2020 Building Performance Modeling Conference and
SimBuild co-organized by ASHRAE and IBPSA-USA

DATA-DRIVEN PREDICTIVE CONTROL FOR COMMERCIAL BUILDINGS WITH MULTIPLE ENERGY FLEXIBILITY SOURCES

Anjukan Kathirgamanathan[1,2], Mattia De Rosa[1,2], Eleni Mangina[2,3] and Donal P. Finn[1,2]
[1]School of Mechanical and Materials Engineering, University College Dublin, Ireland
[2]UCD Energy Institute, O'Brien Centre for Science, University College Dublin, Ireland
[3]School of Computer Science, University College Dublin, Ireland

ABSTRACT

Data-Driven Predictive Control, representing the building as a cyber-physical system, shows promising potential in harnessing energy flexibility for demand side management, where the efforts in developing a physics-based model can be significant. Here, predictive control using random forests is applied in a case study closed-loop simulation of a large office building with multiple energy flexibility sources, thereby testing the suitability of the technique for such buildings. Further, consideration is given to the feature selection and feature engineering process. The results show that the data-driven predictive control, under a dynamic grid signal, is capable of minimising energy consumption or energy cost.

INTRODUCTION

Role of buildings in the future smart grid

The requirement for a Smart Readiness Indicator (SRI) was part of the 2018/844 European Union (EU) directive on the energy performance of buildings (EU 2018). The SRI for buildings is primarily motivated by the need to provide information on the technological readiness of buildings to interact with the emergence of smart energy grids, and more generally their capabilities for more efficient operation and better performance through Information Communication Technologies (ICT). The SRI definition also includes the capability of buildings to adapt their operation to the needs of the occupant. The capability of a building to adapt its operation in reaction to signals from, for example, the electricity grid is of relevance given that the penetration of variable and intermittent renewable energy sources is increasing globally (International Energy Agency (IEA) 2017b; International Energy Agency (IEA) 2017a). The flexibility to manage any mismatch in supply and demand can be provided from the demand side (with buildings making up 40% of the total consumption in Europe (Economidou et al. 2011)) through Demand Side Management (DSM). DSM can be broadly categorised as actions that influence the quantity, patterns of use or the primary source of energy consumed by end users (Hull 2012). Implicitly, buildings possess passive thermal storage capabilities and may often incorporate active thermal storage, active electric storage (batteries), indirect electric storage (electric vehicles) and on-site generation as sources of energy flexibility that allow a shift in energy consumption temporally. Commercial buildings are particularly relevant given their larger thermal mass, requirement for space conditioning and more predictable operation schedules (Aduda et al. 2017).

Data-driven approaches for building energy management

Model Predictive Control (MPC) has been demonstrated in numerous studies as a suitable control strategy for building energy management and in particular exploiting the energy flexibility inherent in a building (Clauß et al. 2017; Afram and Janabi-Sharifi 2014). It is particularly suitable given its inherent optimisation, predictive nature and ability to model the evolution of the building dynamics over time whilst taking into account constraints on occupant comfort and system performance limits. However, buildings are complex systems reacting to changing weather, occupancy and grid signals and capturing these dynamics in a model that can be coupled to an optimisation problem is challenging (Henze 2013; Sturzenegger et al. 2016; Jain, Behl, and Mangharam 2016). The "Internet of Things" revolution has led to the rapid rise and use of sensors in building control and availability of building data. Data-driven approaches show promise where the cost of developing a physics-based control-oriented model of the building dynamics is high and militates against its use. (Schmidt and Åhlund 2018) provides a review of data-driven predictive control approaches treating buildings as Cyber-Physical systems. Challenges of data-driven approaches include ensuring a sufficiently rich training dataset over the building operating envelope and appropriately dealing with large datasets with many features. Development of an accurate and efficient

model often requires feature assessment and this is a part of the model development process that is often overlooked or not given the required importance (Schmidt and Åhlund 2018). With data-driven approaches, the model needs to show adequate prediction performance over the prediction horizon as well as allow optimisation of the control inputs for the model to be a replacement for the traditional models as used in MPC (as data-driven models can be highly nonlinear). One promising technique that has been shown to allow such control synthesis is 'separation of variables' together with tree-based predictors (Behl, Smarra, and Mangharam 2016). This is outlined in the next section in more detail.

Separation of Variables

(Behl, Smarra, and Mangharam 2016) first employed the technique 'separation of variables' where regression trees were built using the training data with the control inputs (variables to be optimised) excluded. This control input data was then used to fit affine models under each leaf of the regression trees. This approach leads to a convex optimisation problem (depending on the affine models fit in the leaves) which can be easily and efficiently solved and was used in the problem of peak shaving for a building. However, this method was not compatible with a receding horizon problem such as MPC, as the regression tree only provided a one-step look ahead prediction. (Jain, Behl, and Mangharam 2016) introduced the concept of multi-variate regression trees with output corresponding to each step of the prediction horizon. They successfully compared the data-driven technique to a traditional linear MPC approach with comparable results and only a small additional cost to the objective function. One of the weaknesses of regression trees is that they are prone to overfitting and hence may perform poorly on unseen data. The work was further extended by (Smarra et al. 2018a) to address this by replacing each regression tree with a random forest, which is an ensemble of regression trees. An average is taken from the ensemble, thereby reducing the variance in prediction. This control technique was termed Data-Predictive Control with Ensemble methods (DPC-En). Both these studies used simulation for validation and the work of (Bunning et al. 2019) demonstrated the suitability of the technique in a real life application (minimising the energy consumption of a residential apartment). Given this novel technique and limited applications of the technique in case studies, there are research gaps present in the robustness of the technique, particularly for different building types, energy systems and climate types. This leads to the motivation of this study which is presented in the next section.

