Department for
Energy Security
& Net Zero

SMETER Validation Methodology

September 2024

Acknowledgements

The work was carried out by Valerie Livina and David Whittaker in the Data Science Department at the National Physical Laboratory (NPL).

We are grateful to Dr Matt Li and Dr Catalina Giraldo Soto (Loughborough University), Dr Frances Hollick (UCL) and SMETER participants for performing blind tests. We are also grateful to Loughborough University colleagues Professor David Allinson and Professor Chris Gorse, as well as UCL colleague Professor Cliff Elwell for their advice on building physics. This work was funded by the UK Government's Department for Energy Security and Net Zero & Department for Science, Innovation & Technology through the UK's National Measurement System programmes.

Any enquiries regarding this publication should be sent to us at:
xxx@energysecurity.gov.uk **[replace with team email address if available, or delete this sentence]**

Contents

Abstract

The report describes a novel validation methodology for heat transfer coefficient (HTC) in built environment. Smart meter enabled thermal efficiency rating (SMETER) technologies aim to estimate HTC values in housing stock. SMETER technologies differ in respect of such factors as the number and temporal range of measurements considered, temporal resolution and uncertainty ranges. To address this complex and diverse backdrop, NPL has developed a surrogate-based methodology with controlled ground truth and workflow, which is supported by a GUI tool for stakeholders. In a series of blind tests, we demonstrate how the surrogate-based validation can be used to compare different SMETER technologies (physics-based, grey-box and black-box machine-learning models). This surrogate-based validation approach affords a technology-agnostic controllable validation tool - subject to further tests on broader data, when available.

1 Executive summary

The UK government aims to assess thermal properties of the domestic built stock. The parameter that characterises the rate of heat loss capacity of built stock (new and retrofit) is heat transfer coefficient (HTC), which describes some thermal properties of the house and is an important metric in energy efficiency of buildings.

The Smart Meter Enabled Thermal Efficiency Ratings (SMETER) Innovation Programme was developed by the Department of Energy Security and Net Zero (DESNZ). Government funded innovations in this area, and at the end of the project in 2020, it faced the challenge of comparing efficacy and 'accuracy'[1] of these solutions. NPL was invited to define a validation methodology of SMETER solutions as an independent third-party partner of DESNZ (November 2022 – November 2023).

For validation of an estimate of a parameter, it is necessary to compare the estimate to a known so-called "ground truth": the most accurate value of the parameter, preferably derived analytically, or measured with precision in a controlled experiment (such as a co-heating test for HTC), or prescribed by construction of a model dataset. The SMETER validation methodology developed here is enabled by *prescribing* controlled ground-truth HTC values (conceptually a user defined heat loss of the building, equivalent to a co-heating test result), from which surrogate data is generated in the form of realistic (but synthesised) measurement data (smart meter and temperature). The surrogate data is derived from physics-based thermal equations and observed patterns in actual historic measured data, including environmental variables, such as external temperature and solar gain. The methodology was developed using existing historic data but aligned to the user-defined heat loss– currently on the BEIS funded SMETER TEST project [Allinson et al 2020] Phase 2 test houses HH01-HH30. The surrogates are then used for blind-testing SMETERS to validate their performance against the prescribed HTC values.

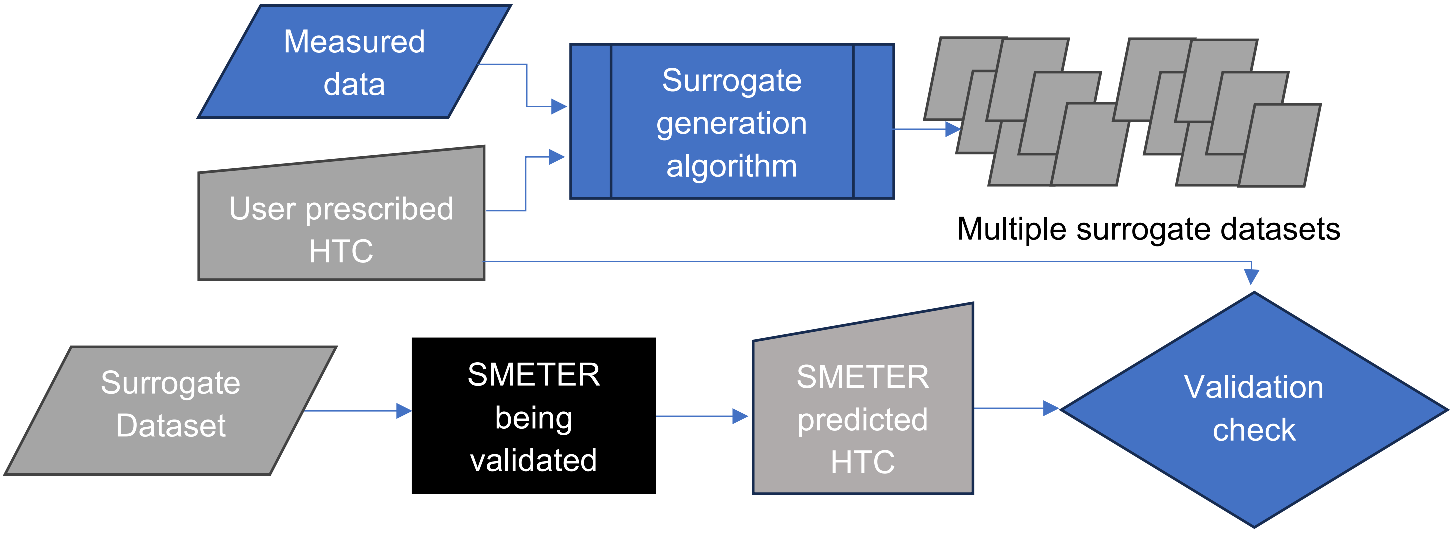

Figure A : Flow chart of SMETER Validation process

[1] Accuracy is not meant here in the absolute sense due to the complexity surrounding the ground truth to which building heat loss can be compared.

Surrogate datasets are artificial data with equivalent statistical properties to the observed real data. Siviour-based[2] SMETERS, using building physics equations for heat loss, can use different surrogates depending on what temporal resolution they require, varying from daily aggregate to half hourly; Machine Learning-based SMETERS and grey-box models can use half hourly surrogates[3]. The advantage of surrogates is that they can be simulated in abundance, with finely tuned features and statistical patterns derived from the observed data, and then used for testing and validation of SMETER models.

The participating stakeholders performed blind tests of their SMETER technologies using the surrogates with controlled HTC values. The results of the blind tests were satisfactory for Siviour-based, grey-box and black-box SMETERs, within 5-10% variation from the ground-truth value.

The surrogate approach promises to be a reliable technology-agnostic tool for SMETER validation. While this is a stochastic modelling approach that does not reproduce detailed physics of a building but rather replaces a part of its fluctuation of temperature with realistic long-range correlated noise, it does capture the variability of energy consumption based on physics principles and addresses the relationship between HTC and energy consumption under default assumptions (regular energy consumption patterns, non-anomalous weather conditions, and regular diurnal cycles of thermal variability). The validation tests do not challenge SMETER modelling algorithms in abnormal conditions such as non-standard heating patterns or extreme weather fluctuations but rather check their ability to identify average HTC correctly from surrogate data based on the current catalogue of measurement data. The validation tests using surrogates could be expanded with additional non-standard data and further work to develop the surrogate generation algorithm further. The surrogate data is constructed to reproduce major physical dependencies between variables (temperature, gas energy input, electrical energy input, weather and solar gain) of monitoring thermal performance of built stock of observed types. As such, the surrogate-based validation can identify accurate and less accurate SMETER algorithms under similar conditions to the originally observed data without exposing the SMETER to the exact same data.

2 Introduction

The heat transfer coefficient (HTC) has emerged as a fundamental metric with which the building fabric performance is assessed. As such, HTC is estimated as heat flux over the temperature gradient between two media and is measured in Watts per Kelvin, thus describing the total heat loss via a wall or roof of a building or the whole building accounting for all heat transfer mechanisms. A key component of the HTC is thermal conductivity, and additional

[2] Siviour equation is a physics based model of heat loss, applicable to buildings.

[3] Three versions of surrogates are described in the methodology covering this temporal range. Grey box and black box models describe the level of transparency of the model.

components include those which are occupant dependent (such as heating pattern, ventilation practise).

The present report focusses on in-use metrics (those measured in dwellings with occupying families), but HTC is not restricted only to these. Co-heating tests produce HTC estimates where fabric response to environmental influences varies but exclude occupant factors, i.e., co-heating tests run in empty houses at fixed indoor temperature (for example, +22 degree Celsius), while external conditions may still vary a lot [Johnston et al 2013]. Co-heating is widely considered to be the ‘gold standard’ of HTC measurement and is generally referred to as the defacto empirical ground truth against which alternative methods can be validated, however co-heating is an involved and time consuming measurement and is not always available therefore the current work develops a validation approach that is agnostic of the SMETER algorithms, can be applied using minimal co-heating measurement and observational conditions during data gathering.

Unlike thermal conductivity, for instance, HTC is not treated in the current context as a physical attribute that is invariant with respect to occupant behaviour, where occupant behaviour describes such activity as opening windows and doors. HTC does vary with many environmental factors, such as wind and humidity, as well as with other variables that affect the heat flux between environment and indoor space (for example, minor wall cavities and faulty insulation layers

HTC describes the average rate of heat loss, normalised by inside-outside temperature difference. It includes the heat loss by conduction through the fabric and by infiltration and exfiltration through ventilation. A lower HTC demonstrates a lower rate of heat loss and therefore better thermal performance. The rate of heat loss will vary considerably depending on internal and external conditions of the environment, and it only makes sense to present the HTC in terms of an average figure (annual, monthly, or weekly). The rate of heat loss from different rooms will also vary and the total loss from whole home is considered, as well as a temporal average. The HTC of a building can vary with weather and season (the winter season in the UK represents the period of interest in terms of heating). The HTC of porous walls changes when wet, for example, and it was observed by Rhee-Duverne & Baker [2013] that the thermal conductivity of bricks is 1.5 to 3 times higher when they are wet than when dry. This reflects the changing composition of building elements (i.e., their increased water content) not the changing thermal conductivity of the material concerned. Furthermore, ventilation heat loss is a significant factor that changes with time, as in cold winters many households reduce or even terminate ventilation via windows to preserve heat.

The HTC combines heat loss by transmission through the fabric with ventilation heat loss. The ventilation heat loss is affected by the weather (windier conditions generate more ventilation) and the occupants (opening windows, opening and closing internal doors, and switching on extractor fans), and the temperature difference between indoors and outdoors. Based on heat transfer principles, we would also expect the rate of heat transfer by transmission through the fabric to depend on wind speed and direction, absolute temperature of the air (inside and outside), cloud cover, and solar irradiance. These relationships have not been studied in detail and, like the moisture content of materials, may or may not have a big impact on the measurement for a specific house.

SMETER technologies that currently exist vary in their modelling approaches. Some of them are based on the Siviour equation (i.e., physics-based), others use machine learning techniques to establish dependencies between thermal conditions and energy consumption; there are also models that obtain probabilistic estimates of HTC using Bayesian ensemble models. Although SMETERs differ, they essentially estimate the same property of the same building, even if they use different sensors and sets of measurements. From the point of view of their comparison, however, this represents a major challenge in their validation due to the often opaque nature of non-physics based methods making validating against an expected benchmark more challenging.

The HTC of buildings has generally been measured by thermally controlled co-heating tests, which are highly resource-intensive in terms of installation and monitoring, or calculated from physics-based models of structures, which require substantial computational resources. Another key barrier is that additional installations for co-heating tests are intrusive and carried out on unoccupied properties. Hence, the mix of resource intensity (cost) and need for unoccupied properties act as barriers to applying co-heating tests at scale. Reduced Standard Assessment Procedure (RdSAP) methodology, which underpins the Energy Performance Certificate (EPC), is not computationally intensive; its potential flaw is its reliance on detailed and accurate input data with multiple parameters, which is often difficult to obtain (e.g., U values of walls, windows and ventilation rates) – often these are assumed at default values.

The growth of smart electricity and gas meters and the increasing consumer acceptance of electronic devices for better control of central heating systems have offered a new approach to evaluating HTC. The data of high temporal resolution collected by these smart sensors and meters can be used to provide rapid and more cost-effective estimates of HTC across usage patterns than traditional methods, because their low-cost (using already measured data from smart meters) in-use (i.e., with inhabitants) operation can be run for a sufficiently long time as to address occasional anomalies and fluctuations. In-use monitoring and analysis of those data is less invasive and requires less resource to capture and analyse in-use data.

SMETER technologies consist of a minimum of the smart meter data (gas and electricity) but can be supplemented with additional sensor, weather and contextual data to improve accuracy of the HTC predictions. SMETERs are an emerging technology area and typologies of SMETERs are still developing, however at this stage remote only (i.e. smart meter data only) and temperature/weather enhanced SMETERs are the most common types. The models underpinning these need to be assessed. The terms immediately below are commonly used in the context of gauging fitness-for-all purpose, and a brief description of what each constitutes is given.

Verification. Assess how complicated the model linking the measured quantities to the quantity of interest is and check the model output for cases where the solution is known exactly; check a mathematical specification of the calculations; check the source code if available; test qualitative behaviour of results (e.g., linearity, monotonicity, sign changes, etc.)

Validation. Compare calculations with reality. Reality could be defined by a trusted different method being used to measure the same object, or by an artefact or material with well-characterised properties. The validation could also include a critical review of the physics that have been included in, and omitted from, the model, along with an estimate of the likely size of terms that have been omitted and their effect on the quantity of interest, but this approach will not include “unknown unknowns”.

Uncertainty quantification. Propagation of uncertainty, whether associated with the measured quantities or other parameters within the model, through the model to look at variability. This aspect informs the decision on whether the data (measured quantities and internal parameters) are good enough. A validation that does not take uncertainty into account is not reliable since the difference between two values is only meaningful when compared with the variation of those values.

In this report, we introduce surrogate data based on patterns in observed building data to facilitate SMETER validation. For validation of an estimate, it is necessary to know the so-called “ground truth”: the most accurate value of a parameter, preferably derived analytically, or measured with precision in a controlled experiment (such as a co-heating test), or prescribed by construction of a model dataset. Surrogate data allows to use such a prescribed ground-truth value of HTC, which is a starting point of building artificial data – thus, by construction, the surrogate data has the ground-truth HTC value. This procedure is then invoked by presenting the SMETER algorithm with synthetic data that it could not have encountered before in order to see if it gives accurate results compared with the ground truth HTC value. This validation approach can be applied to current SMETER technology but should also remain robust as the sensors and algorithms used in SMETERs evolve. We use surrogate data with controlled statistical properties and assigned HTC values for blind testing and validation of SMETERs. The proposed methodology is general and technology-agnostic, which allows to validate and compare SMETERs as black boxes. This is particularly important because of high competition among commercial SMETERS, as many developers do not wish to disclose their algorithms, and SMETERS must be tested on the same datasets for fair comparison. This represents a major challenge for validation, and blind tests with surrogates is a viable solution in this context.

The surrogate data utilised observed data published in Phase 2 SMETER report [Allinson et al 2020], which includes 30 houses with measurements that were used to test several SMETER models in 2020. These houses varied from bungalows to semi-detached and were denoted as HH01-HH30, which notation we use in this report as well. One of the characteristic houses is HH24, and it is a testbed in several experiments discussed in this report, see below.

This report outlines the physical principles of estimating the HTC as a characteristic of the building fabric in terms of heat transfer and introduces three types of surrogates of different complexity, as well as the validation workflow.

