Contents lists available at ScienceDirect

Energy & Buildings

journal homepage: www.elsevier.com/locate/enb

Closing the energy flexibility gap: Enriching flexibility performance rating of buildings with monitored data

Manuel de-Borja-Torrejon [a,b,*], Gerard Mor [c], Jordi Cipriano [c], Angel-Luis Leon-Rodriguez [a], Thomas Auer [b], Jenny Crawley [d]

[a] *Instituto Universitario de Arquitectura y Ciencias de la Construcción, Escuela Técnica Superior de Arquitectura, Universidad de Sevilla, Av. de Reina Mercedes, 2, 41012, Sevilla, Spain*

[b] *Technical University of Munich, TUM School of Engineering and Design, Chair of Building Technology and Climate Responsive Design, Arcisstrasse 21, 80333, Munich, Germany*

[c] *Centre Internacional de Mètodes Numèrics a l'Enginyeria, Building Energy and Environment Group, Pere de Cabrera 16. Office 2G, Lleida 25001, Spain*

[d] *University College London-UCL, The Bartlett School of Environment, Energy and Resources, Central House, 14 Upper Woburn Place, London WC1H 0NN, United Kingdom*

ARTICLE INFO

Keywords:
Energy Flexible Buildings
Energy Flexibility
Flexibility Gap
Performance rating
Demand response
Demand side management
Flexibility indicators
Building labelling

ABSTRACT

Quantifying and rating energy flexibility in existing buildings will become increasingly important as building energy services become electrified. Flexibility ratings based on building design specifications have shown potential to complement energy performance certificates and enable the comparison between buildings. However, relying on physical models and standard boundary conditions could lead to a 'flexibility gap': a difference between predicted and actual flexibility. This article investigates the incorporation of monitored data into design-based flexibility ratings, using an existing rating methodology and two UK case study domestic buildings. We firstly examine whether the current rating methodology can accept monitored data, and find it is able to apart from the final step of rating. We then devise two methods of calculating the metrics required for the flexibility rating, based not on physical models but on data. Using these methods, we examine the impact of the standard operational modelling assumptions on the flexibility metrics compared to using data-informed inputs, which highlights some discrepancies and some concepts in the flexibility rating methodology for which monitored data may be very difficult to obtain (e.g. recovery time). Finally, we suggest how to improve the usefulness of flexibility ratings by incorporating additional information based on monitored data.

1. Introduction

1.1. Energy flexibility ratings for buildings

Energy Flexibility (EF) of buildings refers to the ability of buildings to adapt their energy consumption without adversely affecting functionality or occupants, for example to support load regulation in low voltage electricity grids [1]. Building users and the power sector can take advantage of this ability to react to fluctuating renewable energy availability and energy prices, and improve supply reliability while lowering carbon emissions and energy costs [1–3]. As wind and solar energy and the electrification of the thermal demand of buildings are expanding, the promotion of EF and the interaction between the building and the power sector is growing [4]. In this context, quantifying and rating the EF of buildings is becoming increasingly important, and EF indexes are being used addressing different focus scales such us

Abbreviations: ARX, Autoregressive Models with Exogenous Variables; CV(RMSE), Coefficient of Variance of the Root Mean Square Error; EF, Energy Flexibility; EF Gap, Energy Flexibility Gap; EF Gap $_{\mathrm{MAE}}$, Mean Absolute Error of the Energy Flexibility Gap; EPC, Energy Performance Certificate; EFSI, Expected Flexibility Saving Index; DHW, Domestic Hot Water; DR, Demand Response; E_{DR}, Actual Energy Variation [kWh$_e$]; HP, Heat Pump; FF, Flexibility Function; FI, Flexibility Index; FPI, Flexibility Performance Indicator [-]; GOF, Good of Fit; HA, House A; HB, House B; HB1, House B before user adjustment of thermostat; HB2, House B after user adjustment of thermostat; MAE, Mean Absolute Error; NMBE, Normalised Mean Bias Error; $\dot{P}_{res}$, Committed Power [kW$_e$]; PSS, Peak-Shaving Strategy; RMSE, Root Mean Square Error; SRI, Smart Readiness Indicator; TMY, Typical Meteorological Year; t_{res}, Response Time [h]; t_{rec}, Recovery Time [h]; WD, Weekday; WE, Weekend Day.

[*] Corresponding author.
E-mail address: mdeborja@us.es (M. de-Borja-Torrejon).

https://doi.org/10.1016/j.enbuild.2024.114141
Received 5 September 2023; Received in revised form 1 April 2024; Accepted 2 April 2024
Available online 3 April 2024

neighbourhoods [5,6], single buildings [3,7] and building technologies and control alternatives [7–9].

In order to characterize and incentivize EF in buildings, quantification and rating methods have been proposed [1–3,10]. They can be classified into two main approaches: one assesses the capability to behave flexibly, and the other EF performance. The multi-criteria evaluation scheme for rating buildings based on the Smart Readiness Indicator (SRI) introduced by the European Union [10,11] belongs to the first group. The SRI rating scheme consists of documenting what technology is present in a building to enable a flexible and energy-efficient operation. Based on this information, an SRI score is assigned, representing the readiness of a building for such an operation. While the SRI scheme fosters EF in buildings, it does not assess how much flexibility is achievable nor whether it is achieved in practice. This approach has been criticized[1] [2] and future versions are likely to incorporate the in-use performance of the building [10,12], so it will not be a focus of this article.

The approach that assesses EF performance considers the building's energy demand when it behaves flexibly – flexible profile – compared to a baseline without flexible behaviour [1–3,5,7]. In order to compare time-coincident profiles with identical boundary conditions (e.g. weather and occupancy), energy models are commonly used [2,5]. This approach can be applied at the operational or the design level, to respectively quantify and rate the EF of the building in use or according to design specifications [3,13]. To quantify EF, two perspectives have been identified in the literature [1]: one is based on quantifying the EF that the building can offer, and the other on quantifying the impact of using the EF (e.g. cost savings or CO2-reduction). In the former perspective, three key properties of EF are usually addressed: (i) the temporal flexibility; (ii) the amount of energy or power that can be shifted; and (iii) the associated cost of activating this flexibility [1]. In association with the latter perspective, the Flexibility Index (FI) and the Expected Flexibility Saving Index (EFSI) were proposed by Junkers et al [14]. These indexes are derived from a Flexibility Function (FF), which represents the variation of the building's flexible profile with respect to the baseline, due to a response to a penalty signal (e.g. energy price or CO_2-emissions). FI and EFSI respectively quantify the relative energy and cost savings of the penalty-aware profile compared to the penalty-unaware profile. FF is applicable to evaluate single penalty changes or a varying penalty. Thus, it can assess not only EF associated with a specific demand response (DR) event (e.g. peak shaving strategy - PSS) but also the dynamic change of the EF as it is being used. A potential use of FI and EFSI as the basis for building labelling has been foreseen [2]. However, a corresponding EF labelling methodology and the compatibility with existing certification schemes are yet to be concretised.

A method that addresses these aspects in detail was proposed by Arteconi et al. [3]. Their method combines quantification and rating of EF for assessment at the design level using physical models, and enables the comparison of buildings based on EF. This method is similar in nature to the Energy Performance Certificate (EPC) calculation methodology, which involves the assessment of the building under design conditions and the comparison with the performance of a similar building with minimum required specifications. The methodology by Arteconi et al. is based on the Flexibility Performance Indicator (FPI). FPI is defined as a weighted articulation of the following four parameters addressing the above-mentioned key properties of EF: response time, committed power, recovery time and actual energy variation. Arteconi et al.'s rating method includes an EF labelling system according to EF classes, aiming to extend the EPC with information on EF. The assignment of EF classes is based on the ratio of two FPI values. One FPI value assesses the maximum EF that the building can offer considering design specifications. The other FPI value assesses the maximum EF of the same building but with neglected thermal mass (FPI_{limit}). It is assumed that the maximum EF is obtained when applying a PSS at the time of the year with most unfavourable conditions (in terms of both weather and peak demand in the grid). Given its combination of EF quantification and labelling, and its potential for complementing EPC, Arteconi et al.'s method will be drawn on heavily in the current article and will henceforth be referred to as the "FPI rating method".

When assessing and rating EF at the design level, the outdoor conditions, DR regime and target internal temperatures are set to standard values [3], which can differ from values in existing buildings during operation. In addition, using models allows controlling conditions and testing out different flexibility strategies. However, even though calibration using measured data may improve their accuracy [5,15–17], models do not reproduce with complete precision the actual load profile of an energy system under real operating conditions. Nevertheless, for existing buildings where monitored flexible profiles are available, models are implemented to predict their corresponding baselines [2,5]. There is little work assessing EF of buildings under real operation [5], and there are no examples of assessing EF of existing buildings at the operational level without models. Moreover, there are no examples comparing EF rating at design and operational level using monitored data. The implications of this is that it is not known whether flexibility ratings realistically capture the EF a building actually provides. This is the research gap addressed in this paper, which contributes with the use of monitored data to bring the EF rating closer to the level of EF achieved in practice, and thus to improve the usefulness of the rating system. Before doing so, we consider below the possible errors introduced when monitored data is not used.

1.2. The energy flexibility gap

Rating EF of buildings using models without incorporating actual performance data is likely to lead to a flexibility gap (EF Gap) – a difference between the EF achieved in the model (and used as the basis of the rating) and achievable in practice. Although to our knowledge no studies have previously demonstrated this discrepancy, it is an analogous concept to the energy performance gap, which has been well studied and documented [18,19], and refers to the difference between predicted and actual energy use. Predicted energy use is in many countries the basis for an energy rating such as the EPC, yet there are multiple reasons why the building may perform differently to what has been predicted. These can be grouped into design, construction and operation related factors [18].

In the same way, EF Gap can be defined as follows:

$$EF\ Gap = Expected\ EF\ -\ Actual\ EF \tag{1}$$

In Equation (Eq.) 1, the Expected EF is that estimated applying design specifications. In contrast, the Actual EF is calculated using monitored data of the building in operation. The difference between them is here termed the EF Gap. The EF Gap could in theory be positive or negative and, like the energy performance gap, could be contributed to by a wide range of factors. For example, the design and as-built thermal mass could differ in their ability to store and discharge heat, or the electric heating system may in practice have more or less opportunity to charge a hot water store each day than anticipated in the system design.

The existence of the energy performance gap and, as a result, the discrepancy in information given to a building owner or user on an EPC, has led to a move in some countries towards more data-driven approaches to rating energy performance [19–22]. The problem with approaches based purely on data is that it renders comparison between buildings unfair: this is because occupancy and weather differ between

[1] The expert group working on Energy Flexibility, IEA EPB Annex 67, summarised the problem as, "…there is a need for an approach that takes in to account the dynamic behaviour of buildings rather than a static counting and rating of control devices as proposed by the SRI study" [2]. This is because heating flexibility often relies on dynamic phenomena such as charging and discharging of building thermal mass.

buildings and strongly influence energy performance, which for example can lead to a building with low design energy efficiency consuming less energy than a more efficient building. It is therefore acknowledged that the ideal way to rate buildings is to combine data (which reports real performance) with modelling (which standardises for weather and occupancy and renders buildings comparable) [23,24]. In this study, we explore combining data and modelling to create effective EF ratings while tackling the EF Gap.

1.3. Aim of this study

EF ratings could be used by a number of stakeholders to compare buildings to one another, prioritise buildings for interventions, or give building users an idea of the revenues available from participating in DR programs [25]. Clearly, it is beneficial to give as realistic a picture as possible of the achievable flexibility. For unoccupied or non-built buildings, an EF rating is only possible using physical models and design specifications. However, existing buildings may have monitored data available. Therefore, the purpose of this study is to investigate the incorporation of monitored data into EF rating.

We introduce the novel concept of the EF Gap and explore, using two case studies, how suitable the FPI rating method is for use with monitored data. Combining monitored data and modelling, we also evaluate the EF Gap between the Expected EF and the Actual EF using existing EF metrics. EF Gap is a new concept, and we do not attempt to quantify it in full comparing different types of models (physical and data-driven models) nor addressing construction related factors (e.g. deviation between building component specifications designed and actually built). Instead, we focus on the discrepancy between Expected EF and Actual EF associated with operational factors: for example climate, DR strategies and durations, and how electric heating systems work in practice.

Moreover, we test out a novel method to select recorded baselines from historical data in order to quantify Actual EF in combination with monitored flexible profiles. As predicted baselines are not completely accurate and recorded baselines capture the real performance of the monitored system, we explore how acceptable using the latter is, despite not being time-coincident with the flexible profile. Rather than optimising and validating the method, a preliminary assessment is conducted to evaluate whether this data-only approach is as a viable alternative to using models, so it can be further studied in future works.

Finally, we explore how the Expected EF can be modified to bring it closer to the Actual EF and demonstrate how using monitored data within the EF rating can improve the usefulness of the result for different stakeholders.

1.4. Research questions

The following research questions are addressed:

1) Are existing EF metrics in the FPI rating method suitable for assessing EF using monitored data?
2) What is the difference between Expected EF and Actual EF when calculated using monitored data, and how does this contribute towards explaining the EF Gap?
3) How can data best be incorporated into EF ratings?

1.5. Paper structure

The rest of this paper is structured as follows. **Section 2** presents the methods used in this study. This includes the case studies and monitored data, the indicators selected to rate EF, and the modelling resources used as a support. **Section 3** presents the results, which are discussed in **Section 4**. Finally, **Section 5** summarises the conclusions.

2. Methods

2.1. Case studies and monitored data

The case studies of this investigation are two real occupied houses located in England. This location has a heating dominated climate that belongs to the class Cfb (C: Temperate; f: Without dry season; b: Warm summer) according to the Köppen-Geiger classification [26]. The households were recruited as part of a wider study of the social and technical implications of DR [9,27].

The physical characteristics and configuration of the houses and their heating systems are summarized in Table 1. The houses have contrasting constructions. House A (HA), over 100 years old, is mostly uninsulated but has thick solid walls and thus far higher exposed thermal mass. House B (HB) is a recently built, well insulated (although not to Passive House standard) home with cavity wall insulation separating the internal blockwork from the external brickwork, meaning the latter does not contribute to the effective thermal mass of the dwelling. The heat demand of these two dwellings is very different, as indicated by the heat pump (HP) sizing.

A set of monitored time series data was collected from each house over the winter of 2020/21. During the monitoring period, the houses were both occupied and both had electric HPs as their sole source of space heating. The monitoring took place during the COVID-19 pandemic; lockdowns may have led to higher internal heat gains from occupants being at home more than usual but space heating timing was not affected by the pandemic in these homes since they were continuously heated except for when DR events were occurring. Air temperatures were recorded every 10 min in all the rooms and in one location outdoors. HP electricity consumption was needed at higher resolution to see the behaviour of the HP and was recorded every 5 min in HA and 1 min in HB in line with the logging resources available.

For the purposes of this study, each household had programmed their heating system to operate according to two different daily regimes, as described in Table 2 and illustrated in Fig. 1. On the one hand, on

Table 1
Characteristics and configuration of case studies.

