub-metering of electric heating required.

Use case 5 – Gas heated home, fabric upgrades plus fuel switch to electric with ToU. Gas heating and non-heating model required at daily or monthly level. Non-heating electric model at sub-daily level or post-retrofit sub-monitoring of electric heating.

3. WP2 – D1 Methods for comfort take-back, emissions and bill savings

The goal of metering energy savings is typically to measure not just the energy savings (in kWh), but the financial (bill) savings and the carbon emission savings. The methodology to calculate savings will vary depending on the retrofit use case (These are explained in detail in WP1 findings).

3.1 Bill savings

The metered energy savings (MES) counterfactual would output Avoided Energy Use (AEU) measured in kWh. At a high-level, this avoided consumption, offset by any potential consumption resulting from fuel-switching (for use cases 4 and 5), could be multiplied by the gas or electricity unit price. If a retrofit led to all gas sources within the property being electrified (potentially use cases 4 and 5 if cooking was also electrified), the modelling should also consider the savings made from removing the gas standing charges from householders' bills.

In scenarios where time of use (ToU) pricing is relevant (use cases 3 and 5), peak and off-peak pricing would need to be considered separately. However, if half-hourly or hourly counterfactuals couldn't reliably be produced, it may be possible to aggregate multiple half-hours or hours in line with the tariff periods to avoid the "double penalty" and "spikiness" issues with electricity consumption data identified by previous MES work.

Another consideration for the pay for performance (P4P) model in terms of bill savings would be the fluctuation in household energy prices, currently driven by Ofgem's quarterly energy price cap changes. As retrofit projects may take several months to complete and could have longer lead times, e.g. if additional data collection was required, energy prices could fluctuate quite significantly between the initial design of the measures and associated financial products and the time at which energy savings are being realised. For example, between summer 2022 and winter 2023, Ofgem's household energy price cap more than doubled (not taking into account the temporary Energy Price Guarantee).

Note that this is not a source of uncertainty in metering energy savings (as the metering is backwards looking so prices are always known), but is likely important for a business model that is built on financial savings from retrofit. It will be necessary to better understand the cost distribution and incentive payments proposed under the P4P business model to evaluate the risk distribution associated with the proposed business models, which could be significant.

3.2 Emission Savings

Calculation of emissions savings can be done by converting from calculated energy savings. Conversion factors will depends on fuel type and time of use can also be factored

in if energy savings are calculated on sufficiently short time scales (otherwise average conversion factors can be used).

Use case 1: This covers a home where gas heating is used both before and after retrofit and therefore electricity consumption is expected to remain stable (and so no electricity counterfactual needs to be modelled). Emissions savings can be calculated using a constant conversion factor for the m^3 or kWh-equivalent of gas saved, for example as provided by the Department for Energy Security and Net Zero.

Use cases 2 and 3: This covers cases where electrical heating is already in use in the property and heating remains electric after retrofit. As discussed in previous work on Metered Energy Savings (MES), conversion factors for electricity can be obtained via the Electricity System Operator's API on a more frequent basis, provided that a ToU model is being developed. For daily or monthly counterfactuals, a standard conversion factor could also be used.

Use cases 4 and 5: This covers retrofits where heating is switched from gas to electric over the retrofit. Avoided emissions from avoided gas use should be calculated in the same way as for use case 1. However, the emissions from electricity consumed by the replacement electric heating system, most likely a heat pump, would need to be added to the calculation as gas emissions reductions will partly be offset by emissions generated by the new heating system. Sensors to submeter the heat pump-specific electricity consumption would help to accurately estimate carbon intensity of the heat pumps ToU-specific electricity usage.

Note that if there is embedded generation (e.g. solar PV) then that needs to be submetered and compared to the time of use of the heat pump. If the property includes a battery then emissions estimations become significantly more complex and that is considered outside the scope of this project.

3.3 Comfort / Temperature Takeback

Whilst energy savings from retrofit interventions are usually modelled theoretically based on the existing building fabric and proposed interventions, it has been repeatedly observed and studied that actual, realised energy savings from retrofits fall short of the modelled predictions. This can have several reasons – a major one studied being "comfort take-back" (CTB) or "temperature take-back". Specifically, CTB refers to the situation in which heating demand remains higher than modelled because the building occupants increase the internal target temperature of their property rather than realising the full potential energy cost saving. More generally, this observation, in relation to various energy efficiency measures, not just retrofits, is also called the "rebound effect": where increases in energy efficiency should theoretically lead to reductions in energy demand, these are partially offset as the reduced energy cost resulting from the efficiency improvements allows the consumer to increase their energy demand. Previous research (Galvin and Sunikka-Blank, 2016) has found that the CTB for heating is commonly estimated to lie

between 10%-35% whereas research from the Department of Energy and Climate Change (DECC) (Department of Energy and Climate Change, 2014) assumes 15% CTB. However, the same research also identified a "prebound effect" whereby CTB might be exacerbated further because occupants of inefficient homes reduced their energy consumption below calculated comfort levels prior to the retrofit. According to their calculations on German datasets, this "under-consumption" prior to retrofits was on average 35% meaning that estimates of the impact of thermal retrofits could be exaggerated by 50% or more. Given the recent energy bills crisis, the pre-bound effect could be increased even further although peer-reviewed analysis of its impact on people's heating behaviour has not yet emerged.

Further work has highlighted that the CTB has frequently been studied in the context of social housing (Coyne and Denny, 2021). However, they point out that it might be exacerbated in this context due to higher levels of fuel poverty. According to other work, owner occupiers, might experience lower levels of CTB due to higher levels of wealth which would allow them to heat their homes to an appropriate standard regardless of cost: their research showed CTB at 26.7% for homeowners compared to 41.3% for tenants with household wealth and income being major contributing factors to the observed heterogeneity in CTB (Aydin, Kok and Brounen, 2017). Conversely, other research found that social housing tenants may be less affected by CTB if they were previously able to appropriately heat their home and a "saturation effect" takes place (Calderón and Beltrán, 2018). However, these findings were based on an individual block of flats only and state that individual building properties such as the location of heat pipes could influence these results for different buildings.

Lastly, the time at which the MES takes place should also be considered carefully: in a large panel study of 55,154 homes in the UK using data from the National Energy Efficiency Data-Framework (NEED), research has found that the effects of energy efficiency measures rebound only after a certain time, for example after two to four years for cavity wall insulation and after one to two years following loft insulation (Peñasco and Anadón, 2023). This means that MES that measure the energy consumption immediately after the retrofit may be less affected by CTB; however, longer-term monitoring of the retrofit effects may still be warranted to ascertain whether effects other than installation quality are reducing the impact of the measures. Furthermore, the authors recommend that financial incentives to households benefitting from retrofits could help reduce medium- to long-term CTB.

Based on the literature reviewed, we would expect that the amount of comfort take-back will vary depending on the type of retrofits, dwellings and ownerships chosen for the Pay for Performance (P4P) product. When designing and targeting P4P products there may be benefit to estimating likely scale of comfort take-back for individual homes in advance. Homes likely to have greater comfort take-back may have higher uncertainty on the financial savings, which might need to be considered in the design of the financial product.

Broadly speaking, comfort take-back could either be estimated (or explicitly ignored as is done by CalTRACK) and presented alongside the MES or integrated into the modelling. The latter approach would require additional data collection to be viable. The approach taken to either estimating or modelling comfort take-back will depend on the delivery

model chosen for P4P, e.g. the market segments and outcomes it is targeting. Note that while comfort take-back may be 'bad' from a financial perspective, it can often be a 'good' outcome from a broader perspective – particularly where underheating was occurring. Different approaches for accounting for comfort take-back have been considered and are summarised in the table below.

