Energy & Buildings 193 (2019) 250–266

Contents lists available at ScienceDirect

Energy & Buildings

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

Can semi-parametric additive models outperform linear models, when forecasting indoor temperatures in free-running buildings?

Matej Gustin[a,b,*], Robert S. McLeod[a,b], Kevin J. Lomas[a,b]

[a] School of Architecture, Building and Civil Engineering, Loughborough University, LE11 3TU, UK
[b] London-Loughborough EPSRC Centre for Doctoral Training in Energy Demand, Loughborough University, LE11 3TU, UK

ARTICLE INFO

Article history:
Received 9 January 2019
Revised 4 March 2019
Accepted 26 March 2019
Available online 26 March 2019

Keywords:
Time series forecasting
Generalized Additive Model (GAM)
AutoRegressive model with eXogenous inputs (ARX)
Logistic GAM
Window opening state
Heatwave
Overheating, Indoor temperature

ABSTRACT

A novel application combining semi-parametric *Generalized Additive Models (GAMs)* with *logistic* GAMs was developed to forecast indoor temperatures and window opening states during prolonged heatwaves. GAM models were compared to *AutoRegressive models with eXogenous inputs (ARX)* and validated against monitored data from two case study dwellings, located near to Loughborough in the UK, during the 2013 heatwave. Input variables were selected using backward stepwise regressions based on minimisation of the *Akaike Information Criterion (AIC)* and *Mean Absolute Error (MAE)*, for the ARX and GAM models respectively. Comparison of the models showed that whilst GAMs are capable of improving the forecasting accuracy, the improvements are significant only up to 3–6 h ahead. During heatwaves and over longer forecasting horizons, GAMs were found to be less reliable and accurate than ARX models. The marginal improvement in forecasting accuracy at shorter horizons did not justify the additional computational time and risk of instability associated with more complex GAMs, at longer forecasting horizons. Whilst, *logistic* GAMs were shown to adequately predict the window opening state, incorporating knowledge of the window state did not significantly improve the accuracy of the indoor temperature predictions.

1. Introduction

1.1. Background

Overheating in residential buildings is increasingly acknowledged as an emerging global health risk [1–3]. Climate change projections indicate that the world's most populated regions will experience more frequent and intense heatwave periods over the coming decades [4,5]. The likelihood of events such as the 2003 heat wave (which was responsible for over 30,000 pre-mature deaths across Europe [6]) recurring is projected to increase 100-fold by 2050 [7].

Understanding how individual buildings are likely to respond to extreme climatic events in the future is critical to mitigating their potentially life-threatening impacts. The complexity of this problem originates in the unique time-varying nature of the thermal behaviour of any given building, which is influenced both by its physical characteristics and the unique way in which it is occupied and operated [8].

Fully parametrised *Dynamic Thermal Simulation (DTS)* models have been widely used to assess current and future overheating risks [9–12], however, the results of such studies often reveal a significant gap [13] between the empirically measured and modelled overheating performance of dwellings [14]. This 'modelling-gap' has led some researchers to question the applicability of using white-box DTS models for forecasting overheating [3]. In contrast, the availability of data from large monitoring studies [10,11,15–18] offers the potential to develop empirical models of existing buildings which are capable of making predictions based on the data alone (i.e. machine learning) [19]. In statistical black-box models [20], the time-varying responses of the building fabric, ventilation, etc., are all embedded in the past internal temperature data, obviating the need to make assumptions relating to the building's thermo-physical characteristics.

In a previous study, the present authors [21] have shown that linear *AutoRegressive models with eXogenous inputs (ARX)* are able to forecast indoor temperatures during heatwaves up to 72 h in advance, with reasonable accuracy. However, it was posited that the accuracy of such models may be improved by adopting non-linear methods and modelling the effect of window opening on indoor temperatures [21]. This is because, during spells of hot weather,

* Corresponding author.
E-mail address: m.gustin@lboro.ac.uk (M. Gustin).

https://doi.org/10.1016/j.enbuild.2019.03.048

the likelihood that occupants will open windows may be much higher than during the milder weather upon which the model was trained [22].