Characteristics	House A	House B
Location	Bristol, South West England	Bedfordshire, East England
House type	End terrace	Detached
House age	1905	2011
Floor area	231 m2	153 m2
Thermal mass (Wall construction)	High (30 cm solid walls)	Medium (Cavity insulated walls)
Building fabric efficiency (Details)	Medium (Uninsulated walls; insulated loft and ground floor; some airtightness improvements)	High (Well insulated)
Heat pump type	Ground to water, inverter control	Air to water, inverter control
Heat pump size	15 kW_{th} 6 kW_e	8 kW_{th} 3 kW_e
Heat emitters	Radiators	Radiators
Buffer tank size	90 L	45 L

Table 2
Operation setting during monitoring. DR-days: demand response days. BL-days: baseline days.

Case days	Space heating setpoint and schedule					DHW heating schedule	Monitoring dates	Weekdays	Weekend days	Total days
House A (HA)										
DR-days	20 °C 00:00-11:00	21 °C 11:00-16:00	15 °C 16:00-19:00	19 °C 19:00-20:00	20 °C 20:00-00:00	01:00–03:00	Mon–Sun 04/11/2020–18/01/2021	54	22	76
BL-days	20 °C Constantly					01:00–03:00	Mon–Sun 19/01/2020–28/02/2021	29	12	41
House B, before user adjustment of thermostat (HB1)										
DR-days	18 °C 00:00-05:00	22 °C 05:00-16:00	OFF 16:00-19:00	22 °C 19:00-00:00		21:00–22:00	Mon, Thu, Sat 18/11/2020–05/12/2020	10	3	13
BL-days	18 °C 00:00-05:00	22 °C 05:00-00:00				21:00–22:00	Fri, Sun 18/11/2020–05/12/2020	3	2	5
House B, after user adjustment of thermostat (HB2)										
DR-days	18 °C 00:00-05:00	21.5 °C 05:00-16:00	OFF 16:00-19:00	21.5 °C 19:00-00:00		21:00–22:00	Mon, Thu, Sat 06/12/2020–26/01/2021	30	7	37
BL -days	18 °C 00:00-05:00	21.5 °C 05:00-00:00				21:00–22:00	Fri, Sun 06/12/2020–26/01/2021	7	8	15

‘demand response days’ (DR-days) the space heating was programmed to vary in order to generate flexible energy demand profiles. On the other hand, on ‘baseline days’ (BL-days) the space heating was operated continuously (except for a night set-back in HB). The BL-days and DR-days did not occur simultaneously across the two cases: HB implemented BL-days on Fridays and Sundays and DR-days every other day, whereas HA implemented DR-days for a period 76 days, and then BL-days for 41 days.

Table 2 also shows that the use of real occupied houses introduced complexity in multiple ways. The heating systems in both houses is used not only for space heating but also for domestic hot water (DHW). Therefore, the HP’s load profile is affected in the period scheduled for DHW, during which the heat production may vary from day to day depending on specific user needs (both homes had a hot water tank, which took 1–2 h for the HP to charge). Furthermore, each house presents a different level and schedule of setpoint temperatures, according to the internal conditions desired by the occupants. In addition, the occupant in HB adjusted the main thermostat several weeks into the data collection period, as the household perceived that the house felt too warm, resulting in the need to treat the periods before and after the adjustment separately in the analysis. These periods are named HB1 and HB2. The occupant in HB also implemented a temperature setback at night, which affects the demand profiles. Moreover, in both houses a PSS was implemented between 4 pm and 7 pm, coinciding with the period of high electricity price of the supplier’s tariff, and the period of highest peak demand in the UK [28]. In HA the space heating setpoint temperature was lowered, and in HB the HP compressor was switched off. Nonetheless, the occupant in HA additionally combined the PSS with preheating prior to 4 pm. This was to try to ensure thermal comfort during the peak-shaving period. The occupant in HA also set after the PSS a gradual recovery in to steps of the original setpoint temperature prior to the preheating. This was in case some days had high energy prices in the hour 7–8 pm. All these differences are accounted for in the analysis; they result from the way the occupants wanted to run their homes as well as the type of DR they were prepared to do for the purposes of this study. Ideally, all settings would have been identical between the two homes but the inability to dictate all settings was one limitation of working with occupied homes.

2.2. *Quantification of Expected EF and Actual EF*

To quantify EF we apply the standard method described in the introduction, which involves comparing a flexible load profile with a corresponding baseline. We cover the quantification at the design and operational level, considering them as the Expected EF and Actual EF, respectively.

We use the FPI rating method and its underlying EF parameters (t_{res}, $\dot{P}_{res}$, t_{rec} and E_{DR}) as the EF metrics (see Section 2.3). For each metric, we calculate corresponding values for Expected EF and Actual EF, considering the variant matrix described in Table 3. We support the calculation process with two energy models (one model per case-study house). Each model is used for the quantification of both the Expected EF and Actual EF of the corresponding house. The implemented models are data-driven models (see Section 2.4). We train and calibrate the models using measured data to help improve outcome accuracy, which we evaluate by means of the Root Mean Square Error (RMSE) and a residual analysis including Auto-Correlation Function (ACF), Partial ACF and Cumulative Periodogram (see Section 2.4.2).

To quantify the metrics for the ‘maximum’ Expected EF, we apply the calculation settings specified in the FPI rating method for calculating the ‘maximum’ EF level according to design specifications [3] (see Table 5). Thus, for each case study we make use of the data-driven models to generate the flexible and baseline profiles for the worst-case day (the one in the year with lowest mean outdoor dry-bulb temperature; we exclude the cooling period). The climatic profiles of the worst-

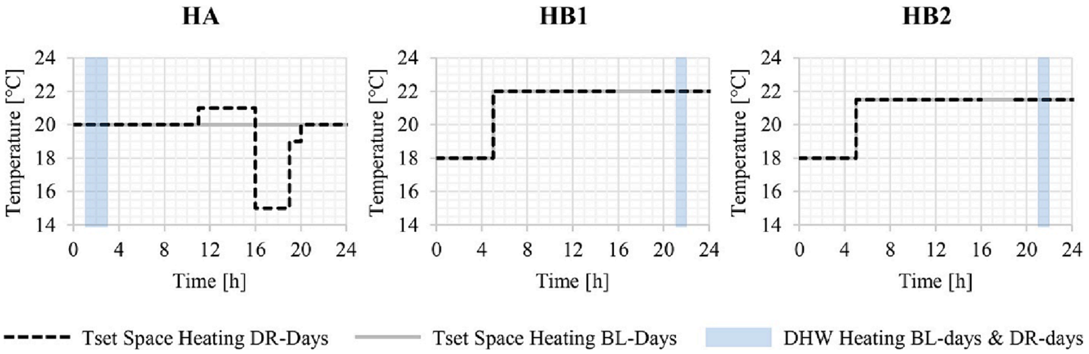

Fig. 1. Graphical representation of heating schedules. DR-days: demand response days. BL-days: baseline days. HA: house A. HB1: house B, before user adjustment of thermostat. HB2: house B, after user adjustment of thermostat. DHW: domestic hot water. Tset: setpoint temperature.

case days for each location are extracted from typical meteorological year (TMY) weather profiles obtained using Meteonorm V.7 [29].

For the Actual EF, we perform two sets of calculations. On the one hand, we use as the flexible profile the monitored time series of each DR-day and use the simulation models to generate the baselines. This is referred to as model-based Actual EF. On the other hand, we explore an alternative way to quantify the Actual EF without the need to use models. This is referred to as data-only Actual EF. For this, we test out a new method to identify a suitable baseline within the profiles of the monitored BL-days for each monitored DR-day profile. This method is based on searching for similar meteorological conditions and distinguishing between weekdays (WD) and weekend days (WE) (see Section 2.5).

After obtaining the results of the quantification process, we evaluate the EF Gap. To do so, we represent the values of the calculated metrics using scatter plots and analyse the discrepancies between the results for the 'maximum' Expected EF at the design level and the values of Actual EF at the operational level, considering the definition of EF Gap as set out in Eq. (1). In this analysis, we pay attention to the effect of several aspects on the results: the numerical method for quantifying the EF metrics; applying monitored data in the quantification; different setpoints; consideration of the preheating phase in HA in addition to PSS as the DR event; and using data-only baselines instead of model-based ones. In addition, aiming at exploring options to contribute to a more robust characterisation of the Expected EF and to limit the EF Gap, we group the metric values of Actual EF by clusters of days according to the daily outdoor temperature, and present the results in form of boxplots. Then, we try out characterizing the Expected EF using the mean values of each cluster as a range of Expected EF values instead of a single 'maximum' Expected EF value. Finally, for these two alternatives, we complete the evaluation of the EF Gap by analysing Mean Absolute Error (MAE) between the Expected EF values and the associated Actual EF as defined in **Eq. (2)** (***$EFGap_{MAE}$***). For the Expected EF resulting from a cluster mean value, the Actual EF values considered in the equations are those from the days belonging to the cluster. In contrast, for the case of a single 'maximum' Expected EF value all calculated results for Actual EF are used. In addition, we evaluate the potential for improving EF Gap when considering the range of Expected EF by taking into account the relative difference of each cluster's $EFGap_{MAE}$ compared to the $EFGap_{MAE}$ obtained considering the 'maximum' Expected EF.

$$EF\ Gap_{MAE} = \frac{1}{n} \cdot \sum_{i=1}^{n} |Expected\ EF - Actual\ EF_i| \tag{2}$$

2.3. Flexibility performance indicator (FPI)

As introduced above, the FPI created in Arteconi et al.'s rating method [3] consists of a weighted articulation of four parameters: response time (t_{res}), committed power ($\dot{P}_{res}$), recovery time (t_{rec}) and energy managed (E_{DR}). These are defined below and graphically represented in Fig. 2:

Response time (t_{res}) [h]: Time that the DR strategy lasts. In the calculation of the 'maximum' Expected EF in the FPI rating method, the response time corresponds to the time necessary to the internal temperature to reach the limit of the comfort band after switching off the thermal system.

Committed power ($\dot{P}_{res}$) [kW$_e$]: Integral of the differences between the HP baseline power demand and the HP power demand during the DR strategy divided by the response time.

$$\dot{P}_{res} = \frac{1}{t_{res}} \int_0^{t_{res}} \left(\dot{P}_{REF} - \dot{P}_{DR}\right) dt \tag{3}$$

Recovery time (t_{rec}) [h]: Duration of the recovery period, consisting on the time that it takes after the DR strategy until the internal comfort conditions are restored as in the baseline.

Actual energy variation (E_{DR}) [kWh$_e$]: Difference in energy use

Table 3
Overview of calculated variants of Expected EF and Actual EF.

Variant	Description / Main role in the study
Model-based maximum Expected EF	Maximum Expected EF as in Table 5, calculated using flexible and baseline profiles predicted by the implemented data-driven models / Principal reference value of the Expected EF to evaluate EF Gap.
Model-based maximum Expected EF*	A variant of model-based maximum Expected EF, using the actual setpoint temperature of the case studies instead of the predefined standard value in the FPI rating method / To support the comparison with the Actual EF values while addressing the impact of the setpoint temperature on the results.
Expected EF range	A set of reference values derived from model-based Actual EF results proposed to complement the representation of the Expected EF / Alternatives for complementing the characterisation of the Expected EF to bring it closer to the Actual EF.
Model-based Actual EF	Actual EF as in Table 5, but exclusively taking into account PSS as the DR strategy as in the maximum Expected EF. It is calculated using monitored flexible profiles and baseline profiles predicted by the implemented data-driven models / Evaluation of the use of the FPI rating method with monitored data and of the discrepancies between the Expected°EF under design conditions and the Actual°EF under operational conditions.
Model-based Actual EF*	A variant of model-based Actual EF, calculated using the settings as in Table 5 / Evaluation of the discrepancies between Expected EF and Actual EF derived from applying the standard DR setting in the FPI rating method compared to DR in practice.
Data-only Actual EF	A variant of model-based Actual EF, using the defined data-only approach to select baselines instead of the implemented data-driven models to predict baselines / Evaluation of the potential for assessing Actual EF using only data instead of building models.

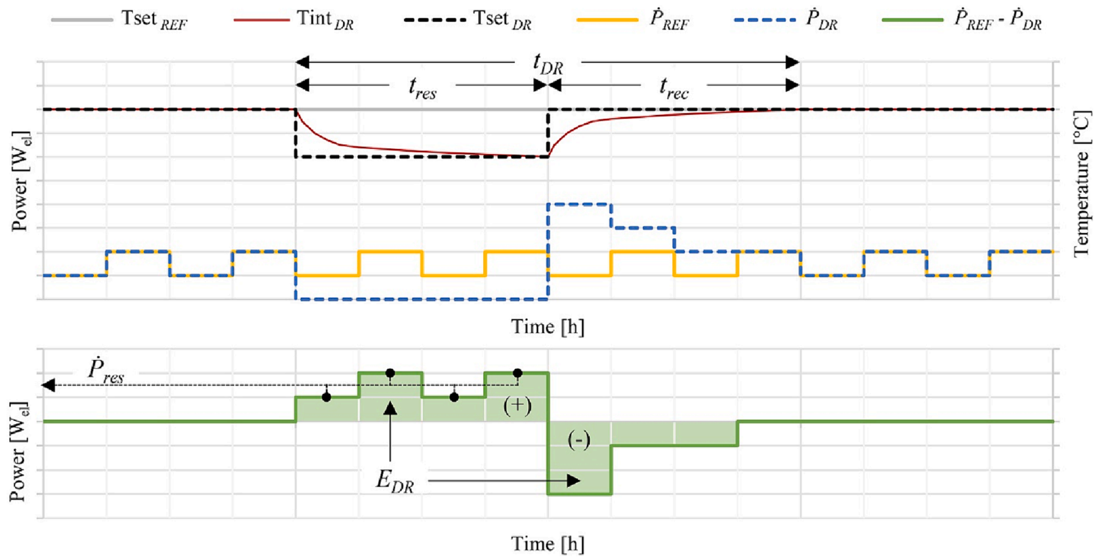

Fig. 2. Graphical schema representing energy flexibility parameters associated with the Flexibility Performance Indicator. t_{res}: response time. t_{rec}: recovery time. t_{DR}: total duration of the demand response event. $\dot{P}_{REF}$: power demand of the heat pump in the baseline case. $\dot{P}_{DR}$: power demand of the heat pump in de demand response case. $\dot{P}_{res}$: committed power. E_{DR}: actual energy variation. $Tset_{REF}$: setpoint temperature in the baseline case. $Tset_{DR}$: setpoint temperature in the demand response case. $Tint_{DR}$: indoor temperature in the demand response case.

between the baseline and the DR profiles during the whole DR event (DR strategy plus recovery period).

$$E_{DR} = \int_0^{t_{DR}} \left(\dot{P}_{REF} - \dot{P}_{DR}\right) dt \tag{4}$$

With:

$$t_{DR} = t_{res} + t_{rec} \tag{5}$$

Flexibility Performance Indicator (FPI) [-]: Dimensionless weighted average of response time, committed power, recovery time and actual energy variation.

$$FPI = \frac{1}{4}\left(p_1 \bullet t^*_{res} + p_2 \bullet \dot{P}^*_{res} - p_3 \bullet t^*_{rec} + p_4 \bullet \eta_{DR}\right) \tag{6}$$

With:

$$t^*_{res} = t_{res}/24 \tag{7}$$

$$\dot{P}^*_{res} = \left|\dot{P}_{res}\right| \Big/ \dot{P}_{rated}, \text{ with } \dot{P}_{rated} \text{ as the HP design power} \tag{8}$$

$$t^*_{rec} = t_{rec}/24 \tag{9}$$

$$\eta_{DR} = \begin{cases} \dfrac{E_{DR}}{\int_0^{t_{DR}} \left(\dot{P}_{REF}\right) dt} \\ 0 \ if \ E_{DR} < 0 \ in \ PSS \end{cases} \tag{10}$$

$p_1 = 60, p_2 = 20, p_3 = 10, p_4 = 10$

According to the method's creators, the first two parameters are more relevant to the grid side while the other two parameters are of more interest for the consumer side. Their weighting factors are respectively 60, 20, 10 and 10, having (t_{res}) the highest impact on the resulting FPI value; the weights were assigned by Arteconi et al. by observing the results of large numbers of simulations [3].