Approach Option	Pros	Cons	Potential application to delivery models
Assume comfort take-back is negligible	• Approach taken by gold-standard models • No requirement to collect additional data	• Likely to work better for commercial property where comfort levels are likely to be more stable pre- and post-retrofit	• Could be applied to delivery models where comfort take-back is expected to be lower, e.g. privately owned property where saturation effect might be taking place or housing where pre-retrofit temperatures were close to standard comfort temperatures (Delivery Model Option 4, see WP3.D8)
Assume comfort take-back does not occur immediately but that behaviour changes take time to settle in	• No requirement to collect additional data	• Method may be dependent on timing of retrofit • Time period over which behaviour changes are likely to take place would need to be assumed / researched	• As above (Delivery Model Option 4)
Estimate comfort take-back (outside of MES models)	• Considerable amount of literature investigating the different levels of comfort take-back • Flexible approach that could be adjusted to specific retrofit project requirements	• No ability to verify whether assumptions match reality	• Could be applied to delivery models where portfolio of properties is targeted, and extreme behaviours could be smoothed out (Delivery Model Option 4)
Estimate comfort take-back and include in MES models	• As above • Assumptions could be applied more consistently rather than	• No ability to verify whether assumptions match reality	• Could be applied to delivery models where portfolio of properties is targeted, and extreme behaviours

	based on individual project team's assumptions	• Assumptions hidden in model would be less transparent to "customer"	could be smoothed out (Delivery Model Option 4)
Survey household comfort levels pre- and post-retrofit (either to include in modelling or for estimation as part of the P4P product)	• Less intrusive than methods that require detailed data collection • Shorter lead-time to start project	• Delivery of surveys (and ensuring they collect good quality data) takes significant effort and may not be practical for many schemes • Could introduce incentive to over-represent internal temperature prior to the retrofit taking place to lower the expectation of comfort take-back and increase the modelled retrofit performance (see below for further exploration)	• Could be applied where installation of additional sensing equipment might be difficult (e.g. Delivery Model Option 7 or 9) • Could be applied where social benefits override performance concerns (Delivery Model Options 1 and 6)
Install detailed monitoring pre- and post-retrofit to measure comfort take-back in terms of temperature changes	• Accurate measurement of comfort take-back	• Increase lead times to allow for data collection • Data collection may be seen as too intrusive by householders • The quality of sensors and the design of the sensor installation could impact the availability and reliability of data captured	• Could be applied to individual home retrofits where extremes are likely to be more pronounced or where social benefits are the overriding concern (Delivery Model Option 1 and 6) • Some platforms to monitor condensation and mould risk already exist which may be adopted for this purpose.

		• System would need to be designed to prevent manipulation (e.g. moving of sensors closer to or further away from heat sources) • Sensor installation quality can be variable on larger schemes introducing measurement errors	
Install detailed monitoring pre- and post-retrofit to enable modelling of comfort take-back	• More accurate modelling of comfort take-back would be enabled	• As above • Translation of temperature increases into avoided energy use (AEU) may be non-trivial	Could be applied to individual home retrofits where extremes are likely to be more pronounced

The following consequences might help guard against incorrectly measuring or estimating the comfort take-back (either based on assumptions or surveys):

- % comfort take-back measured/assumed is too low (for example, assumed 15% but actual was 30%):
    - Avoided Energy Use (AEU) will be over-estimated
    - Performance of retrofit will perform worse than modelled
    - Assuming that social benefits are not relevant to P4P, P4P might pay out more than it should
- % comfort take-back is too high, e.g. due to "saturation effect":
    - Avoided Energy Use (AEU) will be under-estimated
    - Performance of retrofit will perform better than modelled
    - Assuming that social benefits are not relevant to P4P, P4P might pay out less than it should

Due to these consequences, if a survey-based or assumption-based approach was chosen, there might be a perverse incentive to underestimate the comfort take-back to increase the P4P "return".

Taking measurements of internal temperature, as conducted in previous research, would provide the most accurate measurement of comfort take-back (Calderón and Beltrán, 2018). However, it is unlikely to be feasible to install measurement equipment in properties one year prior to the retrofit.

Surveys of temperature set points prior to, and following, the retrofit could be used as a proxy, but they might be inaccurate as temperature set points and achieved target temperatures could vary and households might intentionally misrepresent their current energy usage.

If there is no data available, comfort take-back could be estimated using some of the commonly identified values, e.g. 15%, (as discusses above) - this approach might be suitable if a portfolio approach to retrofits was chosen where extreme observations would likely be smoothened out across multiple properties.

Alternatively, comfort take-back can be estimated by comparing a physics-based methodology that applies post-retrofit temperatures to pre-retrofit fabric with a conventional metered energy savings methodology (see section 4).

4. WP2 – D2 Evaluation Metrics

4.1 Evaluation metrics scope

During the Alpha Phase, all developed Metered Energy Savings (MES) counterfactual algorithms will be evaluated against thresholds previously set to establish whether the results achieve sufficient accuracy in predicting household energy consumption. Only if the MES counterfactuals achieve sufficiently accurate predictions, can they be used in the Beta Phase to calculate avoided energy use (AEU) on homes which have been retrofitted; if the forecasting accuracy was too low, it would not be possible to establish conclusively whether changes in energy demand after the retrofit were due to the retrofit.

In the Beta Phase of the project, AEU calculated from the developed models should also be compared to the energy savings recorded in the "National Energy Efficiency Database" (NEED): the latest NEED report on domestic properties provides the mean and median gas and electricity savings for various measures installed in 2018 (savings were metered in 2019). The figures presented in the report do not account for comfort take-back); hence, the impact of measures may appear to be lower than retrofit designs would assume. Detailed data about retrofitted dwellings is also available that would allow further disaggregation of the findings by administrative region, property age group, floor area, council tax band and index of multiple deprivation[1,2].

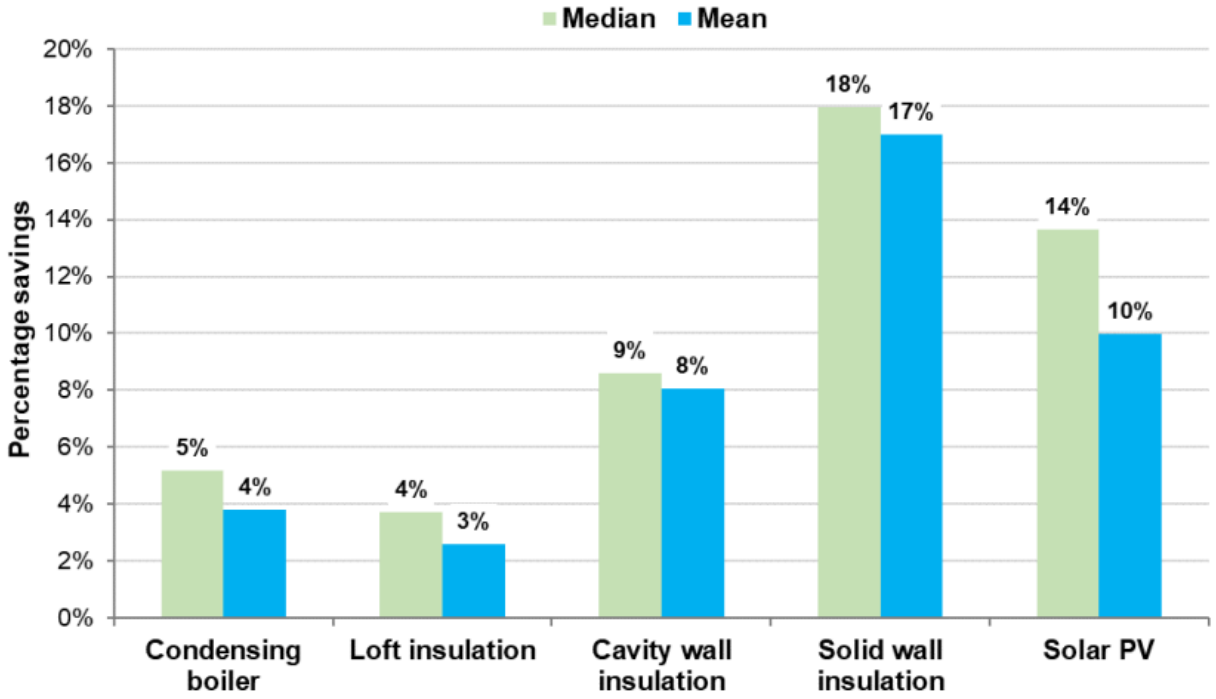


Figure 1: Median and mean energy savings (gas and electricity) from individual retrofit measures implemented in 2018.

[1] For properties in England only: Index of multiple deprivation, England (2019) quintile based on the Lower Layer Super Output Area (LSOA) the property is in.
[2] Gas consumption data is weather corrected.