The main difficulty of including the window opening state in an empirical forecasting model is that occupant behaviour in relation to window control, particularly in residential buildings, is a stochastic rather than a deterministic process [23], requiring stochastic models for its prediction [24–26]. This paper, therefore, explores whether statistical models are capable of adequately emulating window control and crucially whether this additional information can improve indoor temperature forecasts.

1.2. Linear, non-linear and semi-parametric forecasting models

A number of studies [27–29] have shown that non-linear *Artificial Neural Networks* (*ANNs*) such as *Non-linear ARX (NARX)* models outperform linear ARX models for forecasting indoor temperatures. Some researchers [27,28] have posited that the higher forecasting accuracy of NARX models is attributable to their ability to capture the non-linear relationships that govern indoor temperatures. In contrast, Thomas and Soleimani-Mohseni [29], showed that the differences between non-linear NARX and linear ARX models were minimal and Ferracuti et al. [30] found that, both in summer and in winter, more accurate 3 h ahead predictions were obtained with a linear ARX model. Whether or not non-linear models are a better choice than linear models appears to depend on several factors, including the: period of testing, structure of the models, and forecasting horizon. ANNs are also inherently limited by their lack of interpretability [19], which has been referred to as "the Achilles' heel of deep neural networks" [31].

In contrast to ANNs, semi-parametric models, also known as *Generalized Additive Models (GAMs)*, offer transparent interpretability of the results [32] and for some problems, e.g. short-term forecasting of electricity demand [33], they have significantly outperformed ANNs. Because semi-parametric additive models allow non-linear and non-parametric terms to be included within the regression framework, they can readily capture complex non-linear relationships [33].

1.3. Integrating window opening states into forecasting models

During heatwaves, occupants of dwellings are very likely to operate windows to try and stay cool [22]. Indoor and outdoor temperatures have been identified as key predictors for window opening models [24,26], with Yun and Steemers [23] observing that the time of the day is also a crucial factor in characterising window opening behaviour. Others have suggested that window opening is also positively correlated with the CO_2 concentration, solar radiation and illumination level [25]. It is hypothesised herein that the inclusion of window opening states into temperature forecasting models could improve their accuracy.

When knowledge of the transition probabilities between the window opening states (usually modelled as Markov chains) are not required, *logistic regression* (i.e. binary) models based on a single probability are commonly adopted. Because of the binary nature of the output, the dependent variable cannot be described with a Gaussian distribution and is therefore described with a Bernoulli distribution (i.e. as the probability p of the windows being open) [24,26]. Researchers, such as Haldi and Robinson [26] and Schweiker et al. [24], have relied on polynomials to model non-linear effects, but this can lead to inefficient model formulation, correlated terms and counterintuitive results [32]. In contrast, GAMs are far more flexible for modelling non-linear relationships, with predictor functions that are automatically derived during model estimation [32]. This makes them preferable for the *logistic* formulation of stochastic behavioural models.

1.4. Objectives

Despite the potential advantages of GAMs, to the authors' knowledge, they have not been applied to the prediction of overheating; their application to this area is one of the novel features of this work, which addresses three research questions:

1. Can the use of a more complex semi-parametric GAMs significantly improve the accuracy attained by linear ARX models when forecasting over shorter time series (e.g. a single summer season)?
2. Can the hourly window opening state in a residential building be reliably predicted?
3. Does incorporation of the window opening state into ARX and GAM models help to improve the overall accuracy of overheating forecasts?

2. The monitored data set

To stress-test the predictive and generalisation capabilities of a model for overheating forecasting, it is important that it is tested and validated during a period in which temperatures exceed those experienced during the training period. For this purpose, and to test the effect of including window-opening in the model, two rooms from two dwellings, located in close proximity to the town of Loughborough in the English Midlands (and monitored as part of the LEEDR Smart Home dataset [18]). These rooms were selected because of the completeness of the data, their markedly different temperature profiles and frequent use of windows during the 2013 heatwave.[1] This UK-wide heatwave reached a peak temperature of 33.5 °C and lasted from the 3rd to 23rd July 2013 [34], making it the fourth warmest July recorded in the UK, since 1910, in terms of both the mean and mean daily maximum temperatures [35].

