Assessment of Carbon-Aware Flexibility Measures From Data Centres Using Machine Learning

M. Saeed Misaghian, Giovanni Tardioli, Adalberto Guerra Cabrera, Ilaria Salerno, Damian Flynn, *Senior Member, IEEE*, and Ruth Kerrigan

***Abstract*—Potential solutions for unlocking operational flexibility from data centres are investigated. Two types of data centres, in terms of their cooling facilities, are modelled in order to provide an in-depth assessment of potential variations in data centre performance based on their energy consumption and server temperatures. Using synthetic datasets generated from year-long building dynamic simulations under distinct system conditions, four machine learning (ML) models are designed and trained to predict data centre energy consumption and server temperatures. Three different ML algorithms are considered, including random forest, XGBOOST, and multiple linear regression, which all achieve high accuracy ranging from 93.8% to 98.1% for the mean absolute percentage error. The resulting ML models are then utilised to represent a system-wide fleet of data centres, which are integrated within a power system unit scheduling framework for potential demand shifting based on (system) carbon intensity. Five alternative flexibility scenarios are introduced and compared, with the objective of determining the achievable flexibility from a system-wide portfolio of data centres, focusing on concerns relating to server temperatures. On average, a 6.5% load reduction is seen to be achievable for a winter day during periods of high CO_2 intensity.**

***Index Terms*—Artificial intelligence, climate change, CO_2 intensity, data analysis, data centre, demand response, machine learning, renewable energy.**

I. Introduction

A. Background

Recently, the world gathered in Glasgow to reach a new agreement on tackling climate change, where the majority of participating countries agreed to reinforce their commitments, and to intensify activities in the power and energy sector to reduce related carbon emissions. Ireland, which is often seen as a leader in decarbonising its electricity network, introduced a new target of up to 80% of renewables by 2030, as an important measure to tackle climate change, and to achieve climate neutrality no later than 2050, similar to other EU countries [1]. Meanwhile, the world is witnessing increasing demand for computing resources and data centre power due to digitisation, e.g. IT workloads have rocketed sixfold between 2010 and 2018 worldwide, with ensuing environmental consequences [2]. Electricity demand growth in Ireland, for example, over the next ten years is largely expected to come from new data centres, with projections that they will consume approaching 30% of demand by 2030. In 2020, 11% of total Irish electricity demand was dedicated to the data centre industry, contributing 1.85% of Ireland's total carbon emissions in the same year [3]. This rapid digital transformation of businesses is an obstacle to reaching a zero-carbon EU by 2050, unless initiatives are implemented to make data centres carbon-neutral, or more environmental friendly. For example, the "Climate Neutral Data Centre Pact" was agreed by major industry players and the trade association for cloud infrastructure services and data centres in Europe to achieve climate neutrality by 2030 [4]. Recently, the concept of carbon intensity has garnered significant interest, especially regarding how it should be accurately quantified and how to promote its acceptance by end users to change their consumption patterns. Commonly used by TSOs, the total CO_2 emission produced in the system divided by the total system demand is referred to as the "average emission rate" [5]. So, end uses, and particularly large scale ones, such as data centres, could potentially reduce their consumption during periods of high CO_2 intensity, and shift their demand profile towards periods of lower carbon intensity. Given the relative size of data centres, and their overall share of electrical demand, it is not unreasonable for them to provide a greater than proportionate share of "carbon-aware load shifting". Indeed, obtaining equivalent flexibility from a similarly sized block of demand composed of a varied group of small enterprises, public facilities, and residential customers would be more challenging [6].

B. Literature Review

In recent years, considerable effort has focussed on reducing data centre energy costs, improving energy efficiency and/or

Manuscript received 14 March 2022; revised 23 June 2022 and 26 August 2022; accepted 29 September 2022. Date of publication 11 October 2022; date of current version 19 January 2023. Paper 2022-SBSC-0070.R2 approved for publication in the IEEE Transactions on Industry Applications by the Smart Buildings for Smart Cities of the IEEE Industry Applications Society. This work was supported in part by IES Ltd., in part by Science Foundation Ireland (SFI) through the SFI Strategic Partnership Programme under Grant SFI/15/SPP/ E3125, and in part by the UCD Energy Institute. *(Corresponding author: M. Saeed Misaghian.)*

M. Saeed Misaghian is with the School of Electrical & Electronic Engineering, University College Dublin, 4 Dublin, Ireland, and also with the Integrated Environmental Solutions Limited (IES), Glasgow, U.K. (e-mail: mohammad-saeed.misaghian@ucdconnect.ie).

Giovanni Tardioli, Adalberto Guerra Cabrera, and Ruth Kerrigan are with the Integrated Environmental Solutions Limited (IES), Glasgow, U.K. (e-mail: giovanni.tardioli@iesve.com; adalberto.cabrera@iesve.com; Ruth.Kerrigan@iesve.com).

Ilaria Salerno is with the School of Mathematics, University of Edinburgh, Edinburgh EH8 9YL, U.K. (e-mail: I.Salerno@sms.ed.ac.uk).

Damian Flynn is with the School of Electrical & Electronic Engineering, University College Dublin, 4 Dublin, Ireland (e-mail: damian.flynn@ucd.ie).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIA.2022.3213637.

Digital Object Identifier 10.1109/TIA.2022.3213637

reducing carbon emissions, but mostly from an individual data centre perspective [7], [8], [9], [10]. Many challenges still remain in unlocking data centre flexibility, with the most basic approaches relying upon on-site backup generation, uninterruptible power supply units (UPS), or battery energy storage systems. However, utilising UPS systems to provide demand response is typically not possible, unless additional functionalities are incorporated within the power electronics converters [11]. Alternatively, flexibility can be obtained from delay-tolerant IT workloads [12], such that, for example, model predictive control can be applied to optimise data centre operation and increase the system-wide renewables share. A real-time energy management system has also been proposed [9], with the objective of minimising data centre energy costs, but the underlying model only considers limited energy performance details and external power system scheduling influences are not considered [13]. Data centre optimisation within a microgrid has also been investigated, while the potential of using data centres to provide a fast frequency response using UPS and flexible IT workloads has also been evaluated [11].

Similarly, the impact of data centre growth on power system operation has been investigated [10], demonstrating reductions in both electricity price variability and carbon emissions. Microsoft recently introduced a low-carbon scheduling policy for its Kubernetes service [14], while Google has introduced carbon-aware scheduling for their data centres by adjusting virtual server capacity based on system-wide carbon intensity [8]. Although potential solutions involving IT workload re-scheduling have been discussed by big tech companies [8], [14], questions still surround the impact of IT workload re-scheduling on other facilities, such as cooling, and the overall impact on CO_2 emissions, power system operation, and even energy markets.

Analysis of data centres either tends to consider them in great detail [15], or relies upon greatly simplified models [10], [13], which essentially treat all types of data centres as being the same. In practice, however, individual data centres are quite varied, especially in relation to their cooling facilities, such that for relatively mild and humid climates, external airflow is typically used to cool the servers. However, if 24/7 cooling is instead required, then the nature of the cooling facilities will change. Servers represent the major data centre load, with cooling facilities coming next, but they are correlated with server temperatures and IT workload. Ultimately, server temperatures are the main concern for safe data centre operation and, hence, they define the main challenge for demand-side management approaches, due to sensitivity towards secure functioning to deliver requested IT workloads.