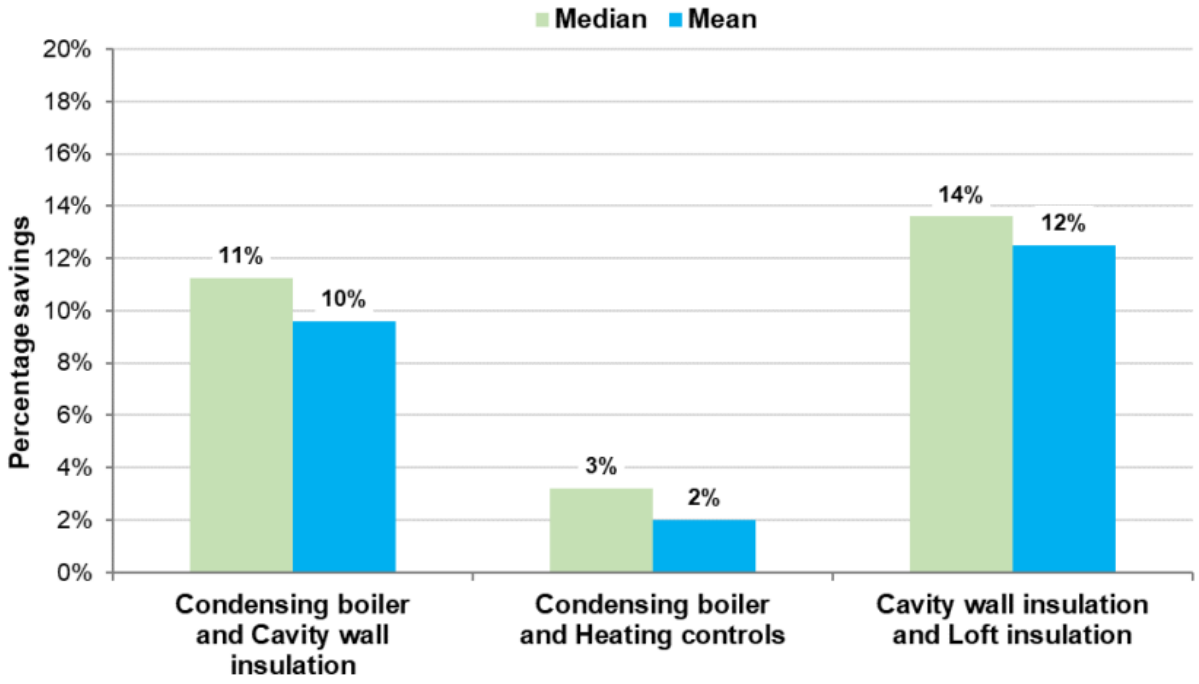


Figure 2: Median and mean energy savings (gas and electricity) from combined retrofit measures implemented in 2018.

There are two other key elements to consider when determining evaluation scope: time resolution and individual property/portfolios. There is a trade-off between these – if savings need to be attributed to individual properties, then a less granular time resolution can be achieved.

4.1.1 Time resolution

Metered energy savings are typically assessed at one of three time resolutions: hourly, daily or monthly.

Hourly (or half-hourly) assessment is typically only required where energy prices vary significantly by time of use and the retrofit is expected to significantly alter when energy is used (e.g. installing a heat pump). In the UK electricity tariffs may have different rates for each half hour (not just each hour), so a half-hourly resolution may be required. Gas tariffs do not vary based on time of use, so this resolution is not required if only gas is being modelled. Achieving a high accuracy with hourly methods is extremely difficult due to variations in occupant behaviour from day-to-day. For example, if the time the evening meal is cooked varies between 5pm and 7pm, the models will struggle to accurately predict that for a specific day.

Monthly assessment is typically used when only monthly bills are available, or when overall magnitude of cumulative savings is the key concern.

Daily assessment evaluates the accuracy of savings on a daily basis, which provides significantly more data points than monthly, and therefore allows a better understanding of confidence in savings, as well as more insight into factors affecting savings (e.g. weather).

For Alpha phase, we will focus on assessing accuracy at a daily resolution.

4.1.2 Individual property vs portfolio

Depending on the business model, metered energy savings can be assessed for a single property or group of properties that had interventions at a similar time.

It is easier to get an accurate assessment of savings on a group of properties because the randomness from individual behaviours in the different properties tends to cancel out, making the effect of the savings clearer.

However, in applications where the value case involves savings being tied to individual properties (e.g. the occupant has a financial product tied to savings in their home) then accuracy on individual properties must be evaluated.

For Alpha phase we will focus on assessing accuracy on individual properties.

4.1.3 Sources of Uncertainty

The Efficiency Valuation Organization (EVO) identifies six sources of uncertainty in the energy savings Measurement and Verification (M&V) process:

Firstly, instrumentation error resulting from inaccurate measurement devices. Whilst this should be relatively small for fiscal gas and electricity meters, low-cost sensor solutions to measure internal temperature and occupancy might suffer from lower accuracy.

Secondly, modelling error introduced by the chosen modelling approach. Commonly used evaluation metrics for establishing the size of the modelling error are summarised in the next section.

Thirdly, sampling error can stem both from limitations on the scope of instrumentation used for M&V (e.g. whole-property vs individual appliance measurements) as well as the temporal resolution of the data collected that might lead to a misrepresentation of the actual energy used.

The remaining three sources of uncertainty are, estimated values, interactive effects that cannot be measured and challenges related to data collection and analysis. However, EVO suggests that these latter three causes of uncertainty should be comparatively small in comparison to the other sources of uncertainty.

Generally speaking, there is a trade-off between M&V rigor and accuracy with the Efficiency Valuation Organization recommending that M&V costs should not exceed 10% of the expected energy cost savings to represent value for money.

4.2 Evaluation Metrics and Justification

4.2.1 CalTRACK and SENSEI methods

The previous evaluation of the SENSEI and CalTRACK methods for calculating metered energy savings (MES) used the Normalised Mean Bias Error (NMBE) and the Coefficient of Variation of the Root-Mean-Square Error (CV(RMSE)) as evaluation metrics which are also recommended by ASHRAE (American Society of Heating, Refrigerating and Air-Conditioning Engineers).

The NMBE measures whether the model consistently over- or underpredicts energy consumption, i.e. whether the model is biased overall. However, as over- and

underprediction errors cancel each other out, it alone cannot be used as a measure of model accuracy. The NBME is defined as:

$$NBME = \frac{\sum(\hat{y}_i - y_i)}{(n-1)} / \bar{y}$$

where $\hat{y}_i$ is the predicted energy consumption at time step i, y_i is the observed value, and $\bar{y}$ is the mean observed usage, and n is the number of time intervals in the data.

The CV(RMSE) is a measure of the size of errors regardless of whether the model is over- or underpredicting consumption. As prediction errors are squared, the direction of prediction error is ignored, and larger prediction errors are penalized more than small errors. The CV(RMSE) normalises the prediction error by the mean observed energy usage and is defined as:

$$CV(RMSE) = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{(n-1)}} / \bar{y}$$

In Alpha we will evaluate point-based methods using NMBE and CV(RMSE). Doing so will provide the benefit that the developed methods will be easily comparable to other literature in this field.

4.3 Physics-based models

Physics-based models, such as the SMETER methodology, calculate the heat transfer coefficient (HTC) of a property, rather than the heating energy consumption (Allinson et al., 2022). However, establishing the HTC through measurements such as a co-heating test is expensive and impractical, and cannot be conducted on scale. As such, it is much more difficult to compare the calculated HTC against a real-world measured value. Large datasets with measured HTC are not easily available.

However, there are two additional ways of measuring the accuracy of HTC estimates when combined with models to convert those estimates into energy use. The first is to evaluate the accuracy of the end-to-end metered energy solution in exactly the same way as above. This makes it harder to identify which element of the modelling is driving errors (e.g. the HTC calculation or the HTC plus temperature to energy usage conversion), but is ultimately likely a suitable approach for this use case and will be used during Alpha.

This can be supplemented by measuring the accuracy of hour-by-hour internal temperature prediction within a home based on observed energy use and HTC. This provides greater insight into where errors might be arising, and so if the internal temperature data are available this approach will be used during Alpha.

4.4 Accuracy thresholds

All candidate MES algorithms should be evaluated against an appropriate threshold that defines whether the algorithm's performance is sufficient to allow usage of the model within a P4P product. Where possible, these evaluation thresholds should be based upon existing methods or findings related to MES to allow comparison of results to established best practices.

For the NMBE and CV(RMSE) evaluation metrics, ASHRAE stipulates an NMBE of 5% and a CV(RMSE) of 15% relative to monthly calibration (baseline) data. If hourly calibration data are used, these requirements shall be 10% and 30% respectively. The guideline does not state any thresholds for models using daily calibration data; however, a threshold between these two values may be appropriate.

In practice the required accuracy often depends on the size of the expected retrofit savings. For a model to be considered valid, the IPMVP standard from the Efficiency Valuation Organization (EVO) also recommends that model accuracy should be considered relative to the level of savings expected from the retrofit: the standard error of the estimate must be <50% of the expected savings at a specified confidence level. This is known as the Fractional Savings Uncertainty (FSU). The confidence level is typically set no lower than 68% although higher confidence levels of 80-90% with lower levels of errors are frequently preferred for M&V analyses.

For commercial property, the Efficiency Valuation Organization (EVO) recommends that retrofits should result in at least 10% of AEU for whole property retrofits where only monthly energy consumption data is available. This is required to accurately distinguish AEU from noise within the data. For retrofits where daily data is available, retrofit options with a predicted saving of 5% or more could also be considered. Based on the NEED data, not all retrofits achieve savings of this level; although the reported savings could be higher if comfort-take-back could be quantified to some extent, e.g. through measurement or surveying of internal temperatures pre-and post-retrofit.

If the business model is based on a portfolio and does not require accuracy on individual properties then more lenient thresholds can be employed because errors in different properties cancel each other out. suggests that a CVRMSE of 100% for individual properties may be suitable for large portfolios as long as portfolio-level thresholds of FSU are respected.

For Alpha, we will look at the proportion of homes meeting different accuracy criteria (e.g. 5% NBME and 15% CVRMSE / 10% NBME and 30% CVRMSE) to understand how suitable the methodology is for different scales of retrofit.