To capture the most pronounced overheating, the internal temperatures (T_{int}) and Window Opening states (*WO*) were logged at one-minute intervals, in the upstairs bedrooms. The weather data, consisting of the external air temperatures (T_{ext}) and Global Horizontal solar Irradiance (*GHI*), was recorded at the nearby Sutton Bonington meteorological station at hourly intervals. For this reason, the data that was recorded in the dwellings was down-sampled for the models by averaging the sub-hourly values to obtain hourly mean values (centred on each hour). For WO, the hourly states were determined by using 0.5 as the state change threshold. The WO states were defined as 0 – closed if $\leq$ 30 min open; 1 – open if $>$ 30 min open.

Outdoor air temperatures during spring and early summer 2013 were considerably below average (Fig. 1). The external air temperature started to rise on the 3rd of July, resulting in a continuous hot spell that lasted until thunderstorms on the 22nd and 23rd of July broke the heatwave. During this extended hot spell, the indoor temperatures (recorded in the bedrooms) were noticeably elevated in both dwellings on 6–7 and 13–19 July. Although indoor temperatures in the two dwellings were very similar on some days, dwelling A warmed up considerably less than dwelling B on most days, with the most pronounced temperature difference (of 6.9 °C) occurring on the 8th July (Fig. 1).

Window opening data indicated that the occupants of both dwellings were consistently operating the windows before and after the heatwave, with the window opening frequency increasing

[1] According to the UK Met Office, based on the World Meteorological Organization definition, a heatwave is defined as, "A marked unusual hot weather (Max, Min and daily average) over a region persisting at least two consecutive days during the hot period of the year based on local climatological conditions, with thermal conditions recorded above given thresholds" [51,52].