C. Research Gap & Main Contributions

1) Dynamic Modelling Approaches of Data Centres: A wide range of numerical models have been proposed for analysing building energy performance [16], [17]. For example, computational fluid dynamics (CFD) simulates air flow and temperature distribution in buildings, such as data centres, to study thermal behaviour [18]. However, computational times are long, and real-time energy management is not feasible [16]. Alternatively, various numerical approaches have evolved, including a Laplacian model to predict temperature distributions [19]. Building and energy simulation models, such as IES-VE, Energy Plus [15], DOE-2, have also been used to design and optimise buildings and HVAC systems, and to determine baselines and probable retrofit energy savings. They can predict a building's thermal performance, but the slow processing time is again a barrier for modelling a fleet of buildings. Resistance-capacitance (RC) network models have also been applied, whereby an electrical analogy is used to simplify building thermal performance and the temperature profile of individual components; however, the modelling simplifications can result in less accurate results. Data-driven methodologies, which have gained much interest lately for examining building energy systems [17], are surrogate models produced from actual data or synthetic profiles, and they can estimate a building's energy use and/or interior temperature more quickly than earlier techniques.

From a power system perspective, and/or implementation of demand response programmes, the thermal performance of each building is generally not seen as being that critical, and, instead, focus is placed on the aggregated electrical power consumption of multiple loads, and simplified models are employed to reduce computational time. However, a model which could predict the energy consumption of individual buildings, while providing reasonably fast computation time, such that a sufficiently large and diverse fleet of buildings could be simulated and aggregated to system level would be preferable. Given their ability to tackle non-linear and complex patterns, while still achieving predictions with high accuracy, machine learning (ML)-based methods, as part of data-driven approaches, offer a potential solution for energy prediction in complex energy systems, which makes them superior to other solutions. However, availability of data can often present an implementation challenge for such methods. The concept of surrogate data centre models for predicting the thermal performance, server temperature and energy usage of data centre using machine learning techniques has not previously been investigated.

ML models require historical training data, and here, synthetic data from different scenarios is leveraged for emulator training. The energy performance of data centres is formulated as a regression ML problem for two data centre types, with discussion on the main technical challenges for flexibility provision from data centres, and their potential impact on power system operation. The dynamic data centre building simulation employed here is merely a tool to support investigation of the feasibility of a system-wide fleet of data centres implementing a demand response programme.

2) Integration of ML Within Optimisation Framework: AI and ML algorithms have attracted interest from research and industrial sectors for providing insightful and sustainable decisions. To date, focus has been placed on classical optimisation frameworks, such as mixed-integer linear programming (MILP) [10], or AI-based decision making methodologies, such as "reinforcement learning" [17], in which ML/deep learning methods are often used to define system components. Reinforcement learning requires knowledge of the system state and understanding of the impacts of individual choices on overall system performance, which can be impractical or computationally costly, particularly for power systems, which are naturally

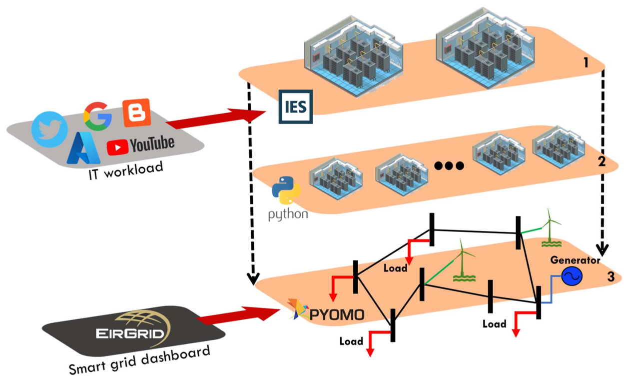


Fig. 1. Model overview, incorporating physics-based data centre models, representative ML models for a system-wide fleet of data centres, and a power system network, for optimal data centre load shifting based on power system CO_2 intensity.

complex and uncertain [20]. Integrating ML algorithms into classical optimisation frameworks represents an alternative method [21]. However, executing multi-period optimisation has not previously been considered, including incorporating an ML end use model within a typical optimisation framework.

The main contributions of the current work can be outlined as: 1) dynamic modelling of data centres, with the aim of assessing their energy performance, before generating a synthetic ML training dataset to efficiently predict their electricity consumption and server temperatures, 2) investigating distinct flexibility solutions for data centres in terms of IT workload rescheduling, cooling equipment, and battery utilisation, 3) bridging ML model and optimisation frameworks for time-shifting of data centre demand based on (system) carbon intensity, and 4) studying the impact of carbon-aware load shifting from a power system perspective. The remaining paper is organised as follows: Section II presents an overview of the modelling structure. Section III presents the data centre dynamic modelling using IES-VE software, and introduces the machine learning models. The optimisation framework, including data centre flexibility, is presented in Section IV, and simulation results for the considered test system with different flexibility measures are illustrated in Section V. Section VI concludes the paper.

II. Model Overview

Data centres are viewed here as flexible loads, such that their electricity consumption can be partially shifted away from high intensity CO_2 periods towards low intensity periods, based upon artificial intelligence and optimisation techniques, as shown in Fig. 1. At the top level, an actual data centre. with two types of cooling systems, is modelled and simulated using IES-VE, a dynamic building simulation tool [22] to determine data centre power consumption and server temperature behaviours. In the middle level, an application programming interface (API), developed in Python, scrapes actual IT workloads from available online platforms, such as Google, to refine the IES-VE simulation model parameters. Subsequently, synthetic datasets (for each cooling type) are formed from the IES-VE outputs, based on simulating year-long variations in IT workloads and server temperature setpoints against time-varying meteorological data. ML models are subsequently trained to predict data centre power consumption and server temperatures, depending on the type of cooling, and with an aggregated system-wide fleet of data centres obtained through combining differing numbers of ML models for each cooling type, with randomised parameters (IT load and temperature setpoints). Finally, in the bottom layer, data centres are supplied with forecasts of (power system) carbon intensity by the TSO, which cause them to adjust (time shift) their load consumption patterns. Individual data centres are connected to different electrical network buses, with different IT workloads and temperature setpoints. In addition, an API, "Smart Grid Dashboard, EirGrid" [23], collects real-time/historical information relating to system demand, wind power, and carbon intensity for the Irish grid. Subsequently, a multi-period optimal DC power flow (OPF) optimisation is solved, incorporating data centres with the ability to shift their loads based on the carbon intensity input.

III. Data Centre Modelling

A. IT Workload

Data centre workload is defined by the volume of arriving jobs submitted by users, with each job assigned to a specific tier [8], such that higher tier jobs permit almost zero tolerance to disruptions and must be completed immediately upon arrival, while lower tier jobs are more delay tolerant and can be completed any time before their deadline. Here, jobs are divided into two categories, namely inflexible (interactive or real-time) or flexible (delay tolerant or batch) jobs [24]. Examples of flexible IT workloads include data compaction and distributed computation for processing videos. In contrast, inflexible IT workloads include streaming of videos on Netflix, YouTube, tweeting on Twitter or posting a photo on Instagram, which represent tasks that must be performed immediately.

Considering the impact of realistic IT workloads, comprising both flexible and inflexible categories, on data centre energy consumption and internal server temperatures, is often neglected [10], [11]. To this end, historical data [25] for the year 2020 in Ireland is normalised here to reflect the national inflexible IT workload, accounting for, for example, utilisation of Gmail, YouTube, Blogger, Websearch, Google Sites, and Google Docs. Regarding the flexible IT workload, a Microsoft Azure Cloud Computing Service dataset for a large company in Ireland has been anonymised. The flexible IT workload data has, subsequently, been randomised to capture a diverse fleet of commercial companies using cloud computing services. The aggregated approach tends to reduce workload forecast uncertainty. Normalised data for the inflexible dataset is illustrated in Fig. 2(a) for one week for various categories, while Fig. 2(b) shows the aggregated inflexible and flexible IT workload. As per [24], it is assumed that the inflexible and flexible workloads are roughly equal.