Motivation

In this work, the relatively novel technique of 'separation of variables' is investigated (through application of DPC-En) for its suitability and robustness for buildings with multiple sources of energy flexibility and ability to allow buildings to participate in a demand side management context. Existing literature utilising 'separation of variables' has not considered feature selection to be an integral part of the model development process, a trend generally seen in studies implementing data-driven approaches for building energy management. Further, another area not investigated by the previous implementations of 'separation of variables', is the relevance of the features used in the data-driven model and how these vary with different climate types. The wider study focuses on the suitability, robustness and scalability of data-driven predictive control for harnessing energy flexibility from commercial buildings with multiple sources of flexibility. This current paper focuses only on the suitability of the technique to harness flexibility when multiple sources of energy flexibility are present and investigates the feature selection and engineering process in model development.

METHODS

Building Model and Energy Systems

The US Department of Energy 'Large Office' archetype white-box model (using EnergyPlus) has been taken and modified to be the testbed building for this case study (Deru et al. 2011). This building is 12 storeys high with a floor area of 46,000 m^2. The building has a gas boiler for heating, two water-cooled chillers for cooling and a multi-zone variable air volume system for air distribution. A thermal energy storage (TES) tank of 100 m^3 volume (chilled), together with a Photovoltaic (PV) and battery system, were added to the existing model to add further flexibility sources to the building. The two chillers are operated in parallel configuration; the primary chiller directly meets the cooling load whereas the secondary chiller is used to charge the TES tank. The reference strategy is for the secondary chiller to charge the TES during the unoccupied hours and the TES discharges during the day, thereby reducing the primary chiller cooling load. The battery is a 1.5 MWh two-hour duration unit with a maximum charge/discharge rate of 0.5 MW. The battery has a charge/discharge efficiency of 85%. The building complies with the minimum requirements of ASHRAE Standard 90.1-2004. The system architecture and HVAC diagram are illustrated in 1. The building operates from 6.00 am

to midnight on weekdays and 6.00 am to 5.00 pm on Saturdays with no occupancy on Sundays. In summary, energy flexibility is provided from the following: building passive thermal mass, building active thermal storage, electric battery and on-site PV generation. Note that the default archetype model uses a Rule-Based Controller (RBC).

The EnergyPlus model uses a simulation time-step of 15 minutes and was simulated for an entire year to generate the training and test dataset. Random excitation of the control variable, i.e., cooling setpoint temperature (indoor zonal air dry bulb temperature), was simulated to generate the training dataset. This was found to be neccessary to ensure the training dataset captured the entire operating envelope of the cooling system and was sufficiently rich.

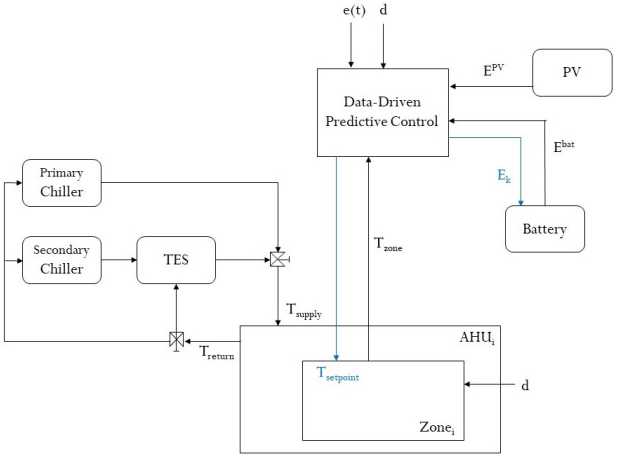

Figure 1: Schematic representation of HVAC and energy system architecture

Feature Selection using Random Forests

As per the methodology used in (Kathirgamanathan et al. 2019), prior to feature selection, feature engineering is performed to generate new potentially relevant features. Proxy schedule variables (s) for the hour of the day and day of the week were generated for the dataset based on the timestamp variable. Similarly, lag terms (autoregressive terms) were also created of the state variables (x) and disturbance variables (d) with an order of 10 (δ_x and δ_d) being used. Once all the predictors are available, they are split accordingly into disturbances, control variables (u), response variables (Y) and schedule variables. As outlined in (Kathirgamanathan et al. 2019), the embedded method of feature selection using the random forest algorithm was used to extract the feature importance data for all models over the prediction horizon.

Building Model Training

The data was split with the months from January to June being used for training and the 1^{st} work week of July used to test the performance of the model. The advantages of using a white-box model to generate synthetic data is that a comprehensive database can be generated for training that can be used without concern for data quality issues that can hinder processing and analysing real world datasets. The control variable (in this case the indoor zonal cooling setpoint temperature) is standardised around 0 as this was found to improve the building model accuracy significantly.

A random forest with 200 trees is used with a minimum amount of 200 samples in each leaf (Bunning et al. 2019). A prediction horizon of 5 hours (20 time-steps) is used. All data processing, feature engineering and feature selection is carried out in Python with the Sklearn package (Pedregosa et al. 2011) being used to create the random forest models. This was carried out on a server machine with an Intel(R) Xeon(R) CPU E5-2697 v2 2.7 GHz and 256 GB of RAM.