2.4. Model-based approach to creating baselines with data

In existing buildings, monitoring systems normally measure only one of the flexible or baseline profiles at a time [2]. Thus, modelling is commonly applied to generate baselines when assessing EF [5]. Models can be detailed physical models (white-box), simplified physical models (grey-box) or data-driven models (black-box) [5,30]. In this study, we lack of key information in order to be able to create plausible physical models of the houses (e.g. actual U value and air tightness of the building envelope). We could have gathered more information, made some assumptions and tried to validate the model using the monitored data. Nonetheless, data-driven models are emerging as a suitable alternative for an extensive deployment of EF and the interaction with the grid [30–32]. This lies in their simplicity with regard to preparation and adaptability to other buildings, in contrast to physical models. In addition, we aim to explore the combination of data with modelling to enrich EF rating and closing the EF Gap. Therefore, in this study we use data-driven models which incorporate our monitored data, based on the definition below.

2.5. Mathematical description of the model

The energy model of each house is built following the methodology demonstrated by Mor et al. [33], which consists of the combination of two autoregressive models with exogenous variables (ARX). Owing to their autoregressive impulse responses, these kind of models enable balancing effectiveness and simplicity while capturing system dynamics [33]. The first ARX in each house's model captures the dynamics affecting the heating system's operation (supply-side ARX) and predicts the HP's electric consumption (Q^e) – see Eq. **(11)**. The second model captures the dynamics influencing the thermal behaviour of the building (demand-side ARX) and predicts indoor temperature (T^i) – see Eq. **(15)**. These ARX models are first trained separately using the monitored data and then articulated through a model coupler. This coupling mechanism plays a pivotal role in simulating the dynamic interplay between the performance of the heating system and the temperature states in the building. Thus, the model coupler ensures the interaction between the individual ARX models, enabling them to generate interconnected predictions based on a specific setpoint temperature. More specifically, the coupled models jointly predict the electricity consumption of the HP based on the indoor temperature's state, while simultaneously forecasting the indoor temperature's evolution considering the anticipated system operation. We use the predicted time series of the coupled ARX models to define the baselines for the quantification of the EF metrics.

The coupled ARX models were implemented in R using the open source library *biggr* [34]. Models of this kind have been previously applied to existing buildings [5,33]. They can be deployed in big data environments [35] to communicate via the cloud with the local monitoring system in real buildings [33]. This enables updates of the models and EF metrics, with a frequency determined by changes in data patterns or system dynamics. Periodic updates could allow for optimisation towards user behaviour and climate variations, while other updates could target adjustments to changes such as new homeowners or building renovation.

The mathematical description of the supply-side ARX model is given in Eq. (11). This model uses the setpoint temperature (T^s), the indoor temperature (T^i) and the outdoor temperature (T^e) as the input variables. In the case of the indoor and setpoint temperature features, an interaction is considered based on the HP's operation status (S^{HP}). Furthermore, the outdoor temperature is subtracted from a fixed reference temperature to build a heating degree factor (Ψ) and help to model non-linearities between electricity consumption and weather conditions.

$$\gamma(B)Q^e_t = \beta_a(B)\mathrm{T}^{s^*}_t + \beta_b(B)\mathrm{T}^{i^*}_t + \beta_c(B)\Psi_t + \varepsilon_t \tag{11}$$

With:

$$T^{s^*}_t = \left(T^s_t \times S^{HP}_t\right) \tag{12}$$

$$\mathrm{T}^{i^*}_t = \left(T^i_t \times S^{HP}_t\right) \tag{13}$$

$$\Psi_t = \left(25 - \mathrm{T}^e_t\right) \tag{14}$$

$\gamma(B), \beta_a(B), \beta_b(B), \beta_c(B)$: Autoregressive terms.
Q^e: Electricity consumption of HP.
T^s: Setpoint temperature.
T^i: Indoor temperature.
S^{HP}: Operation status of the HP (0: off, 1: on).
Ψ_t: Heating degree factor.
T^e: Outdoor temperature.

The mathematical description of the demand-side ARX model is given in Eq. (15). This model uses the outdoor temperature (T^e), HP's electric consumption (Q^e) and solar elevation (I^{el}) as the input variables. Conceptually, it can be considered as a thermal model of the house. To ensure system simplicity and scalability in the methodology, the coefficient of performance is used to connect the power inputs, despite the heat input generated by the HP being the ideal consideration. Multiple interactions between the electricity consumption (Q^e) and other input variables are considered to assess this coefficient of performance: $\mathrm{Q}^{e^*}_t$, H^e_t and H^h_t. The respective input variables included in these interactions are the HP's operation status (S^{HP}), the heating degree factor (Ψ), and the solar elevation (I^{el}). By incorporating these three input features, the coefficient of performance indirectly models the behaviour of the HP. It is worth noting that solar elevation is used instead of solar radiation due to the limited availability of local solar radiation data. By employing the sun position instead of radiation, the methodology remains scalable while considering the effect of solar gains.

$$\phi(B)T^i_t = \omega_a(B)\mathrm{Q}^{e^*}_t + \omega_b(B)\mathrm{T}^e_t + \omega_c(B)I^{el}_t + \omega_d(B)H^e_t + \omega_e(B)H^h_t + \varepsilon_t \tag{15}$$

With:

$$\mathrm{Q}^{e^*}_t = \left(\mathrm{Q}^e_t \times S^{HP}_t\right) \tag{16}$$

$$\mathrm{H}^e_t = \left(\mathrm{Q}^{e^*}_t \times \Psi_t\right) \tag{17}$$

$$\mathrm{H}^h_t = \left(\mathrm{Q}^{e^*}_t \times I^{el}_t\right) \tag{18}$$

$$\Psi_t = \left(25 - \mathrm{T}^e_t\right) \tag{19}$$

$\phi(B), \omega_a(B), \omega_b(B), \omega_c(B), \omega_d(B), \omega_e(B)$: Autoregressive terms.
T^i: Indoor temperature.
Q^e: Electricity consumption of the HP.
T^e: Outdoor temperature.
I^{el}: Solar elevation
S^{HP}: Operation status of the HP (0: off, 1: on).
Ψ_t: Heating degree factor.

The following equation defines the autoregressive terms of the ARX models:

$$f(B) = 1 + f_1B^1 + \cdots + f_nB^n \tag{20}$$

With:

n: Number of lags, or order, of the backward shift operator B, defined as in Eq. (21)

$$B^k y_t = y_{t-k} \tag{21}$$

Where:

y: Considered variable (e.g. the indoor temperature in the case of $\phi(B)$).

2.6. Model validation and accuracy assessment

The models were trained using a sub-set of the monitored data (training data). The supply-side and the demand-side ARX models were trained separately, using the training data corresponding to their respective input and output variables (see **Section 2.4.1**). After the training, a different data sub-set (testing data) was used to evaluate the accuracy of the ARX models both separately and coupled. For this, the testing data associated with the input variables of the models were applied to generate predicted time series of HP's electricity consumption and indoor temperature. These time series were then compared to the monitored ones in the training data using the RMSE. Prior to the accuracy assessment the residuals of the models' predictions were analysed for validation.

The residual analysis of the ARX models (see Appendix A) shows that the residuals follow a Gaussian distribution and are not auto-correlated. Thus, the models achieve the white noise conditions, indicating that they are successfully trained and can be considered as valid. Moreover, the performance accuracy of the models is satisfactory. Table 4 summarizes the RMSE (average of all tested days) of the time series predicted by the individual demand-side and supply-side models and the combined model. However, these values reveal that the accuracy related to energy consumption improves when the time series are evaluated at hourly resolution compared to a 10-minute resolution. This highlights the challenge of reproducing the actual dynamic profile of HP operation.

The trained models perform well when predicting the behaviour of the houses for days with daily temperature similar to the ones of the monitored days. Nonetheless, among the monitored data used to train the model, there is not information about the performance of the HP for days as cold as the one selected for assessing the 'maximum' Expected EF. Consequently, we identified that the models had difficulties in

Table 4
RMSE of timeseries predicted by the energy models (average of all tested days).

House	Time series resolution	Supply-side model	Demand-side model	Coupled model (demand-side and supply-side)	
		Q^e [kW$_e$]	T^i [°C]	Q^e [kW$_e$]	T^i [°C]
HA	10-minute	0.057	0.369	0.156	0.355
	hourly	0.028	0.371	0.068	0.356
HB	10-minute	0.055	0.583	0.083	0.601
	hourly	0.042	0.588	0.070	0.608

Table 5
Differences between settings for 'maximum' Expected EF and Actual EF.

Aspect	'maximum' Expected EF (Arteconi et al. [3])	Actual EF	
		House A	House B
Heating system base setpoint temperature	22 °C	20 °C	22 °C (before user adjustment of thermostat) 21.5 °C (after user adjustment of thermostat)
Heating thermal comfort range	20–22 °C, set in accordance with Fanger's Thermal Comfort	18–25 °C, assumed in accordance with norm EN 16798–1 for an indoor climate category III (An acceptable, moderate level of expectation. Typical for existing buildings)	20–25 °C, assumed in accordance with norm EN 16798–1 for an indoor climate category II (Normal level of expectation. Recommended for new and renovated buildings)
Weather conditions	Day of the year with lowest average outdoor temperature (extracted from a TRY)	Monitored DR-days	Monitored DR-days
DR strategy	Only peak shaving: thermal system switch-off without preheating and switch-on afterwards with setpoint temperature as in the base case.	Preheating, peak shaving and two-step reactivation: 1 K increment of setpoint temperature before reduction to 15 °C and gradual recovery of base setpoint temperature afterwards by setting it to 19 °C and finally increasing it by 1 K.	As in Arteconi et al. [3]
DR starting time	Worst scenario of peak electricity demand	Oriented to the worst scenario of peak electricity demand (4–7 pm [28]): 11am when considering preheating as a part of the DR or 4 pm if only peak shaving is considered.	Oriented to the worst scenario of peak electricity demand (4–7 pm [28]): 4 pm.
Response time	Not pre-set: determine by the time that it takes until the indoor temperature reaches the limit of the base comfort band after initiating the peak-shaving strategy (2 °C variation)	Pre-defined: 5 h preheating (11am-16 pm) + 3 h switch-off (16–19 pm)	Predefined: 3 h switch-off (16–19 pm)
Recovery time	Not pre-set: determined by the time that it takes until the indoor temperature recovers the base setpoint temperature level after the peak-shaving strategy.	Not pre-set but influenced by the DHW schedule (1-3am).	Not pre-set but influenced by the DHW schedule (21–22 pm) and limited by the programmed night set-back (0-5am)

predicting the response for those days. After the DR strategy, the models were able to generate a peak in the heat production by the HP, but only punctually, and during the rest of the recovery time, the HP did not perform at high loads, as it is expected, when the indoor temperature lies far below a given setpoint temperature. For this reason, we included a condition (booster) in the coupled model, which helps generate a prediction with high heat production while the difference between the indoor temperature and the objective setpoint temperature is greater than 1 K. This enabled us to obtain more reliable predictions of the HP's load profile for the design EF assessment.

2.7. Data-only approach to creating baselines

In contrast to a model-based prediction of baselines for a DR flexible profile, in this section we consider the alternative possibility of directly selecting a baseline profile from the profiles monitored in the BL-days. For this purpose, we assume that, on days with similar meteorological conditions and day type (WD or WE), the houses' energy demand present a similar magnitude and profile as well. Thus, we explore a method to select a baseline for each DR-day based on the similarity of their outdoor temperature and day type. We exclude other meteorological parameters, such as solar radiation, as a simplification. Similarity between temperature profiles is measured using three statistical indicators: Normalized Mean Bias Error (NMBE), Coefficient of Variance of the Root Mean Squared Error (CV(RMSE)) and total Goodness of Fit (GOF) [36]. To identify and select the baseline, the ASHRAE criterion for validation of simulation models is used as a reference: $-10\,\% \leq \text{NMBE} \leq 10\,\%$, $\text{CV(RMSE)} \leq 30\,\%$ [37]. The selected baseline corresponds, therefore, to the BL-day that meets this criterion and offers a better GOF (the lowest value).

To analyse the performance of the data-only approach with respect to the model-based approach, we directly compare their respective calculated sets of Actual EF values by calculating the NMBE and the CV (RMSE) between them. In addition, we quantify the EF Gap_{MAE} (Eq. (2) associated with each set of values considering the 'maximum' Expected EF and, then, evaluate the relative difference of the data-only EF Gap_{MAE} compared to the model-based EF Gap_{MAE}. This analysis is carried out for each EF metric and considering only DR-days for which both data-only and model-based results are available.

3. Results

The results are presented as follows. Since it was found that the assumptions used in the 'maximum' Expected EF were different from the real conditions inside and outside the building, we firstly give these differences (**Section 3.1**), which then give context to the rest of the results. Following this, we present the EF metrics underlying the FPI rating, using the standard operational assumptions as well as data-informed parameters, and including model-based and data-only baselines (**Sections 3.2 to 3.4**). Finally, we summarise the effect of different assumptions on the EF Gap (**Section 3.5**).

3.1. Differences between settings for 'maximum' Expected EF and Actual EF

Table 5 describes the differences for 'maximum' Expected EF and Actual EF. In addition to these differences, the FPI rating method includes the above mentioned assignment of an EF class to the building, by comparing the building's FPI to the FPI of the building with neglected thermal mass ($\text{FPI}_{\text{limit}}$). Physical models allow manipulating the thermal capacity specifications of the modelled building in order to calculate $\text{FPI}_{\text{limit}}$. Data-driven models, on the other hand, cannot be adjusted to calculate $\text{FPI}_{\text{limit}}$ as the effect of the thermal mass is implicit in their training data and, consequently, they cannot neglect this effect. As we use data-driven models in our study, we do not quantify $\text{FPI}_{\text{limit}}$ and rate the EF class of the houses according to the labelling scale proposed by Arteconi et al. [3], but we exclusively quantify their FPI.