B. Data Centre Dynamic Model

In order to study the energy performance of data centres, the IES-VE commercial software [22] is used, which incorporates the dynamic thermal physics of data centres. After an initial data

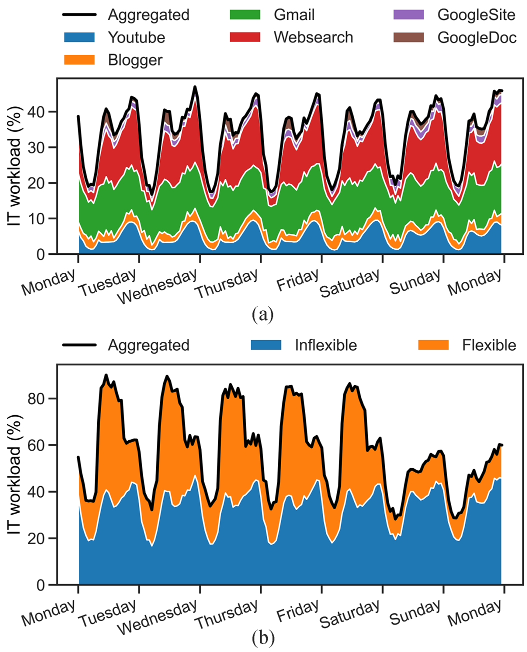


Fig. 2. (a) Normalised inflexible IT workload for a typical week based on historical data in Ireland, consisting of Gmail, YouTube, Blogger, Websearch, Google Sites, and Google Docs (b) Normalised IT workload comprising flexible and inflexible IT workloads based on historical data.

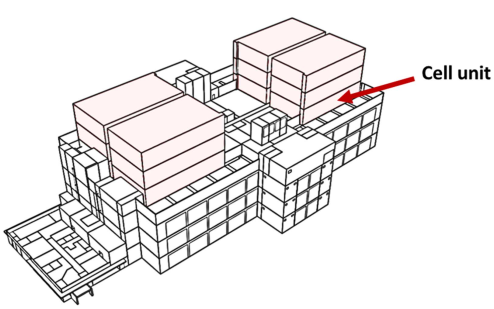


Fig. 3. Data centre building configuration for dynamic simulation using software IES-VE.

gathering campaign, the data centre model is developed using the building dynamic simulation software IES-VE. The simulated facility is a real data centre of approx. 40,000 m^2 with a total volume of approx. 330,000 m^3. The building consists of two main parts: cell rooms, where servers and IT equipment are located (approx. 19,000 m^2) and auxiliary facilities (remaining 21,000 m^2). The cells consist of 24 different units, approx. 800 m^2 each (see Fig. 3), while the auxiliary rooms and facilities include: administrative rooms, offices, backup generation units, electrical units, mechanical and storage rooms, UPS units, amenities and personnel facilities. Each cell unit is modelled assuming a 60 cm steel framing structure for the external walls, with a total U value of 0.22, lighting internal gains modelled considering fluorescent lighting with 7 W/m^2 intensity, and IT equipment and servers modelled with 2500 W/m^2 intensity.

Two cooling strategies, "standard" and "free," are considered with further details provided later in Section III-B1 and III-B2.

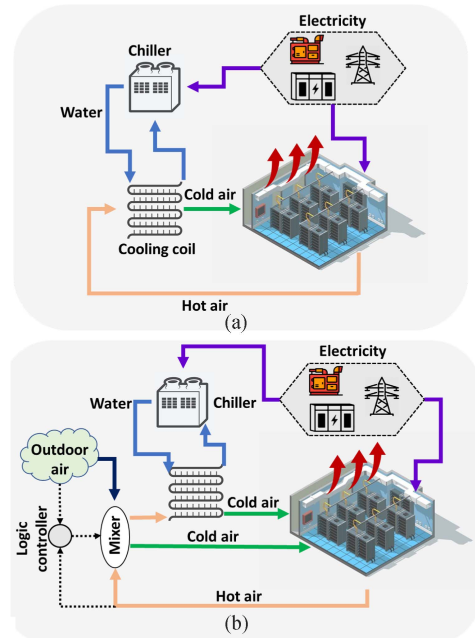


Fig. 4. Data centre configuration (a) standard (b) free cooling.

Generally speaking, for standard cooling, each cell unit is considered as a single thermal zone served by a dedicated air-side heating, ventilation, and air conditioning (HVAC) system. The HVAC is modelled as a packaged single-zone computer room air-handling (CRAH) unit, consisting of a chilled-water cooling coil (no heat), a variable volume fan, and relative humidity zone control. For the free cooling configuration, a similar HVAC system is considered with the addition of an outside air economizer with a dry-bulb temperature high limit. The chilled water loop was modelled considering electric water-cooled chillers with variable speed pumps on the secondary circuit with a total installed cooling capacity of 90 MW. The model considers outdoor temperature, external relative humidity, and IT workloads as input files, before determining the internal server temperatures, average building temperature, cooling system controls and equipment, and server energy consumption. In this study, simulations are performed for a one year period with a time step resolution of 30 minutes. Due to the complexity of the physical model, a year-long simulation for a single data centre takes approx. 4 hours, with standard cooling, and free cooling models running separately on a PC with Core i7-8750H, CPU @ 2.20 GHz, and 16 GB of RAM.

1) Standard Cooling Model: A simplified configuration for a data centre with standard cooling is shown in Fig. 4(a), including backup generators and batteries in addition to the main grid supply connection. Server temperatures must be maintained within a specified range for correct equipment operation and safety reasons. Chillers are used to cool the servers, with circulating cold water (supplied by the chillers) cooling the hot air. (1) represents on/off operation of the cooling system based on the temperature setpoint, T^{ST}, and T^{Server} is the server temperature. T^{ST} must remain within the range $T^{min} \leq T^{ST} \leq T^{max}$ to ensure that the servers operate safely and reliably, where $T^{max/min}$ is the

TABLE I
SYNTHETIC TRAINING DATASET BASED ON IES-VE

Scenarios	Cooling type	Temperature setpoint (°C)
I	Standard cooling	26
II	Standard cooling	$\sim \mathcal{N}(\mu = 26, \sigma^2 = 0.25)$
III	Free cooling	26
IV	Free cooling	$\sim \mathcal{N}(\mu = 26, \sigma^2 = 0.25)$

maximum/minimum allowable temperature range.

$$\Lambda = \begin{cases} On, & \text{if } T^{Server} \geq T^{ST} \\ Off, & \text{if } T^{Server} \leq T^{ST} \\ \text{No change,} & \text{Otherwise} \end{cases} \tag{1}$$

2) Free Cooling Model: In contrast to standard cooling, which is entirely dependent on chillers for cooling down the servers, free cooling can also employ outside air, with a logic controller able to switch between both options. If outside air is being used for cooling purposes, the mixer in Fig. 4(b) combines the hot air from the servers with the outside airflow to cool the servers. Otherwise, the mixer blocks external airflow and only the chillers are used for server cooling. The control logic to utilise outside air is summarised in (2), where T^{amb} represents the outdoor temperature and δ is a safety deadband. The on/off chiller control logic is similar to (1).

$$\Upsilon = \begin{cases} \text{Outdoor air,} & \text{if } T^{amb} \leq T^{ST} + \delta \\ \text{Chiller,} & \text{if } T^{amb} \geq T^{ST} \end{cases} \tag{2}$$

3) Scenarios & Simulation Results: In order to analyse and compare the performance of standard and free cooling, before constructing individual ML models (Section III-C), four different IES-VE scenarios have been considered, as summarised in Table I. The same outdoor temperature and external relative humidity annual profiles are applied for each scenario, as measured at a Dublin weather station for 2020, along with the same IT workload. Year-long simulations are performed assuming, (1) a fixed temperature setpoint of 26 °C and (2) a randomised temperature setpoint with a mean of 26 °C and variance, δ^2, of 0.25, as outlined in Table I. In reality, a random temperature variation is not realistic, but it is applied here to support training of the machine learning model. A training set consisting of $4 \times 8760 \times 2 = 70,080$ rows of data is subsequently generated at a 30-minute resolution.