Data-Driven Predictive Control with Ensem-ble Methods (DPC-En)

The building 'core mid' zone temperature is investigated as one of the response variables as this zone represents the majority of the zonal temperatures, being the largest zone per floor and representing 10 of the 12 floors through symmetry properties of the simulation (the EnergyPlus model only simulates one of these 10 'middle' floors and assumes the other nine floors are identical). Note that the technique of 'separation of variables' has been shown to be suitable for multi-zone buildings and control of multiple zones (Jain 2018). The other response variable is the building total power consumption which is of interest from a DSM perspective.

The predicted electricity consumption at the j^{th} step is given as follows by fitting a regression model on the samples in the leaf with the affine sum of the control inputs being the dependent variable:

$$\hat{Y}_{electricity,j} = \hat{x}(k+j) = \gamma_j[1, u(k), ..., u(k+j-1)]^T \quad (1)$$

Similarly, the predicted temperature at the j^{th} step is as follows:

$$\hat{Y}_{temperature,j} = \hat{x}(k+j) = \alpha_j[1, u(k), ..., u(k+j-1)]^T \quad (2)$$

The γ_j and α_j terms are calculated from the average of the regression coefficients from all the trees of

the forest. The reader is referred to (Smarra et al. 2018b) for an in-depth description of the derivation of the above model. Similarly to what (Bunning et al. 2019) found, the dimensionality of these coefficients is reduced to 2 for each j due to poor prediction performance for larger horizons due to the high dimensionality for the model fitting process. Note that the disturbance part of the model (non-linear random forest) is capable of taking into account implicitly the actual potentially non-linear Coefficient of Performance of the chillers and the need to consider this in the formulation explicitly is removed unlike physics-based approaches.

Once the models are found for describing the building dynamics (temperature and power consumption evolution) over the prediction horizon, they can be integrated into a traditional MPC-like receding horizon problem. Given a grid signal $e(t)$, e.g., real-time pricing, the corresponding linear optimisation problem is:

$$\min_{u,\epsilon} \quad \sum_{j=1}^{N} e(t)u_{k+j} + \lambda\epsilon_{k+j} \tag{3a}$$

$$\text{subject to} \quad x_{k+j} = \gamma_j[1, u(k), ..., u(k+j-1)]^T, \tag{3b}$$

$$t_{min} - \epsilon_{k+j} \leq x_{k+j} \leq t_{max} + \epsilon_{k+j}, \tag{3c}$$

$$u \in U, \tag{3d}$$

$$\epsilon \geq 0, \tag{3e}$$

$$j = 1, ..., N. \tag{3f}$$

where λ is the weighting term used to adjust the relative cost of comfort constraint violations, ϵ is a slack variable for the comfort constraint, t_{min} and t_{max} are the time-varying zonal temperature constraints and U defines the allowed range of control inputs. In this study, the cooling setpoint (indoor zone air dry bulb temperature for cooling seasons) is used as the decision variable as it is easily human interpretable and commonly used in building energy management systems as a feedback to end users/occupants. During occupied hours, the temperature constraints are set at $\pm 1°C$ from a reference temperature of $22°C$ and this is relaxed during unoccupied hours to $\pm 5°C$. A sensitivity study was carried out for the value of λ and it was set to 100 for this study.

This framework was extended with the battery (electrical storage) included in the receding horizon problem. The system model used (Bianchini et al. 2019) is described below. Given that $E^{bat}(k)$ represents the energy or State of Charge (SoC) of the battery at time k, t represents the time in seconds of one time-step, $P_{+}^{bat}(k)$ represents the charge rate of the battery and $P_{-}^{bat}(k)$ represents the discharge rate of the battery (both in kW), then the following equation can be used to model the battery SoC evolution:

$$E^{bat}(k+1) = E^{bat}(k) + \eta P_{+}^{bat}(k) * t - \frac{1}{\eta} P_{-}^{bat}(k) * t \tag{4}$$

where $0 < \eta < 1$ is the battery efficiency. The following additional constraints are introduced to limit the SoC of the battery and the maximum charge/discharge rate.

$$0 \leq E^{bat}(k) \leq \overline{E}^{bat}(k) \tag{5}$$

$$0 \leq P_{+}^{bat}(k) \leq \overline{P}_{+}^{bat}(k) \tag{6}$$

$$0 \leq P_{-}^{bat}(k) \leq \overline{P}_{-}^{bat}(k) \tag{7}$$

In this case, the objective function is slightly modified to account for effect of charging and discharging on the purchased electricity from the grid. Defining the power purchased due to the battery at time-step k as:

$$P^{bat}(k) = P_{+}^{bat}(k) - P_{-}^{bat}(k) \tag{8}$$

and the power drawn from the grid at time-step k as:

$$P^{grid}(k) = Y^{building}(k) + P^{bat}(k) - P^{PV}(k) \tag{9}$$

Then the following new linear optimisation problem is defined minimising the power purchased from the grid:

$$\min_{E,\epsilon} \quad \sum_{j=1}^{N} e(t)P_{k+j}^{grid} + \lambda\epsilon_{k+j} \tag{10}$$

$$\text{subject to} \quad \text{3b, 3c, 3d, 3e, 3f, 4, 5, 6. 7 and 9}$$

In this research, the real-time price used was based on real market data from Italy for 2017 ((GME)). Note that whilst such dynamic prices are often not realised by the end-user themselves currently, it reflects a potential future grid signal and is used to show the capability of the controller to react to such dynamics signals. It should be noted that the controller is agnostic with regards to what $e(t)$ is and this grid signal can be specified per the overall objectives of the controller. As shown in (Smarra et al. 2018b), this signal could be a tracking signal for load following and meeting the regulation needs of the grid. If carbon emissions intensity data is available for generation, a carbon emissions minimisation objective can be specified instead.