3 Uncertainty in HTC estimates

The goal of developing SMETERS for HTC estimation is to assess buildings with respect to their thermal properties. The number of variables for which a generic SMETER may be gathering data may vary from a few to a hundred. This is because they may use different variables to assess heat flow that is affected by material properties of the building as well as by behaviour patterns of the inhabitants. For example, heating in winter will be affected by state and conditions of windows, and it is possible to monitor if a specific window is closed or not with some binary data from a sensor (not available in current SMETERs but may become used in future).

The larger the number of observed variables, the more detailed the underlying model, but there is a certain practical limit on the number of measurements that can be taken in a private household. Using a fully resolved physics-based model with thousands of degrees of freedom for such a purpose is unfeasible because of complexity and limited resources. It is necessary to note that less complex models may still perform with sufficient accuracy. Therefore, the purpose of model validation is not to identify which model is most sophisticated but to assess how accurate each model is and how close to a known ground truth, no matter the approach they are using.

SMETERs vary in their approaches and designs. For example:

- the so-called *steady state methods* that are based on building physics and use daily averaged variables, e.g. Siviour, average and multivariate linear regression method [Jack 2019];
- the so-called *dynamic methods* that are based on building physics and use higher resolution (short time step, 30mins-1hour) data; these methods are usually called grey-box or lumped parameter methods, and there are various ways to solve the equations which are based on the electrical analogy and a resistance capacitance network;
- the so-called *black-box methods* that use machine learning (ML).

HTC values obtained using the Siviour approach are affected by the temporal resolution of the power and solar gains data that are regressed (see Eq. 2). Also, some ML-based SMETERs are often trained on higher-resolution dynamics of measured variables (30 minutes or higher) rather than analyse the aggregated statistics. Similarly to ML-based models, grey-box models, such as lumped thermal capacitance models [Hollick et al 2020] consider subsets of data (as short as 10-days subsets) to reconstruct necessary variables for HTC estimation. The major differences between SMETERS operating at different temporal scales are in conceptual approaches and the statistics/analysis used to output the HTC.

A validation methodology should be:

- controllable, with ground truth that is imposed and varied to assess SMETER accuracy;
- generic in the sense of fitting various types of SMETERs;

- agnostic of the model algorithm, which would ensure that the validation methodology would not act as a barrier to innovation.
- representative of the building stock it is used to assess;
- accounting for the basic measurement uncertainty of the sensor measurements;

HTC is not a function of energy consumption alone. Rather, building energy loss is a function with HTC as a parameter [Jack et al 2018]:

$$\mathrm{HTC} = \frac{\overline{\Delta Q}}{\overline{\Delta T}} \qquad (1)$$

where $\overline{\Delta Q}$ is the net heating power the house receives (W) and $\overline{\Delta T}$ is the averaged difference between indoor and outdoor temperature (K). The HTC describes the average rate of heat loss (normalised by average temperature difference) for a home. At any moment in time the heat loss rate can be much lower or much higher than the average due to the heat stored within the building fabric and its contents and the temperature difference as a key driver.

It should be noted that the energy flow into the building as measured by the smart meters is not the same as the $\overline{\Delta Q}$ in equation 1. For gas heated homes then the gas is converted to space heat and hot water, with an efficiency from the boiler, whereas electric demand (given the constraints around self-generation and usage outside the building envelope e.g. EV charging) will be converted to heat. The key point for SMETERs is that, in the absence of heat meters to measure the actual heat input into the building, then assumptions or inferences about the efficiency of the heating system and the losses associated with hot water.

Factors/quantities other than energy consumption need to be considered for reliable estimation of the thermal performance of a building. For example, a well-insulated flat inside a block of flats, whose owners are away, may have very little or no energy consumption from any appliance within it, but it will still obtain and lose heat passively, via the walls, due to the properties of the adjoining premises. Its HTC cannot be estimated solely on the basis of the recorded power consumption. Such a case study would represent a challenge for a measurement-based HTC estimation, as would many other similarly complex layouts. This can be overcome, for example, by measuring HTC for an entire apartment block rather than individual flat, where possible (in terms of installation of sensors, etc.). The limitation may apply to the individual dwelling, but it does not itself rule out the application of SMETERs on another scale.

Models that estimate HTC in regular (i.e. operational) circumstances can be tested and validated using generic datasets of observed variables in the built environment. HTC estimation can also be more challenging for certain dwelling types, and SMETERs are not proven for all of them; flats are an example of where confident HTC estimates may be difficult to obtain due to added complexity of contributing factors (for example, heat transfer through party walls and the challenge of estimating solar gain). The accuracy of such estimates may vary, and this report discusses sources of uncertainty.

The proposed validation methodology is based on surrogates with an assigned ground truth value of HTC (i.e. the surrogate dataset is generated based on a selected value of HTC, which could be varied for testing purposes in a certain range to reflect realistically thermal properties of a building). Conceptually the methodology takes existing empirical energy/temperature data from a home with a known ground truth HTC (from a co-heating test) and generates (using a statistical transformation) a new set of data around a designated HTC value. This provides abundant and diverse artificial data while preserving the empirical nature of the data, which can be simulated repeatedly for uncertainty quantification and sensitivity analysis, with a known value of HTC.

The core idea of validation is comparison of an estimate under consideration with a reliable value of the parameter - derived using analytical modelling, numerical modelling, or high-precision measurement in fully controlled environment (for example, in a factory). In the case of HTC validation, the complexity of the system is very high, and even measurement that are considered most reliable, such as the co-heating tests, may be affected by the complex factors of the high-dimensional system of domestic housing. These may include environmental variables (weather extremes, such as excessive precipitations), building fabrics anomalies (cavities, capillary effects), operational conditions (ventilation), and others. The choice of surrogates as the basis for validation methodology was dictated by the need for ground truth values of HTC, which otherwise were not available. This is not the only way of validation, but one of the most straightforward and controllable.

We first quantify the uncertainty in the existing HTC estimates, using parametric bootstrapping of data in the Siviour equation, following [Jack 2019], with input data from thirty houses [Allinson et al 2020]. Parametric bootstrapping is a technique of sampling random numbers from a probability distribution that is fitted to real data. This allows one to obtain repeatedly synthetic data with statistical properties that are equivalent on aggregate to those of measured variables. Instead of a single time series of a variable, we obtain an ensemble of values through parametric bootstrapping that provides an uncertainty range for the variable concerned.

Siviour equation, unlike ML-based models, provides analytical description of thermal properties of a building in relation to HTC, and thus allows one to propagate the uncertainty through this model. This is why the parametric bootstrapping can be applied to quantify the uncertainty of HTC in this layout. It is necessary to mention that uncertainty quantification of ML-based models is possible when the models are provided for examination and application; unfortunately, these were not available during the course of the project.

There are equivalent techniques of HTC estimation, such as multiple linear regression (MLR) and simple averaging method (AVG), and they provide very similar estimates and uncertainties – see section 7.1.

HTC was estimated using input datasets of the daily total heat (power) input Q, solar radiance S (Wm-2) with solar aperture g (m2), and temperature difference (internal and external) ∆T (K) according to

$$\frac{Q}{\Delta T} = -g\frac{S}{\Delta T} + \mathrm{HTC} \quad (2)$$

as the intercept of the least-square fit of the Siviour plot (see Fig.1), where HTC and g are derived by means of an optimisation procedure from the data.

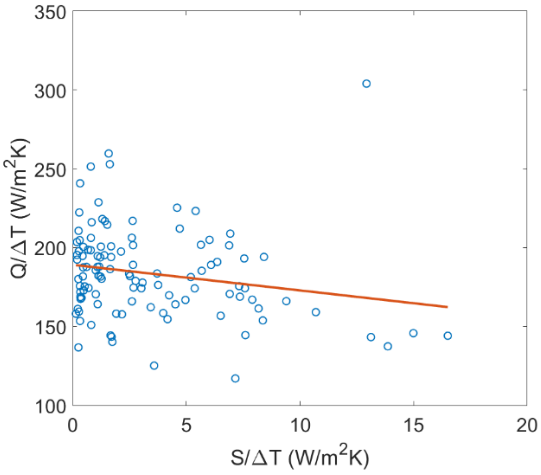

Figure B Estimation of HTC as an intercept, using least-square fit of variables scaled with temperature differences, for HH24 dataset over the winter period 2020-2021

Uncertainty quantification uses random numbers as a statistical tool. The input variables were first sampled independently from uniform distributions following established practices for quantification of uncertainties [GUM1995; GUM2008] with mean values and variances estimated from the input data of houses HH01-30 [Allinson 2020]. This means that instead of a single value of measurement we obtained a set of close values with homogeneous properties, similar to real data. These uncertainties were propagated through the Siviour equation (1) to obtain the combined uncertainty of HTC. This approach takes no account of correlation between variables, and it does not reflect time-dependent variability, but it can serve as a first crude estimate of HTC with uncertainty and the basis for “go-no go” initial assessment of an SMETER. When propagation of uncertainty is applied, it requires a model describes the system, and Siviour equation is one of the examples of such a model; other SMETERS can be tested in the same way if their equations are disclosed.

Currently, the co-heating test [Allinson et al 2020] is considered as ground-truth estimate in industry. The methodology comprises a particular experimental set up, including methods to limit the impact of solar and wind, the latent heat from materials that may be drying out, the heat exchanges as a result of party elements, and the thermal capacity of ground floors etc. While variations in the method do exist, the method was considered sufficiently robust, with some adaptations to ensure uncertainty quantification in measurements. Adaptations may include some corrections for heat delivered via party walls, which may explain why the co-heating test HTC estimates will be higher (indicative of the property in question having poorer thermal performance). This raises the question of the extent to which environmental factors should be considered when estimating HTC. It is also necessary to consider the artificial nature of co-heating tests, as the properties are maintained at an approximately uniform temperature of 25 degrees Celsius using electrical heating only (so relative heating efficiency can be excluded) without occupants, and ventilation apertures are sealed to prevent air infiltration (particularly if weather is windy), and exfiltration. If, during occupancy of the building, ventilation losses are significantly greater than during co-heating, then HTC values obtained through a co-heating test would tend to be lower.

4 Challenge of temporal scale separation in HTC dynamics diurnal vs sub-diurnal HTC estimates.

When applying a linear fit to the variables of the Siviour equation, it is necessary to remember that there exists a scale separation in dynamics of high-resolution observed data and its low-resolution (daily) averages. This effect of variable resolution on HTC estimation using the Siviour method is clear from Fig. 2, where solar gains and power have been considered for the same period using data averaged over 24 hours and 30 minutes, respectively. When diurnal averaging is applied, it reduces variability by smoothing the data with removal of short-term variability; this produces a more reliable averaged estimate. It also follows the assumptions of the model (approximately) by removing most of the effect of thermal mass.

This is because the Siviour model is not (and is not expected to be) a good approximation over short time periods, and high-resolution data would require a different physics-based model. The Siviour method is based on the assumption that thermal mass has no net impact over the averaging period. This is valid because the daily average of heat input and temperature difference accounts for most of the energy stored in fabric and furnishings. The Siviour method cannot be reliably applied at finer than daily resolution, as we assume the amount of heat stored in the thermal mass is the same at the start and end of each period. For higher resolution data this assumption is not valid and so the Siviour method cannot be applied.

In Fig.2, we show the effect of daily averaging of surrogate data, for illustration. The surrogate methodology is explained in detail in section 7, where we provide algorithms of generating surrogates that mimic real data with a known, prescribed value of HTC. In this experiment, we use surrogates V3 generated at 30-minute step, with realistic diurnal patterns, as described in subsection 7.3. We then perform daily average of the surrogate data, and for both datasets we estimate HTC using the Siviour approach to demonstrate the effect of data average on the data and the HTC value: HTC diminishes when daily average is applied.

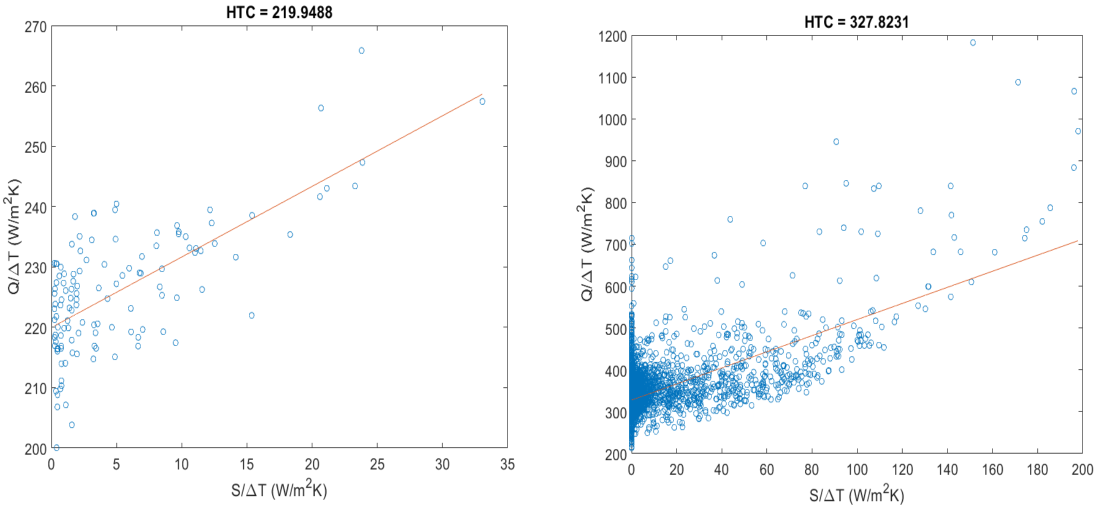

Figure 2 The effect of changing the resolution of the power and solar gains data regressed in the Siviour equation as applied to surrogate data with prescribed HTC = 220 W/K: left panel shows correct estimate in daily-averaged data; right column shows overestimated value of HTC derived from data of 30-minute resolution, which produces a much denser cloud of points.

One can see how the mean value of solar data time series of Fig. 3 changes when considering data at daily or 30-min timescales. This figure illustrates how daily averaging affects the pattern of intermittent time series: 30-minute-averaged data has lots of zero values (as at night there is no sunlight), whereas in daily-averaged data almost all values are non-zero. Zero values at the 30min timescale highlights the importance of the assumption behind the Siviour equation, that the energy flows are balanced over the time period which it is applied. At the 30min timescale the solar insolation is highly intermittent , but the temperatures do not respond accordingly, distorting the HTC as delivered from the Siviour equation, SMETERs based on SIVIOR or similar steady state algorithms. Daily averaging greatly reduces this dynamic effect.