For the same operating conditions, and in an Ireland context, free cooling requires approx. 11% less electricity annually, due to the available opportunities for outside air cooling. However, standard cooling enables tighter control of the server temperatures, most noticeably when IT workloads are higher. Fig. 5(a) shows a comparison of the power consumption from both free and standard cooling arrangements for a one week winter period, with an average outdoor temperature of 7.9 °C and an average external relative humidity of 89%, considering only inflexible IT workloads. It can be seen that the IT workload is the main driver for power consumption, but, as expected, free cooling helps to reduce dependency on the chillers. For the time period shown in Fig. 5(a), data centres with free cooling consume an average/maximum of 20.5/24.6 MW, which increases to 23.0/31.4 MW with standard cooling. Fig. 5(b) shows a pie chart of energy consumption for the same period, showing that chillers account for 3.7% using free cooling, but 12.1% with standard cooling.

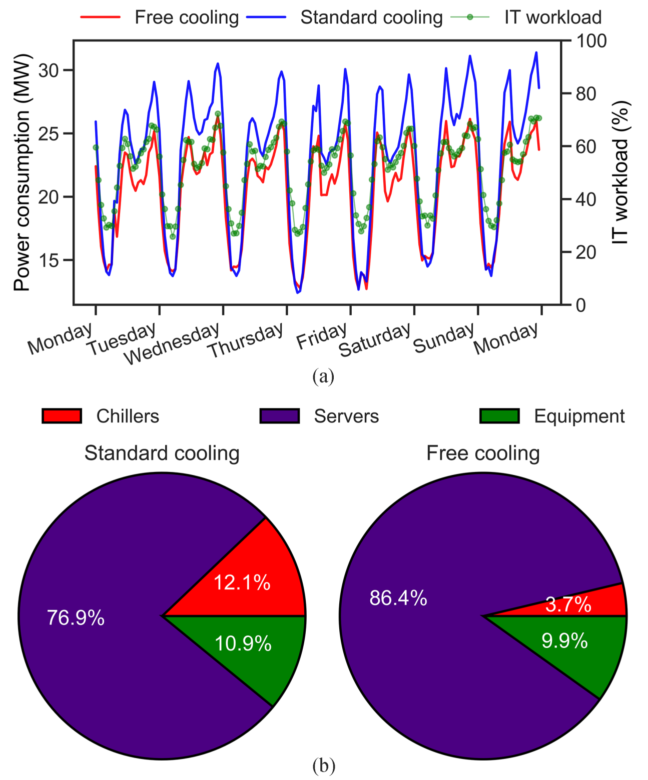


Fig. 5. Data centre (a) power consumption, (b) energy consumption for free and standard cooling, for one week period.

TABLE II
INPUT FEATURES AND TARGET VARIABLES (GREY HIGHLIGHT)

Feature description	Size	Features
Hourly meteorological information	2	External relative humidity (RH) Outdoor temperature (°C)
Calendar information	4	Year, month, day, hour
Holiday (one-hot code)	2	Weekend and public holiday Weekdays
IT workload	1	Flexible workload + Inflexible workload
Chiller setpoint	1	T^{ST} at t
Hourly server temperature	4	T^{server} at $t-1, t-2, t-3, t-4$
Hourly power consumption	1	Total consumption (chillers + equipment + servers)
Server temperature	1	T^{server} at t

C. Machine Learning

Machine learning algorithms are now introduced with the aim of predicting data centre power consumption and server temperatures.

Typical load and temperature forecasting features are incorporated to predict the target variables of interest, i.e. power consumption and server temperature, as outlined in Table II . IT workload is considered as an input variable due to its large impact on both target variables. Outdoor temperature, external relative humidity, temperature setpoint, and calendar information are also considered. Temporal and holiday features are transformed from categorical data to numerical data through a one-hot encoding procedure to capture the impact of normal working days, and otherwise. The thermal inertia of the building, and its impact on the temperature and power consumption of the servers, is recognised by including lagged values of the server temperatures for 4 hours (i.e. $t-1$, $t-2$, $t-3$, and $t-4$).

For training and testing purposes, the dataset has been split 70:30, with each data feature normalised using the MinMaxScaler approach [26], based upon the training dataset. Three distinct ML regression models have been employed, namely multiple linear regression (MLR), random forest regression (RF), and extreme gradient boosting (XGBoost) to predict the target variables of interest. The choice of algorithms is based on industry deployment and consideration of some principles: 1) complex models to be avoided, while also reducing the number of variables; 2) following the international performance measurement and verification protocol (IPMV) [27], all measurement and verification activities should be transparent and adaptable to future changes; and 3) lightweight loading on cloud system services, e.g. Microsoft Azure, to enable organisations to focus on business operations rather than complicated IT infrastructure maintenance. For reasons of transparency, interpretability, flexibility to future changes, and accessibility across all cloud system services, MLR would be anticipated to be a preferred option, subject to meeting modelling performance requirements. For similar reasons, the RF approach is also popular, and unlike other ML models, such as k-nearest neighbours, its default settings appear to work well across a wide variety of datasets [28]. XGBOOST is selected as it offers one of the fastest ML approaches with high resilience and flexibility, as seen in Kaggle ML contests [29]. In each case, 4 separate ML models are required, representing power consumption and server temperature, for standard and free cooling data centres.

Multiple Linear Regression: MLR is a well-established ML approach that aims to find the relationship between several independent variables and one dependent variable, (3),

$$Y = \beta_0 + \beta_1 X_1 + +\beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \qquad (3)$$

where, for each observation, Y is the dependent variable, X_k is a vector of independent variables, and β_0, β_k, and ε represent a constant, slope coefficients, and the model error. MLR can be understood, in an ML context, by considering input X as the training dataset and input Y as the target variable, with β_k to be suitably selected to represent the influence of the considered features in the dataset [30].

Random Forest Regressor: Random forest proceeds in an iterative manner to perform binary feature splits, and, in so doing, create a decision tree structure. By randomly selecting feature subsets, rather than investigating all possible splits, individual nodes are split based on the "best" feature, until further splits are no longer possible. RF employs a minimum number of "leaves" in order to reduce the likelihood of problem overfitting [28].

Gradient Boosting Regressor: Finally, gradient boosting (GB) attempts to achieve a "good" prediction from an ensemble of "poor" prediction models, which is achieved by combining "poor learners," based on the prediction errors, to create a "good learner". In particular, XGBRegressor (XGBR) is an efficient open-source representation of the GB algorithm, capable of achieving high computational performance [29]. Table III summarises a performance evaluation for the considered ML models, where common forecasting evaluation metrics are considered, including mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE) and R-squared value. In general, it tends to be easier to predict the server temperature for the standard cooling over the free cooling model, while the reverse is true for the power consumption, which follows from the nature of the cooling mechanism in each case. A range of approaches are available to improve performance, with hyperparameter tuning playing a vital role. A hyperparameter is a parameter whose value is established prior to the learning process. A number of tuning methods, such as grid search [30], are available, while FLAML [31] has been used here to optimise the hyperparameter values, e.g. number of estimators, maximum number of leaves. Given the similar performance of the three ML approaches, the MLR approach is ultimately chosen as it avoids additional non-linearities and complexities as part of the later optimisation problem (Section III).