The objective functions are linear, thereby guaranteeing a convex program and a tractable solution.

The linear program is solved using the COIN-CBC solver (Forrest et al. 2018) and the Pyomo (Hart, Watson, and Woodruff 2011) optimisation modeling framework using Python.

Co-Simulation Framework

In the absence of a real building, the virtual and high-fidelity E nergyPlus m odel o f t he ' Large Office' archetype building is used to test the data-predictive control in a closed-loop simulation through the use of co-simulation. The PyEp python module is used for communication between EnergyPlus and Python through the use of an Open Platform Communications (OPC) bridge as outlined in (Jain et al. 2018). The overall co-simulation framework is described graphically in Figure 2. The performance of the controller is compared with the reference RBC. This reference control modulates the indoor zonal cooling setpoint and TES charge/discharge based purely on time with fixed schedules.

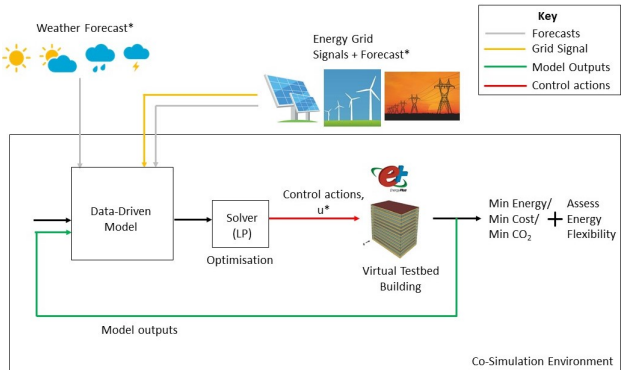

Figure 2: Co-simulation framework

RESULTS AND DISCUSSION

Feature Selection

Figure 3 plots the variable importance measure (how much a feature contributes to the decision making in the model, measured using the "mean decrease impurity", see (Breiman 2001)) as output from the random forest predictor, as a heatmap, for all the predictor variables (y-axis) used to train the 'N' models over the prediction horizon (x-axis). The figure shows that the 1^{st} lag term ('CORE MID:Zone Air Temperature $[C]$()' which is the value of response variable one time step in the past, i.e. 15 minutes prior) is the most relevant variable for the temperature prediction with the importance declining as the prediction step increases (i.e., the 1^{st} lag term is more relevant for predicting 1 time-step away as opposed to predicting 2 time-steps away) as would be expected. The next most significant variable is the proxy variable, hour of the day, with the opposite trend being shown here, i.e. this variable becomes more relevant the further away we are predicting.

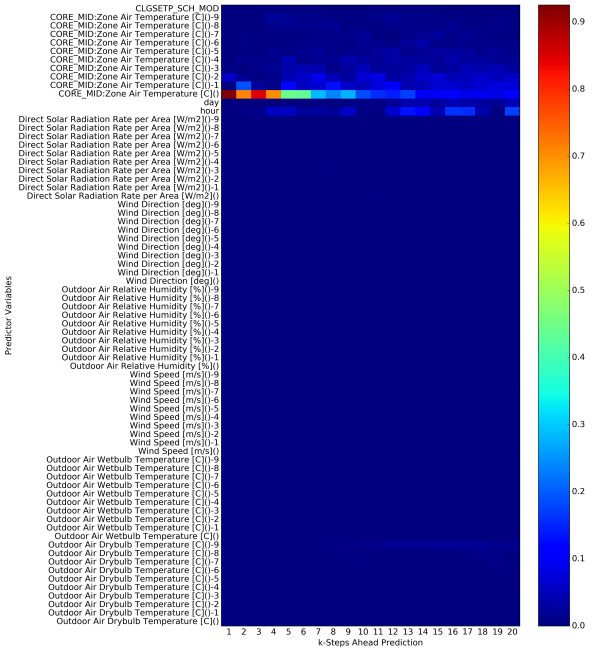

Figure 3: Heatmap of highest ranking (highest feature importance) predictors for n-step ahead predictions of core mid temperature

In order to investigate the relevance of the other variables, Figure 4 shows the variable importance for the remaining variables (with 1^{st} lag terms, hour of day and day of week hidden). The further lag terms for the response variable are the next more useful variables with the order of the autoregressive term and prediction step interacting. This pattern is similar when repeated for the power consumption prediction (results not shown for conciseness). This analysis was also performed for several training datasets with different climate data used. Although the most relevant variables did not change significantly, there were differences in the lesser relevant variables with climate, e.g., dry bulb temperature was a more useful predictor than the wet bulb temperature for the climate of Rome, Italy (ASHRAE climate zone 3C), whereas the opposite trend was observed for the climate of Dublin, Ireland (ASHRAE climate zone 4C).