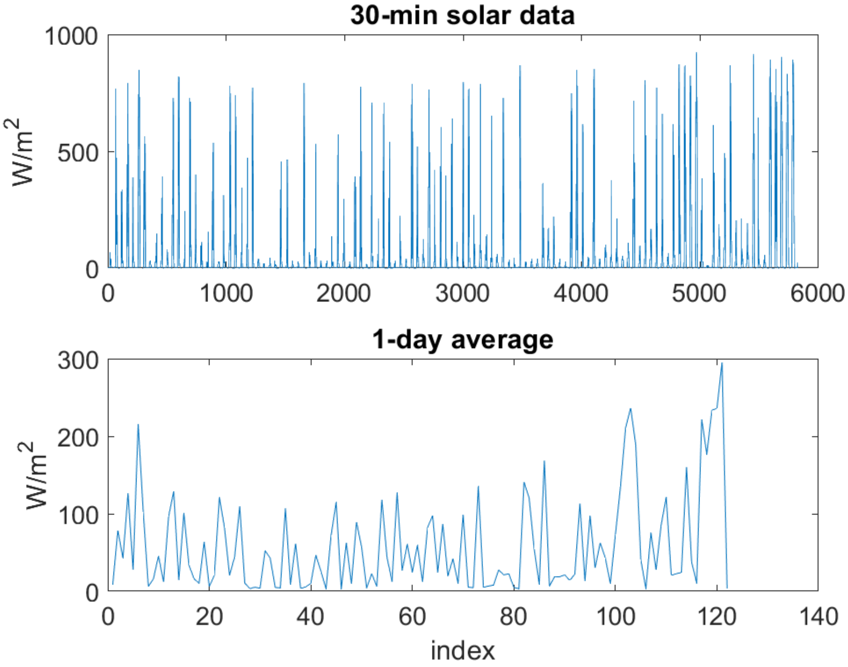

Figure 3 The amplitudes of solar radiation: top panel – 30-minute data plotted vs index, lower panel – the same data averaged over a day that contains 48 datapoints, vs index (denoting 30-min time step in the upper panel and 1-day time step in the lower panel). Data are significantly lower when they are averaged over 24 hours rather than 30 minutes. The solar diurnal cycle is reflected in 30-min data with zero values at night, whereas in daily data values rarely reach zero value.

HTC estimations using daily-averaged data are reliable, as reported in literature [Alinson et al 2020; Jack et al 2018; Jack 2019], but the challenge of validating the SMETERs based on sub-diurnal data requires development of a separate type of surrogate data. The surrogates with sub diurnal resolution must be suitable for both dynamic SMETERS (ML-based or grey-box) and for Siviour-based SMETERS, which aggregate data with daily averages. This is why we have developed several types of surrogates, which combine these properties (see section 7).

The significant change of the HTC estimate between daily and sub-daily data is caused by the diurnal variability of the contributing variables: for example, solar irradiance is zero at night and has high peak values at noon; its daily average is non-zero and is smaller than the peak value (see Fig. 3). The inter-day variability is much smaller, and further averages, for instance, over a week, do not create such a separation of scales in HTC estimation. This difference is the reason why several types of surrogates have been developed, see section 7.

In principle, such an effect can also be observed due to seasonal variability (daily solar irradiance in winter and summer differ a lot), but, due to the nature of the SMETER project, which is concerned with thermal performance of built stock in the cold season, we simulate surrogates for the period from 1st November to 31st of March, when intra-seasonal variability is small. Some SMETERs may use smaller datasets (e.g., two weeks). This can be addressed in surrogates, as they can be generated of any length, as necessary for a specific SMETER. Estimating HTC in winter reduces uncertainty by maximising Q and ΔT while minimising S (see Eq.2). Geometrically, when Q increases (due to added gas heating in winter) and S decreases (due to less sunlight), the dispersed point cloud is more compact and shifted up along the y-axis – closer to a shape like straight line; the variability of the intercept HTC, therefore, becomes smaller, although the HTC value increases. Also, a rather typical temperature measurement error of 0.5K has a relatively small effect in winter compared with late spring, as the temperature difference is bigger in winter, and this is the variable that affects HTC estimate. In these subperiods, the home is more likely to be operated in a state that we are interested in, i.e., in the primary heating state.

It is noteworthy that late autumn (Oct and Nov) typically has more sunshine than winter (Dec and Jan), due to the duration of daylight and the varied atmospheric conditions. In such sub-seasons, SMETERS produce different HTC estimates (as observed in historical data of the Phase 2 report), as seasonal variability includes heat losses to the ground and occupant practices (the latter is not included in co-heating tests).

It is important to remember that the HTC is the average rate of heat loss and so it would not be feasible to evaluate it accurately over a small time step (as currently there are no practical ways to achieve high accuracy of HTC estimate without large-scale monitoring with sophisticated measurement instruments being installed in the home). The diurnal variation in solar irradiance is one aspect, but fundamentally it is determined by the change in heat that is stored in the building fabric and furnishings. This includes heat from the sun, occupants, lights, appliances, heating system, etc.

5 Effect of intermittent gas usage

Gas heating often runs at specific user-defined hours or activates when external temperature becomes too low. This means that there may be very large energy use during certain times of the day but very little (only due to electric appliances) outside these times. Also, there may be heat loss through hot water cylinders and large swings in gas consumption within the heating periods due to oversizing [Bennett et al 2019]. Such energy outliers separate the point cloud of Siviour plot into clusters, which affects the estimate of HTC as an intercept of the linear fit.

The following figures (Figs. 4 to 7) illustrate how the HTC value changes when periods of zero gas use records are excluded, these are presented here to illustrate the importance of capturing all aspects of the original empirical data in a surrogate dataset for validation in order to test both the core SMETER algorithm and its data filtering. In both 30-minute and daily-averaged data, in real and in surrogate datasets for house HH24 (Phase-2 report), for which co-heating tests estimated HTC to be about 150 W/K. Zero-gas values are observed in 30-min data, and those values were excluded for that time stamp along with all other variables, and then daily averages were obtained over subsets with non-zero gas use. This means that instead of 48 records per day, a fewer number is considered, without those with zero gas use, and they are averaged to obtain a daily mean. This is because the average over a set with multiple zeros is smaller than the average over fewer non-zero values. When zero gas use is not taken into account, the HTC value is systematically higher than in the complete data. This effect is an example of bias that may occur in heterogeneous data and crude SMETER algorithms. We present this example as a demonstration and do not suggest that this should be required preprocessing of data.   However, in the context of SMETER validation, we address this bias in the construction of V3 surrogates, which have diurnal intermittent gas supply following the observed patterns in real data HH01-HH30. We ensure that each generated surrogate is back-tested to reproduce the simulated HTC value within 2-3% error (which is due to the stochastic effects and is negligible).

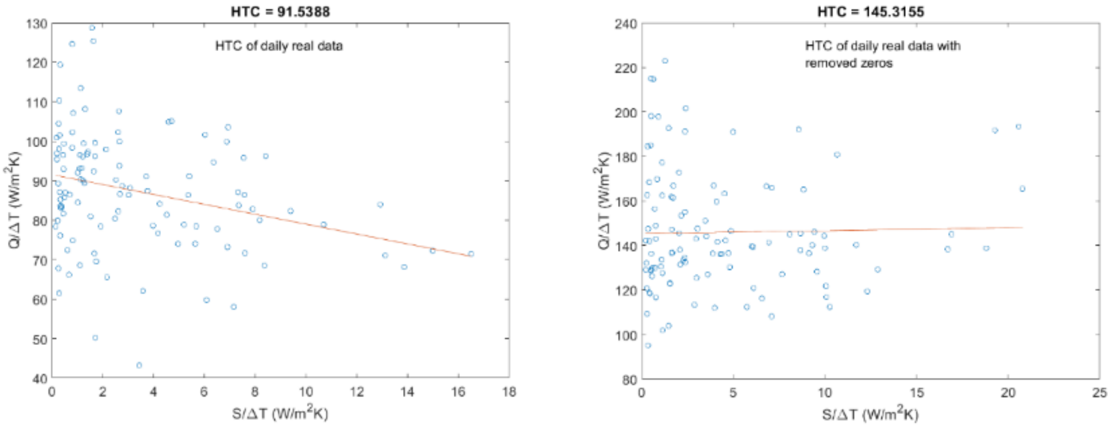

Figure 4C HH24 house real data: daily averages without and with zero-gas removal: HTC estimate is lower with zero-gas values, because the average value of power is lower.

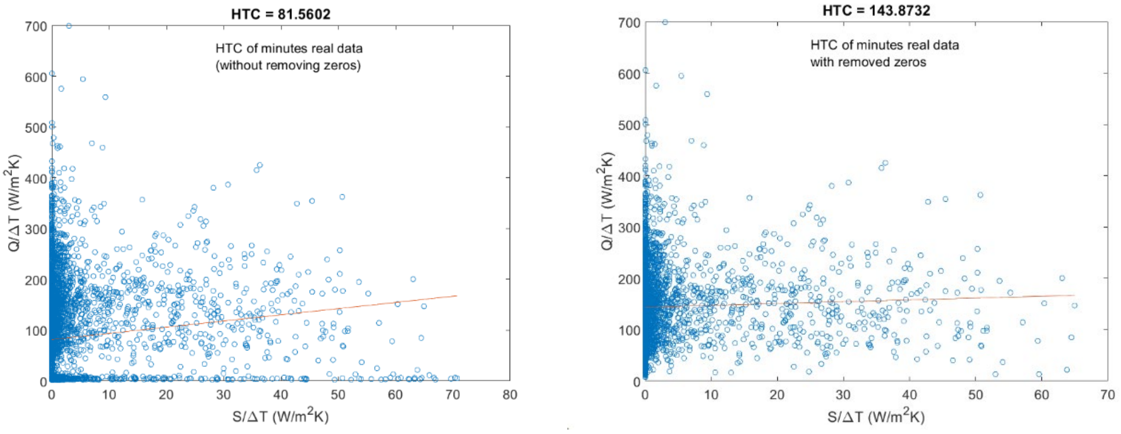

Figure 5 HH24 house real data: 30-minute data without and with zero-gas removal: HTC estimate is lower with zero-gas values, because the average value of power is lower.

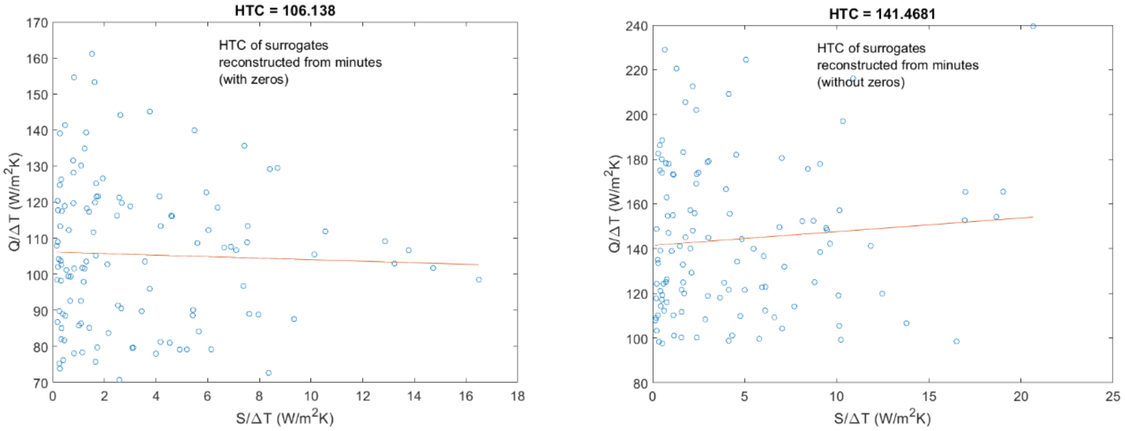

Figure6. HH24-based surrogate data: daily averages without and with zero-gas removal (similar to real data in Fig. 4).

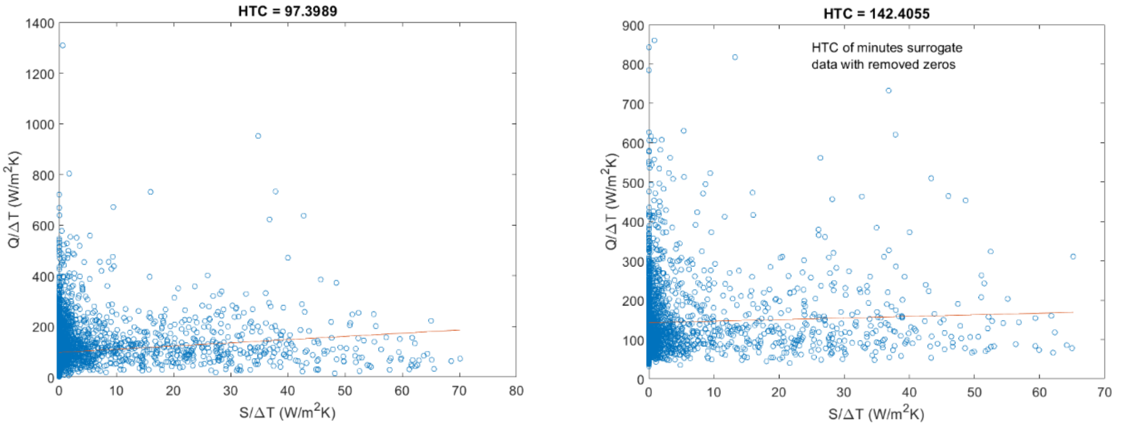

Figure 7 HH24-based surrogate data: 30-minute data without and with zero-gas removal (similar to real data in Fig. 5).

These experiments (Figs. 4-78) are for illustration purposes of the possible sources of bias in HTC estimation, as we do not suggest removing zero-gas records for HTC estimation. Such experiments helped during the course of the project to design complex V3 surrogates.

6 Validation framework

Validation of a model requires a trustworthy ground-truth estimate. This can be done using a combination of a co-heating test (as ground truth) and measured energy demand and thermal response of a building which can be fed to a SMETER algorithm. However, in order to avoid reusing the same data across multiple validation exercises (some of which could be with the versions of the same SMETER), a method is needed which can provide sufficient 'fresh' data for validation. For this purpose, we propose using surrogate data [Schreiber and Schmitz 2000] with controlled values of HTC based on the properties of the observed variables and the building thermal performance model . The controlled HTC can be selected by the user, in this case the person or organisation carrying out the validation, this is in effect a prescribed, but artificial, co-heating test result. Using the minimal information obtained in the observed data (external temperature, solar gain) and prescribed HTC values, relevant variables (temperature difference, consumed power) are then generated as inverse estimates. This is sufficient for generating daily-resolution V1 surrogates. V2 surrogates are obtained from diurnal V1 surrogates by using randomly distributed high-resolution artificial data (without realistic diurnal patterns). For complex 30-min resolution V3 surrogates with realistic diurnal patterns, observed data is used reproducing temporal autocorrelations for realistic patterns of internal temperature and power use in the surrogates. Environmental variables, such as solar radiation and external temperature, are copied from the input datasets without changes.

The simplest way to generate surrogates is to sample a suitable probability distribution independently (i.e., reproduce probability density of the real data but not autocorrelations which are dictated by diurnal patterns), whereas more advanced surrogate data can be generated in such a way as to preserve temporal dependencies and correlations of variables to reflect reality more fully. An overview of how these surrogate data were generated is given below.

Because large-scale (at the level of a region or bigger, including different houses that are exposed to different environmental conditions) validation requires certain abstraction and simplification of the model to address various types of houses, we propose to use stochastic modelling. Variability is approximated by stochastic components with controlled properties such as scaling exponents [Hasselmann 1976; Livina et al 2013] when using such modelling. Scaling exponents describe the strength of auto-correlations in the data, which define the data patterns and drifts.

We stress that the goal of validation here is not the development of digital replicas of all possible houses, as this is time- and resource-consuming and is not feasible for large-scale application. In testing diverse SMETER solutions, it is expected that some simplifications must be assumed when generating surrogate data, and this may require adjustment of surrogates (such as temporal resolution and additional variables – e.g., heat measurements that may appear in newly developed SMETERS), and continued support throughout testing may be recommended to address this in further versions of surrogates. For example, adding heat measurements would be required to produce new surrogates with added variable that is currently not included in V1-V3 surrogates, which are currently based on the available Phase-2 data and the SMETER technologies that took part in the blind tests.