TABLE III
PERFORMANCE EVALUATION FOR IMPLEMENTED ML MODELS

	ML Models	Target variable	MAE	RMSE	MAPE	R^2
Standard cooling model	MLR	Temp	0.65 °C	0.82 °C	97.81%	0.93
		Power	1.34 MW	1.75 MW	93.82%	0.89
	RF	Temp	0.57 °C	0.74 °C	98.10%	0.94
		Power	1.12 MW	1.52 MW	94.91%	0.92
	XGBOOST	Temp	0.61 °C	0.79 °C	97.96%	0.93
		Power	1.11 MW	1.48 MW	94.92%	0.92
Free cooling model	MLR	Temp	1.35 °C	1.78 °C	95.81%	0.76
		Power	1.19 MW	1.53 MW	93.53%	0.88
	RF	Temp	1.11 °C	1.5 °C	96.56%	0.82
		Power	0.88 MW	1.21 MW	95.30%	0.93
	XGBOOST	Temp	1.21 °C	1.65 °C	96.25%	0.80
		Power	0.86 MW	1.15 MW	95.41%	0.93

Fig. 6 shows predictions of the server temperature and power consumption using standard cooling, based on the inputs defined in Table II. Considering the MLR coefficients, the predicted server temperatures for the free cooling models tend to be much more sensitive to IT workload compared to the standard cooling models ($\sim 45\%$), since the server temperature is quite dependent on previous hours. In contrast, the power consumption in standard cooling models is more influenced by the IT workload (18%) relative to the free cooling models.

So far, ML models of individual data centres have been developed to predict electrical demand and server temperatures. Subsequently, an aggregated system-wide flexible load is estimated through simulating individual stochastic devices, formed

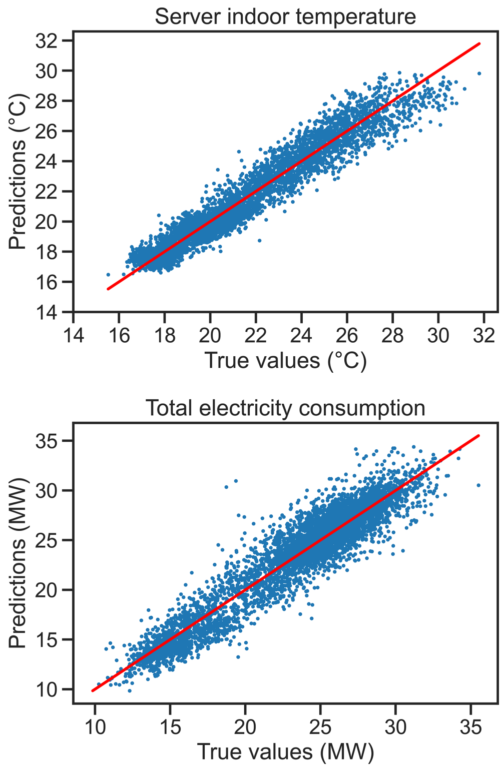


Fig. 6. Server temperature and power consumption MLR prediction for data centre with standard cooling.

from differing numbers of data centre types (standard or free cooling), with randomised parameters (IT load and temperature setpoint). The aggregated ML approach adopted provides an opportunity to obtain insight over critical components of each data centre within the considered system-wide fleet from a high-level perspective, while ensuring fast, yet accurate, prediction of their behaviour.

IV. Optimisation Framework

A. Power System Scheduling

The optimisation framework involves a multi-period optimal power flow, which aims to minimise the total operating (fuel) cost, subject to technical constraints for generating units, battery storage, system non-synchronous share [10], and power flow network limits. Additional constraints are introduced in Section IV-B to capture carbon-aware scheduling of the data centres, which bridges to the ML model, Section III-C. The optimisation framework and DC power flow constraints are based on [32]. The objective function is defined in (4):

$$Min.\left\{\sum_{g,t} C(P_{g,t}) + \sum_{i,t} VOLL.LS_{i,t}\right\} \tag{4}$$

where, $C(P_{g,t})$ is the operational fuel cost of unit g at time t, generating power $P_{g,t}$. $VOLL$ is the value of lost load (due to load shedding) at bus i, $LS_{i,t}$. Technical constraints, such as maximum/minimum power and ramp up/down rates for generating units, battery energy storage systems and their

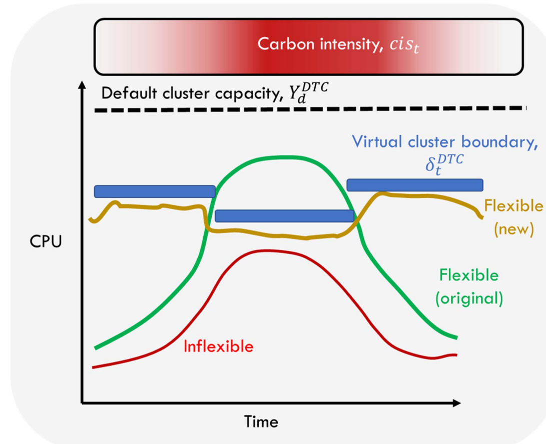


Fig. 7. Virtual boundary to regulate cluster-level capacity.

associated constraints, and generation-demand balancing are also considered, as per previous studies [32].

B. Assessment of Carbon-Aware Flexibility Measures

Given the commitment of major technology companies towards achieving sustainable development goals and reducing carbon emissions, an understanding of flexibility measures that can be provided by data centres, as an increasingly significant load sector, is required. Furthermore, the EU emissions trading system provides an additional motivation for companies to increase their flexibility. Consequently, various technological options are now investigated, based on following a carbon intensity signal, to increase the operational flexibility available from data centres. As stated in Section III-C, multiple linear regression is adopted here to predict the power consumption and server temperature of the data centre population. Data centres are typically connected to the power system via a number of power distribution units (PDUs) on medium-voltage feeders. The PDUs support IT and cooling equipment, with each PDU supplying a few thousand machines, within an overall cluster. An operating system handles IT jobs within a cluster, and assigns tasks across available machines, in terms of allocating resources, e.g. CPU/RAM, job start times [8].

1) IT Virtual Boundary: A virtual boundary can be defined to reduce cluster-level CPU capacity (by adjusting availability for incoming flexible tasks) when (power system) carbon intensity is high, and then raising the virtual boundary (making more CPUs available) when carbon intensity is low. Consequently, flexible IT workloads can be delayed until resources become available, as demonstrated in Fig. 7, where, cis_t represents the time-varying carbon intensity for hour t, being higher here during the middle of the sample day. Given an upper bound on cluster-level capacity, $Y_d^{DTC,max}$, for a given day, $d \in D$, the (initial) flexible IT workload is subsequently reduced beneath the imposed virtual boundary, based upon a time-varying coefficient δ_t. The coefficient is inversely related to the system carbon intensity, i.e. $\delta_t \sim$ k/cis_t, where k is a constant. If the IT workload, IT_t^{DTC}, in each data centre, DTC, accounts for both flexible and inflexible categories at time t, (5) forces the flexible IT workloads to be shifted towards low carbon intensity periods, subject to all tasks being completed by the end of the day, (6).

$$\sum_{t=1}^{t=24} IT_t^{DTC} \leq \delta_t^{DTC}.Y_d^{DTC,max} \qquad \forall d \in D \tag{5}$$

$$\sum_{t=1}^{t=24} \Delta\delta_t^{DTC} = 0 \tag{6}$$

The power consumption, P_t^{DTC}, and equivalent server temperature, T_t^{server}, are calculated using the MLP approach for each data centre server. Since changes in IT workload can significantly affect the server temperatures, minimum/maximum thresholds are imposed, $T^{DTC,min,server}$ and $T^{DTC,max,server}$. Similarly, data centre power consumption is limited to $P^{DTC,max}$, and the chiller setpoints are maintained within minimum/maximum boundaries, $T^{DTC,min,ST}$ and $T^{DTC,max,ST}$, to ensure safe operation.

2) Flexibility From Chillers: Given that the temperature setpoint, T^{ST}, affects the power consumption and server temperature, adjusting the max/min temperature boundaries, based on carbon intensity, can be applied, whereby higher temperatures are permitted during periods of high carbon intensity (with a consequent reduction in power consumption), subject to server safety limits.