Next, the predictive accuracy of the ensemble models is verified over the prediction horizon. Figure 5 presents the predictions of the total power consumption made n-steps ahead over the prediction horizon (note that ground truth represents the synthetic test data). The prediction accuracy generally declines for

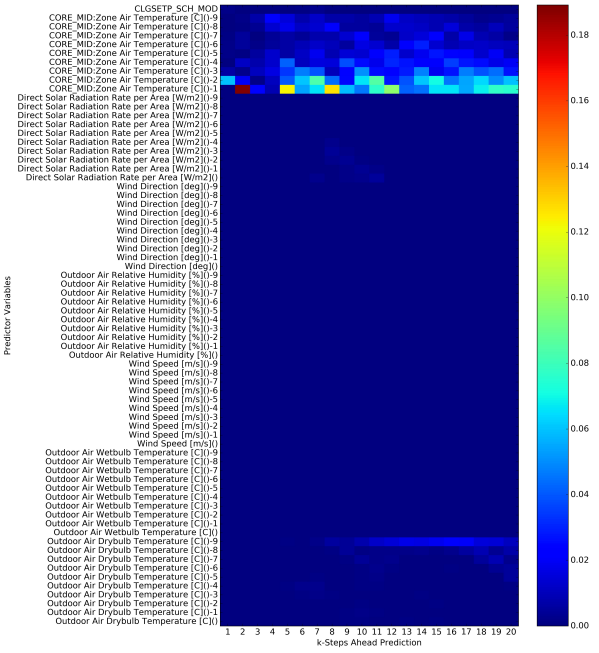

Figure 4: Heatmap of highest ranking (highest feature importance) predictors for n-step ahead predictions of core mid temperature (with 1-step lag and hour and day proxy predictors removed)

larger step ahead predictions, as expected. Note that the accuracy of longer-term predictions is not critical in a receding horizon problem such as MPC or data-driven predictive control. This is because only the first control input is applied to the system and at every time step, new input is collected correcting for any model errors. Figure 6 plots the Normalised Root Mean Squared Error (NRMSE) over the prediction horizon.

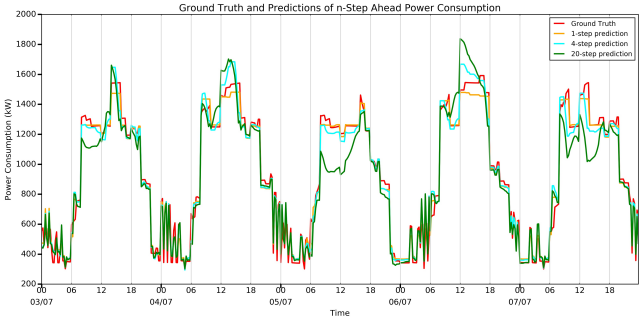

Figure 5: Ground truth and predictions of n-step ahead power consumption for test work week

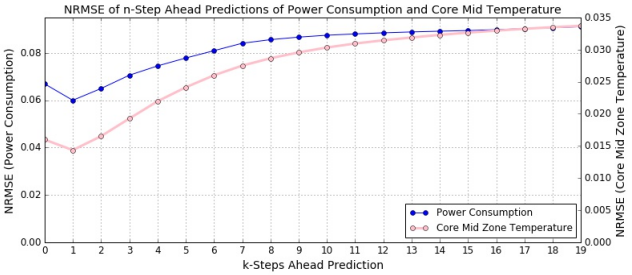

Figure 6: NRMSE of predictors over n-step prediction horizon for test period

Controller Performance

Figure 7 shows the results when the data-predictive controller is utilised for a 5-day working week in July for the test office building in a closed-loop simulation. Table 1 provides a summary of the controller performance (for both DPC-En and reference RBC) specified i n t erms o f e nergy c onsumption, e nergy spend and discomfort hours. Figure 7a shows the real-time electricity price used as input by the controller which shows the general trend of having lower prices during night hours (midnight to 06.00 am) and higher prices during the evening peak (06.00 pm to midnight). Figure 7b illustrates the ambient environmental conditions through plotting the environment dry bulb temperature and direct solar radiation rate. Figure 7c shows the control variable, in this case, the optimal scheduled indoor zonal cooling setpoint (in blue, compared to pink representing the reference RBC control). It can be seen that the controller is able to pre-cool the building during the early morning taking advantage of lower real-time prices and hence requiring less cooling over the work day when higher prices are prevalent. The controller is also able to reduce the cooling demands during the hours prior to office closure (evening peak for the grid) and setback by taking advantage of the thermal mass compared to the existing rule-based controller. This is also illustrated in Figure 7f showing the total power consumption of both control strategies.

Table 1: Comparison of performance of DPC-En compared to reference RBC control

Control Type	Energy Purchased (kWh)	Energy Spend (€)	Discomfort Degree-Hours
RBC (Ref)	115337	5702	0.60
DPC-En	104405	5081	0.04

Figure 7d shows the zonal temperature evolution over the analysed period and shows lower comfort constraint violations compared to the reference RBC con-