The performance of different types of houses (with different appliances inside them) may be quite different from each other to warrant the production of surrogates that reflect them. This is because:

- different homes have different gains/losses associated with construction. The relative importance of these changes with construction type and model accuracy depends on its ability to represent such factors. A simple example would be a home with south facing windows compared to one with a large south facing facade: the model's ability to account for solar gains is tested much more in the second than the first case (and there are direction issues as well).

- the magnitude of unmetered gains matters compared to the efficiency of the fabric since the less heating is required the lower the HTC.

Note that by construction the surrogates would generate different patterns in the data in different rooms with south-facing window or façade. This happens because surrogates mimic properties of the real measured data.

The current validation framework is based on the surrogates derived from the HH01-HH30 data, and further input data is necessary to broaden the validation framework.

6.1 Limits of validation

Once validation is performed and HTC estimates are obtained by a SMETER, for prudent decision making, their uncertainties must be quantified and compared with the SMETER provider’s estimates. How large should be the error to classify a SMETER as inacceptable in accuracy? At this stage of research, NPL can report the errors, but only a few tests on a limited number of house types have been performed so far, and this may be not sufficient for identifying classifying tiers of accuracy. This may be a topic of future research.

6.2 Assessment of input data

There are several data quality dimensions that must be addressed in SMETER input data: Accuracy, Completeness, Consistency, Uniqueness, Timeliness, and Validity.

Although electricity and gas meters are inspected and replaced in case of inaccurate measurements, other SMETER measurements may be obtained with low-cost sensors which are not expected to be monitored or calibrated. The sensor quality may vary, and at this stage it is not clear what calibration requirements would be imposed on sensors. This may be a subject of further discussion in follow-up projects.

The accuracy of the input data is likely to be assessed mainly from the documentation of manufacturers. When validating SMETERs, such documentation should be provided alongside HTC estimates and a description of the measurement network.

Furthermore, observed records might have missing numbers (due to failure of sensors, lost connection, etc.). There are various approaches to resolve this issue in real data, depending on its severity in time (e.g., missing hour or days) and in space (e.g., if the problem is with temperature measurements, then one can use other sensors). In real data, missing values can be addressed either by entering a mean value of a time series or by simulating artificial data. It is necessary to take into account what type of SMETER will be applied to this data, which is an integral part of the SMETER product: if the SMETER uses bulk data for Siviour regression, then mean values would suffice, as they would not affect the intercept of the regression with bias towards higher or lower values. If the SMETER is based on a dynamical model or ML approach, though, it is important to reproduce patterns in data: constant values instead of real fluctuations would be unsuitable. In that case, suitable random numbers can be used together with a sorting procedure that would reconstruct temporal correlations [Livina et al 2014]. This approach would address completeness of the input data.

Input data may include various measurements, different for each SMETER. Typically, electricity and gas consumption will be included, as these are common smart meter measurements, and they directly impact heat flow into a household. Furthermore, if a sensor is replaced in a device, and no expert calibration is performed, there could be discrepancies between subsets of data before and after such replacement. Building characteristics, such as number of rooms, are reflected in surrogate data obtained from the observed variables; other parameters, such as house area, are stored in the auxiliary files of houses HH01-HH30 and assumed to be the same for surrogates. Data may also be affected by changes in formatting or time step output. Repeatability of HTC is affected not only by building properties but also by environmental factors, which may require multi-year observations. Not every winter (or its sub-period) is average; for example, excessive precipitation will affect building properties and heat loss (leaving aside what metric is used to estimate it). In a very wet winter, it is likely to get estimates of building performance that are different from an average winter, but this means that validation attempted in such a year might be incorrect, and repeatability will not be achieved but not because of the validation procedure. These issues challenge consistency of the input SMETER data.

To ensure uniqueness of data (i.e., absence of repeated data rows with the same time stamps), routine checks of the time stamp variable using the differencing operator (discrete differentiation) can be run. This will produce a time series that should have constant values, which is easy to visualise or test using coded conditions. Any deviation from constancy will reveal repeated or missing measurements (thus, also revealing data incompleteness). Some loggers can get stuck in a fixed state (i.e., generating timeseries data but the data that are wrong), and an additional test is needed to check that the data is varying at a sensible rate. What is far more common is gaps in smart meter data (a timestamp is present but without associated recorded data) and then a catchup reading where all the energy use is aggregated into one "spike". Such outliers require investigation and adjustment, and could be generated intentionally with a view of assessing the ability of a SMETER methodology to account for them (i.e., the extent to which an SMETER invokes these interventions automatically).

Timeliness of the input data may need to be assessed in real-time SMETERS, which are currently not considered. Some, for instance, might use the time of day to determine solar gain more accurately from solar inclination. Timeliness may be affected by the quality of connection in data transfers (for example, over 3G network) and pre-processing algorithms of large subsets. It is necessary to ensure that historic data remains meaningful, and if any replacements have taken place, those should be delivered in time. In future, when the HTC estimates may be required in real-time, this aspect of data gathering may become more important.

With smart meters, the sampling rate can be quite high (48 measurements per day corresponding to a 30-minute sampling frequency). Attention is necessary in processing data that requires resampling and averaging since the last time step and use of non-standard units, such as KWh, since the last time step. To obtain daily power average for Siviour estimates, a relevant SMETER averages a set of 30-minute measurements of electricity over one day. The daily power average must remain the same, irrespective of sampling rate. The calculation of daily average power must be carried out in units of kW, and so “kWh since the last time step” must be converted into kW, which can be thought of as kWh since the last time step, by multiplying by 2 in the case of a sampling period of 30 minutes. Similar rescaling must be applied when measurements are provided at different sampling rates. The conversion to average power (watts) must be handled carefully and take into account the time step. This preprocessing ensures validity of the input data and should be performed by every SMETER that uses daily averaged data; a part of validation of such SMETERS is to ensure that preprocessing is performed correctly for specific temporal resolution and sampling rate.

6.3 Assessment of sensor accuracy

Sensor accuracy is documented by manufacturers, and instruments are expected to function normally within their reported limits. Sensor calibration can require substantial resources (in the form of remuneration to qualified technicians and engineers), and with growing use of sensor networks it is unlikely that such calibration would be common; yet assessment of the combined uncertainty using information from sensor documentation will remain feasible. Input data accuracy is dependent not only on sensor accuracy but also where it is placed, as well as what is not measured but assumed.

Not every variable’s uncertainty will be contributing equally to the combined uncertainty, due to very different scales of measurements. For example, gas consumption, which is the major source of heat throughout winter for homes with gas boilers, when converted into power, constitutes several thousand watts in an average household. Documented instrument error of up to 10% in direct heat measurements may equate to several hundred watts; at the same time, an error in temperature measurement may be a fraction of degree Celsius. In the case of the Siviour equation, it uses delta T because this is a smaller value, errors in temperature measurement have greater impact. The influence of these errors may be of the same reported percentage but the impact of the heat error on the HTC estimate may be much larger than that of temperature, because power in Watts may be bigger than temperature difference in Kelvin by orders of magnitude. This impact is defined by the combination of the uncertainty in the effect with the sensitivity of the measurand with respect to that effect, and it is in this way that the impact of each effect can be assessed [GUM 1995; GUM 2008].

6.4 Quantification of combined uncertainty

Combined uncertainty, which is obtained using standard method for the combination of variances and expressed in the form of standard deviations, is based on model accuracy, data completeness, measurement uncertainties, and operational errors [GUM1995; GUM2008]. Models for HTC estimation vary in accuracy, and this is to be assessed using the developed surrogates with an assigned ground truth. Measurement errors are estimated based on instrument documentation. Data completeness is analysed during pre-processing of the input data. Operational errors in this framework are minimal due to the automated data gathering and processing.

One of the biggest uncertainties is associated with measurement of the whole space temperature and this can dominate the HTC error (e.g., a home where temperature is recorded in spaces that do not represent the whole house well). Reduction of uncertainty may require expert interrogation and a judgment call over what it is reasonable to expect. In some cases, it was observed that a small measurement error over a temperature sample led to a large error in HTC on individual homes [Gori and Elwell 2018].

6.5 Effect of system noise on SMETER algorithms

The total metered energy use in a house is not equal to heat energy, as some appliances may consume electrical energy in heat pathways that bypass the fabric/ventilation or are stored over long enough periods to cause problems to the SMETER models, e.g., photovoltaic installations, batteries, or heat pumps without heat measurements; or there may be heat loss due to wastewater or electrical uses outside the home. This introduces uncertainty in HTC estimation, as demonstrated in the following experiments, which illustrates how adding/subtracting a small stochastic component (Gaussian white noise) affects HTC estimates. This noise denotes some unaccounted energy source/sink or a persistent measurement error that may lead to systematic bias in HTC estimation. The experiment is for illustration purposes, as we do not discuss here the source of the increase/decrease of energy use. This small perturbation is not visible by eye, as shown in the top panels, but in the lower panels it is clear that the HTC estimate is affected as higher HTC estimates with added noise and lower with subtracted noise are observed: HTC increases from 181 to 199 with added noise and decreases from 181 to 164 with subtracted noise. Such added small-scale noise may not be contributing to heat transfer in the house, but it will contribute to the estimate of daily power.

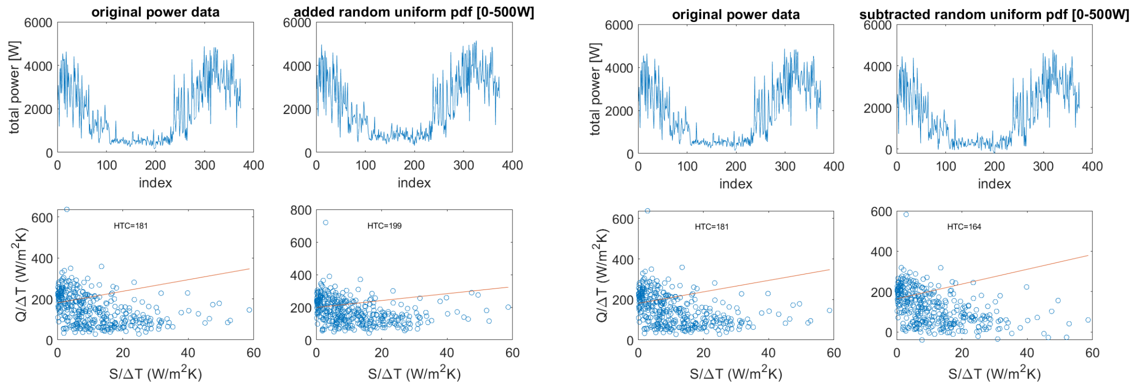

Figure 8 Left four panels illustrate how added power data noise of small amplitude increases HTC; right four panels illustrate how subtracted noise reduced HTC estimate.

This experiment illustrates the scenario where some background level of energy is underestimated or overestimated systematically. Underestimation may be due to unaccounted external appliances or technologies, overestimation may be due to measurement error.

Due to the increasing role of non-heating electricity (external appliances, such as a charging electric car, etc), wastewater, and non-metered heat (solar, metabolic, wood-burning stoves, etc), domestic dwellings may pose a problem for accurate HTC estimation because of its relatively greater significance. It is possible to model total unmeasured heat energy in surrogates based on direct measurements of electricity and gas consumption, and then adding metabolic input as extra parameters (for average number of inhabitants). This can be accounted for in further development of surrogate-based validation. Explicitly trying to create surrogates that would model unaccounted variables and absent measurements is very difficult. One way to address this would be to obtain data from more houses with different numbers of occupants, solar gains, fabric efficiency, etc, and extrapolate the observed variables for broader ranges of the UK domestic stock.

7 Surrogate data for SMETER validation

The core concept of the proposed methodology is the design and simulation of surrogates (datasets that contain the same number of days and variables as required by real-data tests, in which real time series are replaced by artificial yet realistic time series) for validation of SMETERs. This ensures a controlled ground-truth value of HTC. Although surrogates are not universal in capturing the full complexity of thermal performance of buildings, they can identify poor performance of tested algorithms, and they make possible comparison of SMETERs that use different technologies.

Because of the diversity of SMETER solutions and problem complexity, there is a need for several different types of surrogates for validation of SMETERs.

The first version of surrogate data, **V1**, comprises the basic variables (energy and temperature) having a resolution of 24 hours. The second version, **V2**, builds on V1, with further separation of dependent variables to give gas consumption, electricity consumption, indoor and outdoor temperature, solar gains but with a temporal resolution of 30 mins. **V3** surrogates add additional complexity from the source data.

Some SMETERs estimate HTC from data using the Siviour approach and equivalent methods applicable in quasi-steady state, while others use higher-resolution measurement data but still obtain the Siviour estimate from averaged daily data. These two types of SMETER use essentially the same methodology, but for their validation different types of surrogate data can be used (different temporal resolution as input, if the SMETER requires so, although at daily time step the approaches are equivalent). We denote these surrogates by V1 and V2. This type of surrogates reflect equivalent dynamics but with different temporal resolution, as there may be SMETERs that require this kind of input.

Although HTC is an averaged value, some SMETER approaches using machine learning (ML) require realistic training data having both temporal and spatial resolutions which capture thermal dynamics throughout a building (in particular, diurnal variability). For such SMETERs, and other building physics methods, i.e., the so-called dynamic methods, we developed V3 surrogates for validation, following the same inverse approach that starts with a prescribed HTC value and then creates surrogate variables to obtain total power. V3 surrogates are generated at high temporal resolution (30 minutes or higher), and they use a controlled the HTC value, diurnal patterns of simulated variables, and keep the observed independent variables of solar radiation and external temperature of the Phase-2 houses HH01-H30 as a part of the surrogate dataset. These models also use observed dynamical patterns of internal temperatures of adjacent compartments (between 5 and 8 compartments, depending on house type). The total power is obtained using the prescribed HTC and surrogates. Gas and electricity consumption are then obtained from the total power, following the ratio of electricity and gas consumption mean values in the observed data.

Surrogate versions V1 and V2 of surrogates comprise the three variables required to estimate HTC using the Siviour equation: solar radiation (in watts per square metre), the difference between outdoor and indoor temperatures (in kelvin), and consumed power (watts). Solar gain and external temperature are independent variables in this physical system, and for realistic surrogates we use the real time series of solar radiation and external temperature from houses HH01-30. Internal temperature in multiple rooms can be simulated using power-law correlations [Makse et al 1996] with a scaling exponent of 1.3 following analysis of datasets of houses HH01-30, i.e., the data will be realistically power-law correlated with persistence, or memory effects (not uncorrelated white noise). This modelling provides red-noise power-law correlated indoor temperature equivalent to observed datasets in aggregated statistical properties. Power-law correlations in the temperature data mean that there is some linkage between consequent temperature values (memory), which addresses thermal inertia: after a high value of temperature, a sudden change to a low value of temperature is unlikely - this is observed in random white noise, where fluctuations are completely independent. When there are power-law correlations, high temperature value is likely to be followed by a similar high temperature value. By this effect, the surrogate temperature follows observed trends in data that reflect slow response of the building fabric to changes of the external temperature, which less correlated (has much lower scaling exponent) than the internal temperature conserved by the building.

Thus, external temperature will be the same as in the input data, internal temperature will be realistic surrogate, and the temperature difference will be close to the observed but not equal to it, due to stochasticity in the internal surrogate temperature.

The temperature difference ΔT is anti-correlated with solar gains (i.e., when solar gain is low, in winter, the difference between internal and external temperatures, is largest). For prescribed values of HTC, consumed power is then obtained as

$$Q = -g \cdot S + HTC \cdot \Delta T + \zeta \tag{3}$$

where ζ is an additive stochastic component that addresses the thermal mass.