We will also quantify how different sizes of portfolio affect accuracy and uncertainty.

4.5 Other methodological consideration related to model evaluation

"WP1-D1 Data Requirements and Sample Size" discussed that clustering of collected data prior to model development may help in identifying types of properties to which the methodology can be applied with higher levels of certainty. If this suggestion was followed, the evaluation metrics should be disaggregated by this dimension (provided the sample size for each cluster is sufficiently large) to compare which types of properties may be most suitable for a P4P product.

The end-to-end methodology should not only report on model-uncertainty but also other areas of uncertainty, e.g. from incomplete data or measurements errors or missing input variables (e.g. for the calculation of comfort take-back).

5. WP2 – D3 Methodology Options Review

5.1 Goals of methodologies / challenges

The fundamental objective of the metered energy savings solution under consideration is "to calculate the financial, carbon, and energy savings that can be attributed to a heating-related retrofit." The term "attributable" here is crucial. There are various factors that can affect energy usage which should not be attributed to the retrofit, and ideally, a metered energy savings solution should account for these. These factors include:
1. Variations in weather conditions, such as a warmer winter.
2. Alterations to household composition, for instance, the birth of a child or a child moving out.
3. Changes in occupancy patterns that are not connected to the retrofit, such as working from home or going on a vacation.
4. Adjustments in energy usage due to variations in financial circumstances or fluctuating energy prices.

A significant consideration is how changes in behaviour post-retrofit that could reasonably be attributed to the retrofit should be included in this solution. Specifically:
5. Changes in heating demand due to the possibility of different budget/comfort trade-offs post-retrofit, commonly referred to as 'comfort take-back', as explained above in section 3.3.

It's clear that isolating the impact of a retrofit from these other factors for an individual home is an incredibly complex problem, requiring a pragmatic approach that delivers a 'good enough' solution for specific applications. Typically, this involves:

- Adjusting for weather, which is relatively simple given the availability of weather data.
- Aggregating results from a large number of properties, with the hope that some of the other changes will cancel each other out.
- Evaluating savings at a less-granular time resolution, such as daily instead of hourly, in the expectation that some of the other changes will even out over time.

The Covid-19 pandemic and the energy crisis have underscored the need to adjust for large-scale changes in behaviour that aren't related to retrofitting. Some approaches now purposefully aim to address this by using comparison groups. Other strategies seek to control for other drivers of change by gathering and integrating larger volumes of data, either over extended periods (12+ months) or through additional data types, such as internal temperature. However, gathering more data poses significant practical challenges when implementing actual retrofit projects.

It's important to note that we are explicitly interested in the impact of heating-related retrofit, so it's crucial to separate out heating energy use. A key question is whether models need to achieve half-hourly resolution to provide accurate estimates of energy cost, for instance, due to a time-of-use tariff.

Thus, when evaluating each approach below, six points should be considered:
1. What data does it require?
2. How does it separate heating from non-heating energy usage?
3. How does it adjust for weather?
4. How does it account for changes to heating demand potentially related to the retrofit (comfort take-back)?
5. How does it account for changes to heating demand resulting from macro-scale external factors like energy price fluctuations?
6. How does it account for changes in occupancy patterns that are unrelated to both the retrofit and macro-scale external factors?

5.2 Approaches

In this section, we're going to delve into the currently used strategies to evaluate the effects of energy efficiency initiatives. The insights garnered will be pertinent to the RetroMeter project, and we discuss the potential benefits and drawbacks of integrating similar approaches within our project. Furthermore, we highlight possible enhancements to each method and specify the data that would be necessary for their implementation.

5.2.1 CalTRACK 2.0

CalTRACK refers to a set of procedures that allows for the calculation of energy savings, primarily focusing on the estimated energy consumption in a building post-intervention as though the intervention never occurred, this is known as the counterfactual.

CalTRACK's processes are developed under an open-source methods charter, and each method is assigned a number and version for reference. The development of these methods is a collaborative effort hosted on Github, a platform well-suited for open-source projects. These methods are typically required to undergo empirical testing, with the results shared in the CalTRACK Methodological Appendix. Various member organizations contribute to the development and approval of these methods.

CalTRACK is currently on version 2.0 and references to CalTRACK imply CalTRACK 2.0, unless otherwise stated. Note that the community is currently working on an update (CalTRACK 2.1) which may be available by the start of Alpha.

5.2.1.1 CalTRACK 2.0 - Daily method

The CalTRACK daily method relies on a linear model with three energy use regimes: heating, cooling, and baseload, which is fitted to the year's data. The CalTRACK methods attempt to identify external temperatures at which heating and cooling begin to be required in the home (these are known as heating and cooling balance point temperatures). The heating (and cooling) requirements on different days are then accounted for using the difference between external temperature and the balance point temperature and the duration of the difference (this concept is known as heating degree days). This then provides the basis by which the model can adjust energy use for external temperature. This method does not account for comfort take-back, external factors or changes in occupancy patterns.

Necessary data:

- Daily energy usage
- Hourly external temperature

These are required before retrofit for 1 year (for model training) and after retrofit (for energy savings evaluation).

CalTRACK prescribed data sufficiency criteria dictate that the number of days with missing consumption and temperature data should not exceed 37 days (10%) for daily methods.

Potential improvements to CalTRACK daily:

1. Identify groups of homes for which CalTRACK performs better/worse. This would improve the uptake of CalTRACK in UK metered energy savings projects for the homes for which CalTRACK works better and could focus work to improve CalTRACK for the other homes. (This would require significant analysis of lots of home energy data, but not actually an improvement to CalTRACK)
2. Improve heating energy estimates using internal temperature with external temperature to find the true temperature delta. (We expect slightly higher heat loss from homes when they are hotter e.g. 18 vs 21).
3. Use additional weather parameters. (e.g. wind, solar irradiance and rain may impact heating requirements)

5.2.1.2 CalTRACK 2.0 - Hourly method

The CalTRACK hourly method produces hourly estimates of metered energy savings. It fits a different model for each calendar month using data from the corresponding month and previous and following months from the previous year (surrounding months are given half weighting in the fit). It models occupancy using the energy use data on a time-of-week basis with 168 states, represented as either 0 or 1 to indicate occupancy. The model also incorporates binned external temperatures to fit coefficients. Thus, the hourly model serves as a time-of-week temperature (TOWT) model (originally proposed here). This model does not split out energy use for heating and other uses. This method does not account for comfort take-back, external factors or changes in occupancy patterns (between pre- and post-retrofit).

Necessary data:

- Hourly energy usage
- Hourly external temperature

These are required before retrofit (for model training) and after retrofit (for energy savings evaluation). The duration of required data depends on the required prediction duration (if only single months need modelling), but in practice a year of pre-retrofit data is required to estimate savings for a year post-retrofit.

The data sufficiency criteria for the hourly method dictates that no minimum baseline period length is necessary, but baseline consumption data must be available for over 90%

of hours in the same calendar month and in each of the corresponding, previous and following calendar months of the previous year.

Potential improvements to CalTRACK hourly:

1. Does CalTRACK work well on certain types of homes? Identify these groups to improve the uptake of CalTRACK in UK metered energy savings projects. (This would require significant analysis of lots of home energy data)
2. Train two separate models (occupied and unoccupied), and then use the same approach to estimating occupancy that is applied pre-retrofit to flag occupancy post-retrofit in order to select which model to use for each hour.
3. Improve estimated occupancy during pre-retrofit period (e.g. with internal temperature data or other sensor/survey data).
4. Disaggregating heating before using CalTRACK hourly model. Could be done programmatically or with sub-meter. (Unclear how/if this would improve accuracy of the model, we would then have two models to train, baseload and heating).
5. Adapt it to work on half-hourly data. (If required for increased granularity of cost calculations)

5.2.1.3 Implementation and usage

OpenEEMeter is an open-source implementation of the CalTRACK framework. OpenEEMeter is currently on CalTRACK v2.0, which was assessed for use in UK homes during the Metered Energy Savings project. Carbon Coop has submitted changes to CalTRACK which have been accepted under OpenEEMeter CalTRACK v2.1. This includes improvements to accuracy of the daily model using separate winter/summer/shoulder season models with an additional split between weekday and weekend. This also includes working with temperatures in Celsius and improvements to weather data access outside the US.

Evaluation of CalTRACK/OpenEEMeter on UK homes as part of the previous Metered Energy Savings project demonstrated acceptable performance for aggregations of properties (>15) but poor performance for individual properties, particularly for hourly methods.

Advantages:

1. CalTRACK methods are already in use in the US and have a proven business case there.
2. Open implementations of CalTRACK methods are available, so would make a good starting point for modifications and improvements.
3. Designed to meet guidelines established by The American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE Guideline 14) and the Uniform Methods Project (Chapter 8 - Whole Building Methods). Compliance with whole site International Performance Measurement and Verification Protocol (IPMVP) requirements.
4. The methods are subject to ongoing review and improvement by an international technical working group.