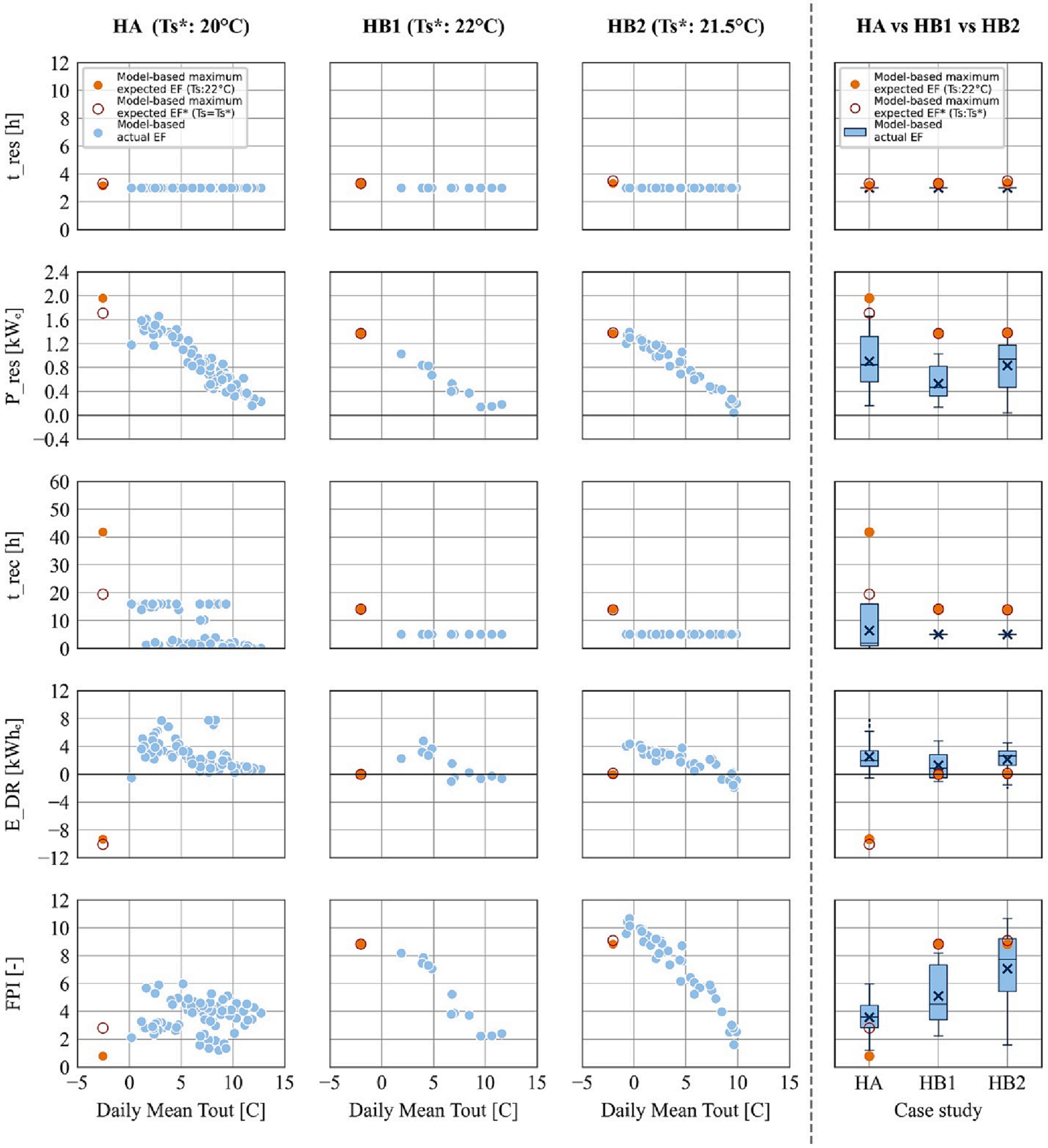

Fig. 3. Results for Expected EF and Actual EF of the indicator FPI and its associated parameters (t_{res}, $\dot{P}_{res}$, t_{rec} and E_{DR}), using energy models and exclusively considering PSS as the DR strategy. Model-based 'maximum' Expected EF: calculation for the day of the year of lowest average temperature and with 22 °C as the base setpoint temperature. Model-based 'maximum' Expected EF*: same as before but using Ts* (base setpoint temperature in BL-days) as the reference. Model-based Actual EF: calculation for each monitored DR-day. HA: house A. HB1: house B, before user adjustment of thermostat. HB2: house B, after user adjustment of thermostat.

3.2. House A and B considering only PSS

Fig. 3 depicts the results obtained making use of the energy models and considering exclusively the PSS DR strategy. In the matrix of subplots, the rows correspond to the calculated indexes. From left to right, the three first columns correspond to the specific values for the three cases HA, HB1 and HB2, represented by means of scatter plots, where single index values are plotted according to daily mean outdoor temperature. In the fourth column, the three cases are represented together and the single values are combined into boxplots in order to facilitate the comparison and analysis of results. Each subplot contains the model-based Actual EF values obtained for all the monitored DR-days and values for both 'maximum' Expected EF and 'maximum' Expected EF*. The former is calculated applying the standard setting from Table 5, in which a base setpoint temperature of 22 °C is specified. In the latter, the standard settings are also implemented but using the base setpoint temperatures of the houses for the BL-days (Ts*): 20 °C, 22 °C and 21.5 °C for HA, HB1 and HB2, respectively.

The results in Fig. 3 show that the response time (t_{res}) associated with the PSS remains constant at 3 h in each house for the Actual EF at all DR-days, as predefined by the operating schedules. This value of 3 h is very close to the response time obtained for 'maximum' Expected EF. Therefore, in these case studies, the impact on a certain EF Gap when looking at the resulting FPI values is low despite the differences in terms of calculation settings and the higher weight of t_{res} compared to the other parameters. Furthermore, t_{res} is higher in 'maximum' Expected EF* than in the 'maximum' Expected EF. This is probably due to the fact that, when applying the 2 °C reduction in the Expected EF*, the indoor temperature is to be reduced to a lower level (e.g. 18 °C in HA) compared to the Expected EF (20 °C). As the indoor temperature in the Expected EF* is closer to the outside temperature, the heat losses are smaller and the rate at which the inside temperature decreases is slower than in the Expected EF.

With regard to the committed power ($\dot{P}_{res}$), there is a negative linear correlation between this metric and the outdoor temperature in all three cases. The values in HA are generally higher with an associated 'maximum' Expected EF value of 2 kW$_e$, compared to 1.3 kW$_e$ in HB. This is due to the lower energy efficiency of HA and higher heated floor area and the correspondingly higher sizing of its HP compared to HB, which results in a higher electrical load being deactivated when the HP is switched off. In HA the committed power approaches zero at higher outdoor temperature values in contrast to HB. This is also due to the lower energy efficiency of HA compared to HB, which might result in HB having lower heating demand than HA or even no heating demand at all on days with outdoor temperature values above 10 °C. The committed power values for Expected EF and Expected EF* are well related to those for Actual EF, following the linearity of the correlation. The values for Expected EF and EF* in the case of HB2 are practically identical, indicating that the 0.5 °C difference between the setpoints of these parameters have almost no impact on the results of this case study. This can also be due to the higher energy efficiency of HB. On the contrary, in the case of HA a larger difference is observed, with values of almost 2kW$_e$ for Expected EF and about 1.7 kW$_e$ for Expected EF*. This indicates that the lower the setpoint temperature, the lower the committed power. This is logical considering that the thermal demand and setpoint temperature are positively related. In general, the results for committed power clearly show a potential EF Gap associated with this parameter that is more pronounced the bigger the difference between the daily outdoor temperature of the DR-day and of the "worst-case" day.

In terms of recovery time (t_{rec}), in HA two bands of constant values are obtained around 1 h and 16 h. In addition, several intermediate values with an upward linear trend and negative slope are observed in the direction of the value associated with the Expected EF. The Expected EF* value is around 20 h, confirming the negative upward trend of the intermediate values of Actual EF between the 1 and 16 h range observed in the graph.

The constant bands of values indicate difficulties in quantifying this metric in HA when real building operation data is used. Thus, the values of t_{rec} in HA around one hour are due to two factors. Firstly, the indoor temperature at the beginning of the PSS strategy is, due to preheating, higher than the setpoint temperature programmed after 7 pm in the baseline. Consequently, even though the indoor temperature drops after starting the PSS, it does not have to increase very much before it meets the value of the baseline setpoint, rendering t_{rec} very low. Secondly, the indoor temperature in the houses during real operation does not remain constant at the level of the setpoint temperature. On the contrary, the indoor temperature fluctuates around the value of the setpoint level due to the actual operation of the HP itself, which consists of cycles of compressor operation. As a result, sometimes the indoor temperature at the baseline is lower than the setpoint and the indoor temperature in the flexible profile reaches the baseline temperature earlier, after receiving the power peak implemented by the HP at the end of the DR strategy to adjust to the new setpoint (see Fig. 5).

Regarding the higher values of t_{rec} around 16 h in HA, these derive from the DR-day heating schedule being set to end the PSS at 7p.m. and start preheating at 11 a.m. the following day. This results in a maximum possible recovery period of 16 h. However, the fact that t_{rec} for Expected EF* is also around 16 h, indicates that this could be the building's own t_{rec} for a setpoint temperature of 20 °C. Thus, the scheduling of preheating at 11 a.m. would not affect the result. Nevertheless, this again demonstrates that the t_{rec} metric may be constrained by the actual operation of the building. This is also indicated by the results in HB, for which the values related to Actual EF remain constant at 5 h. This is due to the beginning of the night setback in this house being scheduled at midnight. Thus, when applying the t_{rec} quantification from the end of the DR strategy at 7 pm, the indoor temperature in the flexible profile reaches the value of the indoor temperature in the baseline at midnight, resulting in the t_{rec} being 5 h in all the DR-days independently of the outdoor temperature values.

Furthermore, the results show that the values of recovery time for the Expected EF strongly differ from the values obtained for the Actual EF. The Expected EF values are approximately 41 and 15 h respectively for HA and HB. Thus, the gap between the expected and actual recovery time ranges between 25 and 40 h in HA, while in HB it is around 10 h.

Similar to what is observed for committed power, the Expected EF and Expected EF* values are practically identical in the case of HB and differ in HA. However, in HA, the relative discrepancy derived from the differences between these values for recovery time (21 h: ~50 % variation) is notoriously higher than in the values for committed power (0.3 kW$_e$: ~15 % variation).

With respect to energy variation (E_{DR}), the graphs show that the values increase from lower outdoor temperature values up to reaching a peak and then decrease again. The highest values are related to the Actual EF and lie in the range of mean outside temperatures between 0 and 5 °C. Thus, energy variation in HA reaches up to 8 kWh, while the highest value in HB is just over 6kWh. Moreover, the Expected EF value in HB1 and HB2 is approximately zero while in HA it is negative and slightly higher than the positive peak in the Actual EF values. This indicates that HA is more likely to experience a gap between the energy variation estimated at the design level and the actual energy variation during operation than HB. This can be related to the lower energy efficiency of HA and the higher size of its HP, but also due to the differences between the user's operational settings with respect to the standard values implemented for Expected EF. In addition, the Expected EF and Expected EF* values of energy variation are very similar in all case studies, showing a limited impact of the differences between the applied setpoints on this parameter.

The FPI values of HA lie between 1 and 6, while in HB they lie between 1 and 11. Therefore, HB reaches a higher EF in terms of FPI, despite the higher $\dot{P}_{res}$ values of HA with respect to HB. This seems to be mainly a consequence of the method of not considering directly $\dot{P}_{res}$ in the FPI calculation but the normalized $\dot{P}^{*}_{res}$ by means of a division by the nominal electrical power of the HP. As this nominal power in HA is twice that of HB (6 kW$_e$ in HA and 3 kW$_e$ in HB), the normalisation results in a greater transformation in HA with respect to HB of the relationship observed in $\dot{P}_{res}$ and FPI between these values and outdoor temperature. The values t_{res} in HA and HB are identical and the effect of this parameter on the differences in FPI between both houses can therefore be discarded. The values of t_{rec} and E_{DR} are different in HA and HB, but their effect plays only a minor role as they are taken into account with a factor of 10 in the FPI equation compared to the factor 20 of $\dot{P}_{res}$.

The values of Actual EF related to FPI in HA do not show a clear correlation with the outdoor temperature. The same occurs if these values are considered together with the value of 'maximum' Expected EF*. However, in combination with the values of 'maximum' Expected EF, there could be a weak positive correlation. In addition, these maximum values are located in the lower range of calculated values for FPI, indicating that the highest FPI for this case study might not occur in the worst-case day but in days with more moderate outdoor temperatures. On the contrary, in HB the 'maximum' Expected EF values related to FPI do appear to be in the higher part of the calculated results, being in line with the assumption that the highest FPI in this case study is more likely to be reached in the worst-case day. In addition, the values of Expected EF and Expected EF* in combination with the Actual EF values show a stronger relationship between FPI and outdoor temperature values in HB with respect to HA. This is negative and tends to be exponential, indicating that the lower the average outdoor temperature, the lower the FPI.

Considering what has been said above about the expected FPI and the ranges of actual FPI values, the results show that the gap between these values can reach a higher level in HB than in HA. This indicates an opposite trend to what is observed in the parameters committed power,

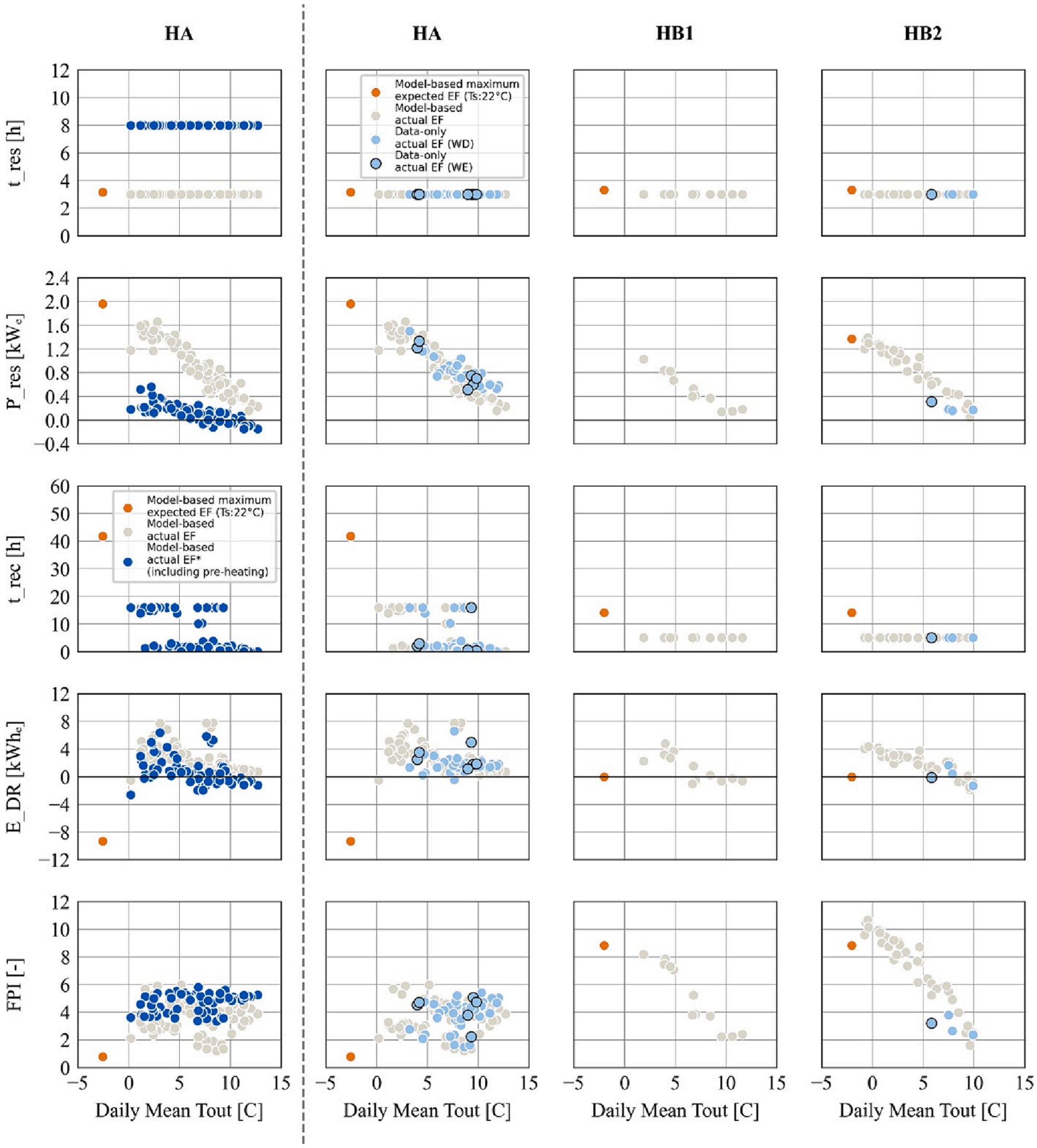

Fig. 4. Comparison of results for Actual EF applying different calculations. Model-based 'maximum' Expected EF: calculation for the day of the year of lowest average temperature and with 22 °C as the base setpoint temperature. Model-based Actual EF: values obtained considering exclusively PSS as the DR strategy and using energy models to generate the baseline profiles. Model-based Actual EF*: values obtained considering the combination of preheating and PSS as the DR strategy and using energy models to generate the baseline profiles. Data-only Actual EF: values obtained considering exclusively PSS as the DR strategy and applying profiles of monitored BL-days as the baselines instead of using predicted baselines using the models. WD: weekdays. WE: weekend days. HA: house A. HB1: house B, before user adjustment of thermostat. HB2: house B, after user adjustment of thermostat.

recovery time and energy variation.