The first version of surrogate data, V1, comprises the three variables having a resolution of 24 hours. The second version, V2, builds on V1, with further separation of dependent variables to give gas consumption, electricity consumption, indoor and outdoor temperature, solar gains but with a temporal resolution of 30 mins. V3 surrogates are more complicated, reproducing diurnal patterns in the data, and they are explained in detail in section 6.3.

For the purposes of validation, we have generated multiple reference datasets [Kok et al 2016] for internal temperatures. These datasets can be used in the surrogates where indoor temperature needs to be simulated.

The list of observed variables in the datasets HH01-HH30 of Phase 2 report includes electricity and gas consumption, solar gain, internal and external temperatures and humidities.

7.1 V1 surrogates

This is the basic surrogate that can be used for validation of SMETERs that use daily values of observed variables to produce HTC estimates. The V1 surrogate contains the temperature difference, gas and electricity consumption, and solar irradiance.

Algorithm of surrogates V1

1- X_i^j	Load the matrix of SMETER measurements of Phase 2 report (houses HH01-30) with necessary processing: selecting the winter subset (November-March), conversion of cubic metres gas and kWh of electricity into average watts of power per day ; i is time index, j is index of variables, and the required quantities are solar irradiance, internal and external temperature (as required by different classes of SMETERs , e.g. remote only would not receive internal temperature data), gas and electricity consumption.

2- HTC value selection	Prescribe HTC value for surrogate (the current range of values in the supplied reference dataset has HTC values 100:50:500 W/K, i.e., from 100 to 500 with step 50, so there are nine datasets).
3- $Q = HTC \cdot \Delta T + S + \zeta$	Generate daily power Q time series using HTC (prescribed), the difference between internal and external temperature ΔT (simulated), and solar irradiance S (from the real dataset), with added noise ζ. The noise in V1 surrogates is Gaussian and represents the daily variability of power consumption due to weather conditions. Its amplitude defines the dispersion of the data cloud in the Siviour plot but does not affect the HTC estimation (the slope of the least-square fit remains the same). For realistic representation of the data properties in the surrogates, we generate power noise amplitude at about the same level as in real data, about 10% of standard deviation of the total power.
4- $Q = E + G$	Using single-value power ratio of the real data (ratio of mean values of time series), generate daily time series of electricity E and gas G surrogates. Because V1 surrogates are simple daily simulations that reproduce aggregated thermal performance, we assume that there is no loss of heat for the SMETERS that aggregate daily data for Siviour-based HTC estimates.
5- Y_i^j	Write a surrogate dataset simulated data (E, G, solar irradiance and temperature difference indoors-outdoors)

V1 is suitable for SMETERs that require daily-resolution input without diurnal variability. Currently, the existing SMETERS operate with sub-diurnal data (30-min resolution), but in case of a generic SMETER that may operate their own measurements with their own resolution, V1 surrogates may be more convenient to use for validation.

V1 surrogates use HH01-30 input with further modification of data based on varied HTC, and they can be considered to be as agnostic to the type of a building because of their minimal requirements on the input data. When they are simulated with different HTC values (varying from 100 to 500), they simulate expected power use based on the Siviour equation, and this implicitly assumes different thermal properties because of different HTC – this addresses different properties of the building, such as different glazing amount or composite materials.

V1 surrogates can be used as a “go-no go” test at the initial stage of SMETER assessment (if the SMETER requires higher-resolution data, it can be tested further with V2 and V3 surrogates). Surrogates are using environmental data from the house datasets and some dependencies of power use, but they are not equal to the observed 30 houses. By imposing different HTC and modelling power consumption, we create a modified model of the input house with different thermal properties. When we impose another HTC value, we obtain more surrogate data based approximately on real data. These data have patterns similar to those observed, but they are defined by the prescribed HTC with different level of energy use, assuming different thermal properties of the house.

V1 surrogates are illustrated in Fig. 9:

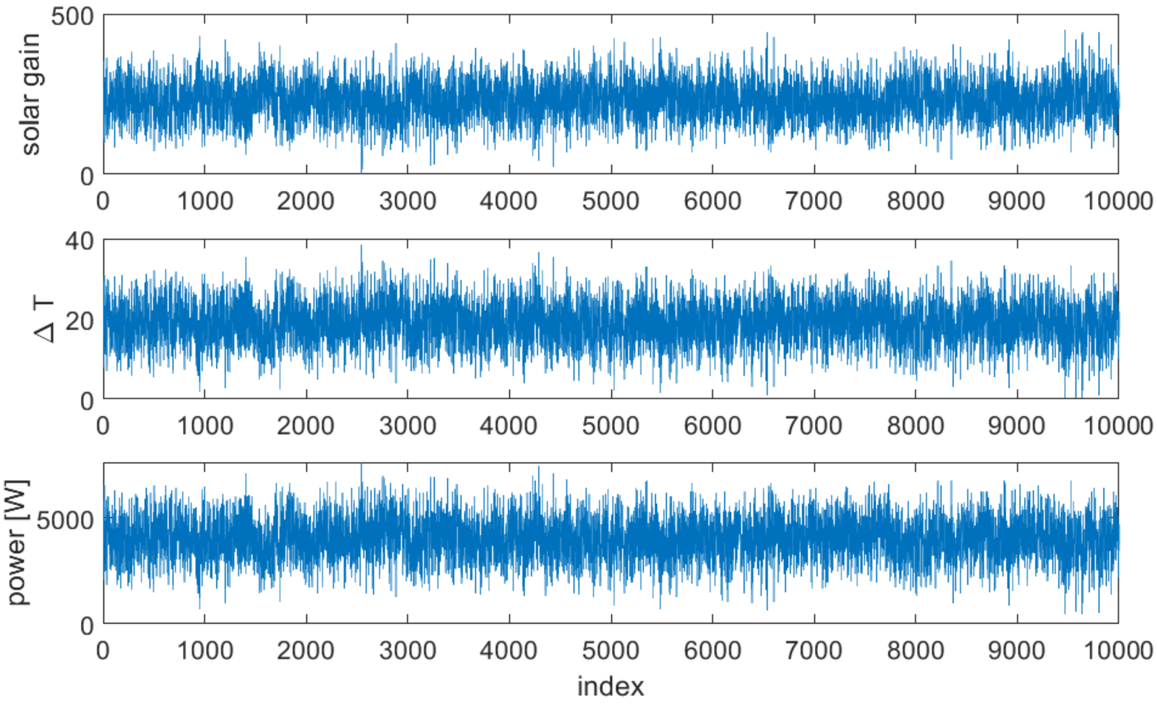

Figure 9 V1 surrogate with three main variables for HTC estimation. Here index represents a time step, which is expected to be one-day.

V1 surrogates are provided as datasets with daily time stamp, averaged daily gas and electricity consumption, solar irradiance and averaged temperature difference (indoor-outdoor). This is the minimal dataset required to estimate HTC using the Siviour approach.

7.2 V2 SURROGATES

This type of surrogate provides 30-minute resolution for those SMETERs that require such input without diurnal correlations in the data.

Algorithm of surrogates V2

1- X_i^j	Load the matrix of SMETER measurements of Phase 2 report (houses HH01-30) with necessary processing: selecting the winter subset (November-March), conversion of cubic metres gas and kWh of electricity into watts of power per day; i is time index, j is index of variables, and the required quantities are solar irradiance, internal and external temperature, gas and electricity consumption.
2- HTC value selection	Prescribe HTC value for surrogate (the current range of values in the supplied reference dataset has HTC values 100:50:500 W/K).
3- $Q = HTC \cdot \Delta T + S + \zeta$	Generate daily power Q time series using HTC, difference of internal and external temperature ΔT, and solar irradiance S, with added noise ζ. The temperature fluctuations of V2 surrogates have to be linked to variable external temperature in the way that reflects house heating response: when external temperature is lower, internal temperature has to remain the same, and thus the temperature difference ΔT must be higher – this means that observed external temperature and ΔT must be anti-correlated (i.e. increase in one variable corresponds to decrease in another). For realistic representation of the data properties in the surrogates, we generate power noise amplitude at about the same level as in real data, about 10% of standard deviation of the total power
4- E, G	Generate normally distributed 30-min (sampling rate $f = 48/\mathrm{day}$) time series of electricity and gas surrogates. Standard deviations are obtained from the ratio of variables in real data.

5- $Q = E_f + G_f$	Using power-ratio (gas/electricity) adjust variables to get modelled power as a sum (simple constrained optimisation)
6- sY_i^j	Write a surrogate dataset with replacement of corresponding columns with simulated data (E, G, and internal temperatures, whose number depends on the number of rooms in the input data).

V2 surrogates do not reproduce time series diurnal patterns. These surrogates can be suitable for those SMETERs that aggregate daily data but require high-resolution input, which is later averaged. V2 surrogates are derived from V1 surrogates to obtain higher-temporal resolution datasets without the need to reconstruct time series temporal organisation, i.e., dynamical auto-correlations and cross-correlations between variables. Both V1 and V2 surrogates are designed to work well for the steady state methods but not for dynamic methods. Neither have diurnal patterns, and the starting point of the day (midnight or whenever) are not relevant in these surrogates (in other words, the “beginning” of the day may be arbitrary).

We create a portfolio of reference datasets that allows one to test different types of SMETERS. There may be SMETERs that do not need diurnal complexity because of their algorithms. As we already know, V3 surrogates were successfully tested with simple models, and in future one can use only V3 surrogates. However, V1 and V2 surrogates can have their use, too: for example, V1 and V2 surrogates can be generated as much longer datasets, because for them external temperature may be simulated independently, too, and this may be useful for uncertainty quantification of suitable SMETERS.

V2 surrogates are illustrated in Fig.10, where one can see that the lowest outdoor temperature corresponds to highest energy use, according to the surrogate stochastic model.

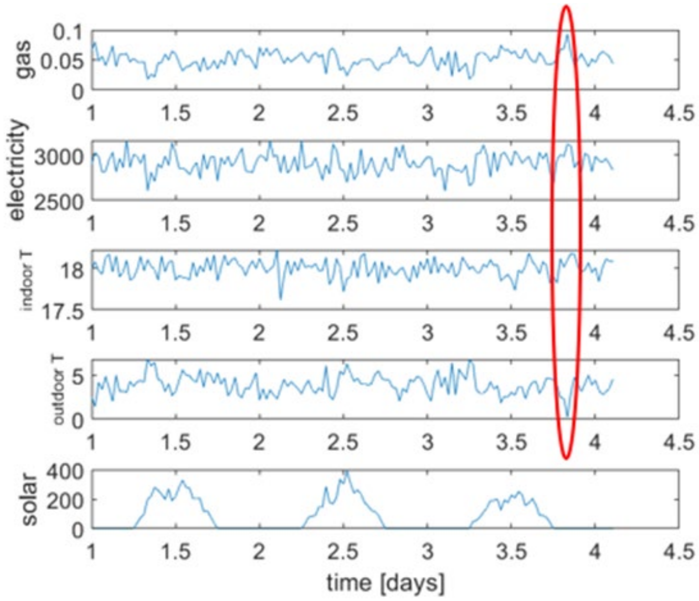

Figure 10 V2 surrogate with five variables generated with 30-min resolution. Red line shows abnormal event reflected in other variables using the stochastic model for surrogates.

V2 surrogates are more complex than V1 surrogates: they include all variables of the input datasets, where surrogate time series are internal temperatures in different rooms (according to the house HH01-HH30), and consumed gas and electricity. Other variables of the input dataset are not changed. The format of the input file is preserved, so that surrogate test could be used by SMETERS in the same way as the real data.

7.3 V3 SURROGATES

Algorithm of surrogates V3

1- X_i^j	Load the matrix of SMETER measurements of Phase 2 (houses HH01-30) with necessary processing: selecting the winter subset (November-March), conversion of cubic metres gas and kWh of electricity into watts of power per day; i is time index, j is index of variables, and the required quantities are solar irradiance, internal and external temperature, gas and electricity consumption.
2- HTC value selection	Prescribe HTC value for surrogate (the current range of values in the supplied reference dataset has HTC values 100:50:500 W/K).

3- $Q = HTC \cdot \Delta T + S + \zeta$	Generate daily power Q time series using HTC, difference of internal and external temperature ΔT, and solar irradiance S, with added noise ζ.
4- E, G	Generate log-normally distributed 30-min (sampling rate $f = 48/\mathrm{day}$) time series of electricity and gas surrogates.
6- $Q = E_f^s + G_f^s$	Apply sorting (see Fig. 11) to reproduce realistic diurnal fluctuations of the variables and rescale them using power ratio (gas/electricity) and optimisation to obtain the modelled daily power as sum.
7- $\mathrm{s}Y_i^j$	Write a surrogate dataset with replacement of corresponding columns with simulated data (E, G, and internal temperatures, whose number depends on the number of rooms in the input data).

The Siviour equation is used in all three surrogates, V1-V3, to generate the initial daily data, as this is the underlying physics of building performance at slow temporal scale (days and longer). The initial daily data is then further granulated in V2 and V3 for 30-min resolution, with added diurnal patterns in V3.

V3 surrogates are the most realistic that are presented in this report and are suitable for use by several types of SMETERs, including Siviour-based, grey-box, and black-box. Siviour-based SMETERS can use V1 or V2 surrogates (depending on what temporal resolution they require), and they also can use V3 surrogates after diurnal averages; ML-based SMETERS and greybox models can use V3 surrogates. These surrogates have realistic diurnal patterns of internal temperatures, gas and electricity consumption. Because gas and electricity usage are non-negative with large peak values, they are modelled using log-normally-distributed normal numbers on the sub-diurnal temporal scale.

For a defined computational problem, the characterization of the solution to that problem can be used to generate data for which the solution is known exactly. Often, there is redundancy in the generation of the data that can be exploited to generate data that is “close to” available real data. As an example, consider the (overdetermined) least-squares problem $Ax = y$ for a given design matrix A. For a prescribed (reference) solution x^* a corresponding set of (reference) data can be constructed trivially as $y^* = A \cdot x^*$. However, $y = y^* + e$ will also be a solution if $Ne = 0$ where N satisfies $A^T N = 0$, i.e., as a linear combination of the columns of N that define the null space. The multipliers in the linear combination can be chosen randomly (giving many sets of data having the same reference solution) or to minimise the distance between $y^* + e$ and y^0, where y^0 denotes an instance of real data.

This principle can be applied when, from a single time series of power (obtained as in V1 surrogates), one needs to obtain separate time series of electricity and gas consumption (gas consumption is converted to energy using the approximate calorific parameter 11.2 MJ/m3). The sum of these two time series has to be equal to total power generated as in V1; the ratio between gas and electricity can be taken from a real house (over the observed period, as a ratio of mean values of time series), and then rescaled in the constrained optimisation problem. This approach proved successful in V3 surrogates, with HTC estimated from generated data (using the Siviour approach) within 2-3% of the ground truth HTC value.

Due to the separation of scales in diurnal and sub-diurnal data and thermal mass effects described above, the Siviour approach works either if data are daily-averaged or if high-resolution data are used to calculate a daily average. ML-based and dynamic physics-based models, instead, use dynamical changes of high-resolution data to obtain information about dependencies of power consumption on temperature differences at sub-daily temporal scales. Because of this, V3 surrogates must be designed to reflect the realistic dynamic patterns of high-resolution variables. Diurnal averages of V3 surrogates must also satisfy the Siviour equation, in which the HTC value has to be controlled as ground truth.