Limitations

1. CalTRACK methods do not account for external factors like energy price changes. (See Comparator section for methods to overcome this)
2. CalTRACK methods do not account for comfort take-back. (See Physics-based section for methods to overcome this)
3. CalTRACK hourly methods do not perform sufficiently accurately for individual homes

5.2.2 Comparator

Comparator models aim to address a key issue in metered energy savings: attributing changes in energy use to interventions rather than external factors. These models compare a home that has undergone an intervention to similar homes without interventions, providing a more effective baseline for external factors than comparing pre- and post-intervention energy use within the same home.

A comparator model matches a home either to a single best comparison home or to an aggregate of multiple homes. Using a single home is simpler but leaves the method vulnerable to changes impacting only the comparison home, and it also has increased data privacy challenges. Aggregating data from multiple homes avoids these issues but may result in artificially smoothed energy usage data and require access to more data.

To implement a comparator model, data is required from both participating homes and comparison homes (where no energy efficiency measures will be implemented during the trial). The data includes energy data, weather data, and home attribute data (used to segment and match homes). Various methods can be employed to predict energy use. Energy use of comparison homes in the same category can be used directly as the counterfactual energy use of the participating home, or a modelled counterfactual can be produced for both participating and comparison homes, adjusting the counterfactual for the participating home based on the error seen in matched comparison homes. This latter method is known as the difference-of-differences method.

The main downsides of comparison methods are the need for large datasets for comparison groups and the question of incentivizing comparison groups to provide their data. GRIDmeter is a method used in the US that combines CalTRACK methods and comparison groups to model energy savings and account for potential external impacts on energy use. It uses the comparison group data to calculate an additional counterfactual which is then used in a difference of difference method. Any changes in energy usage across the comparison group could then be attributed to external factors, allowing for a correction to the energy savings calculations for participants who underwent an intervention.

Using GRIDmeter, the data requirements start from the same point as CalTRACK, but for many homes. Data is needed for sufficient homes to find a similar single or group of

homes. We expect there will be categorical data required about the homes, for example building age, size and occupancy, which will be used to group/match the homes. This model would not explicitly disaggregate heating from other energy use, although if CalTRACK daily were used, this would be done internally to the model. Adjustments for weather are as described for the CalTRACK models and would not account for comfort take-back. The comparison between similar homes would account for external factors but not for changes in occupancy.

According to Recurve, portfolio and comparison group sizing are crucial for reducing uncertainty. They suggest that using a sample size of 3,000 non-participant buildings results in a mean error of 0.3-0.7% and a mean absolute percentage error (MAPE) of 1.4-1.5%. The stability of comparison groups is also an important consideration, as all groups will eventually diverge due to factors outside the trial's control (for example unrelated retrofits may occur in the comparison group).

Matching of homes is a very important consideration for these methods. Homes can be matched using various features, such as energy use, external temperature, and internal temperature. Different strategies can be employed to find the best match within the sample population, including evaluating carefully chosen metrics for the participating home and finding the N comparison homes that are “closest” or within a certain threshold “distance”. Other methods involve segmentation of homes into groups using certain features and then matching homes within those groups using other features.

The Pacific Gas and Electric Company conducted a study that investigated various methods of modelling energy savings, with a focus on comparator methods integrated with existing CalTRACK methods or entirely new methods. They recognized the need for comparator methods due to population-wide impacts, such as the COVID-19 pandemic, that investigated various methods of modelling energy savings, with a focus on comparator methods integrated with existing CalTRACK methods or entirely new methods. They recognized the need for comparator methods due to population-wide impacts, such as the COVID-19 pandemic.

The key findings from the study were as follows:

1. Existing CalTRACK methods cannot account for the effects of the COVID-19 pandemic.

2. Existing CalTRACK methods show upward bias even before the pandemic.

3. Comparison groups improve the accuracy of the CalTRACK method.

4. The choice of segmentation and matching characteristics matter more than the method of matching customers.

5. Synthetic controls may perform well but are subject to extreme bias in some cases.

6. Using aggregated granular profiles in the CalTRACK Difference-in-Differences approach yields comparable results to using individual customer matched controls.

7. Accuracy and precision depend on the number of sites aggregated together.

8. No method is completely free of error.

Based on these findings, the study recommends the use of a framework to test, certify, and estimate savings for NMEC methods. Certification criteria for accuracy and precision should be based on sample size. The best-performing models were found to include CalTRACK models with control groups and their alternative Pre-Post models with control groups. Furthermore, the study revealed that Euclidean distance matching outperformed propensity score matching and stratified random sampling. However, it was noted that the choice of matching segments played a more significant role in producing a good matching group.

Advantages:

1. Accounts for population wide changes in energy usage (and therefore improves accuracy where/when these have taken place).
2. Builds on CalTrack method. Any improvements made there will filter through and improve this method.

Limitations:

1. Requires very large dataset of non-participating homes including energy usage and contextual data. This currently doesn't exist in a way that would allow wide adoption.
2. Does not account for comfort take-back.
3. Does not account for changes in occupancy.
4. Previous trials of this method have used larger datasets that we will have available to us in the Alpha phase.

5.2.3 Physics-informed model

The power required to heat a home to a fixed temperature can most simply be modelled by the following equation:

Power = (Internal Temperature - External Temperature) * Heat Transfer Coefficient

The Heat Transfer Coefficient (also known as the Heat Loss Coefficient) represents the rate at which heat is lost from the home. This can be used in estimating energy savings.

A key limitation of the CalTRACK approach is that it assumes occupant behaviours (particularly heating patterns) do not change after the retrofit. The counterfactual to which actual energy usage is compared represents energy use for pre-retrofit behaviours occurring in the post-retrofit property with post-retrofit weather. Comfort take-back is therefore ignored. The counterfactual corresponds to "How much would it have cost to heat the home the way I used to when I hadn't had the retrofit?"

See an example in the table below:

	Pre-retrofit	Post-retrofit	Counterfactual
External Temperature	9°C	**11°C**	11°C
Internal Temperature	**18°C**	19°C	18°C
Fabric Heat Loss	**200 W/K**	100 W/K	200 W/K
Heating Power Usage	1800W	800W	1400W (savings: 600W)

An alternative would be to create a counterfactual that consists of post-retrofit behaviours applied to the pre-retrofit property under post-retrofit weather. This intrinsically corrects for any changes to behaviour (e.g. comfort take-back). The counterfactual corresponds to: "How much would it have cost to heat the home the way I currently do if I hadn't had the retrofit?" (see example below).

	Pre-retrofit	Post-retrofit	Counterfactual
External Temperature	9°C	**11°C**	11°C
Internal Temperature	18°C	**19°C**	19°C
Fabric Heat Loss	**200 W/K**	100 W/K	200 W/K
Heating Power Usage	1800W	800W	1600W (savings: 800W)

This would involve three components:

1. Create a model of pre-retrofit energy usage required to achieve different internal temperatures in different weather conditions (probably by estimating whole-home Heat Transfer Coefficient).

2. Measure post-retrofit heating patterns (via internal temperature monitoring) and weather.

3. Run post-retrofit heating patterns and weather through the model to produce counterfactual energy usage.

Advantages:

1. Solutions exist for estimating whole-home Heat Transfer Coefficient (HTC) from measured data (see SMETER), and these are more mature than direct metered energy savings solutions. They can involve different levels of monitoring – one solution demonstrated good accuracy using only smart meter data and weather, whilst others required internal temperature data, humidity, and more. See Appendix A for more synergies between SMETER and RetroMeter.

2. The end-to-end metered energy solution can be modular, allowing any Heat Transfer Coefficient (HTC) estimation tool to be used within the methodology. This means that

homes with pre-existing internal temperature sensors could benefit from increased accuracy, but the methodology could be used on homes without temperature sensors as well, allowing broad applicability.

3. Direct comparison of pre and post-retrofit HTC can be a part of the retrofit assurance process.

4. Using actual occupancy patterns (based on post-retrofit internal temperature) rather than assumed occupancy may provide a more accurate sub-daily estimate of savings.

5. Some HTC estimation solutions require only short periods of intensive data collection (rather than 12 months), which would allow a metered energy solution to work for homes where collecting 12 months of data pre-retrofit was impractical (e.g. those without smart meters).

6. May account for external population wide changes if these result in changes to the internal temperature readings. (e.g. energy price rises result in reduced heating, meaning measured internal temperature is lower).

Limitations:

1. In the SMETER project, some of the better entries managed typical uncertainty/error in the HTC estimation range between 12-21%. This would propagate through the methodology, leading to at least this level of uncertainty in the estimation of savings.