Finally, the boxplots in Fig. 3 show that, in general, design-based Expected EF values are far from the mean values of operational-based Actual EF and even reach the level of outliers. Furthermore, the Actual EF values usually differ from the 'maximum' Expected EF values, highlighting the risk of an EF Gap. This is further evaluated in Section 3.5.

3.3. House A considering pre heating and PSS

The plots in the left column in Fig. 4 combine the Actual EF values of HA obtained by taking into account the combination of preheating and PSS as the DR strategy, with those from Fig. 3 (only considering PSS). These results show that when the quantification of EF is applied considering preheating as part of the DR, the Actual EF values for t_{res} increase from 3 to 8 h, resulting in a difference with respect to the value of Expected EF of approximately 5 h. This is due to the heating schedule in HA, which starts preheating at 11:00 until 16:00, when the PSS starts at 16:00 and lasts until 19:00.

Furthermore, the values for $\dot{P}_{res}$ of Actual EF including preheating as a part of the DR strategy in the calculation are lower with respect to those of considering only PSS. This is because during preheating, the load demand profile is higher than that of the baseline, resulting in negative values when subtracting both profiles, in contrast to what occurs during the PSS. The total committed power is calculated as the average of all these values.

In terms of t_{rec} there are no differences as the effect of the preheating was captured in the monitored data from HA. Thus, even though the results of Fig. 3 were calculated without considering the preheating time in the calculation, its effect was implicit in the results of recovery time. On the contrary, Actual EF values for E_{DR} decrease when preheating is

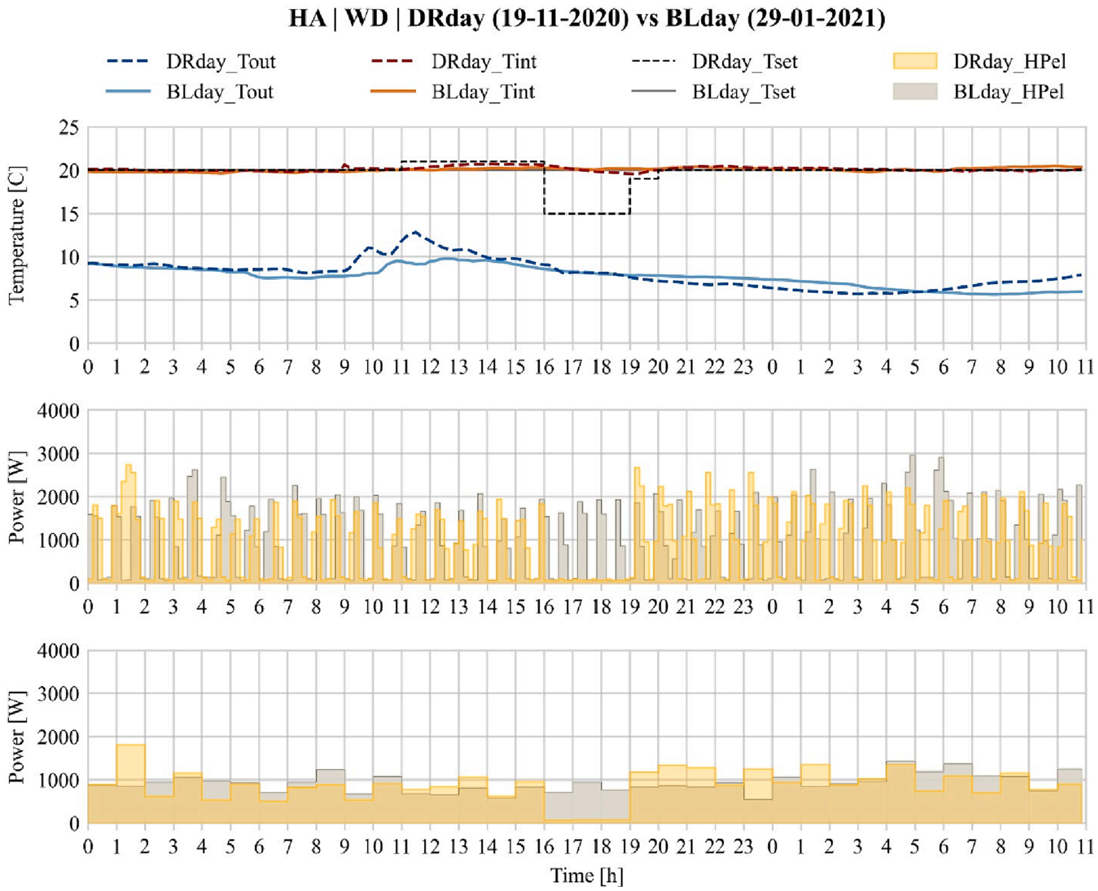

Fig. 5. Comparison between a DR-day and its baseline in HA using data-only baseline. Upper graph: temperatures. Middle graph: electrical load in 10-minute resolution. Bottom graph: electrical load in hourly resolution. HA: house A. WD: weekday. DRday: demand response day. BLday: baseline day. Tout: outdoor temperature. Tint: indoor temperature. Tset: setpoint temperature. HPel: electricity consumption of the HP.

taken into account in the DR strategy. This is due to the compensation of the lower demand during PSS with the additional demand in the preheating at the DR-day with respect to the baseline.

With regard to FPI, the Actual EF values are more concentrated between the range of 3 to 6, thereby increasing the average FPI value of HA with respect to the case without preheating. This occurs despite the smaller values of $\dot{P}_{res}$ and E_{DR}, due to the higher values of t_{res}, which have a large impact on the FPI calculation as they are computed in the equation with a weight factor of 60 (compared to 20 and 10 for $\dot{P}_{res}$ and E_{DR}, respectively). Compared to not considering preheating, the higher average of the actual FPI values including preheating lead in turn to a higher average gap between these operational values and the design-level Expected EF. This points out limitations of the quantification of FPI and suggests the need of incorporating further DR events when estimating the Expected EF to cover a broader range of strategies, which are likely to be implemented in a building in operation (such as the combination of preheating and PSS).

3.4. Model-based vs data-only baselines

By applying the baseline selection method with the monitored data, DR-days and BL-days could be matched. However, it was not possible to assign a baseline to all monitored DR-days. In the case of HA, 44 DR-BL pair days were formed (58 % of its 76 monitored DR-days). For HB, 4 DR-BL pair days could be formed in the case of HB2 (11 % of its DR-days) and none in the case of HB1.

As shown in Fig. 4, the values of data-only Actual EF obtained are in the range of the values of model-based Actual EF. Furthermore, no clear separation is observed between the values obtained for weekdays and weekend days, which are integrated among all values without building specific clusters. The approach of matching weekday baselines to weekday DR-days and weekend-day baselines to weekend-day DR-days

Table 6
Comparison between data-only and model-based Actual EF values obtained for house A.

Compared values	**Error index**	EF metrics				
		t_{res}	$\dot{P}_{res}$	t_{rec}	E_{DR}	FPI
		[h]	[kW$_e$]	[h]	[kWh$_e$]	[-]
data-only Actual EF and model-based Actual EF	NMBE	0 %	11 %	0 %	13 %	6 %
	CV(RMSE)	0 %	27 %	0 %	72 %	15 %
data-only Actual EF and ‘maximum’ Expected EF	data-only $EF\,Gap_{MAE}$	0.17	1.27	37.53	11.55	3.00
model-based Actual EF and ‘maximum’ Expected EF	model-based $EF\,Gap_{MAE}$	0.17	1.16	37.53	12.03	3.21
data-only $EF\,Gap_{MAE}$ and model-based $EF\,Gap_{MAE}$	Relative difference	0 %	−7%	0 %	3 %	8 %

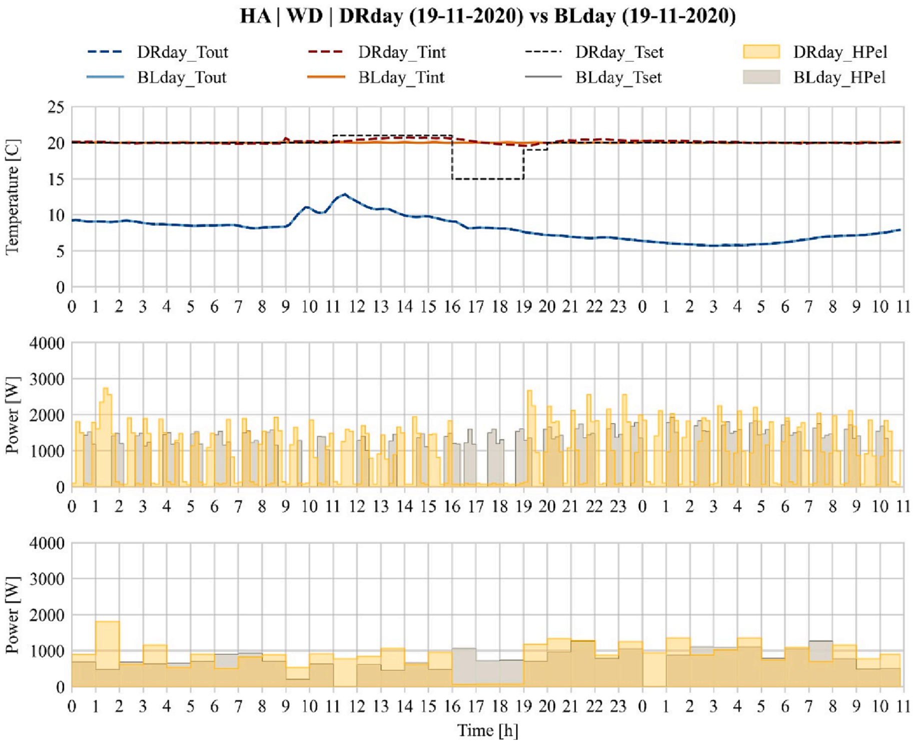

Fig. 6. Comparison between a DR-day and its baseline in HA using model-based baseline. Tout: outdoor temperature. Upper graph: temperatures. Middle graph: electrical load in 10-minute resolution. Bottom graph: electrical load in hourly resolution. HA: house A. WD: weekday. DRday: demand response day. BLday: baseline day. Tout: outdoor temperature. Tint: indoor temperature. Tset: setpoint temperature. HPel: electricity consumption of the HP.

is therefore shown to be helpful. Table 6 additionally summarises the results of comparing the data-only and model-based Actual EF values. The comparison concentrates on HA, due to low availability of data-only values obtained for HB. Thus, the results show no differences between both sets of Actual EF values for t_{res} and t_{rec}. Furthermore, a similar average EF Gap results in each approach, with a relative difference of data-only $EF\,Gap_{MAE}$ compared to model-based $EF\,Gap_{MAE}$ of −7%, 3 % and 8 % for $\dot{P}_{res}$, E_{DR} and FPI, respectively. The results of NMBE and CV (RMSE) do not show strong discrepancies between the data-only and model-based Actual EF values, except for E_{DR} with a CV(RMSE) of 72 %. Aspects that can play a role in this result are the constrained ability of the model to reproduce the operation in practice of the HP, and a non-optimal data-based baseline selected for the DR-day. In general, these results indicate that the data-only method is feasible and it could be an alternative to the use of models. It offers the potential of using actual operational data instead of a model-based prediction. However, the outcome of the data-only approach is limited by the presence of baseline data closely resembling the weather conditions on DR-days.

An example of the profiles of a DR-day and its associated baseline from HA is represented in Fig. 5 and Fig. 6 (see also an example from HB in Appendix B). In Fig. 5, the baseline profiles of temperature and electricity consumption corresponds to the data of the monitored BL-day selected to build the DR-BL pair days (data-only baseline). In Fig. 6, the baseline profiles are the time-coincident ones generated by the energy models (model-based baseline). In each of these figures, the upper graph represents the temperature profiles, and the middle and bottom graphs represent the electrical load profiles respectively in 10-minute and hourly resolution.

The indoor temperature in the DR-day rises above 20 °C during the preheating period (11:00–16:00), falls below 20 °C during the PSS (16:00–19:00), and stabilises again in the recovery period (in this case 19:00–20:00). With respect to the internal temperature of the baselines, the model-based profile generally remains constant at 20 °C, while the data-only profile shows slight fluctuations around this value. This is a reflection of the actual operation of the HP, with an intermittent on–off profile.

The ten-minutely graph of HP electrical load highlights the complexity of actual HP operation. The consumption is intermittent, with peak loads generally around 2000 W_e in the data-only case. The model-based baseline also reproduces this intermittent behaviour, although it has longer consumption intervals with lower peak demand compared to the monitored data.

The graphs with hourly consumption values help corroborating the expected consumption peak after the PSS strategy in the DR-day profile compared to the BL-day. The hourly representation also shows that the total consumption during the response and recovery period in the model-based and data-only baselines are very similar, suggesting that if the former is normally applied, the latter could also be used as an alternative to the use of models. The model manages to reproduce the intermittent operation of the HP and its hourly consumption, but has difficulties in predicting peak demand at a higher resolution. The data-only baseline, on the other hand, is not a time-coincident profile with the DR-day profile, but it does correspond to a real HP operation profile with intervals and demand peaks more similar to those of the DR-day in the hours before and after the DR event.