V3 surrogates are complex and may require tailoring (i.e., further adjustment of the input variables, temporal resolution, such as 5-min instead of 30-min, and adding new variables for newly introduced measurements, such as heat metering) for a specific SMETER to take into account its set of input variables and to reflect the real building physics of house archetypes, where temporal organization of SMETER measurements would be different. The goal of an SMETER is to identify the response of the building's fabric from dynamical patterns of the data, and these patterns must therefore be reproduced in the surrogate time series. For that purpose, we use a surrogate sorting algorithm [Livina et al 2013]. This algorithm is illustrated in Fig.11, where a random dataset (middle panel) has to be rearranged to obtain the sine wave pattern (top panel). This is achieved by rearrangement and "sorting" its values according to the sine wave (20 points of about two periods of data with periodicity 12, which corresponds in diurnal data with sampling rate of 2 hours) —the result can be seen in the bottom panel, where random values are rearranged to follow the pattern of the sine wave.

In the sorting procedure, the time series must have the same number of data points —20 in this illustrative example, and the sine wave has two full cycles, i.e., $y(x) = \sin\frac{2\pi x}{10},\ x = 1..20$. The main goal of the sorting procedure is to modify the random, uncorrelated sequence (middle panel) into a correlated sequence of the same set of numbers (bottom panel). This means that sorting preserves the probability distribution of the input time series but changes the temporal correlations in such a way that they reproduce correlations in the target time series. This allows us to create realistic dynamic patterns in V3 surrogates that are similar to observed data. Fig. 12 shows a typical example of sorted surrogate data. V3 surrogates are the most realistic among the three types by construction.

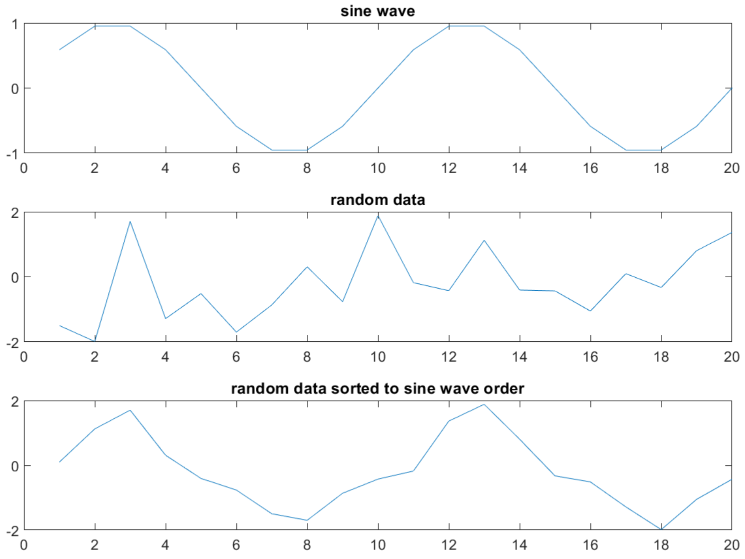

Figure 11 Experiment with artificial data to illustrate the surrogate sorting algorithm.

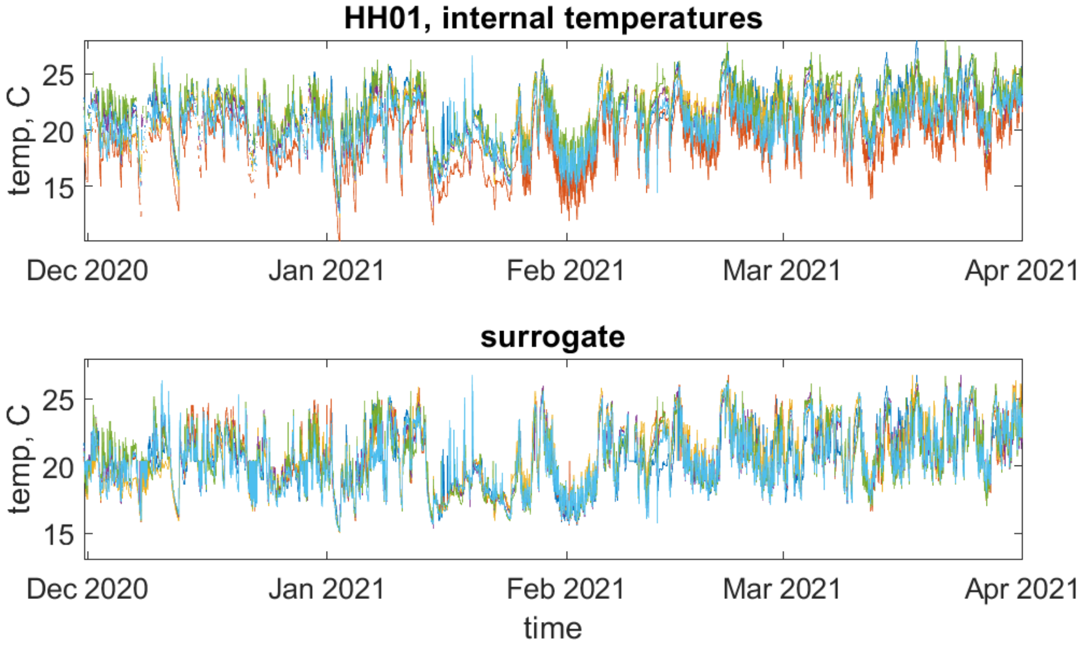

Figure D Comparison of observed temperatures in six rooms and derived sorted surrogates.

If V2 or V3 surrogates are averaged over a day, they will be similar to V1 (but not identical). V1 has a smaller number of variables, which may be convenient for some SMETERS. V2 has simplified diurnal patterns compared with V3, which may also be useful for some SMETER types. V1 and V2 surrogates can be simulated using random number generators without being based on real data, but the most realistic V3 surrogates use Phase 2 data (houses HH01-30) and for further development they need more real input.

7.4 V* SURROGATES

It is expected that in future SMETERS may add other variables that would help improve HTC estimates. There are plans for developing tailored surrogates that would incorporate currently unobserved variables, such as internal heat gains.

The flowchart in Fig.13 gives an overview of the SMETER validation process with the three different types of surrogates that are currently under consideration. The flowchart states 20% error as a preliminary "decision rule", but in future, based on negotiations with policymakers, this rule can be substituted.

This flowchart represents the currently available surrogates (provided as supplementary materials) and does not reflect, for example, a potential change of occupants or higher use of hot water or heating. However, it is not an immediate purpose of surrogates to compare performance of the same SMETER with different number of occupants in the same house; rather, the purpose is to test the accuracy of an arbitrary SMETER HTC estimation compared with the prescribed value in idealised conditions (i.e., assuming average occupancy with regular behavioural patterns in energy use). It is possible that an SMETER would pass a surrogate test but become inaccurate when applied across different real homes with unusual energy use. This stresses the need to develop more surrogates on a broader range of houses with more diverse usage profiles.

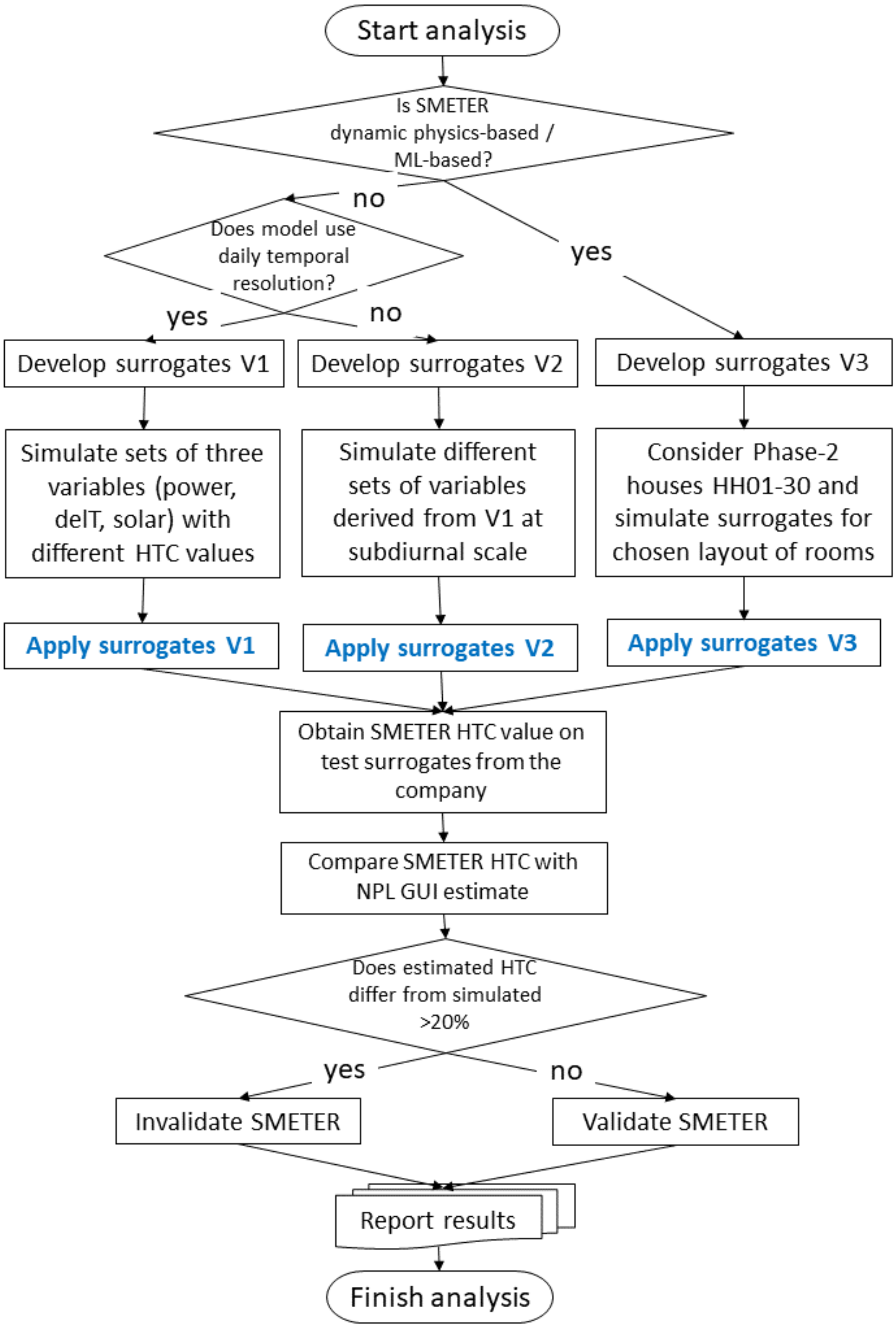
Start analysis
Is SMETER dynamic physics-based / ML-based?
no
yes
Does model use daily temporal resolution?
yes
no
Develop surrogates V1
Develop surrogates V2
Develop surrogates V3
Simulate sets of three variables (power, delT, solar) with different HTC values
Simulate different sets of variables derived from V1 at subdiurnal scale
Consider Phase-2 houses HH01-30 and simulate surrogates for chosen layout of rooms
Apply surrogates V1
Apply surrogates V2
Apply surrogates V3
Obtain SMETER HTC value on test surrogates from the company
Compare SMETER HTC with NPL GUI estimate
Does estimated HTC differ from simulated >20%
yes
no
Invalidate SMETER
Validate SMETER
Report results
Finish analysis

Figure 13 Workflow chart of SMETER analysis

8 Testing surrogates

8.1 Blind tests of V1 surrogates: Loughborough university

Three datasets of V1 surrogates were successfully blind tested twice by Loughborough University, estimating HTC as shown below. Each data set consisted of a 10,000-day sample of variable values. The prescribed values of HTC in the three datasets were 125, 97 and 145 W/K, and the estimates from the first Loughborough University blind test were:

DATA 1: HTC= 122.97 W/K

DATA 2: HTC= 96.29 W/K

DATA 3: HTC= 144.92 W/K

The second blind test experiment reported their results with uncertainty. Three linear regression methods were applied on a rolling basis to all available 21-day windows (a total of 9980 windows, i.e. SMETER estimations, from each 10 000-day data set).

SIV — The Siviour analysis approach, as defined in Eq. (2).

SIVP — Siviour plus regression (SIVP), for which the regression model was

$$Q + g \cdot A_{SIV} S = HTC \cdot \Delta T,$$

where gA_{SIV} is the solar aperture obtained through Siviour analysis.

MLR — Multiple linear regression (MLR) for which the regression model was

$$Q = HTC \cdot \Delta T - g \cdot A_{SIV} S.$$

A simple averaging method (AVG) was also applied, in which the HTC estimate H_AVG (W/K) was given by

$$H_{AVG} = \frac{\sum Q + g \cdot A \cdot \sum S}{\sum \Delta T} \tag{4}$$

where the sums are taken over consecutive days, and $g \cdot A$ is a constant solar aperture calculated for those days. Here, solar apertures were as obtained using the SIV and MLR methods as applied to daily mean data from the *n* relevant days; the corresponding HTC estimates are denoted $H_{AVG,SIV}$ and $H_{AVG,MLR}$ respectively. AVG method analyses were deemed valid when the following criteria were satisfied:

1 The mean indoor-outdoor temperature $\frac{1}{n}\sum \Delta T$ was no less than 10 K.

2 The magnitude of the estimated solar gains $g \cdot A \cdot \sum S$ was greater than 10 % of the total estimated heat gains $\sum Q + g \cdot A \cdot \sum S$.

Results failing either of these criteria were excluded from the analysis. The summary is given in Table 1 and Fig. 14. It is clear that the mean value estimates are very accurate, and different statistical estimates are in agreement between the applied techniques. ‘SD’ denotes standard deviation.

Table 1 Descriptive statistics for HTC estimates (W/K) of surrogates V1 obtained through analysis of 21-day long data intervals.

Dataset	HTC estimate (1)	Count	Mean (W/K)	SD (W/K)	Min (W/K)	Q1 (W/K)	Median (W/K)	Q3 (W/K)	Max (W/K)
data1	H_{SIV}	9980	125.0	2.6	114.1	123.2	124.9	126.7	135.4
	H_{SIVP}	9980	124.4	2.4	116.1	122.8	124.3	125.9	134.0
	H_{MLR}	9980	125.0	2.4	115.5	123.4	124.9	126.4	134.5
	$H_{AVG,SIV}$	8230	125.6	2.3	119.4	123.9	125.4	127.0	135.3
	$H_{AVG,MLR}$	8304	125.5	2.1	119.3	124.0	125.3	126.7	134.5
data2	H_{SIV}	9980	96.9	1.5	90.7	96.0	96.9	97.8	102.8
	H_{SIVP}	9980	96.3	1.3	91.1	95.4	96.2	97.0	101.5
	H_{MLR}	9980	97.0	1.1	92.8	96.2	97.0	97.7	103.1
	$H_{AVG,SIV}$	4293	97.7	1.2	94.3	96.8	97.6	98.4	102.6
	$H_{AVG,MLR}$	4258	97.5	1.1	94.3	96.8	97.4	98.1	103.0
data3	H_{SIV}	9980	145.0	3.0	134.1	143.2	145.1	146.9	157.6
	H_{SIVP}	9980	144.8	3.5	130.1	142.6	145.0	146.9	156.6
	H_{MLR}	9980	145.1	2.0	136.0	143.8	145.1	146.4	154.1
	$H_{AVG,SIV}$	7518	145.9	2.2	139.7	144.4	145.7	147.3	156.3
	$H_{AVG,MLR}$	7758	145.5	1.8	139.6	144.3	145.5	146.7	154.3

(1) Subscript indicates HTC estimation method applied: SIV = Siviour; SIVP = Siviour plus regression; MLR = multiple linear regression; AVG,SIV = average method with solar aperture from Siviour analysis; AVG,MLR = average method with solar aperture from multiple linear regression analysis.