2. Changes to heating technology as part of the retrofit (e.g. moving from a boiler to a heat pump) will normally change the internal temperature profile of the home due to differences in the way those technologies operate. Requiring one technology to operate under the preferred mode for another technology when estimating the counterfactual might not be a fair comparison. (Boiler controls are often set on/off throughout the day, heat pumps are operated most efficiently when left on constantly with a small setback overnight)

3. The measured internal temperatures do not correspond directly to the target internal temperatures of the occupants, which might lead to overstating the energy required for the alternative technology to achieve the occupant's target temperature. For example, a home may be at 17°C at 5am because it is slowly warming up to reach the target temperature of 18°C at 7am, rather than because it needs to be at 17°C at that time. This is a fundamental limitation of this approach and would need appropriate consideration before employing it in retrofits where the heating technology has changed. Target temperatures could be used rather than actual internal temperatures, but these are much harder to obtain as they require interfacing with different thermostat systems, and they would also make the modelling required significantly more complex, so are not really practical at scale.

5.2.4 Probabilistic

The significant differentiating feature of probabilistic models is their output. Rather than making a single prediction of energy use for each timestep, probabilistic models give predictions of the likelihood that the energy use will be within a certain range. There are many such probabilistic methods but here we will focus on quantile regression.

In a quantile regression the model predicts the quantiles for energy use at each timestep. For example, using ventiles, 20 points would be output for each timestep. These would predict the energy usage for quantiles with 5% probability spacing. Such models are fit in a similar way to linear regression but separate fits must be done for each quantile.

The benefit in using such a quantile regression does not come from improvements in accuracy or precision of the model predictions, but the additional understanding of the model uncertainty that comes from analysis of its outputs. For example, finding periods of higher model uncertainty and quantifying that uncertainty.

One major issue with point modelling of metered energy savings is that of the double penalty effect. This was identified in the previous metered energy savings project as a major contribution to the model error. This method would not overcome this issue but it might allow further analysis of it, through better understanding of model uncertainties.

CalTRACK documentation gives recommended approach for calculating uncertainty for portfolio aggregated studies. This also gives some indication of how this might be done for a single site.
Sensei documentation gives an approach for estimating the uncertainty in the model. It is not clear if either of these models has built in uncertainty calculations.

Data requirements are flexible, but energy and external temperature data are a requirement for both training (1 year prior) and evaluation period. Internal temperature could be used to improve the model and would be required to account for comfort take-back. This method would not explicitly separate the heating from the other energy use. It would not account for external changes or changes in occupancy.

5.3 Comparisons

Summary and comparison of potential approaches is given in tabular form below:

Methodology	Comfort take-back	Time-resolution	Systematic NREs	Individual property NREs	Data requirements	Individual homes	Existing tools	Likely error?	How hard to develop?
CalTRACK daily	No	Daily	No	No	1 year pre-retrofit, daily energy, and hourly external temperature	Maybe	Yes	Existing CalTRACK V2.0 ~ 30%	Pre-existing, fairly simple to improve
CalTRACK hourly	No	Hourly	No	No	Corresponding months pre-and post. Hourly energy and external temperature	Unlikely	Yes	Existing CalTRACK V2.0 ~ 90-100%	Pre-existing, fairly simple to improve
Comparator (GRIDmeter)	No	Daily or hourly (based on CalTRACK)	Yes	No	1 year pre-retrofit, energy and external temperature. Large dataset of non-retrofit homes required.	Maybe	Yes	Should be better than CalTRACK	Pre-existing, improving matching likely to be complex
Physics-informed (SMETER type)	Yes	Daily or hourly	No	No	Energy data, external and internal temperature for a period (~1 winter month) or 1 year pre-retrofit energy and external temperature	Yes	Partial	10-30%	HTC calculation could be pre-existing, but still significant work to integrate
Probabilistic	No	Half-hourly	No	No	Half-hourly energy and weather data for 1 year prior	Maybe	No	Prediction uncertainty quantified by model	Quite hard

5.4 Options to consider

In the table below we outline different methodology options considered, along with their advantages and limitations. All of the options outlined are valid for retrofits where the heating system starts as gas, and either remains gas or is changed to electrical heating with sub-metering of the electrical heating system post-retrofit. They all could produce property-level counterfactual with confidence intervals, but focus will be on portfolio level accuracy (as this is more feasible).

The following section discusses the selected methodologies (highlighted in green) in more detail.

OPTION	COMBINED OPTIONS	DATA	MODEL	ADVANTAGES	LIMITATIONS
1 – CalTRACK daily	No	SERL or Hildebrand or ESCs Living Lab	CalTRACK daily for gas (and electric)	Pre-existing and therefore low risk.	Only works in smart metered homes.
2 – CalTRACK daily + internal temperature	No	ESCs Living Lab	CalTRACK daily with internal-external temperature added to model for gas (and electric)	Comfort take-back via changes in average daily internal temperature.	Only works in smart meter homes. 1 year of internal temperature required pre-retrofit.
3 – CalTRACK hourly + internal temperature	No	ESCs Living Lab	CalTRACK daily for gas and hourly for electric with internal-external temperature added to model	Comfort take-back via changes in internal temperature.	Only works in smart meter homes. 1 year of internal temperature required pre-retrofit.
4 – CalTRACK hourly + observed occupancy	No	SERL or Hildebrand or ESCs Living Lab	CalTRACK daily for gas and hourly for electric. Occupancy post-retrofit used in determining counterfactual.	Occupancy changes accounted for.	Only works in smart meter homes.
5 – Comparison based	No	SERL or Hildebrand	Adapted GRIDmeter, daily for gas, hourly for electric	External factors accounted for by comparison with groups of similar homes.	Only works in smart meter homes. Requires ongoing availability of large quantities of comparison group data.

6 – Physics based, smart meter only	No	SERL or Hildebrand or ESCs Living Lab	Open-source HTC calculation that uses only smart meter and external temperature data, combined with model to map HTC + internal and external temperature to energy usage	Intrinsically corrects for comfort take-back. Can use alternative HTC measurements to avoid need for 1 year smart meter data.	Only works in smart meter homes. Smart meter-based HTC may be harder/less accurate than HTC with internal temperature
7 – Physics based including internal temperature	No	ESCs Living Lab	Pre-existing solutions (or in-house equivalent) used to calculate HTC.	Intrinsically corrects for comfort take-back. Works in homes without 1 year smart meter data.	Requires kit installation pre-retrofit.
8 – Physics based (no internal temp) + CalTRACK variant	Combines options 1 and 6	SERL or Hildebrand or ESCs Living Lab	N/A	Intrinsically corrects for comfort take-back. Could account for occupancy.	See options 1 and 6.
9 – Physics based (no internal temp) + comparison based	Combines options 1, 5 and 6	SERL or Hildebrand	N/A	Intrinsically corrects for comfort take-back. Could account for occupancy. External factors accounted for.	See options 1, 5 and 6.
10 – Physics based (with internal temp) + CalTRACK variant	Combines options 3 and 7	ESCs Living Lab	N/A	Intrinsically corrects for comfort take-back. Could account for occupancy.	See options 3 and 7.
11 – Physics based (with internal temp) + comparison based	Combines options 3, 5 and 7	ESCs Living Lab	N/A	Intrinsically corrects for comfort take-back. Could account for occupancy. External factors accounted for.	See options 3, 5 and 7.

5.5 Selected methodologies

The requirement for 1 year of internal temperature prior to retrofit is impractical in many cases, and would dramatically restrict the applicability of the methodology to real world retrofit programmes. As a result, options that require this have been eliminated from consideration.

The requirement for ongoing availability of large quantities of comparison group data is also a key challenge. This data is not currently available (unlike in the USA), and as such this methodology may not be practical in the short term. However, the comparison-based methodology is probably the best way to account for external effects which, as Covid-19 and the energy crisis have shown, threaten to fundamentally undermine other methodologies.

As a result, we propose taking forward option 9 to Alpha - a combination of options 1, 5 and 6. This is because each option has significant downsides in isolation which are addressed through the combination, but also because the failure of one or two of the options due to insurmountable barriers (e.g. data availability for comparisons) would still leave the project with a viable solution.

These solutions also synergise nicely because they all only require smart meter data plus weather pre-retrofit - i.e. they do not require pre-retrofit internal temperature data. This is an advantage, but it also means they do not work for non-smart meter properties (which make up ~50% of homes). This needs to be considered, and is partially mitigated by the fact that option 6 could work with other existing HTC estimation methodologies that aren't dependent on historical smart meter data.

The below diagram illustrates how the solutions layer together.

Option 1 – CalTRACK daily

This option utilizes the existing CalTRACK daily methodology. As such there is low delivery risks, given that the algorithm and code are already largely in existence. Despite this, the method is not without its limitations, as it is unable to account for external factors or comfort take-back. However, any improvements made to this methodology will also enhance Option 5.