3.5. Contribution to the EF gap

This part complements the evaluation of EF Gap in Section 3.2 by analysing the $EF\,Gap_{MAE}$ between the values obtained for Expected EF and Actual EF, as explained in Section 2.2. Thus, we use two different interpretations of ‘Expected EF’:

- The ‘maximum’ Expected EF obtained by applying Arteconi et al.’s method based on the design approach method.
- A Expected EF ‘range’ of values obtained by, first, clustering the results using the monitored data (Actual EF values) according to daily outdoor temperature and, then, calculating the cluster mean.

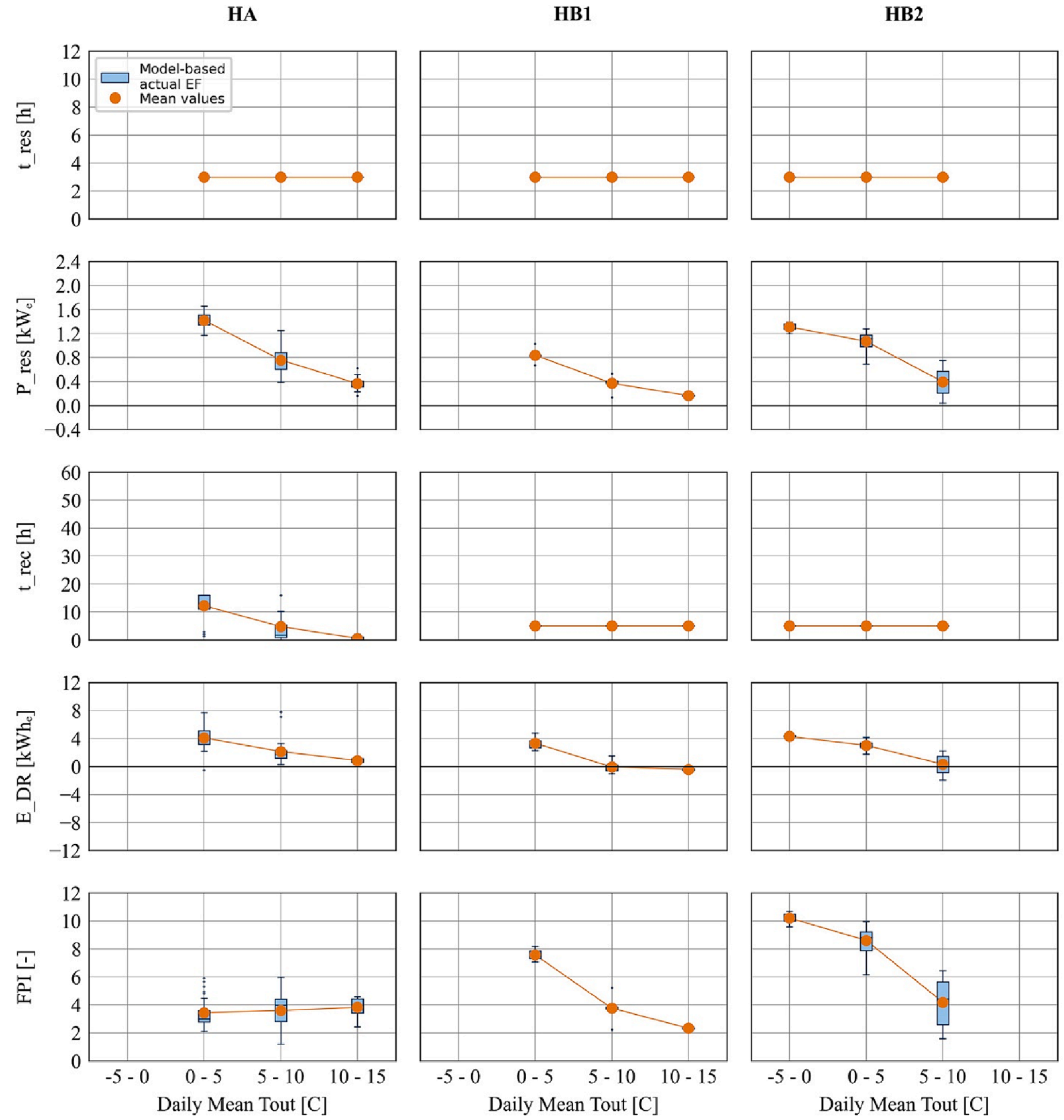

Fig. 7. Calculated model-based Actual EF, clustered by daily mean outdoor temperature. HA: house A. HB1: house B, before user adjustment of thermostat. HB2: house B, after user adjustment of thermostat.

Table 7
Expected EF values and error between Expected EF and (model-based) Actual EF. HA: house A. HB1: house B, before user adjustment of thermostat. HB2: house B, after user adjustment of thermostat. Tout: outdoor temperature.

Case study	Interpretation of Expected EF		Expected EF values					(model-based)*EF Gap*$_{MAE}$				
			t_{res}	$\dot{P}_{res}$	t_{rec}	E_{DR}	FPI	t_{res}	$\dot{P}_{res}$	t_{rec}	E_{DR}	FPI
			[h]	[kW$_e$]	[h]	[kWh$_e$]	[-]	[h]	[kW$_e$]	[h]	[kWh$_e$]	[-]
HA	‘maximum’ Expected EF		3.17	1.96	41.83	−9.31	0.78	0.17	1.09	35.97	11.81	2.89
	Expected EF ‘range’	0 <= Tout < 5	3.00	1.39	11.54	3.65	3.46	0.00	0.12	5.65	1.40	0.93
		5 <= Tout < 10	3.00	0.77	5.00	2.39	3.61	0.00	0.14	5.24	1.41	0.93
		10 <= Tout < 15	3.00	0.47	0.59	1.21	4.14	0.00	0.14	0.67	0.42	0.58
HB1	‘maximum’ Expected EF		3.33	1.37	14.17	−0.02	8.83	0.33	0.84	9.17	1.78	3.72
	Expected EF ‘range’	0 <= Tout < 5	3.00	0.84	5.00	3.32	7.57	0.00	0.08	0.00	0.71	0.35
		5 <= Tout < 10	3.00	0.37	5.00	−0.07	3.76	0.00	0.09	0.00	0.76	0.63
		10 <= Tout < 15	3.00	0.16	5.00	−0.42	2.34	0.00	0.02	0.00	0.19	0.08
HB2	‘maximum’ Expected EF		3.33	1.37	14.17	−0.02	8.83	0.33	0.60	9.17	2.29	2.69
	Expected EF ‘range’	0 <= Tout < 5	3.00	1.07	5.00	3.01	8.61	0.00	0.13	0.00	0.48	0.80
		5 <= Tout < 10	3.00	0.35	5.00	0.29	3.93	0.00	0.18	0.00	1.07	1.37
		10 <= Tout < 15	–	–	–	–	–	–	–	–	–	–

Fig. 7 shows the Actual EF values from Fig. 3 clustered into three groups according to the daily outdoor temperature value and represented as boxplots. The mean values of Actual EF of these clusters differ from the 'maximum' Expected EF values. Table 7 shows the errors between the different Expected EF values and their corresponding set of Actual EF values. In general, the error is reduced when considering the data-based mean values as the Expected EF. For example, in HA, the $EF\ Gap_{MAE}$ for committed power related to the reference design-based 'maximum' Expected EF is around 1.09 kW_e, (~55 % of the Expected EF value of 1.96 kW_e). Conversely, $EF\ Gap_{MAE}$ is constrained to 0.12 kW_e, 0.14 kW_e, and 0.14 kW_e when respectively considering the three data-based mean values as the Expected EF. These values show an improvement (i.e. relative difference) in $EF\ Gap_{MAE}$ of 87–89 % compared to the $EF\ Gap_{MAE}$ associated with the design-based 'maximum' Expected EF. In HB, this improvement for committed power falls within the range of 70 % to 98 %. With respect to FPI the enhancement ranges are 68–80 % in HA and 49–98 % in HB.

4. Discussion

In this study we defined the EF Gap in the introduction as Expected EF minus Actual EF. We have investigated ways to incorporate monitored data in Expected EF to align it more closely with Actual EF. Two approaches based on data have been used in the FPI rating method. This process is reflected on below.

4.1. Reflection on research questions

- *Are existing EF metrics in the FPI rating method suitable for assessing EF using monitored data?*

It was found that the EF metrics within the FPI rating method can be calculated with monitored data, apart from the final step of calculating the FPI_{limit} and then deriving the building label. This is due to the need to neglect the thermal mass in the reference building, which is not possible without a physical model of the building. Therefore, currently the use of this type of models is the only option for labelling the building's EF. Our recommendation is to re-think this final step so that it can be created using approaches based on data.

Despite the suitability of most of the EF metrics to incorporate monitored data, the pre-set boundary conditions used in these metrics were found not to reflect actual DR within the case study buildings, for example, the setpoint temperatures, the DR strategy, and the outdoor temperatures. The effect of this on the calculated EF and the EF Gap are discussed below in answer to the third research question. However, here we reflect specifically on one parameter – the response time. According to the FPI rating method, this is calculated by turning the heating off and waiting until the building cools down to a certain temperature. This is probably not possible to implement in reality, therefore using monitored data the response time will usually be capped at the length of time the heating is off according to real constraints such as the length of the DR period, in turn determined by external factors such as a time of use tariff. This particular metric, the response time, accounts for 60 % of the weighting in the FPI, so this inability to capture what is really meant in the FPI rating method due to practical constraints is problematic. We recommend complementing the rating of EF using the 'maximum' response time by additionally considering fixed DR periods (e.g. 1–3 h for PSS). Instead of using metric values that represent a 'maximum' state, the EF rating could then consist of a matrix of values, which represent the EF at a higher resolution covering different situations.

Despite the benefits of using monitored data, there are also several difficulties. Firstly, we showed the difficulty of obtaining days with similar enough weather to construct baselines. Secondly, the occupants in both houses made adjustments to the setup which made the analysis more difficult. Thirdly, as mentioned above, the real schedules used led to fixed response and recovery times for all cases. For example, the buildings were not allowed to take as long as they needed to recover to their pre-DR temperature – either the occupants used the HP for DHW after the end of the DR period, or they used a lower temperature setpoint overnight.

- *What is the difference between design and operational EF when calculated using monitored data, and how does this contribute towards explaining the EF Gap?*

We found the model-based Expected EF calculated by applying the FPI rating method to deviate from the Actual EF, in some cases by a high proportion (e.g. 55 % on average in Committed Power for HA). Four reasons for this are identified: firstly, the indoor temperature setpoints and outdoor conditions are different for the design and operational level. The Expected EF under design conditions is specific to the worst case day, and the indoor setpoints are standardised. Secondly, and related to this point, building physical models are able to maintain the indoor temperature at the setpoint, whereas in reality the indoor temperature fluctuates as the HP cycles. Thus, comparing the indoor temperature under DR conditions to a baseline of no DR (either generated using a model based on data, or directly measured during real operation) involves contrasting two inherently variable profiles. This can lead to the recovery time being highly variable, since in different situations but with similar conditions, the indoor temperature under DR can meet the one of the baseline at different times simply because of the HP's cycles.

Thirdly, the concept of the recovery period was found not to fully exist in the two real cases, due either to prioritisation of other energy services (DHW) or to a lower space heating temperature setpoint being used in the evening/night than before the DR period. Thus, even though the recovery period is a useful construct in the FPI rating method to express the building's ability to restore original internal conditions, it may not be a valuable aspect of EF performance in evening peak shaving in some countries.

Finally, the existing FPI rating method is based on a PSS, while in reality this may be accompanied by preheating – for example in the UK domestic tariffs are encouraging this combined approach.[2] Clearly, applying the metric for situations in which it was never intended will not lead to insightful results. However perhaps the FPI rating method could be extended to include other types of DR strategy for which monitored data may be more likely to be available. This could be of use for both the users and network owners to better plan and implement DR programs.

- *How can data best be incorporated into existing EF metrics and ratings?*

Two approaches based on data were investigated in this paper to derive the variables to enter the FPI rating method: a model-based and a data-only method. Both approaches were found to give a good estimate of EF, and both were found to have strengths and weaknesses as described below.

A model-based approach allows generation of baseline profiles which are time-coincident with flexible profiles, and make use of these profiles to quantify EF values. It also allows creation of a more robust set of reference values that enable characterising EF, constraining the risk of an EF Gap. The model, however, had limited capacity to reproduce the real operation of energy systems at high temporal resolutions, as seen in Section 2.4.2 for our data-driven model; better results are obtained by aggregating hourly. Whether this is a problem or not depends on the application: hourly resolution is adequate for hourly dynamic electricity pricing, whereas higher time resolution is needed for use cases such as optimising the self-consumption rate of fluctuating renewable energy production. Another drawback of data-driven models is their limitation of predicting conditions for which training data is lacking or missing.

Alternatively, it is possible to quantify EF using timeseries data only,

[2] See for example https://octopus.energy/smart/cosy-octopus/.

without making use of energy models. Real load shifting events and counterfactual periods are used and as such the EF predictions are based on real operation of the system. This relies on using baselines which are not time coincident with DR events, but instead which are selected due to similar weather conditions and capture the real operation of the energy system. The results obtained are in line with those from the model-based approach, however the method seems to require a large amount of data in order to match most/every real DR event to a suitable baseline period – this was not possible in our case studies in which data was collected for around 3 months.

The data-only approach has potential to be further developed by refining the selection of baseline days. For example, incorporating additional weather variables – such as solar radiation – into the selection could improve the matching process, as it does differentiating between weekdays and weekend days.

Both the model-based and the data-only approaches have the benefit that characterisation of uncertainty is automatically output with the EF metrics, since estimates of EF are accompanied by confidence intervals. The potential financial benefits of quantifying uncertainty in EF are described in previous literature [38]. Both approaches also present the ability to evolve and develop them over time by extending the underlying data on which they are based, to capture changes in the building's use or configuration. This is discussed in Section 4.2. Conversely, the model is only as good as the training data, and if as was the case here there is a lack of cold days in the dataset, steps have to be carried out to predict the HP's operation during such times.

In the next section we move from calculating the underlying metrics to discussing how these would be presented within a EF rating.

4.2. Application to EF rating systems

In this section we explore how to combine the benefits of the standardised EF rating system using a building physical model with the more realistic predictions from our data-based methods.

EF rating systems, being used to compare buildings to one another, must incorporate a certain degree of standardisation - or normalisation - to enable this comparison. For example, energy performance ratings are standardised for weather and occupant behaviour to render buildings comparable to one another independent of their location or occupant behaviour [39]. The FPI rating method is designed for this purpose.

EF quantification based on models informed by data can standardise weather conditions by fixing a certain outdoor temperature and other weather variables. It cannot however remove the effects of occupant behaviour. Thus, our proposal is not to replace the FPI rating method but to present additional information with it to give a more contextualised and likely picture of available EF. On the certificate would be two sets of values:

1. Arteconi et al.'s 'maximum' Expected EF – allows comparison between buildings independent of occupant behaviour but unlikely to give realistic values
2. Our set of Expected EF values – a contextualised picture of the likely EF achievable using different DR strategies. Given at several standard weather conditions (see Fig. 7) which really occurred during the monitoring period, and incorporating real setpoints and building operation strategies. This would be useful for certain stakeholders, for example distribution grids and aggregators seeking knowledge of how buildings respond to different types of DR events. By incorporating monitored data, when the building occupants change, the rating may also change. This opens the possibility to better characterise the Actual EF.