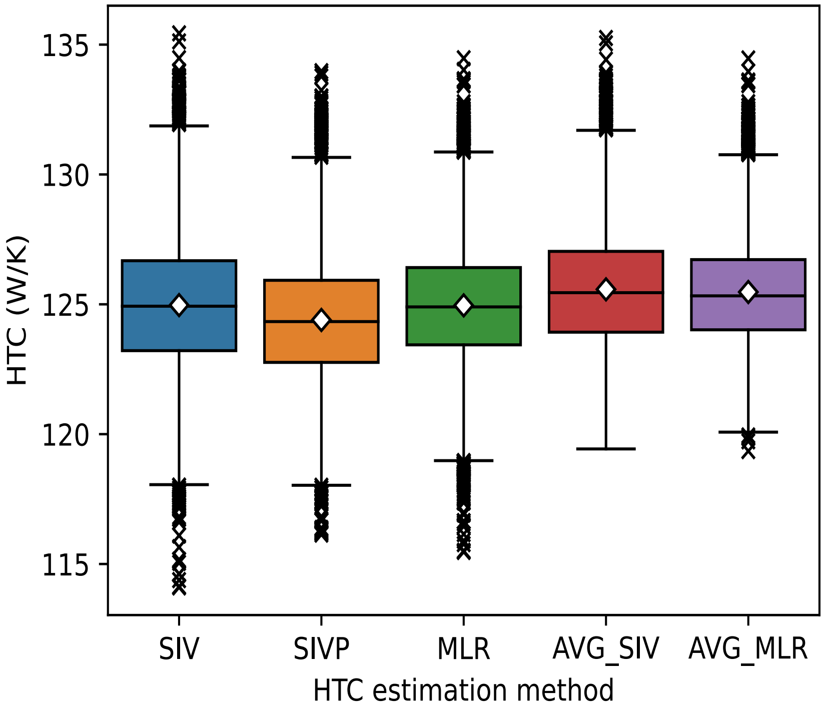

Figure 4 Distributions of HTC estimates calculated for dataset 1 using each HTC estimation approach. (SIV = Siviour; SIVP = Siviour plus regression; MLR = multiple linear regression; AVGSIV = average method with solar aperture from Siviour analysis; AVGMLR = average method with solar aperture from multiple linear regression analysis.) Box bounds indicate lower quartile Q1, median Q2 and upper quartile Q3; whiskers indicate minimum and maximum values excluding outliers (values falling outside of [Q1 - 1.5(Q3-Q1), Q3 + 1.5(Q3-Q1)]); ◇ = mean; × = outlier.

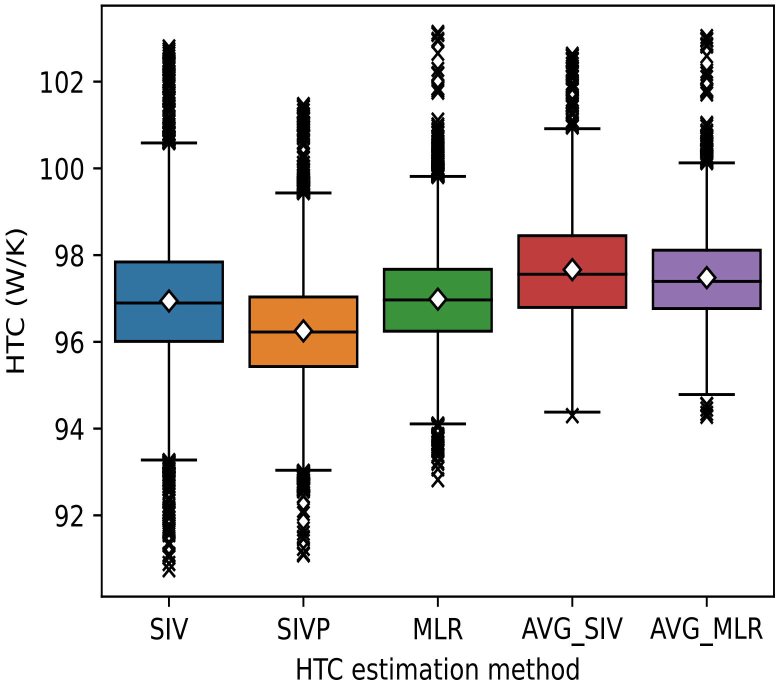

Figure 15 Distributions of HTC estimates calculated for dataset 2 using each HTC estimation approach. Methods and markers are as in Fig.14

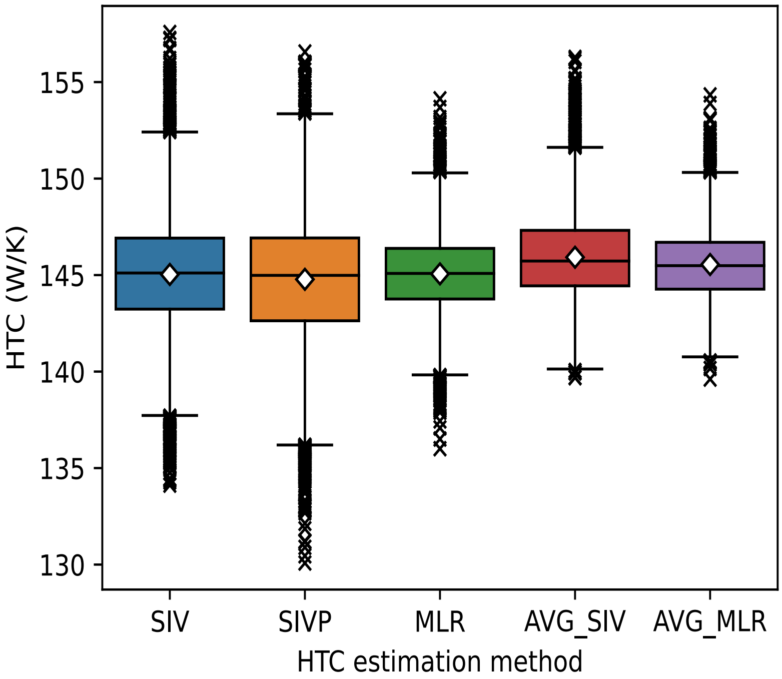

Figure 16 Distributions of HTC estimates calculated for dataset 3 using each HTC estimation approach. Methods and markers are as in Fig.14.

8.2 Blind Tests of V3 surrogates: Loughborough university

A dataset of V3 surrogates was successfully blind tested by Loughborough University using the approach described above. House HH03 was considered, and the prescribed value of HTC was 220 W/K. The blind test results were reported (Table 2), and mean value estimates of HTC were satisfactory, given the complexity of V3 surrogates. They applied several techniques of HTC estimation as follows.

Table 2 Descriptive statistics for HTC estimates (W/K) of surrogate V3.

HTC estimate (1)	Mean (W/K)	Min (W/K)	Median (W/K)	Max (W/K)
H_{SIV}	229.1	202.6	224.7	269.1
H_{SIVP}	238.2	218.1	239.2	250.1
H_{MLR}	217.8	182.4	216.2	254.5
$H_{AVG,SIV}$	224.3	198.6	221.6	262.8
$H_{AVG,MLR}$	222.8	161.8	221.0	258.8

(1) Subscript indicates HTC estimation method applied: SIV = Siviour; SIVP = Siviour plus regression; MLR = multiple linear regression; AVG,SIV = average method with solar aperture from Siviour analysis; AVG,MLR = average method with solar aperture from multiple linear regression analysis.

8.3 Blind test of V3 surrogates

The SMETERs mentioned in this section use different approaches: Loughborough colleagues apply Siviour and equivalent techniques; SMETER C has an extended Siviour-based portal that uses detailed input of a building and advanced solar gain analysis; SMETER B uses dynamic models (grey box), and SMETER D uses an ML-based approach.

Within the term of the project, after several discussions with stakeholders, further tuning was applied to V3 surrogates to address intermittent gas usage and rescaling due to the sampling rate of the available data. Data from this surrogate type with a value of HTC = 155 W/K were provided to several stakeholders for a blind test. Because these surrogates were derived from observed data in a specific winter season (2020/2021), the size of the dataset was much smaller than in the experiment with V1 surrogates: the number of days in the cold period was 151, which produced surrogates with 30-minute step of 7248 datapoints. The result was then estimated from data using the Siviour approach as 153.9. The participants returned the following results:

SMETER A	SMETER B	SMETER C	SMETER D
~173±3 W/K	~170±17 W/K	160 [-13,+12] W/K	154 W/K

SMETER A, provided detailed estimates of HTC:

Table 3 Descriptive statistics for HTC estimates (W/K) of adjusted surrogate V3 by SMETER A, using techniques described above.

HTC estimate	Count	Mean (W/K)	SD (W/K)	Min (W/K)	Q1 (W/K)	Median (W/K)	Q3 (W/K)	Max (W/K)
H_{SIV}	102	176.8	17.3	149.7	166.2	172.7	180.6	224.2
H_{SIVP}	102	171.2	16.8	147.5	161.8	167.7	173.0	221.0
H_{MLR}	102	170.1	16.1	146.8	161.5	167.4	173.1	218.4
$H_{AVG,SIV}$	99	173.9	17.3	148.6	163.7	170.3	176.3	222.6
$H_{AVG,MLR}$	93	172.9	17.1	147.8	163.1	169.6	176.2	220.0

The model used by SMETER B is a “grey-box” model that combines Bayesian methods and lumped thermal capacitance model [Hollick et al 2020]. The model of SMETER D is a ML-based model trained on real data from the Phase 2 report. Interestingly, on this surrogate dataset, as the blind test shows, all these models of different types provide the result that is within 10-15% of the imposed ground truth values.

As this blind test shows, then, the developed surrogates provide common grounds for validation of diverse SMETERS.

9 Data collection plan for future SMETERS

9.1 Data quality requirements for future SMETERS

Future SMETERs may differ from current ones in various ways. First, they may use new variables - in particular, those related to behavioural patterns (binary variables for open/closed window, presence/absence of people in the room; continuous variables for human heat, cooking, etc.). Second, SMETERs may gather data at higher temporal resolution, such as 5 minutes or higher. Third, they may deploy new algorithms such as novel deep neural networks.

The physical principles of heat transfer remain the same, however, and increased SMETERs a complexity or novelty will not be advantageous if conventional SMETERs perform with equivalent accuracy. Validation of future SMETERs must be based on the approaches that are as neutral and flexible as the surrogates proposed here.

The currently available data is rather small in terms of building type and duration of representative (winter) records. Because V3 surrogates require realistic input for reproducing diurnal patterns, such input data needs to be abundant and diverse. For gathering future records, ideally it would be useful to obtain data:

For house archetypes that have significant features in the HTC context that are not currently captured in the Phase 2 datasets and would be expected to impact on the SMETER algorithm, e.g., flats (given their coupling with surrounding properties)

- At high temporal resolution, which will provide sufficient statistics for analyses
- Which is as complete as possible, without a large number of missing numbers
- With aligned periods of observations and - if possible - the same regional location, such that houses can be distinguished based on their material properties rather than on different environmental conditions
- Over a period without weather anomalies, which may cause atypical behaviour of buildings (for example, abnormal precipitations in a particular year). Anomalous weather may cause different response of the building fabric (for example, when the temperature drops below zero and a building has capillary effects in its fabric) – to model such fluctuations in the surrogates may require adjustment of probability density of generated random numbers.
- With additional measurements that may help understand heat transfer (for example, wall temperature measurements from adjacent properties, heat dissipation from appliances, windows thermal leakage)
- With realistic efficiencies of appliances, including gas boilers and additional heaters, or distributed heat measurements
- With additional variables and metadata specific to a household (for example, presence of pets that add occupancy heat, cooking patterns using gas or electricity, exercising and physical activities, etc.)

Given such additional data, it would be possible to expand the surrogate approach by extending the dimensions of the input data flow with more variables, and with modelling specific surrogate variables based on real data, similar to modelling in V3 surrogates (power-law correlations and sorting of new measurements). Such an extension is straightforward mathematically, and specialists in stochastic modelling could undertake such work in subsequent DESNZ projects on HTC as necessary. The impact of some of these variables on energy use, however, is not fully understood and modelled, and therefore care in development of surrogates that represent realistic situations is necessary.

9.2 Possible solutions for incomplete or inaccurate data

It was noted during the project that missing data may affect HTC estimates, which are encountered in 5-10% of the currently available real data. To alleviate this effect, it is possible to replace missing values with either mean values of time series or with realistic random values. When surrogates are generated, by construction, missing numbers are reproduced in synthetic data in the same locations as they were observed in the real data. For improving such surrogates, close work with specific SMETERs is essential, because surrogates must be tailored to their layout and input requirements. Missing data is common in this field. SMETERs must be able to deal with it with some proprietary solutions, and therefore it may be necessary to test how they handle it. By construction, surrogates reproduce gaps in data according to input, so missing values should be handled before generating surrogates if there are large gaps.

Empirical data collection should be accommodated by gathering the full documentation on sensors to ensure correct uncertainty propagation. A large component of uncertainty is due to how the measurements represent the real building. This error in the conceptualisation of the system/measurands needs to be considered. When such documentation is missing, measurement uncertainty may be assumed less accurate than when it is documented by sensor manufacturer.

Co-heating tests are expensive and cumbersome, and therefore they should be kept to a necessary minimum. As such, co-heating must be scheduled in representative houses (in terms of building type and environmental conditions): one per type – to minimize the effort and cost. At the same time, co-heating test data must be representative and sufficient to provide accurate estimates of HTC in typical housing stock, even if some marginal types of building are not covered. Careful consideration of co-heating is required to manage the cost, and the complexity of the later operation of tested homes needs to be considered.

10 Errors in SMETER validation

The sources of errors in SMETER validation may have various origins. Input data errors due to instrumental issues, incomplete observations, and poorly sampled empirical distributions may lead to discrepancies between observed time series values and SMETER model determinations. There may be also discrepancies due to unobserved biases and unaccounted heat losses/gains. All these will affect HTC estimation; some errors are reducible, and others are irreducible. The validation procedure assesses this based on provided documentation about instruments (see the example in Fig. 17, where documented instrumental errors are used for sensitivity experiment).

Energy consumption depends on several components. In addition to the thermal inertia of the building, there may be other powerful devices within the building such as gaming computers or plasma TVs [PowerDB]. Unlike boilers or hot water cylinders, they are not insulated and contribute to immediate heat exchange.

Dynamic methods can cope with various heat inputs that are measured. The bigger problem is usually the unmeasured things, e.g., the heat from people, the heat vented straight outside by a tumble drier, the heat in hot water that goes down the drain, assumed efficiencies of conversion to heat, etc.

Behavioural patterns of heating use may affect energy consumption (individual choices of indoor temperature and ventilation), thus affecting HTC estimation based on energy input. For example, some people may want to have a warmed house but with continuous supply of fresh air by natural ventilation – this will require increase of heating in winter, even though the building fabric is not responsible for such heating setup. Furthermore, it could also be the case that the HTC of a specific dwelling could change markedly in response to even a single change to its configuration. By configuration here is meant the entire set of permutations with respect to how many doors and/or windows are open and blinds and/or shutters are closed. For instance, a building with a low HTC when a certain door is closed could have a significantly greater HTC when it is left open if the door in question is constructed from a very thermally insulating material and leads into a room having external walls constructed from material which is a good thermal conductor.