3) On-Site Generation: Since most data centres include some form of on-site generation, if only with limited storage capacity, individual data centres could alternatively be disconnected from the main grid, and operated with on-site battery storage/UPS, or even renewables.

V. Case Study

The impact of data centre carbon aware scheduling on power system operation is now studied, based on a modified and simplified model of the Irish power system, including the Republic of Ireland and Northern Ireland [33], which consists of 25 conventional units, 11 (aggregated) wind farm clusters, and three battery energy storage systems. A Python-based energy API scrapes data from the EirGrid smart grid dashboard [23], to obtain real-time/historical data regarding system demand, wind generation, and CO_2 intensity. A total of 10 data centres is assumed, each connected to a specific bus, accounting for $\sim 11\%$ of the total electrical demand. Since IT workload is the main driver for data centre power consumption, a national TSO-level perspective is assumed. IT workloads and temperature setpoints are slightly randomised to capture the variability associated with individual data centres, while hourly profiles for the outside temperature and relative humidity for Dublin are obtained from Met Éireann. Finally, the optimisation problem is coded in Python and solved via Pyomo.

Six scenarios are defined, as part of assessing the potential flexibility from a system-wide fleet of data centres: (1) a "Base" scenario assumes that data centre power consumption is fixed and inflexible across the entire year, (2) '100% Flex" permits the entire data centre load to be flexible, subject to the total aggregated load being unchanged at the end of each day, (3) "FlexIT-NoTemp" is similar to [8], whereby flexible IT loads can be shifted/re-arranged within each 24 hr period, but server temperature constraints are not imposed, (4) "FlexIT+Temp" is similar to (3), except that server temperatures must be restricted within safety margins, where $T^{min,server} = 15$ °C, $T^{max,server} = 32$ °C, (5) "FlexChiller" assumes that the IT workload is fixed, but chiller temperature setpoints can be adjusted based on carbon intensity, as described in Section IV-B2, and finally (6) "FlexChiller+FlexIT" is analogous to (4), but with the additional ability to adjust chiller temperature setpoints. Scenario (4) differs from scenarios (5) and (6) in that the temperature setpoint can fluctuate between fixed boundaries, while (5) and (6) enable the setpoint boundaries to vary based on the carbon intensity.

A representative winter (January) day is now considered, with summary optimised scheduling results, based on the six flexibility scenarios, presented in Fig. 8 and Table IV. For scenarios (2)–(4), the IT virtual capacity of each data centre is adjusted for each timestep based on the estimated CO_2 intensity. Fig. 8 shows the system demand, with load reductions seen from high intensity CO_2 periods towards low intensity periods, relative to the "Base" scenario. For the given day, the highest CO_2 period occurs between 9 am to 5 pm, leading to load

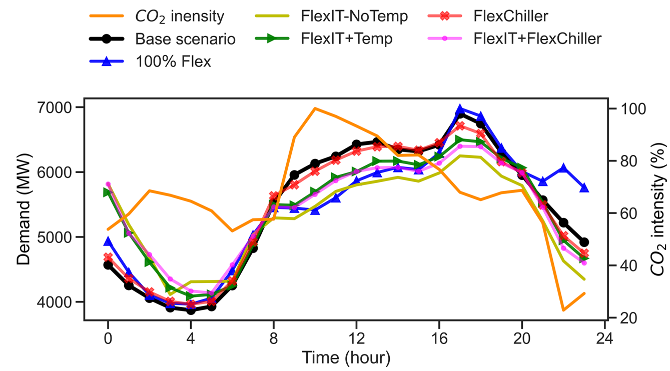


Fig. 8. System electrical demand, before and after applying carbon aware signals to data centres, a winter day.

TABLE IV
Carbon Aware Flexibility Measure Scenarios

Scenarios	Operating cost (pu)	Aggregated demand (GW)	Demand standard deviation (MW)
Base	1	131.1	1002
100% Flex	0.95	131.1	879
FlexIT -NoTemp	0.90	127.4	669
FlexIT +Temp	0.94	130.6	777
FlexChiller	0.98	130.7	938
FlexChiller +FlexIT	0.92	130.4	710

TABLE V
COMPARISON OF DATA CENTRE MODELLING COMPLEXITY APPROACHES

		Current Paper	Ref. [8]	Ref. [10]	Ref. [11]	Ref. [13]
Data centre modelling		Data-driven	Piecewise linear model for CPU and power consumption	✗	✗	Component based approximation
Aggregation approach		Bottom-up	✗	% total demand	% total demand	✗
Flex measures	On-site generation	✓	✗	✓	✓	✗
	IT	✓	✓	✗	✓	✗
	Cooling system	✓	✗	✗	✗	✗
Incentive		Carbon	Carbon	Reserve	Reserve	✗
Power System network		✓	✗	✗	✓	✗

reductions of 6.5%, 8.3%, 4.5%, 0.5%, and 5.5% for the "100% Flex," "FlexIT-NoTemp," "FlexIT+Temp," "FlexChiller" and "FlexChiller+FlexIT" scenarios, respectively. For "100% Flex," the daily power consumption is unchanged, based on how the scenario is defined. However, for "FlexIT-NoTemp" and "FlexIT+Temp," daily load reductions of 2.8% and 0.5% are achieved, due to re-scheduling IT workloads towards times when the cooling requirement is reduced (lower IT loads and lower ambient temperatures), particularly for "FlexIT-NoTemp," given that server temperatures do not restrict the ability to reschedule the IT workload. Similarly, 0.6% and 0.3% load reductions are observed by adjusting the chiller temperature setpoints. On average, a 5.2% operating cost reduction is achieved due to the additional flexibility provided by the data centres compared to the "Base" scenario.

Comparing the various carbon aware approaches against the "Base" scenario, it can be seen that the load profile becomes somewhat flatter, achieving standard deviation reductions of 12.2%, 33.2%, 22.5%, 6.4%, and 29.1% for the '100% Flex," "FlexIT-NoTemp," "FlexIT+Temp," "FlexChiller" and "FlexChiller+FlexIT" scenarios. The reduced variability also implies that the ramping requirements for online (conventional) generators are noticeably reduced. "FlexIT-NoTemp" is most effective at reducing the total operational cost, and flattening the load profile, with "FlexChiller+FlexIT" next in line, but, by scenario design, excessive variations in server temperature are not considered.

For the same winter day, Fig. 9 and Fig. 10 show the flexible IT workload and server temperatures for the "FlexIT-NoTemp" and "FlexIT+Temp" scenarios. Data centres 1–5 employ standard cooling, while 6–10 utilise free cooling. In both cases, the IT workload is shifted from the middle of the day to early morning or late night periods, avoiding the carbon intense periods. As seen in Fig. 9(b), server temperatures are kept well below 32 °C. Not surprisingly, the free cooling data centres reach higher server temperatures, while standard cooling achieves lower temperatures, due to 24/7 operation of their cooling systems. Fig. 10(a) shows that IT workloads are significantly shifted from the middle of the day, but for a few data centres, server temperatures approach 150 °C since temperature constraints do not apply, which is, of course, not acceptable. However, bounds on server temperatures, as outlined in Fig. 9(b) will restrict the ability to shift load at certain times, which will further depend on the nature of the cooling (standard or free) employed by individual data centres, and the external power system / time of year conditions on ambient temperature, IT workload, etc.