The data requirements for our Expected EF values would be: type of DR strategy, setpoint temperature schedules, monitored energy demand of the thermal system, indoor temperature and outdoor conditions.

5. Conclusion and further work

There is a move towards data-driven approaches to reduce inaccuracies associated with models and standard boundary conditions not representing the actual energy performance of existing buildings for a variety of reasons. In this article, we applied this to EF by introducing the novel concept of the EF Gap, the difference between the building's Expected EF (the EF at the design level) and Actual EF (the EF at the operational level). We then explore this concept by using monitored data from two case studies in combination with an existing EF quantification and rating method.

Thus, we calculate several EF metrics which in previous literature are ascertained using building physical models and design specifications: response time, committed power, recovery time, actual energy variation and FPI. We determine the value of these metrics for the worst-case day (i.e. coldest day) under predefined DR conditions to represent the Expected EF, and for the monitored days under user-defined DR conditions to represent the Actual EF. The EF quantification derives from comparing the flexible profile of the building under DR conditions to a baseline without DR. For the Actual EF values, the flexible profiles used in the calculation correspond to monitored profiles. To obtain the baselines we applied two method: a data-based method in which a data-driven model is trained with monitored data to predict baselines time-coincident with the flexible profiles; and a new data-only method in which baselines are selected from monitored data from similar days as the DR day. For the Expected EF, the data-driven model is used to determine both the flexible and the baseline profiles.

The results show that most metrics can be calculated using monitored data consisting of HP electricity consumption and indoor/outdoor temperature data of 1–10 min resolution associated with winter days in two UK locations. Other days and/or locations might need to incorporate data on total solar radiation as well. On the other hand, practical constraints governing actual DR mean that some of the results do not properly correspond to the intended metrics. For example, the Expected EF values for response time and recovery time are not meaningful, since the DR in practice does not resemble the theoretical DR conceptualised for the purposes of EF rating. Furthermore the standard assumptions used regarding setpoints and outdoor temperature, while useful for comparing buildings to one another, lead to discrepancies with the calculated values of Actual EF. All of this contributes to the EF Gap.

In addition, the data-only approach shows promising results as an alternative to the model-based approach for calculating Actual EF values: the EF Gap values calculated with the data-only approach show a low relative difference of approximately ± 8 %, compared to the model-based values. Moreover, the results using the model-based approach show MAE reductions of 50–98 % of the EF Gap when considering Expected EF values derived from Actual EF values instead of the design-based Expected EF.

We conclude that EF ratings can be improved by incorporating information on Actual EF performance but that the existing rating method may need to be tweaked to more readily incorporate monitored data. In contrast to only using a single Expected EF, we propose the use of different reference EF values covering a range of bands of outdoor temperature, setpoints and DR types to produce a more effective EF rating, which can also be updated over time (e.g. regularly to better represent the current state and use of the building and/or after significant changes such us new occupants or renovation).

The next steps involve setting out in more detail how the values calculated using data-driven approaches are turned into a rating, given that it is not possible to complete the last step of the rating method (comparing the building's FPI to an $\mathrm{FPI}_{\mathrm{limit}}$ of the same building with neglected thermal mass) without a building physical model. Further details such as which external temperatures, setpoints and DR strategies should be covered also require defining, and further work on the selection methods of baselines for the data-only approach could be undertaken.

Incorporating monitored data into assessments of EF can provide more realistic estimations of what can be achieved, helping flexibility business models, local area strategies and consumers to implement flexibility and thus deliver a reliable and low carbon electricity system.

Funding

This work was funded by the UK Research and Innovation trhough the UK Centre for Research on Energy Demand (EPSRC Reference: EP/R035288/1) for the collection of monitored data.

CRediT authorship contribution statement

Manuel de-Borja-Torrejon: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Gerard Mor:** Writing – review & editing, Validation, Software, Formal analysis, Data curation. **Jordi Cipriano:** Writing – review & editing. **Angel-Luis Leon-Rodriguez:** Writing – review & editing, Supervision. **Thomas Auer:** Writing – review & editing, Supervision. **Jenny Crawley:** Writing – review & editing, Writing – original draft, Supervision, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

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

Data availability

The authors do not have permission to share data.

Appendix A

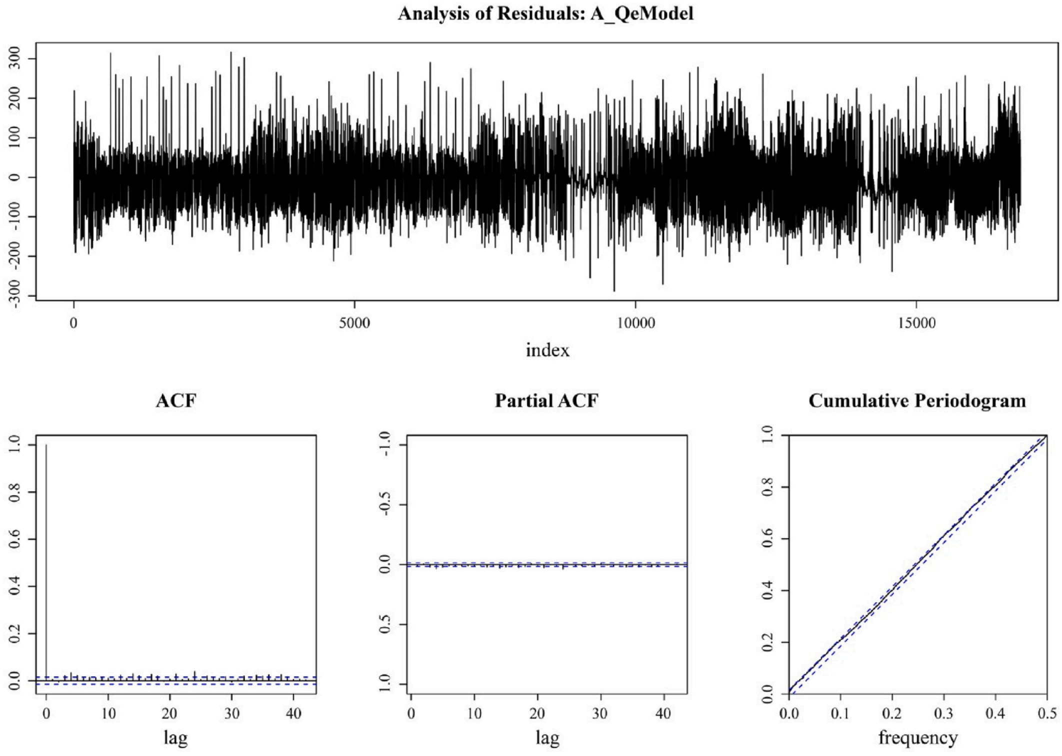

Fig. A1. Residual analysis of data-driven demand-side model of House A.

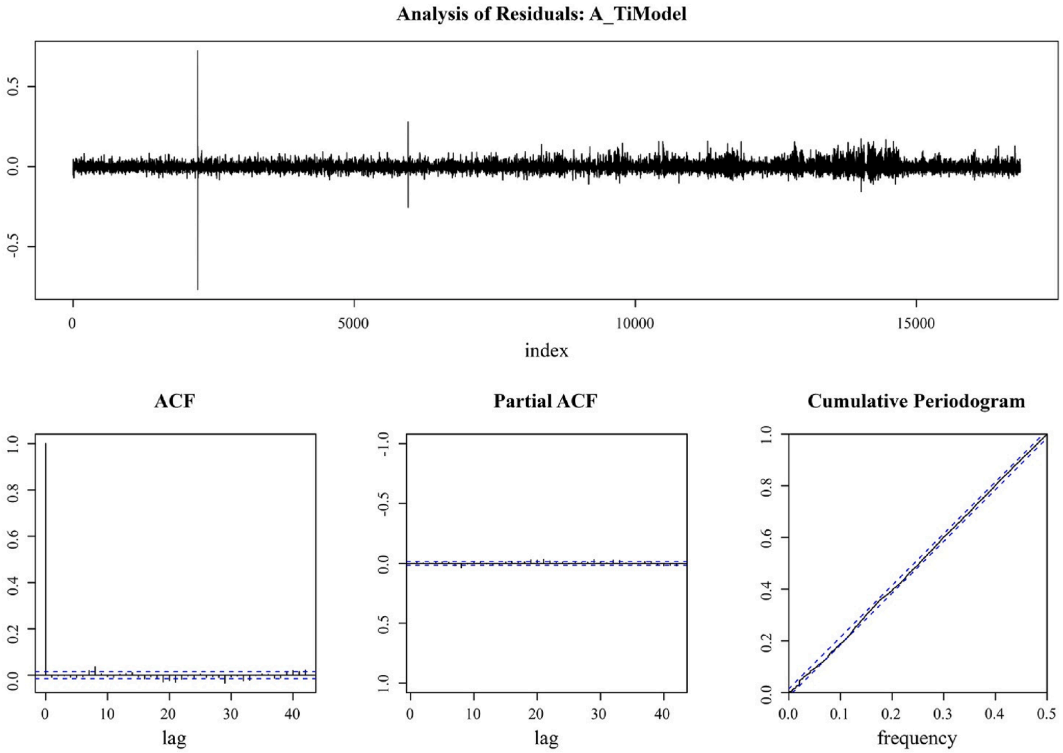

Fig. A2. Residual analysis of data-driven supply-side model of House A.

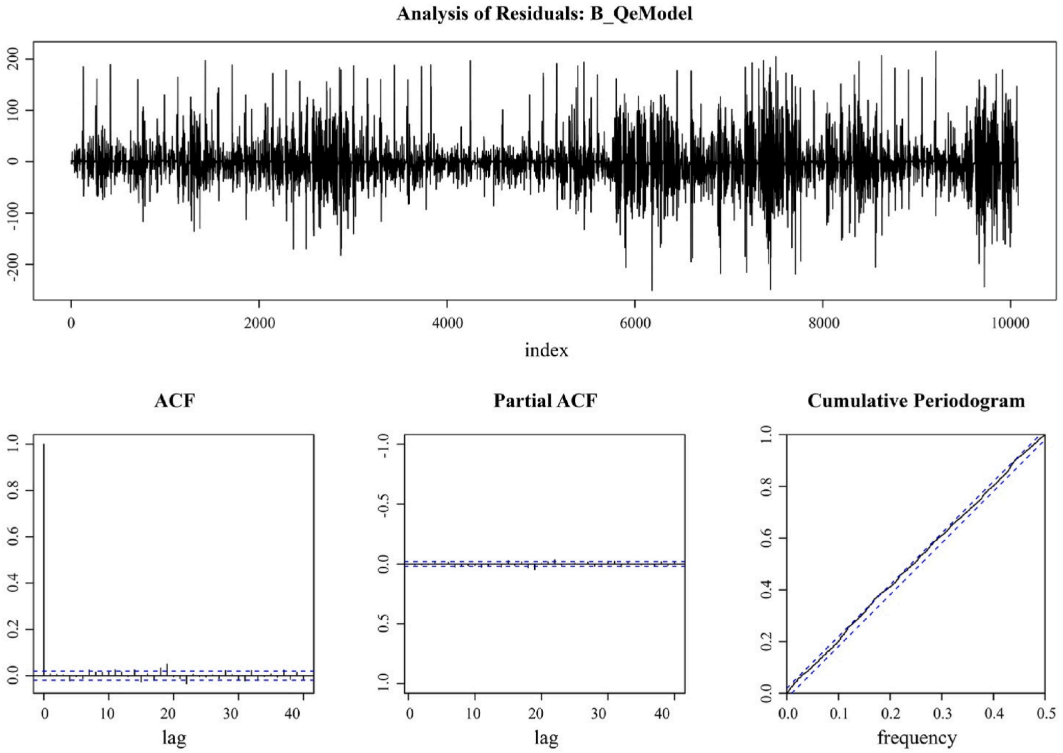

Fig. A3. Residual analysis of data-driven demand-side model of House B.

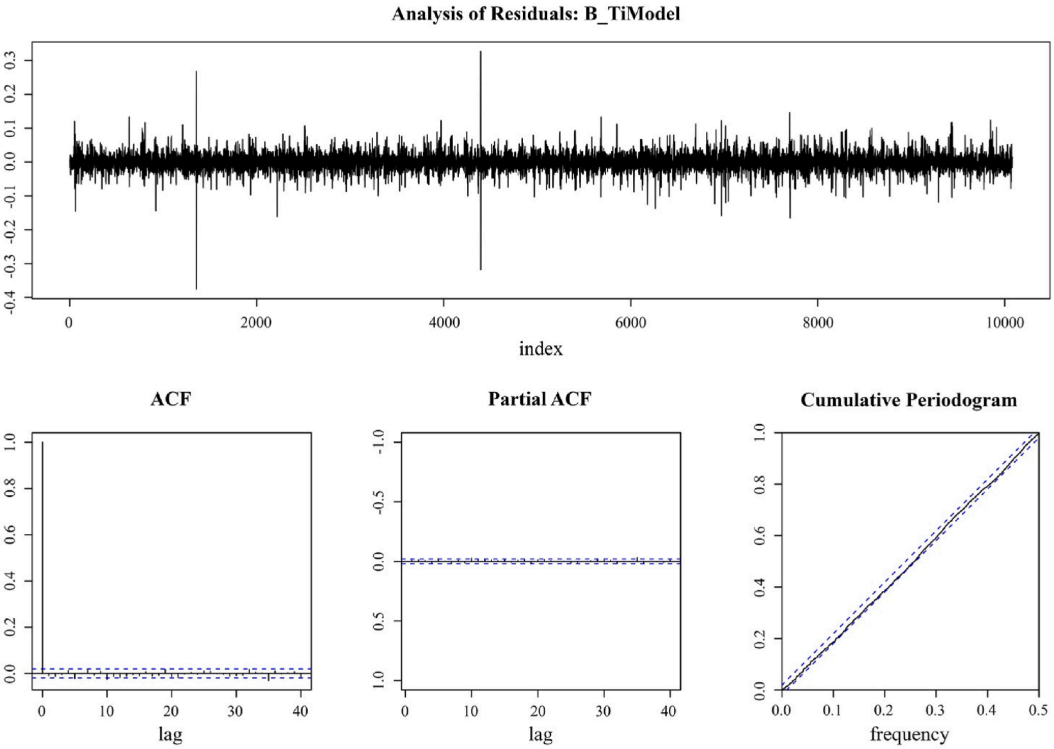

Fig. A4. Residual analysis of data-driven supply-side model of House B.