Houses can have complex dynamics, made even more complex and less predictable by the presence of people living in them: something as simple as opening or closing a door can drastically change the way air moves into, out of, and within the house, as can many other aspects of residents' behaviour. If this is an occasional event, from the point of view of HTC estimation, this fluctuation will be averaged out when considering a dataset of sufficiently long duration (or not, if the door is open continuously over long period of time). In terms of the number of observables, it is not reasonable to expect acquisition of data which allows one to fully model every factor of potential relevance inside a house: in fact, this is usually simply unfeasible. Instead, we suggest analysing factors such as ventilation rate, temperature gradient and energy consumption supported by sensitivity analysis for uncertainty propagation. When using energy consumption as a proxy variable for HTC estimation, we implicitly assume that most if not all electricity and gas is used for heating. This is an irreducible error.

11 Uncertainty and sensitivity analysis

We propose controlled experiments with varying input variables (these experiments are linked to the practical sources of variability listed above) for sensitivity analysis as the basis for acceptable uncertainty in HTC estimates.

We performed an experiment with induced instrument error of various magnitudes up to 10 % of the input variables (such error is usually reported in instrument documentation, but with time any instrument may drift and require calibration). We can conclude that such instrumental error can affect HTC estimates discernibly:

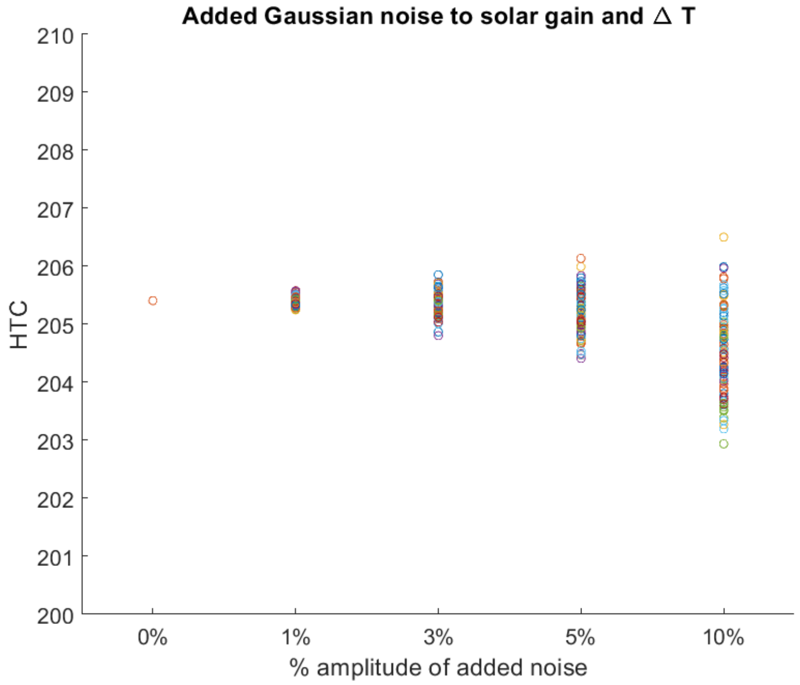

Figure 7 Effect of instrumental noise in input variables on HTC estimate in artificial data with simulated HTC = 205 W/K.

Fig.17 shows that an input error due to instrument variability up to 10 % may cause deviation in the HTC estimate up to 5 %. In this example, the input errors are assumed to be uniformly distributed, and estimates of HTC are performed using Monte Carlo simulations based on the Siviour equation. This experiment with artificial data demonstrates the effect of input variability for a single representative HTC value, and the same approach may affect other HTC values differently—this may be further investigated in future, and experiment can be conducted on real data for some of the houses. Probability distribution of instrumental errors significantly affects the propagated uncertainty, which has to be taken into account in designing data gathering and analysis.

This experiment demonstrates the increased uncertainty in HTC estimate due to instrumental error within the documented sensor variability. If this variability is systematic, it will persist over any period of observation, and this can be reduced only by sensor replacement or recalibration.

Uncertainties on input values of both physics-based and ML models must be propagated correctly to arrive at a valid final estimation of the uncertainty in HTC. These analyses facilitate

- identification of variables that contribute significant uncertainty to HTC in order that they can be further investigated with a view to constraining them in a physically legitimate way
- the identification of correlations between variables to reduce redundancy
- possible model simplification following the identification of any correlations and discarding any variables that make an insignificant contribution.

We have developed the following:

- A large collection of reference datasets of V1, V2 and V3 surrogates suitable for testing a large range of SMETERS
- Uncertainty analyses using available input variables in comparison with test experiments using surrogates.

For uncertainty quantification and surrogate tests, we have developed a GUI-driven software tool. The first version of this was demonstrated to DESNZ in July 2023.

12 Future work

As the blind tests have shown, the developed surrogates provide common ground for validation of diverse range of SMETERS. As the next step, it would be interesting to conduct similar blind tests at large scale (with more surrogate datasets and with more diverse SMETERS).

V1 and V2 surrogates can be generated in abundance, as they do not require detailed realistic input (although may use it). V3 surrogates are based on real data with diurnal patterns, and the greater the set of domestic dwellings, the greater their diversity (greater diversity is desirable). V3 surrogates are most suitable for ML-based SMETERS, which may require more data for validation, as more SMETERS of this type are expected to appear in the market with growth of AI technologies.

We understand that future SMETERs may introduce new measurements, from which they will obtain HTC estimates. For example, heat meter measurements can be used instead of total power. To develop surrogates V* for such SMETERS, as an alternative experimental approach (in addition to Siviour and equivalent methods), one can use Fourier's law of thermal conduction to approximate heat variable in surrogates:

$$\frac{\Delta Q}{\Delta t} = UA(-\Delta T), \qquad (5)$$

where U is conductance and A is cross-sectional area. Alternatively, dynamic models can be used – this is a subject of further investigation. Direct measurement of heat by a SMETER (using an additional heat meter, as implemented by some SMETERs on the market) could be compared with heat estimates derived from temperature variability and prescribed HTC, and this would constitute a basis for validation of such SMETERs. Given that heat meters may have measurement error up to 10 %, the instrumental error can also be used for sensitivity analysis [Venton and Wright 2023]. All SMETERs that differ from conventional ones in their sets of measurements will have to provide their measurement data and related documentation to be used for generation of validation surrogates.

As possible future work, NPL suggests the following:

- Requirements during data collection, advising on data quality requirements (missing values, documented instrumental errors, arrangement of sensors, their number and necessary measurements)

- Collation of existing data sets that can be used to assess the accuracy and repeatability of SMETER methods (good candidate data sets with 100's of homes)
- When co-heating test are planned, they should be adapted to better represent the boundary conditions (regions of measurements) of a home specifically for the validation of SMETERs. This may include redesigning the sensor network inside the building to avoid any biased measurements in building's subdivisions.
- Gathering data at higher spatial resolution (for example, having a temperature sensor in every room would be a significant advantage)
- Collection of new primary data to test the accuracy and repeatability of SMETERs in a wider proportion of the housing types e.g., flats and new-build houses, heat metering for gas boilers and heat pumps. Ideally, heat meter measurements would replace gas consumption data where gas boilers are in use; and a proportion of the electricity consumption. (immediate extension potentially with Loughborough University)
- Obtain, where possible, probability distributions of instrumental errors
- Gather longitudinal data to understand how HTC changes over time and due to changing boundary conditions
- Collect new primary data (various sizes, types and use profiles) to test the accuracy and repeatability of SMETERs used before and after retrofit
- The dependencies of variables that affect the rate of heat transfer (wind speed and direction, absolute temperature of the air inside and outside, cloud cover, solar irradiance), as well as the moisture content of materials may or may not have a big impact on the HTC estimates, and this lack of knowledge should be addressed in future data gathering and analysis
- Integration of SMETER-type data collection into any suitable government-funded field trials to continue the gathering of evidence and the assessment of benefits
- Identifying reducible and irreducible errors in input variables based on the specification of instruments provided by SMETERs developers
- The continued development of robust methods to estimate the uncertainty of SMETER methods, maintaining digital surrogates as a product for future use for blind tests via NPL procedures
- Tools (policy and consumer) for cost/benefit analysis for heating/energy intervention including uncertainty effects.

To realise data gathering described above, carefully set up experiments are necessary.

The surrogates approach is flexible and can be further modified to address new measurements based on the statistical properties of the observed data. Currently, V3 surrogates are based on 30 datasets of the Phase-2 SMETER report. The homes used in the project did not represent fully the UK housing stock due to funding limitations; there are further plans, however, to perform measurements in other types of buildings, and that data will be analysed in due course. When other variables are added in future, surrogates can be tailored in the same way as V3 surrogates were developed from the initial V1 approach.

The surrogates can also be offered to the stakeholders for use in self-validation exercises, which later could be assessed by NPL.

13 Summary

The aim of the project was to develop a generalised approach to SMETER validation with respect to HTC estimation. By analysing the available data and developing the surrogate methodology, NPL has succeeded in this aim with blind tests carried out by academic colleagues and those SMETERs developers that were prepared to participate. We pursued the approach of stochastic modelling with surrogates, in which HTC values are controlled for tests. These surrogates can be tailored for specific SMETERs, including possible future innovations, while they remain based on ground-truth and building physics aspects.

The conceptual difference between SMETERs is that they may assess the thermal properties of a building in different ways, and this is addressed in the validation approach by the creation of different types of surrogates. HTC, however, is a parameter that must be independent from the approach that is used to estimate it, and therefore ground-truth validation based on surrogate datasets must be introduced. V1 surrogates are based on building physics principles, and V2 extend that to the same type of SMETERs that require higher temporal resolution. V3 surrogates address heat storage and dynamics using observed data in a controlled way. Although the resulting surrogates vary in the number of variables and temporal resolution, this approach achieves the main goals of the validation project: it is based on building physics principles and reproduces observed properties of measured variables in real houses.

HTC estimation and validation have high impact in policymaking. The immediate challenge of validating HTC is how to map the HTC uncertainty to retrofit schemes. In the blind tests we performed with SMETER partners, the error of HTC on controlled surrogates was within 5-10 %. These included ML-based and grey-box models, which means that surrogates are agnostic to the SMETER technology. A large error in an HTC estimate may bring a house from one retrofit requirement level (light to deep retrofit) to another, and it is therefore important to define the permissible uncertainty ranges for specific types of houses in the context of policymaking. This will require further experiments and data gathering in the context of identifying tiers of HTC for retrofit recommendations.

14 Supplementary materials

14.1 Graphical user interface for HTC estimation

NPL has developed a GUI for HTC estimation that can be useful for policymakers who would like to quickly assess HTC from some input dataset. The GUI was developed in Python and is accompanied by a Jupyter Notebook that can be deployed in a web-browser, for example, in Google Colab without installed Python on a specific computer. The software package has been audited against NPL's software quality procedures to a high integrity level and is supplied with associated quality documentation.

14.2 Reference surrogate datasets

The described surrogates V1-V3 have been provided as a reference dataset in the form of csv-files, which include simulated HTC values in the range 100:50:500 W/K. Because surrogates are generated using random numbers, it is possible to produce multiple time series for each selected value of HTC.

The reference dataset with undisclosed HTC values can be used by governmental policymakers for blind tests of SMETERs. Furthermore, surrogates with disclosed HTC values can be provided to SMETER companies for self-validation if they choose to make use of them.

15. References

[Allinson et al 2020] Allinson et al, Technical evaluation of SMETER technologies (TEST) project, Phase 2 field trial report, 2020.

[Bennett et al 2019] Bennett G, C. Elwell, T. Oreszczyn, Space heating operation of combination boilers in the UK: The case for addressing real-world boiler performance, Building Services Engineering Research and Technology, 40(1), 75-92 (2019).

[Gori and Elwell 2018] Gori V. and C. Elwell, Estimation of thermophysical properties from in-situ measurements in all seasons: Quantifying and reducing errors using dynamic grey-box methods, Energy and Buildings 167 (15), 290-300 (2018).

[GUM 1995] Guide to the Expression of Uncertainty in Measurement, JCGM 100:2008(E), Bureau International des Poids et Mesures, https://www.bipm.org/en/committees/jc/jcgm/publications

[GUM 2008] Evaluation of measurement, data — Supplement 1 to the "Guide to the expression of uncertainty in measurement" — Propagation of distributions, JCGM 101:2008, Bureau International des Poids et Mesures, https://www.bipm.org/en/committees/jc/jcgm/publications (accessed on 30th October 2023)

[Hasselmann 1976] Hasselmann K., Stochastic climate models, Tellus XXVIII 6, 473-485 (1976).

[Hollick et al 2020] Hollick F., V. Gori, C. Elwell, Thermal performance of occupied homes: a dynamic grey-box method accounting for solar gains, Energy & Buildings 208 (2020)

[ISO22989] International draft standard ISO DIS 22989:2021, Information technology — Artificial intelligence — Artificial intelligence concepts and terminology (2021)

[Jack et al 2018] Jack R., D. Loveday, D. Allinson, and K. Lomas, First evidence for the reliability of building co-heating tests, Building Research & Information 46 (4), 383-401 (2018).

[Jack 2019] Jack, Richard. 2019. Building Diagnostics: Practical Measurement of the Fabric Thermal Performance of Houses. PhD thesis, Loughborough University, https://hdl.handle.net/2134/19274 (accessed on 30th October 2023)

[Johnston et al 2013] Johnston D., D. Miles-Shenton, D. Farmer, J. Wingfield, Whole House Heat Loss Test Method (Coheating), Leeds Beckett University, 2013, https://www.leedsbeckett.ac.uk/-/media/files/research/leeds-sustainability-institute/coheating-method-for-whole-house-heat-loss/lsi_cebe_coheating_test_method_june2013.pdf (accessed on 30 November 2023)

[Kok et al 2016] Kok G., P.M. Harris, I.M. Smith, and A.B. Forbes, Reference data sets for testing metrology software, Metrologia 53, 1091-1100 (2016).

[Livina et al 2013] Livina V. et al, Forecasting the underlying potential governing the time series of a dynamical system, Physica A 392 (18), 3891-3902 (2013).

[Makse et al 1996] Makse H., S. Havlin, M. Schwartz, and H. Stanley, Method for generating long-range correlations for large systems, Phys. Rev. E 53, 5445 (1996)

[NPL 2023] Validation of SMETERs: Intermediate Milestone Report, NPL Report COM 2529, 2023.

[PowerDB] Power consumption database: plasma TVs, https://www.tpcdb.com/list.php?type=7 (accessed on 30th October 2023)

[Rhee-Duverne and Baker 2013] Rhee-Duverne S. and P. Baker, Research into the thermal performance of traditional brick walls, English Heritage Research Report, 2013, https://historicengland.org.uk/research/results/reports/6935/ResearchintotheThermalPerformanceofTraditionalBrickWalls (accessed on 30th October 2023).

[Schreiber and Schmitz 2000] Schreiber T. and A. Schmitz, Surrogate time series, Physica D 142, 346-382 (2000).

[Smirnov 1948] Smirnov N., Table for estimating the goodness of fit of empirical distribution, Annals of Mathematical Statistics 19 (2), 279–281 (1948).

[Venton and Wright 2023] Venton J. and L. Wright, Good practice in sensitivity analysis for metrology, NPL NMS report 2245, March 2023.

This publication is available from: www.gov.uk/desnz [replace with direct URL if known]

If you need a version of this document in a more accessible format, please email alt.formats@energysecurity.gov.uk. Please tell us what format you need. It will help us if you say what assistive technology you use.