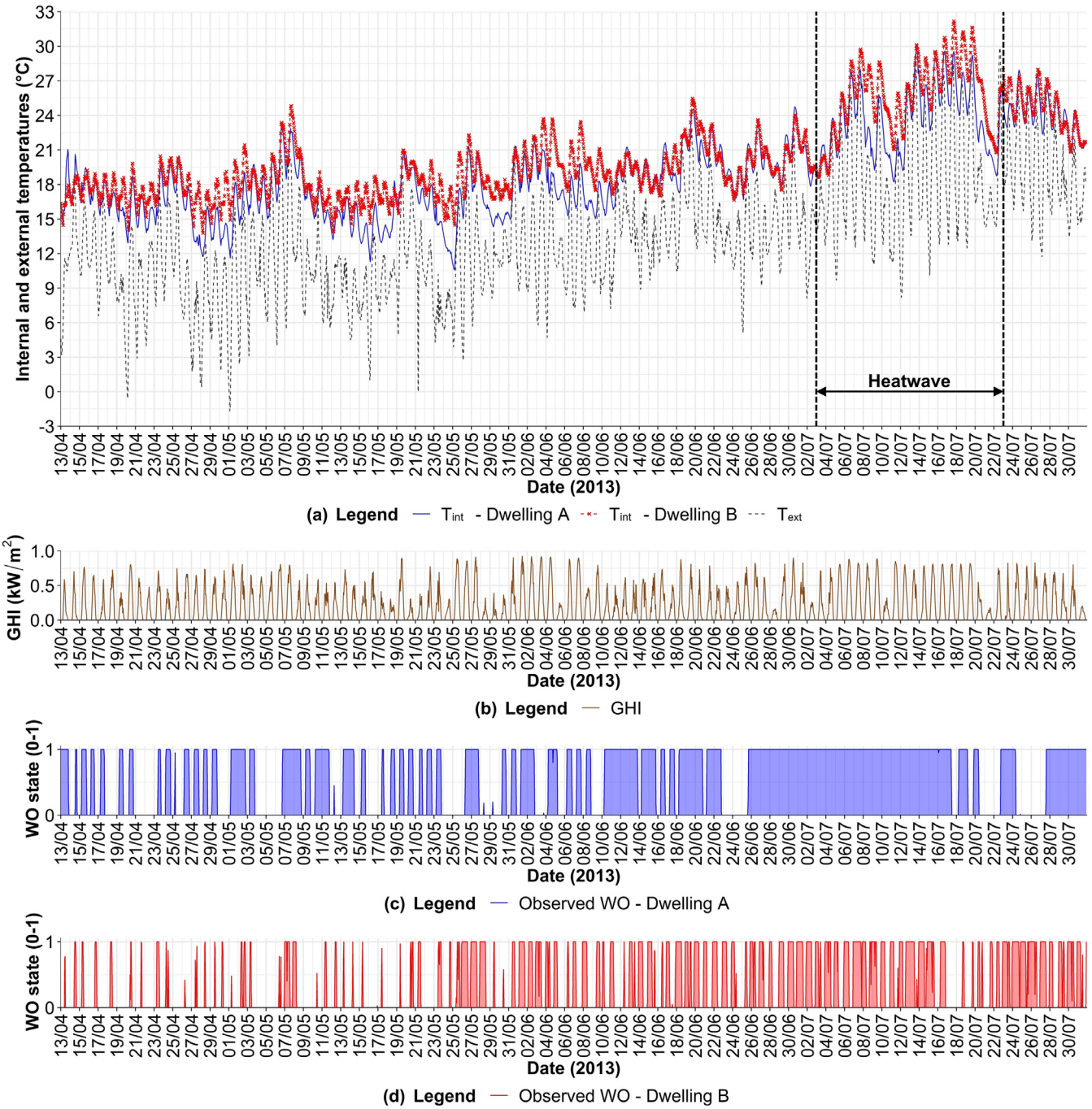

Fig. 1. (a) Hourly averages of the observed internal temperatures (T_{int}) and external air temperatures (T_{ext}) from the 13th April 2013 to the 31st July 2013; (b) Global Horizontal solar Irradiance (GHI); (c) Window Opening state (WO) in dwelling A; (d) Window Opening state (WO) in dwelling B.

as the external temperatures rose. The main difference between the operation of the windows in the two dwellings was that in dwelling B, the occupants reacted to the heat with more frequent window opening but with windows that were never left open for more than 23 h in a row, whereas the occupants in dwelling A left their windows opens for longer periods of time before eventually leaving them open for almost the entire duration of the heatwave (from 26 June to 17 July). Although leaving the windows open overnight (when the outdoor temperatures are low) can lower the indoor temperatures, having them open during the day (when the outdoor temperatures are high) can have the opposite effect. It can be observed on multiple occasions before and during the heatwave (16 May, 24–25 May, 28–29 May, 21–22 July), that even though the windows in dwelling A were closed, the indoor temperature was markedly lower than in dwelling B. Therefore, the cause of the temperature difference between the dwellings cannot be solely attributed to the operation of windows.

3. Methods

3.1. Structure of the models

In the previous work [21], indoor temperatures were forecasted by the ARX model based on the lagged effects of the internal temperature (T_{int}), external air temperature (T_{ext}) and Global Horizontal solar Irradiance (GHI). Here, additional predictor variables are

considered for inclusion in both the newly developed GAMs and ARX models alongside those adopted in the previous study. These new additional variables were chosen based on inputs adopted by Fan and Hyndman [33,36]: hour of the day (H), the indoor temperature at the same time on the previous day ($T_{int}(t-24)$), minimum and maximum indoor temperatures in the past 24 h (T_{int}^{-} and T_{int}^{+}), and the 24-hour means of the indoor temperature ($\overline{T}_{int(24h)}$), outdoor temperatures, ($\overline{T}_{ext(24h)}$) and Global Horizontal solar Irradiance ($\overline{GHI}_{(24h)}$). These additional inputs were iteratively recalculated at every time step.

In GAMs the relationships between the dependent (output) and independent (input) and variables are represented by two-dimensional smooth functions.[2] The only exception is the hour of the day, which was modelled as a cyclic cubic regression spline, which is a smooth function with a constrained relationship at either extreme (i.e. the first and last hours of the day, 00 and 23, adopt the same value). The hour of the day cannot be discretised as a single variable in a linear ARX model, because the relationship would be fixed as a constant for every hour.[3] To perform the forecasts at a specific time-step (t) and forecasting horizon (h), the models are first fitted on the training data, a process which estimates the relationships (parametric for the ARX model and semi-parametric for the GAM) between the independent and dependent variables.