Appendix B

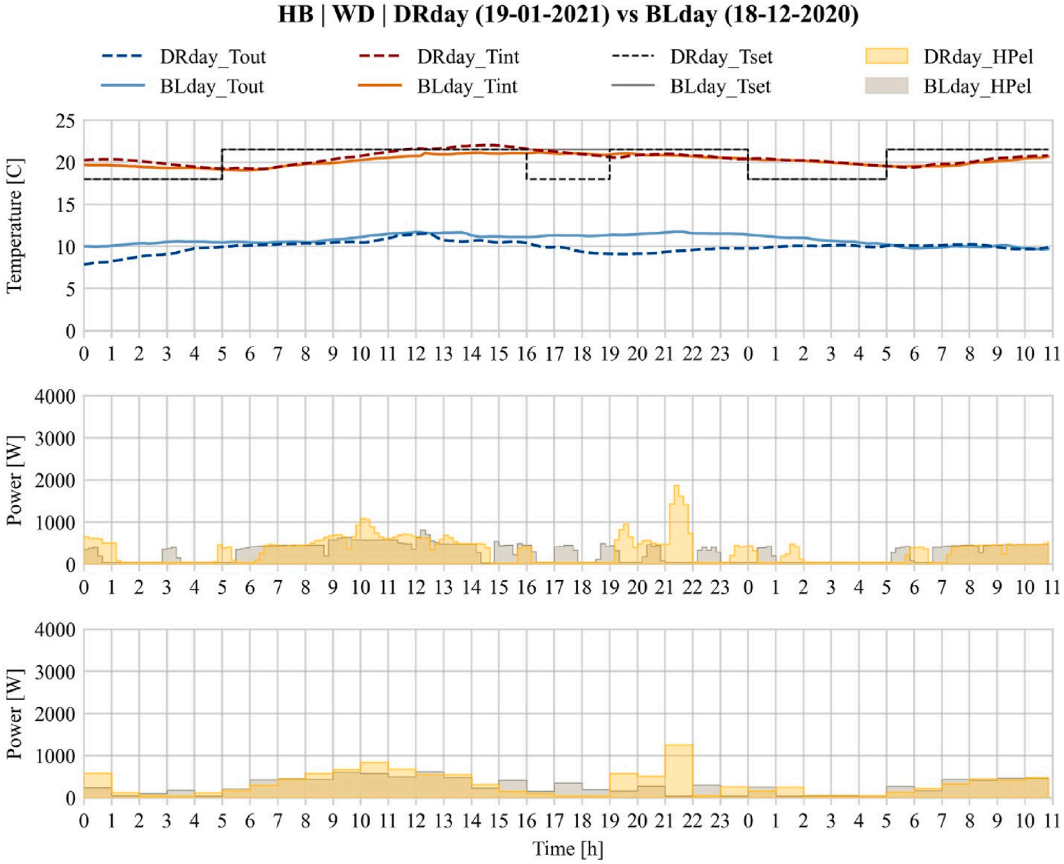

Fig. B1. Comparison between a DR-day and its BL-Day in HB using data-only baseline. Upper graph: temperatures. Middle graph: electrical load in 10-minute resolution. Bottom graph: electrical load in hourly resolution. HB: house B. WD: weekday. DRday: demand response day. BLday: baseline day. Tout: outdoor temperature. Tint: indoor temperature. Tset: setpoint temperature. HPel: electricity consumption of the HP.

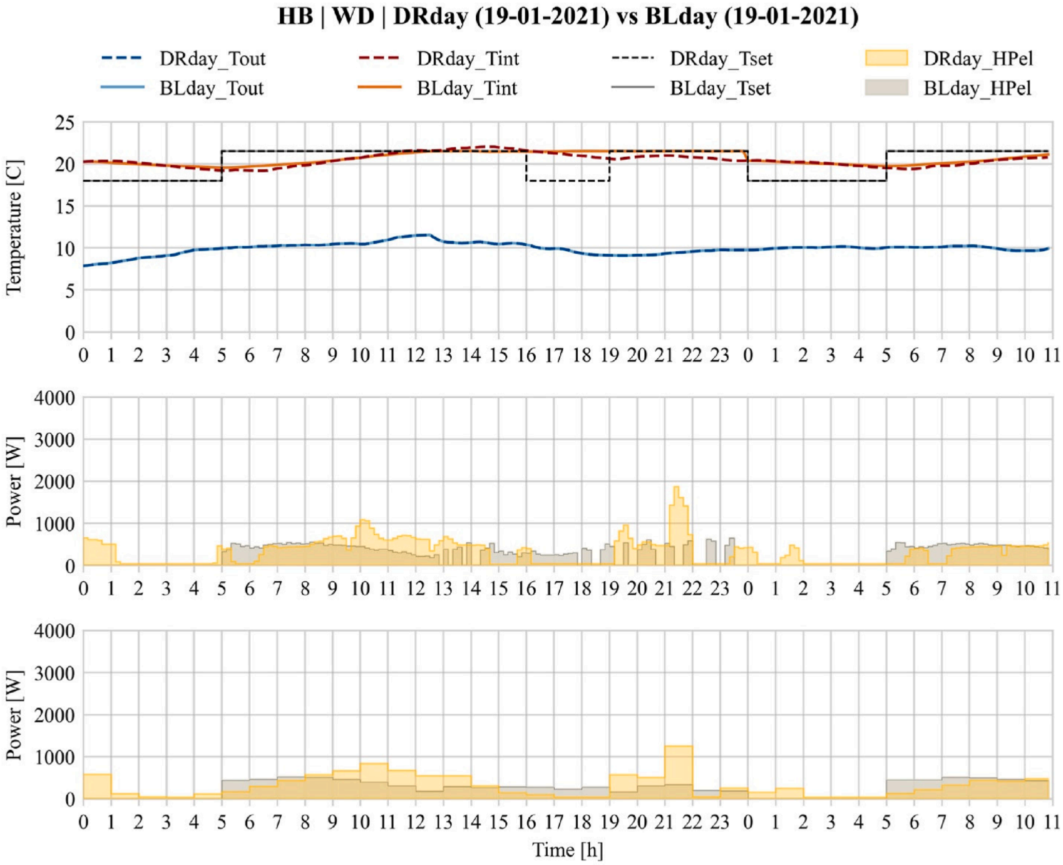

Fig. B2. Comparison between a DR-day and its BL-Day in HB using model-based baseline. Upper graph: temperatures. Middle graph: electrical load in 10-minute resolution. Bottom graph: electrical load in hourly resolution. HB: house B. WD: weekday. DRday: demand response day. BLday: baseline day. Tout: outdoor temperature. Tint: indoor temperature. Tset: setpoint temperature. HPel: electricity consumption of the HP.

References

[1] G. Reynders, R. Amaral Lopes, A. Marszal-Pomianowska, D. Aelenei, J. Martins, D. Saelens, Energy flexible buildings: an evaluation of definitions and quantification methodologies applied to thermal storage, Energ. Buildings 166 (2018) 372–390.

[2] A.J. Marszal, H. Johra, T. Weiss, A. Knotzer, S.Ø. Jensen, H. Kazmi, I. Vigna, R. Pernetti, J. Le Dréau, K. Zhang, R.G. Junker, H. Madsen, R. Amaral Lopes, D. Aelenei, K. Arendt, G. Reynders, A. Hasan, M. Lu, International Energy Agency, Danish Technological Institute, Denmark, 2019.

[3] A. Arteconi, A. Mugnini, F. Polonara, Energy flexible buildings: a methodology for rating the flexibility performance of buildings with electric heating and cooling systems, Appl. Energy 251 (2019) 113387.

[4] European-Union, Directive (EU) 2018/844 of the european parliament and of the council of 30 May 2018 amending Directive 2010/31/EU on the energy performance of buildings and Directive 2012/27/EU on energy efficiency, in: T.E. P.A.T.C.O.T.E. UNION (Ed.), European Union, Official Journal of the European Union, 2018.

[5] G. Mor, J. Cipriano, B. Grillone, F. Amblard, R.P. Menon, J. Page, M. Brennenstuhl, D. Pietruschka, R. Baumer, U. Eicker, Operation and energy flexibility evaluation of direct load controlled buildings equipped with heat pumps, Energ. Buildings 253 (2021) 111484.

[6] Z. You, M. de-Borja-Torrejon, P. Danzer, A. Nouman, C. Hemmerle, P. Tzscheutschler, C. Goebel, Cost-effective CO2 abatement in residential heating: a district-level analysis of heat pump deployment, Energ. Buildings 300 (2023) 113644.

[7] G. Hausladen, T. Auer, J. Schneegans, K. Klimke, H. Riemer, B. Trojer, L. Qian, M. de Borja Torrejon, Lastverhalten von gebäuden unter berücksichtigung unterschiedlicher bauweisen und technischer systeme. Speicher-und Lastmanagement-potenziale in gebäuden (load performance of buildings under consideration of different constructions and technical systems. Storage and Demand-Side Management Potential in Buildings), Fraunhofer IRB Verlag, Stuttgart, 2014.

[8] A. Arteconi, D. Costola, P. Hoes, J.L.M. Hensen, Analysis of control strategies for thermally activated building systems under demand side management mechanisms, Energ. Buildings 80 (2014) 384–393.

[9] J. Crawley, A. Martin-Vilaseca, J. Wingfield, Z. Gill, M. Shipworth, C. Elwell, Demand response with heat pumps: Practical implementation of three different control options, Build. Serv. Eng. Res. Technol. 44 (2) (2022) 211–228.

[10] S. Verbeke, D. Aerts, G. Reynders, Y. Ma, P. Waide, Final report on the technical support to the development of a smart readiness indicator for buildings, Publications Office of the European Union, Luxembourg, 2020.

[11] European-Commission, Commission Delegated Regulation (EU) 2020/2155 of 14 October 2020 supplementing Directive (EU) 2010/31/EU of the European Parliament and of the Council by establishing an optional common European Union scheme for rating the smart readiness of buildings, in: E. Commision (Ed.), Official Journal of the European Union, n.p., 2020.

[12] n.a., SRI implementation tools, in, European Commission, 2023.

[13] A. Arteconi, F. Polonara, Assessing the demand side Management potential and the energy flexibility of heat pumps in buildings, Energies 11 (7) (2018).

[14] R.G. Junker, A.G. Azar, R.A. Lopes, K.B. Lindberg, G. Reynders, R. Relan, H. Madsen, Characterizing the energy flexibility of buildings and districts, Appl. Energy 225 (2018) 175–182.

[15] J. Lizana, M. de-Borja-Torrejon, A. Barrios-Padura, T. Auer, R. Chacartegui, Passive cooling through phase change materials in buildings. a critical study of implementation alternatives, Appl. Energy 254 (2019) 17.

[16] C.M. Calama-González, R. Suárez, Á.L. León-Rodríguez, Thermal comfort prediction of the existing housing stock in southern Spain through calibrated and validated parameterized simulation models, Energ. Buildings 254 (2022) 111562.

[17] E. Burman, N. Jain, M. de-Borja-Torrejón, Towards net-zero carbon performance: using demand side management and a low carbon grid to reduce operational carbon emissions in a UK public office, J. Phys. Conf. Ser. 2069 (1) (2021) 012150.

[18] P.X.W. Zou, X.X. Xu, J. Sanjayan, J.Y. Wang, Review of 10 years research on building energy performance gap: life-cycle and stakeholder perspectives, Energ. Buildings 178 (2018) 165–181.

[19] R. Cichowicz, T. Jerominko, Comparison of calculation and consumption methods for determining energy performance certificates (EPC) in the case of multi-family residential buildings in Poland (Central-Eastern Europe), Energy 282 (2023) 128393.

[20] L. Wederhake, S. Wenninger, C. Wiethe, G. Fridgen, D. Stirnweiß, Benchmarking building energy performance: accuracy by involving occupants in collecting data - a case study in Germany, J. Clean. Prod. 379 (2022) 134762.

[21] n.a., The Government's methodology for the production of Operational Ratings, Display Energy Certificates and Advisory Reports, in, Department for Communities and Local Government, London, 2008.

[22] n.a., Der Energieausweis. Hintergründe, Daten und Empfehlungen zum Energiebedarfs- und Energieverbrauchsausweis, in, Deutsche Energie-Agentur GmbH (dena), 2023.

[23] J. Crawley, E. McKenna, V. Gori, T. Oreszczyn, Creating domestic building thermal performance ratings using smart meter data, Buildings and Cities (2020).

[24] Y. Li, S. Kubicki, A. Guerriero, Y. Rezgui, Review of building energy performance certification schemes towards future improvement, Renew. Sustain. Energy Rev. 113 (2019) 109244.

[25] J. Crawley, D. Manouseli, P. Mallaburn, C. Elwell, An empirical energy demand flexibility metric for residential properties, Energies 15 (14) (2022) 5304.

[26] H.E. Beck, N.E. Zimmermann, T.R. McVicar, N. Vergopolan, A. Berg, E.F. Wood, Present and future Köppen-Geiger climate classification maps at 1-km resolution, Sci. Data 5 (1) (2018) 180214.

[27] A. Martin-Vilaseca, J. Crawley, M. Shipworth, C. Elwell, Living with demand response: Insights from a field study of DSR using heat pumps, in: ECEEE 2022 Summer Study Proceedings, ECEEE, Hyères, France, 2022.

[28] J. Morley, K. Widdicks, M. Hazas, Digitalisation, energy and data demand: the impact of internet traffic on overall and peak electricity consumption, Energy Research & Social Science 38 (2018) 128–137.

[29] J. Remund, S. Müller, S. Kunz, B. Huguenin-Landl, C. Studer, D. Klauser, C. Schilter, R. Lehnherr, Meteonorm. Global Meteorological Database. Version 7. Software and Data for Engineers, Planners and Education. Handbook Part I: Software. Version 7.1, METEOSET, Bern, 2016.

[30] Y. Chen, M. Guo, Z. Chen, Z. Chen, Y. Ji, Physical energy and data-driven models in building energy prediction: a review, Energy Rep. 8 (2022) 2656–2671.

[31] J. Jungwirth, Lastmanagement in gebäuden, doctoral thesis, Technische Universität München (2015).

[32] Z. Wang, R.S. Srinivasan, A review of artificial intelligence based building energy use prediction: contrasting the capabilities of single and ensemble prediction models, Renew. Sustain. Energy Rev. 75 (2017) 796–808.

[33] G. Mor, J. Cipriano, E. Gabaldon, B. Grillone, M. Tur, D. Chemisana, Data-driven virtual replication of thermostatically controlled domestic heating systems, Energies (2021).

[34] R library of the BIGG AI toolbox, in.

[35] G. Mor, J. Vilaplana, S. Danov, J. Cipriano, F. Solsona, D. Chemisana, EMPOWERING, a smart big data framework for sustainable electricity suppliers, IEEE Access 6 (2018) 71132–71142.

[36] T. Reddy, I. Maor, Procedures for reconciling computer-calculated results with measured energy data, ASHRAE Research project 1051-RP, American Society of Heating, Refrigerating and Air-Conditioning Engineers, Atlanta, GA, USA, 2006.

[37] n.a., ASHRAE Guideline 14-2014: Measurement of Energy, Demand, and Water Savings, in, ASHRAE, 2014.

[38] M.M. Hu, F. Xiao, Quantifying uncertainty in the aggregate energy flexibility of high-rise residential building clusters considering stochastic occupancy and occupant behavior, Energy 194 (2020).

[39] J. Crawley, E. McKenna, V. Gori, T. Oreszczyn, Creating domestic building thermal performance ratings using smart meter data, Buildings & Cities 1 (1) (2020) 1–13.