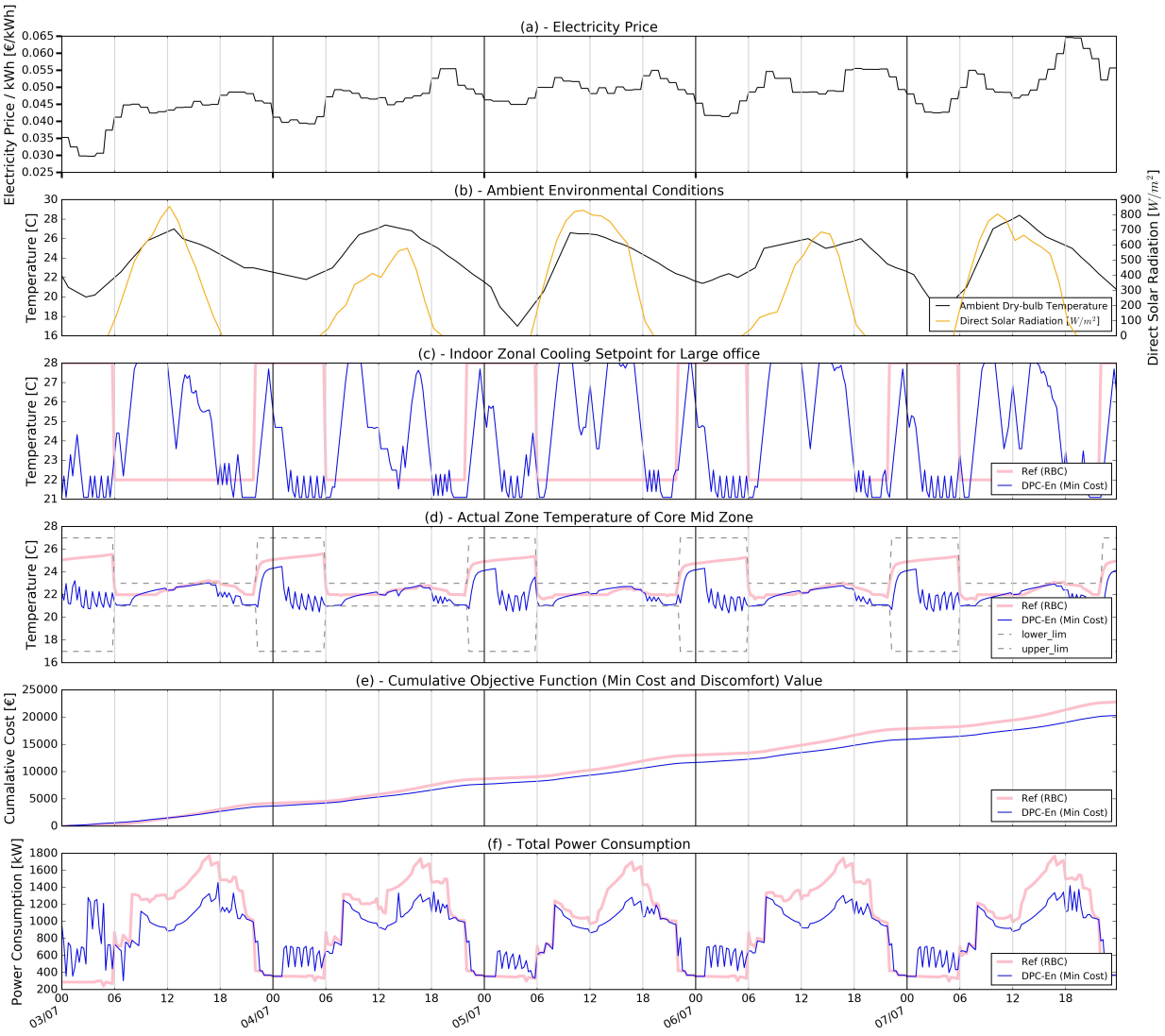

Figure 7: DPC-En results over work week in July for test climate of Rome

trol. Finally, Figure 7e shows the effect on the objective function (in this case total cost) and the savings that the DPC-En controller is able to achieve. For this case, the DPC-En controller is able to manage, compared to the baseline RBC, a reduction in energy consumption of 9.5 %, a reduction in energy spend of 12.2 % and a reduction in constraint violations of 93.4 % (measured in degree-hours that a comfort band of $\pm 1°C$ is exceeded during occupied hours). Note that the comfort band temperatures can be freely adjusted to real scenarios and user requirements.

Figure 8 illustrates the energy flexibility provided by both the battery and the TES. Plot (a) shows the state of charge given the grid signal in Figure 7a. There is a period during the first day of the test data in the early morning where the price is sufficiently lower than the following hours that the battery charges and discharges. Given the efficiency losses of charging and discharging, the variance in the real-time price during the prediction horizon has to be greater than this loss factor for the battery use to be optimal. The prediction horizon is currently limited to five hours or 20 time-steps as the accuracy of the random forest ensemble model for temperature and power predictions for longer prediction horizons is current work. This limited prediction horizon coupled with the low variance in grid signal seen may explain why greater use of the battery is not seen.

The thermal energy storage tank is also used to shift cooling loads from on-peak times to off-peak times.