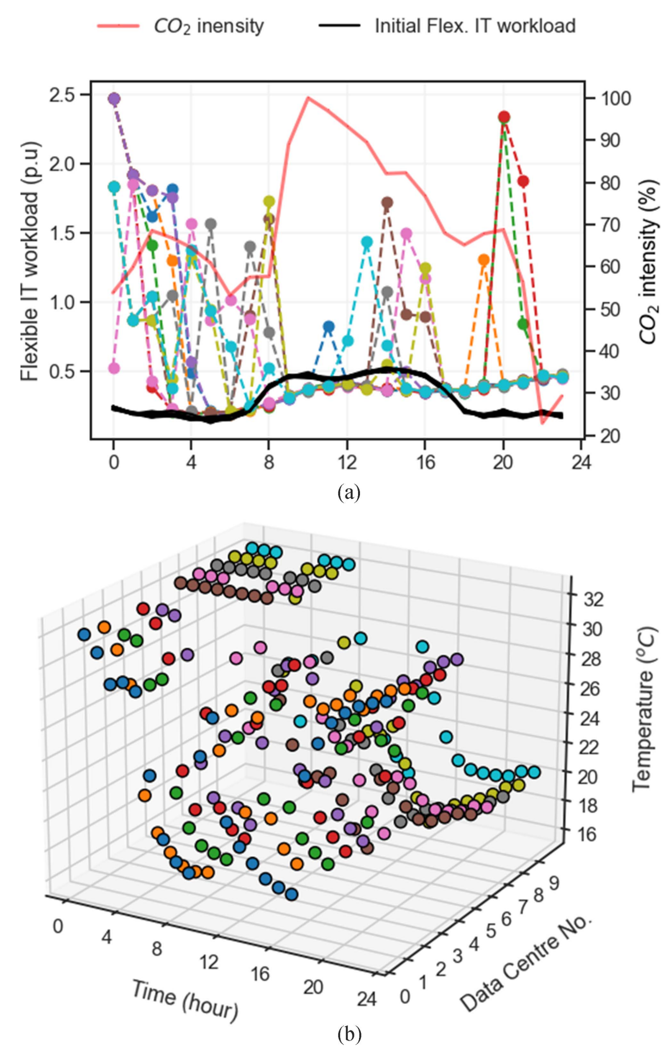
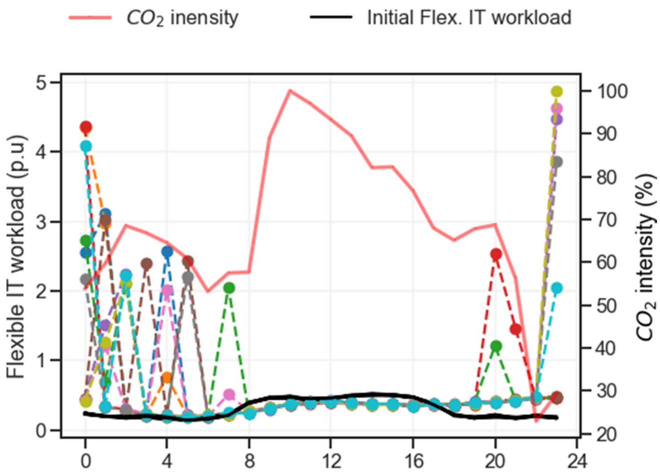
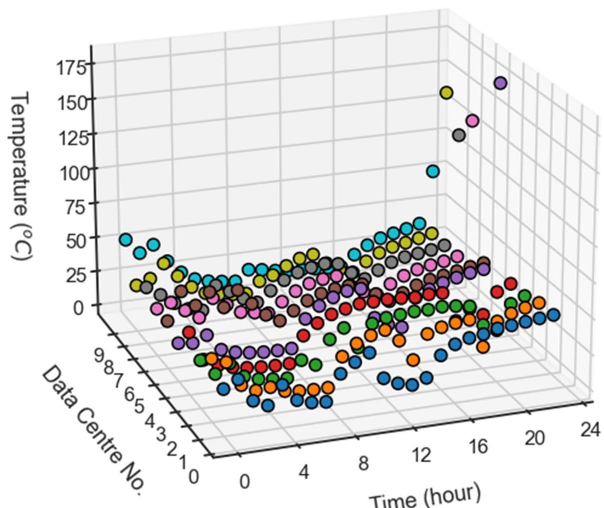


Fig. 9. "FlexIT+Temp" scenario: (a) IT flexible workload, (b) server temperatures, where each dotted line / coloured circle represents an individual data centre.



(a)



(b)

Fig. 10. "FlexIT-NoTemp" scenario: (a) IT flexible workload, (b) server temperatures, where each dotted line / coloured circle represents an individual data centre.

A. Modelling Complexity & Results Comparison

A range of different modelling and analysis approaches have previously been proposed for assessing the available flexibility from data centres. The present work is compared here against four alternative approaches [8], [10], [11], [13], and five indices are considered relating to modelling complexity, as summarised in Table V. The present research is the only option that applies a data-driven method, taking into account not only physical dynamic models of data centres but also ML techniques for reducing the computational time. Furthermore, for system-wide aggregation, a bottom-up strategy is implemented here, whereby a cluster of data centres is represented as multiple ML models with stochastic parameters. Although [10], [11] study a system-wide fleet of data centres, it is fairly simply assumed that data centres account for a given percentage of overall grid demand, and specific data centre models are not included. In contrast, models of individual data centres are the focus of [8], [13].

In terms of measures for adjusting the power consumption of data centres, each study is different: for example, the present study and [8] propose avoiding consumption during carbon intensive periods, while [10], [11] focus on how data centres can support power systems with high renewable shares. The results presented here align with those provided in [10], in terms of data centres supporting system-level flexibility needs, and how IT workload rescheduling can reinforce available flexibility [8]. However, notably, the present study also considers how restrictions on server temperature (and other) variations limit the available flexibility at certain times, while measures such as adjusting chiller temperature setpoints are shown to improve the available flexibility.

B. Replicability, Scalability, and Generalisability

The generalisability of the proposed model is not affected by the modelling approach applied in this study. Fundamentally, its generalisability and scalability are constrained by data availability, which fast digitalisation of the industry is helping to avoid. In addition, the integrated modelling approach adopted here has limited weighting placed on building archetypes, as the main objective was to create a system-wide heterogeneous fleet of data centres by considering 1) different types of cooling systems, 2) distinct weather conditions, in terms of temperature and relative humidity, and 3) randomised IT workloads. Indeed, the uncertainty associated with modelling multiple end uses can be mitigated by randomising the primary inputs and associated parameters. Representing the data centre response in the form of a large and diverse fleet helps to reduce the uncertainty surrounding the available demand response [34].

VI. Conclusion

The potential flexibility from data centres, in order to shift their energy consumption based on a carbon intensity signal is presented. Using detailed thermodynamic simulations of free cooling and standard cooling data centres, a representative dataset of data centre dynamic behaviour is created. Machine learning models are subsequently designed and trained to predict the power consumption and server temperature of individual (free or standard cooling) data centres. A fleet of data centres is represented by a suite of ML models, which is then integrated within a power system unit scheduling framework. Five flexibility scenarios are introduced and compared, with the aim of determining the achievable flexibility from a system-wide portfolio of data centres, focusing on concerns relating to server temperatures. The consequent impacts on the system demand load profile are assessed, including reduced flexibility requirements, and maximum upward ramp rates.

As part of future work, it is intended to consider how best to define carbon intensity and carbon footprint for a system, and specifically how the network location of data centres should play into reducing the observed carbon intensity. A much wider range of system conditions should also be considered, as part of judging the cost effectiveness and utilisation of individual implemented strategies. The work here has also focussed on demand shaping, but it is recognised that data centres can potentially provide various reserve products, including a fast frequency response, which would be particularly valuable in supporting system stability as part of the transition towards low inertia power systems.

References

[1] Department of the Environment, Climate and Communications, Government of Ireland, "Climate Action Plan 2021, Securing Our Future," Nov. 2021. [Online]. Available: https://www.gov.ie/en/publication/6223e-climate-action-plan-2021/

[2] E. Masanet et al., "Recalibrating global data center energy-use estimates," *Science*, vol. 367, no. 6481, pp. 984–986, 2020.