Option 5 – Comparison-based

This option incorporates comparison or matching methods into the existing CalTRACK daily framework. This augmentation allows the system to correct for external factors. However, it would require significant effort to adapt and validate the GridMeter code and methodology for UK homes. The delivery risk for this option is moderate during the Alpha phase, due to potential insufficiencies in the matching methodology given the limited number of homes available (~10k). Consequently, this could lead to this solution performing worse than Option 1, rather than better. During the Beta/rollout phase the delivery risk is high due to the uncertainty around the ongoing availability of data for comparison groups (as the project would no longer be research, access to some datasets may be restricted). Despite this, if successful, Option 5 would replace Option 1, resulting in

a methodology that is more robust to external factors, such as energy price changes. In the event of failure, we would revert to Option 1.

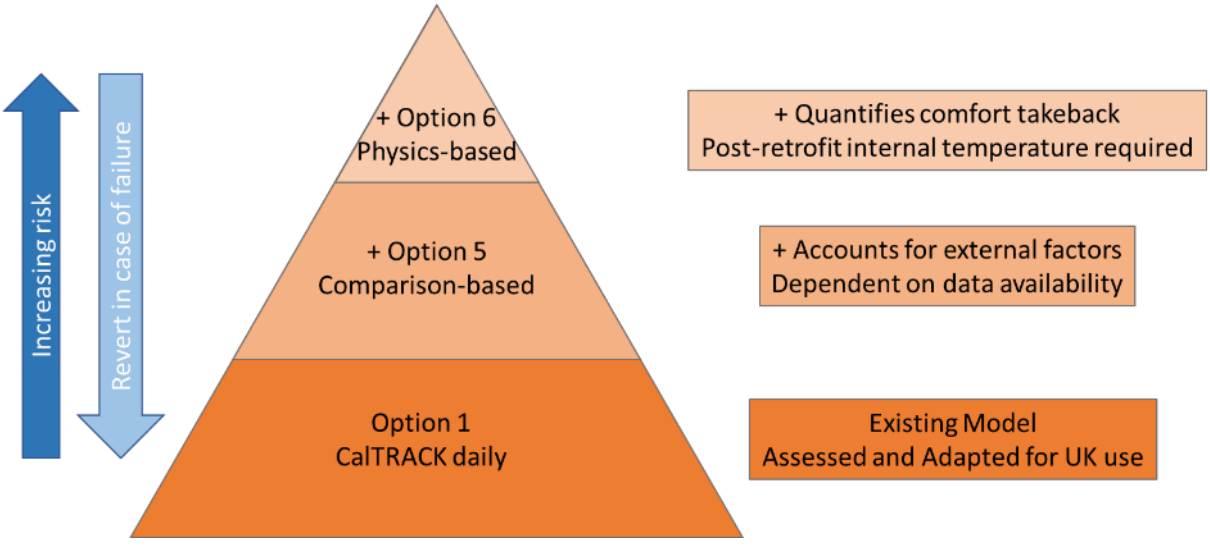
Option 6 – Physics-based, smart meter only

Option 6 offers a physics-based, smart-meter-only solution. This method combines a data-driven estimate of pre-retrofit heat loss (HTC) with post-retrofit internal temperatures to produce a counterfactual that adjusts for comfort take-back. This counterfactual is then compared to the counterfactual from Option 5 (or 1) to split the metered energy savings into 'realised energy savings' and 'energy savings taken as comfort'. However, significant effort is required to develop both the smart meter only HTC calculation and the model to translate HTC plus internal temperature into energy usage. The delivery risk for Alpha is high, as the smart meter only HTC calculation may not perform adequately or may require more effort than is available within the Alpha timelines. For Beta, the risk is moderate as it requires the collection of internal temperature post-retrofit, which could increase the monitoring complexity and fragility of the data pipeline. If successful, Option 6 can be used in combination with Option 5 (or 1) to quantify comfort take-back, or it can serve as a standalone measure when smart meter data does not exist. In the case of failure, we would revert to Option 5 (or 1).

This approach means that at the end of Alpha we can expect to have:

Worst Case: A daily counterfactual methodology that is integrated into an end-to-end data pipeline and works adequately for gas heated homes (or gas to electrical heating retrofits) which have 12 months of pre-retrofit smart meter data – but which cannot account for external factors (e.g. energy price changes).

Best Case: A modular counterfactual methodology that is integrated into an end-to-end data pipeline and works well for gas heated homes (or gas to electrical heating retrofits) which have 12 months of pre-retrofit smart meter data (correcting for external factors and estimating comfort take-back). It will also have the ability to utilise alternative HTC measurement tools to sidestep the need for 12 months of pre-retrofit smart meter data (but the resulting metered energy savings estimates would not separate out comfort take-back).



5.6 Additional Considerations

5.6.1 Data Pipelines

Whilst the focus of the above is on evaluation of the underlying algorithms, the surrounding data engineering and software interface are equally key. Building an end-to-end data product that includes pipelines for retrieving, cleaning and combining smart meter and weather data will be important. We will also need to consider processes for detecting and resolving broken data connections or other monitoring issues.

5.6.2 User Interface and Platform

Successful deployment of this as a final product will require software development to produce an effective platform with user interfaces for registering homes, connecting smart meters and other monitoring devices, reviewing savings on individual homes and portfolios etc.

5.6.3 Model Confidence/Uncertainty

In order to develop robust pay-for-performance business models (as well as individual retrofit assessments), quantification of uncertainty and confidence in model results will be required. The level of scientific rigour required will be dependent on stakeholder needs and business models. Clearly articulated indicators of uncertainty and confidence may be of more value to many stakeholders than more complex (but more robust) assessments. More examination of this will be required in Alpha and Beta phases.

5.6.4 Availability of Comparison Groups

Data privacy means the public availability of smart meter data for individual properties for use as comparators is likely to remain extremely limited in the near to medium term. There are two likely routes to resolving this for Beta phase and afterwards:

1. RetroMeter users submit their smart meter data to a trusted intermediary with access to large quantities of smart meter data who then performs the matching and returns the anonymised counterfactual. This would require significant investment in setting up and running a secure and performant system for this (and leaves the matching in the hands of a single party).
2. A larger number of aggregated counterfactuals (e.g. 20-50 homes) are published openly by an organisation with access to smart meter data (e.g. SERL) to enable public matching to comparator groups. This is likely to be easier to arrange and so is probably to be preferred (as long as it performs similarly). This will be explored further during Alpha.

Bibliography

Aydin, E., Kok, N. and Brounen, D. (2017) 'Energy efficiency and household behavior: the rebound effect in the residential sector', *The RAND Journal of Economics*, 48(3), pp. 749–782. Available at: https://doi.org/10.1111/1756-2171.12190.

Bjerregård, M.B., Møller, J.K. and Madsen, H. (2021) 'An introduction to multivariate probabilistic forecast evaluation', *Energy and AI*, 4, p. 100058. Available at: https://doi.org/10.1016/j.egyai.2021.100058.

Calderón, C. and Beltrán, M.R. (2018) 'Effects of fabric retrofit insulation in a UK high-rise social housing building on temperature take-back', *Energy and Buildings*, 173, pp. 470–488. Available at: https://doi.org/10.1016/j.enbuild.2018.05.046.

Coyne, B. and Denny, E. (2021) 'Retrofit effectiveness: Evidence from a nationwide residential energy efficiency programme', *Energy Policy*, 159, p. 112576. Available at: https://doi.org/10.1016/j.enpol.2021.112576.

Department of Energy and Climate Change (2014) *Estimated impacts of energy and climate change policies on energy prices and bills: 2014 - GOV.UK*. Available at: https://www.gov.uk/government/publications/estimated-impacts-of-energy-and-climate-change-policies-on-energy-prices-and-bills-2014 (Accessed: 19 April 2023).

Galvin, R. and Sunikka-Blank, M. (2016) 'Quantification of (p)rebound effects in retrofit policies – Why does it matter?', *Energy*, 95, pp. 415–424. Available at: https://doi.org/10.1016/j.energy.2015.12.034.

Haben, S., Holderbaum, W. and Voss, M. (2023) *Load forecasting: core concepts and methods with applications in distribution networks*. Springer Nature.

Peñasco, C. and Anadón, L.D. (2023) 'Assessing the effectiveness of energy efficiency measures in the residential sector gas consumption through dynamic treatment effects: Evidence from England and Wales', *Energy Economics*, 117, p. 106435. Available at: https://doi.org/10.1016/j.eneco.2022.106435.

Young, S. *et al.* (2022) *Metered Energy Savings: Unlocking home retrofit financing by reliably measuring energy savings - Energy Systems Catapult*. Available at: https://es.catapult.org.uk/report/metered-energy-savings/ (Accessed: 18 April 2023).

Ziel, F. and Berk, K. (2019) 'Multivariate Forecasting Evaluation: On Sensitive and Strictly Proper Scoring Rules'.