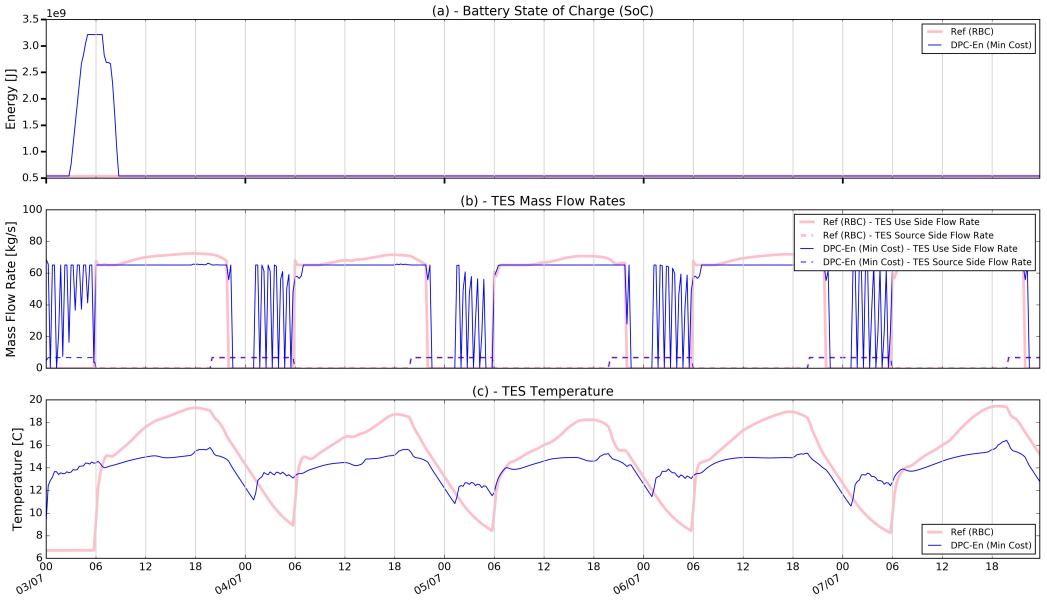

Figure 8: DPC-En results over work week in July for test climate of Rome showing Battery and TES behaviour

Chiller 2 (secondary chiller) is utilised during the night to charge the TES, where the TES is subsequently discharged during the day. Currently the charge and discharge times and hours are not part of the optimal control problem. Ongoing work is addressing this. However, as Figure 7c and d shows, DPC-En is able to take advantage of the extra energy flexibility provided by the TES to shift more of the cooling load to off-peak times.

The average time taken per time-step (over the co-simulation run period) to use the data-driven model to make predictions and then perform the optimisation is approximately five seconds which is significantly smaller than the 15 minute sample time-step used. This shows the efficiency of the data-predictive controller and hence suitability of this technique for real-time application and even offers the possibility that a smaller time-step may be used.

CONCLUSION

The Data Predictive Control (DPC-En) technique using 'separation of variables' together with random forests was shown to minimise energy costs (by 9.5%) through activation of building energy flexibility whilst also improving thermal comfort for occupants for a work week. It was shown that this technique is also capable of being used in buildings featuring multiple sources of energy flexibility, such as active thermal energy storage, active electrical storage and on-site generation. While training the random forest ensemble models, it was found that the feature importance depends on prediction horizon and climate type of the data. The process of selecting the features used as predictors and generating new features through feature engineering (such as creating lag terms) should utilise some domain knowledge. The approach developed can be easily modified for minimum carbon emissions or energy consumption objectives. The DPC-En approach has a low computational burden with an average run time of only around five seconds per time step allowing the approach to be applied in real-time.

Limitations include the use of synthetic data for training and perfect forecasts being used. There have been a very few examples using this novel technique on a real case study building. The robustness of the 'separation of variables' technique using random forests is still largely unproven. Given that the efficacy of the data-driven controller depends upon the quality and quantity of the training data available, this is an issue needing to be addressed. The ability of a controller to react to data that it has not seen in the training data is one such interesting question to investigate. The scalability of the approach over multiple diverse buildings is yet to be verified and subject of future work.

ACKNOWLEDGMENT

This work has emanated from research conducted with the financial support of Science Foundation Ireland under the SFI Strategic Partnership Programme Grant Number SFI/15/SPP/E3125.

REFERENCES

Aduda, K.O., T. Labeodan, W. Zeiler, and G. Boxem. 2017. "Demand side flexibility coordination in office buildings: A framework and case study application." *Sustainable Cities and Society* 29:139–158.

Afram, Abdul, and Farrokh Janabi-Sharifi. 2014. "Theory and applications of HVAC control systems - A review of model predictive control (MPC)." *Building and Environment* 72, no. February 2014.

Behl, Madhur, Francesco Smarra, and Rahul Mangharam. 2016. "DR-Advisor: A data-driven demand response recommender system." *Applied Energy* 170 (January): 30–46.

Bianchini, Gianni, Marco Casini, Daniele Pepe, Antonio Vicino, and Giovanni Gino Zanvettor. 2019. "An integrated model predictive control approach for optimal HVAC and energy storage operation in large-scale buildings." *Applied Energy* 240 (January): 327–340.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45(1):1–33.

Bunning, Felix, Benjamin Huber, Philipp Heer, Ahmed Aboudonia, and John Lygeros. 2019. "Experimental demonstration of data predictive control for energy optimization and thermal comfort in buildings." *Preprint submitted to Elsevier*.

Clauß, John, Christian Finck, Pierre Vogler-finck, and Paul Beagon. 2017. "Control strategies for building energy systems to unlock demand side flexibility – A review." *Proc. of BS2017: 15th Conference of International Building Performance Simulation Association, San Fransisco, USA, Aug 7-9.* San Fransisco.

Deru, Michael, Kristin Field, Daniel Studer, Kyle Benne, Brent Griffith, Paul Torcellini, Bing Liu, Mark Halverson, Dave Winiarski, Michael Rosenberg, Mehry Yazdanian, Joe Huang, and Drury Crawley. 2011. "U.S. Department of Energy commercial reference building models of the national building stock." Technical Report February 2011, NREL.

Economidou, Marina, Jens Laustsen, Paul Ruyssevelt, and Dan Staniaszek. 2011. "Europe's Buildings Under the Microscope." Technical Report, Buildings Performance Institute Europe (BPIE).

EU. 2018. Directive (EU) 2018/844 of the European Parliament and of the Council amending Directive 2010/31/EU on the Energy Performance of Buildings and Directive 2012/27/EU on Energy Efficiency.