[3] Host In Ireland, "Ireland's data hosting industry," *Biannual Rep.*, 2021. [Online]. Available: https://www.hostinireland.com/report

[4] "Climate neutral data centre pact," Accessed: Dec. 2021. [Online]. Available: https://www.climateneutraldatacentre.net/

[5] H. He et al., "Using marginal emission rates to optimize investment in carbon dioxide displacement technologies," *Electricity J.*, vol. 34, no. 9, 2021, Art. no. 107028. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1040619021001196

[6] Government of Ireland, "Government statement on the role of data centres in Ireland's enterprise strategy," 2022. [Online]. Available: https://www.gov.ie/en/publication/5e281-government-statement-on-the-role-of-data-centres-in-irelands-enterprise-strategy/

[7] Y. Yao et al., "Data centers power reduction: A two time scale approach for delay tolerant workloads," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, 2012, pp. 1431–1439.

[8] A. Radovanovic et al., "Carbon-aware computing for datacenters," *IEEE Trans. Power Syst.*, May 2022, doi: 10.1109/TPWRS.2022.3173250.

[9] Y. Wu et al., "Real-time energy management of large-scale data centers: A model predictive control approach," in *Proc. IEEE Sustain. Power Energy Conf.*, 2020, pp. 2695–2701.

[10] R. Lahon et al., "Impact of wide-scale data centre growth on power system operation with large share of renewables," in *Proc. IEEE 17th Int. Conf. Eur. Energy Market*, 2020, pp. 1–6.

[11] D. Al Kez et al., "Potential of data centers for fast frequency response services in synchronously isolated power systems," *Renewable Sustain. Energy Rev.*, vol. 151, 2021, Art. no. 111547.

[12] D. Paul, W. -D. Zhong, and S. K. Bose, "Demand response in data centers through energy-efficient scheduling and simple incentivization," *IEEE Syst. J.*, vol. 11, no. 2, pp. 613–624, Jun. 2017.

[13] R. Rahmani et al., "Modelling and optimisation of microgrid configuration for green data centres: A metaheuristic approach," *Future Gener. Comput. Syst.*, vol. 108, pp. 742–750, 2020.

[14] A. James and S. Daniel , "A low carbon kubernetes scheduler," in *Proc. ICT4S*, 2019. [Online]. Available:http://ceur-ws.org/Vol-2382/ICT4S2019_paper_28.pdf

[15] K. Sun et al., "Prototype energy models for data centers," *Energy Buildings*, vol. 231, 2021, Art. no. 110603.

[16] E. Oró et al., "Energy efficiency and renewable energy integration in data centres. Strategies and modelling review," *Renewable Sustain. Energy Rev.*, vol. 42, pp. 429–445, 2015.

[17] G. Pinto et al., "Data-driven district energy management with surrogate models and deep reinforcement learning," *Appl. Energy*, vol. 304, 2021, Art. no. 117642.

[18] B. Fakhim et al., "Cooling solutions in an operational data centre: A case study," *Appl. Thermal Eng.*, vol. 31, no. 14-15, pp. 2279–2291, 2011.

[19] V. Lopez et al., "Heat transfer modeling in data centers," *Int. J. Heat Mass Transfer*, vol. 54, no. 25-26, pp. 5306–5318, 2011.

[20] Z. Zhang et al., "Deep reinforcement learning for power system applications: An overview," *CSEE J. Power Energy Syst.*, vol. 6, no. 1, pp. 213–225, 2019.

[21] F. Ceccon et al., "OMLT: Optimization & machine learning toolkit," 2022, *arXiv:2202.02414*.

[22] Integrated environmental solutions limited, "IES VE," 2011-202. [Online]. Available: https://www.iesve.com/products

[23] "Smart grid dashboard," accessed Feb. 2022. [Online]. Available: https://www.smartgriddashboard.com/#all'

[24] Z. Ding, Y. Cao, L. Xie, Y. Lu, and P. Wang, "Integrated stochastic energy management for data center microgrid considering waste heat recovery," *IEEE Trans. Ind. Appl.*, vol. 55, no. 3, pp. 2198–2207, May-Jun. 2019.

[25] Google, "Google's transparency report," accessed Dec. 2021. [Online]. Available: https://transparencyreport.google.com/

[26] F. Pedregosa et al., "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[27] Efficiency Valuation Organization, "IPMVP Generally Accepted M&v Principles," 2018. [Online]. Available: https://evo-world.org/images/corporate_documents/IPMVP-Generally-Accepted-Principles_Final_26OCT2018.pdf

[28] D. N. Cosenza et al., "Comparison of linear regression, k-nearest neighbour and random forest methods in airborne laser-scanning-based prediction of growing stock," *Forestry: An Int. J. Forest Res.*, vol. 94, no. 2, pp. 311–323, 2021.

[29] W. Zhang et al., "Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization," *Geosci. Front.*, vol. 12, no. 1, pp. 469–477, 2021.

[30] N. Zarbakhsh, M. S. Misaghian, and G. Mcardle, "Human mobility-based features to analyse the impact of COVID-19 on power system operation of Ireland," *IEEE Open Access J. Power Energy*, vol. 9, pp. 213–225, 2022.

[31] Microsoft Research, "A fast library for automated machine learning & tuning," 2022. [Online]. Available: https://microsoft.github.io/FLAML/

[32] M. Misaghian et al., "Tri-level optimization of industrial microgrids considering renewable energy sources, combined heat and power units, thermal and electrical storage systems," *Energy*, vol. 161, pp. 396–411, 2018.

[33] A. Soroudi et al., "Power to air transportation via hydrogen," *IET Renewable Power Gener.*, vol. 14, no. 17, pp. 3384–3392, 2020.

[34] M. Aunedi et al., "Economic and environmental benefits of dynamic demand in providing frequency regulation," *IEEE Trans. Smart Grid*, vol. 4, no. 4, pp. 2036–2048, Dec. 2013.

M. Saeed Misaghian is currently working toward the Ph.D. degree in electrical engineering with University College Dublin, Ireland. His research interests include renewables integration, demand response, optimisation, applied machine learning, and data analysis.

Giovanni Tardioli received the Ph.D. degree in mechanical engineering from University College Dublin, Ireland, in 2019. He is currently a Research Lead with Integrated Environmental Solutions Ltd. His research interests include physics-based building energy modeling, optimisation, data-driven modelling, machine learning, and data analysis.

Adalberto Guerra Cabrera received the master's degree in sustainable design and bio-climatic architecture from the Universitat de Barcelona, Barcelona, Spain, in 2016, and the Master of Science (with distinction) degree in data science, technology and innovation from University of Edinburgh, Edinburgh, U.K., in 2020. He is currently a Research Lead with Integrated Environmental Solutions Ltd. His research interests include developing scalable tools for physics-based and data-driven modelling, control optimisation, automated fault detection, diagnostics, measurement, and verification.

Ilaria Salerno is currently working toward the Ph.D. degree in mathematics with University of Edinburgh, Edinburgh, U.K. She is currently a Research Modeller with Catapult Energy System, Birmingham, U.K. Her research interests include smart buildings, heating and cooling optimisation, energy management systems and sustainability.

Damian Flynn (Senior Member, IEEE) is currently an Associate Professor of power system operation and control with University College Dublin, Ireland. His research interests include power system analysis and control, integration of renewable generation within electrical networks, and advanced modeling and control techniques applied to power plant.

Ruth Kerrigan received the Ph.D. degree in civil engineering from Trinity College Dublin, Ireland, in 2005. She is currently a Chief Operating Officer for Integrated Environmental Solutions Ltd. Her research interests include dynamic simulation modelling, construction methods, control system engineering, energy efficient buildings, sustainable cities, and local energy systems.