6. Appendix A

Synergies between SMETER and RetroMeter:

RetroMeter potential synergy streams	SMETER
Data Warehouse	• Opens prospect of incorporating thermal performance improvement (pre and post retrofit) as another performance measure into the national database of domestic home efficiency and EE project outputs. This will leverage on data collected from the SMETER sensors used for indoor air temperature, gas and electric demand and relative humidity. • When targeting homes with specific energy improvement assets or offers, the HTC may be a key factor in determining the suitability of specific technologies, such as heat pumps. The "baseline" or "improved" HTC at the time of electrification of heating technologies will also be a key factor for determining the potential load profile changes that arise from fuel switches and energy improvements. • Connecting HTC to qualitative assessments of occupant comfort and the "value" they derive from their residential setting en-masse may also help with the valuation of personal comfort as well as health outcomes.
Standard Methodologies	The HTC is predicted as part of an Energy Performance Certificate (EPC) for new homes (using the SAP method) and for existing homes (using the RdSAP method). There is a potential opportunity to use the CALTrack methodology alongside HTC for existing homes to help to disaggregate changes to heating behaviour and other energy end uses. This can be helpful for modelling disaggregated savings or for estimating the savings "absorbed" by comfort take-backs where additional data is present.
Baselining CalTRACK	• UK Data Archive: 29 houses with gas meters and possible SMETER follow up data (100 homes part of GHG dataset) can be used as data points in improving CalTRACK baseline. • Understanding HTC will improve the correlation of internal air temperature to outdoor air temperature

	(i.e. Heating Degree Days) with the heating component of energy consumption, which could be used to improve CalTRACK baseline.
Improving CalTRACK Calculation	SMETER's shared measured HTCs, dwelling characteristics and ancillary measurement for homes can be used as part of the data being collected to improve the CalTRACK method: • SMETER can help to understand the impact of energy improvements on load profiles. • Although HTC alone cannot inform us of comfort take-backs, when used in conjunction with internal temperatures data / set points it can assist retrofit evaluators in determining the magnitude and timing of savings "offset" by comfort take-backs, and whether these would have been higher or lower had the intervention not occurred. • From discussion: doesn't consider "real" events in the home (comfort take-back, occupant behaviour)

7. Appendix B

This was included in section 3 above, but as we are no longer considering a probabilistic method, it has been removed for clarity but is kept here in an appendix for completeness.

Probabilistic methods

Since behaviour in a home varies significantly on an hourly basis, existing hourly methodologies typically struggle to accurately model energy usage (as found by the MES project(Young *et al.*, 2022)). One option is to develop a new type of counterfactual that leverages probabilistic forecasting methods. Because probabilistic forecasts produce more than one estimate, the evaluation metrics mentioned above would not be applicable.

The probabilistic forecasts seek to accurately represent the probability distribution of the energy consumption at any given time-step, rather than predicting a specific value. However, in the observed data against which the forecast will be evaluated, there is only one true observation available. Therefore, probabilistic forecasts are evaluated based on a scoring function whereby the minimum score can only be achieved by the true distribution, as discussed here(Haben, Holderbaum and Voss, 2023).

Other previous work(Bjerregård, Møller and Madsen, 2021) provides an overview of evaluation methods for multi-variate probabilistic forecasts which remains an active field of research. They evaluated three major scoring functions and came to the following findings:[3]

Advantages and disadvantages			
	LogS	CRPS	VarS
Calibration of mar-ginal distribution Correlation structure	+ Mean and variance are evaluated. + Fully evaluated.	+ Mean and variance are evaluated. ÷ Very poorly evaluated, models can not be separated in practice.	÷ Mean is **not** evaluated. + Indirectly but effectively evaluated.
Run time at the k'th dimension Viability for multivariate problems	÷ Increases exponentially with k. +/÷ Theoretically useful but too computationally demanding at higher dimension.	÷ Increases exponentially with k. ÷ Practically useless and computationally demanding.	+ Increases quadratically with k. +/÷ Partially useful and completely computationally feasible.
Summarized properties			
General characteristics	**LogS** Evaluates all information about the forecast distribution, but penalizes hard in the tails. Theoretically useful for any dimension, but not practically applicable at higher dimension.	**CRPS** Evaluates the overall shape of the forecast distribution including mean and variance, but fails to evaluate correlation in multivariate scenarios. Useful for one dimension, useless for multivariate problems.	**VarS** Evaluates correlation and variance indirectly, but fails to evaluate the mean. Partially useful for evaluation of multivariate forecasts but cannot stand alone.
Numerical methods required for implementation	Estimation of PDF, e.g. using a kernel density estimate, cf. Section 3.1.	1) Estimation of CDF, e.g. using the estimated CDF. 2) n-dimensional integration, e.g. using importance sampling, cf. Section 3.2.	Simple arithmetic operations, cf. Section 3.3.
Computational performance	Fast and accurate for lower dimension, computation time and memory footprint quickly increases with higher dimension.	Fast and accurate for lower dimension, computation time and memory footprint quickly increases with higher dimension.	Very fast for lower dimension, increase in run time and memory footprint is limited.

Due to these findings, they recommend applying multiple rather than just a single scoring rule, specifically to apply VarS to the full, multivariate forecast, while simultaneously evaluating its marginal densities by either univariate CRPS or LogS, depending on whether the shapes of the tails are considered important (LogS) or not (CRPS). Finding from other work (Ziel and Berk, 2019) brings a similar conclusion, recommending that the "energy

[3] Vars: Variogram score of order p
LogS: Logarithmic score
CRPS: Continuous ranked probability score

score" which performed best in their simulation is combined with other scoring functions such as the CRPS or the bivariate copula energy scores to prevent heavy reliance on single evaluation metric only.

Due to the ongoing development of appropriate evaluation methods for probabilistic forecasts it is to be expected that the literature applying probabilistic forecasting to the problem of MES will be extremely limited. Therefore, opportunities to compare such a model's performance to any existing literature will be very limited.

8. Licence/Disclaimer

Energy Systems Catapult (ESC) Limited Licence for RetroMeter – Methods Development
ESC is making this report available under the following conditions. This is intended to make the Information contained in this report available on a similar basis as under the Open Government Licence, but it is not Crown Copyright: it is owned by ESC. Under such licence, ESC is able to make the Information available under the terms of this licence. You are encouraged to Use and re-Use the Information that is available under this ESC licence freely and flexibly, with only a few conditions.

Using information under this ESC licence
Use by You of the Information indicates your acceptance of the terms and conditions below. ESC grants You a licence to Use the Information subject to the conditions below.

You are free to:
- copy, publish, distribute and transmit the Information
- adapt the Information
- exploit the Information commercially and non-commercially, for example, by combining it with other information, or by including it in your own product or application.

You must, where You do any of the above:
- acknowledge the source of the Information by including the following acknowledgement:
- "Information taken from RetroMeter – Methods Development, by Energy Systems Catapult"
- provide a copy of or a link to this licence
- state that the Information contains copyright information licensed under this ESC Licence.
- acquire and maintain all necessary licences from any third party needed to Use the Information.

These are important conditions of this licence and if You fail to comply with them the rights granted to You under this licence, or any similar licence granted by ESC, will end automatically.

Exemptions
This licence only covers the Information and does not cover:
- personal data in the Information
- trademarks of ESC; and
- any other intellectual property rights, including patents, trademarks, and design rights.

Non-endorsement
This licence does not grant You any right to Use the Information in a way that suggests any official status or that ESC endorses You or your Use of the Information.

Non-warranty and liability
The Information is made available for Use without charge. In downloading the Information, You accept the basis on which ESC makes it available. The Information is licensed 'as is' and ESC excludes all representations, warranties, obligations and liabilities in relation to the Information to the maximum extent permitted by law.

ESC is not liable for any errors or omissions in the Information and shall not be liable for any loss, injury or damage of any kind caused by its Use. This exclusion of liability includes, but is not limited to, any direct, indirect, special, incidental, consequential, punitive, or exemplary damages in each case such as loss of revenue, data, anticipated profits, and lost business. ESC does not guarantee the continued supply of the Information.

Governing law
This licence and any dispute or claim arising out of or in connection with it (including any noncontractual claims or disputes) shall be governed by and construed in accordance with the laws of England and Wales and the parties irrevocably submit to the non-exclusive jurisdiction of the English courts.

Definitions
In this licence, the terms below have the following meanings: 'Information' means information protected by copyright or by database right (for example, literary and artistic works, content, data and source code) offered for Use under the terms of this licence. 'ESC' means Energy Systems Catapult Limited, a company incorporated and registered in England and Wales with company number 8705784 whose registered office is at Cannon House, 7th Floor, The Priory Queensway, Birmingham, B4 6BS. 'Use' means doing any act which is restricted by copyright or database right, whether in the original medium or in any other medium, and includes without limitation distributing, copying, adapting, modifying as may be technically necessary to use it in a different mode or format. 'You' means the natural or legal person, or body of persons corporate or incorporate, acquiring rights under this licence.

Energy Systems Catapult
7th Floor, Cannon House
18 Priory Queensway
Birmingham
B4 6BS

es.catapult.org.uk