Forrest, John, Ted Ralphs, Stefan Vigerske, LouHafer, Bjarni Kristjansson, jpfasano, EdwinStraver, Miles Lubin, Haroldo Gambini Santos, rlougee, and Matthew Saltzman. 2018, July. coin-or/Cbc: Version 2.9.9.

(GME), Gestore Mercati Energetici. Online database.

Hart, William E, Jean-Paul Watson, and David L Woodruff. 2011. "Pyomo: modeling and solving mathematical programs in Python." *Mathematical Programming Computation* 3 (3): 219–260.

Henze, Gregor P. 2013. "Model predictive control for buildings: a quantum leap?" *Journal of Building Performance Simulation* 6 (3): 157–158.

Hull, Linda. 2012. "DSM / DSR: What, Why and How?" Technical Report November, EA Technology.

International Energy Agency (IEA). 2017a. "Market Report Series: Renewables 2017, analysis and forecats to 2022." Technical Report, IEA.

International Energy Agency (IEA). 2017b. "Snapshot of global photovoltaic markets." Technical Report T1-31:2017, IEA.

Jain, Achin. 2018. "Data-Driven Model Predictive Control with Regression Trees—An Application to Building Energy Management." *ACM Trans. Cyber-Phys. Syst* 2 (21): 1–21.

Jain, Achin, Madhur Behl, and Rahul Mangharam. 2016. "Data Predictive Control for Building Energy Management." *Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments - BuildSys '16*, no. May:245–246.

Jain, Achin, Derek Nong, Truong X Nghiem, and Rahul Mangharam. 2018. "DIGITAL TWINS FOR EFFICIENT MODELING AND CONTROL OF BUILDINGS AN INTEGRATED SOLUTION WITH SCADA SYSTEMS 1 Flexergy AI , Philadelphia , PA 2 Northern Arizona University , Flagstaff , AZ." *2018 Building Performance Modeling Conference and SimBuild, co-organized by ASHRAE and IBPSA-USA Chicago, IL, September 26-28, 2018.* Chicago, IL, USA.

Kathirgamanathan, Anjukan, Mattia De Rosa, Eleni Mangina, and Donal P Finn. 2019. "Feature Assessment in Data-driven Models for un-

locking Building Energy Flexibility." *IBPSA BS 2019. 2-4 September, 2019, Rome, Italy.* Rome.

Pedregosa, Fabian, Ron Weiss, Matthieu Brucher, Gaël Varoquaux, Alexandre Gramfort, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Vincent Dubourg, Jake Vanderplas, and Alex Passos. 2011. "Scikit-learn : Machine Learning in Python." *Journal of Machine Learning Research* 12:2825–2830.

Schmidt, Mischa, and Christer Åhlund. 2018. "Smart buildings as Cyber-Physical Systems : Data-driven predictive control strategies for energy efficiency." *Renewable and Sustainable Energy Reviews* 90 (April): 742–756.

Smarra, Francesco, Achin Jain, Rahul Mangharam, and Alessandro D Innocenzo. 2018a. "Data-driven Switched Affine Modeling for Model Predictive Control." *IFAC Conference on Analysis and Design of Hybrid Systems ADHS 2018*. Oxford, United Kingdom.

Smarra, Francesco, Achin Jain, Tullio De Rubeis, Dario Ambrosini, Alessandro D Innocenzo, and Rahul Mangharam. 2018b. "Data-driven model predictive control using random forests for building energy optimization and climate control." *Applied Energy*, no. February:1–21.

Sturzenegger, David, Dimitrios Gyalistras, Manfred Morari, and Roy S. Smith. 2016. "Model Predictive Climate Control of a Swiss Office Building: Implementation, Results, and Cost-Benefit Analysis." *IEEE Transactions on Control Systems Technology* 24 (1): 1–12.

Nomenclature

α Leaf Coefficients for Temperature Model

δ_d Autoregressive Order for Disturbance Variables

δ_x Autoregressive Order for State Variables

ϵ Comfort constraint slack variable

η Charging/discharging efficiency of battery

γ Leaf Coefficients for Power Model

$\hat{y}_i$ predicted value of response variable

λ Comfort Weighting Term

$\overline{E}^{bat}(k)$ Maximum charge of battery (J)

$\overline{P}_{+}^{bat}(k)$ Maximum charge rate of battery at time-step $k(kW)$

$\overline{P}_{-}^{bat}(k)$ Maximum discharge rate of battery at time-step $k(kW)$

$\overline{y}_i$ mean value of response variable

d Disturbance Variables

$e(t)$ Grid Signal, e.g. Real-Time Price (Euros/kWh)

$E^{bat}(k)$ State of Charge (SoC) of battery at time-step $k(J)$

j $\in 1, ..., N$ steps in prediction horizon

N Prediction Horizon

n Number of trees in Ensemble

$P^{grid}(k)$ Power purchased from grid due to battery at time-step k

$P^{PV}(k)$ Power output of PV at time-step k

$P_{+}^{bat}(k)$ Charge rate of battery at time-step $k(kW)$

$P_{-}^{bat}(k)$ Discharge rate of battery at time-step $k(kW)$

s Schedule Variables

$T_{ref}(k)$ Reference Set-point Temperature $(^{\circ}C)$

u Control Inputs

X Measured Data - Inputs

Y Measured Data - Outputs

y_i actual value/ground truth of response variable