To evaluate whether the Window Opening (WO) state improves the forecasting accuracy, both models were deployed with and without the inclusion of the WO state. Firstly, in order to establish the maximum possible benefit of including a window opening model, the actual WO state was adopted in the models. This approach explicitly determines the net contribution that the WO parameter could make by excluding the uncertainty associated with the auxiliary window state forecasting model.

The general equation of the ARX model can be written in the form shown in Eq. (1).

$$\begin{aligned} T_{int}(t+h) = {} & c + \sum_{i=1}^{n} p_{\Phi,i}\, T_{int}(t+h-i) + p_{\Phi,24}\, T_{int}(t+h-24) \\ & + \sum_{j=0}^{n} p_{\alpha,j}\, T_{ext}(t+h-j) + p_{\beta,j}\, GHI(t+h-j) \\ & + p^{-}\, T_{int(24h)}^{-} + p^{+}\, T_{int(24h)}^{+} + p_{\mu,1}\, \overline{T}_{int(24h)} + p_{\mu,2}\, \overline{T}_{ext(24h)} \\ & + p_{\mu,3}\, \overline{GHI}_{(24h)} + p_{wo}\, WO + e(t+h) \end{aligned} \quad (1)$$

where:

$T_{int}(t+h)$	forecasted hourly internal temperature at the time step t for the forecasting horizon h (°C)
t	hourly time step (h)
h	forecasting horizon, hourly time steps ($h = 1, \ldots, 72$) (h)
c	intercept (°C)
n	maximum lag (previous n time steps) of the input variables that are being considered in the model
i	lag count (1-5) for autoregressive inputs (i.e. previous time steps of the output variable)
j	lag count (0–5) for exogenous inputs, where count 0 is weather data at the forecasted time step
$T_{int}(t+h-i)$	observed or forecasted hourly internal air temperature at lag i before the forecasting horizon h (°C)
$p_{\Phi,i}$	parametric coefficients of the lagged (previous n) T_{int}
$T_{int}(t+h-24)$	observed or forecasted hourly internal air temperature 24 h before the forecasting horizon h (°C)
$p_{\Phi,24}$	parametric coefficient of the T_{int} on the previous day at the same hour (t-24)
$T_{ext}(t+h-j)$	observed or forecasted hourly external air temperature at lag j before the forecasting horizon h (°C)
$p_{\alpha,j}$	parametric coefficients of the lagged (previous n) T_{ext}
GHI (t + h − j)	observed or forecasted Global Horizontal Irradiance at lag j before the forecasting horizon h (W/m^2)
$p_{\beta,j}$	parametric coefficients of the lagged (previous n) GHI
$T_{int(24h)}^{-}$	minimum internal air temperature in the past 24 h (°C)
p^{-}	parametric coefficient of the minimum T_{int} in the past 24 h
$T_{int(24h)}^{+}$	maximum internal air temperature in the past 24 h (°C)
p^{+}	parametric coefficient of the maximum T_{int} in the past 24 h
$\overline{T}_{int(24h)}$	mean internal air temperature in the past 24 h (°C)
$\overline{T}_{ext(24h)}$	mean external air temperature in the past 24 h (°C)
$\overline{GHI}_{(24h)}$	mean Global Horizontal Irradiance in the past 24 h (W/m^2)
$p_{\mu,1}, p_{\mu,2}, p_{\mu,3}$	parametric coefficients of the mean values in the past 24 h of T_{int}, T_{ext} and GHI respectively
WO	Window Opening state (0 – closed; 1 – open)
p_{wo}	parametric coefficient of WO
$e(t+h)$	forecasting error: hourly difference between the forecasted and observed temperatures at the time step t (°C)

The general equation of the GAM can be written in the form shown in Eq. (2).

$$\begin{aligned} g(T_{int}(t+h)) = {} & c + \sum_{i=1}^{n} s_{\Phi,i}\, T_{int}(t+h-i) + s_{\Phi,24}\, T_{int}(t+h-24) \\ & + \sum_{j=0}^{n} s_{\alpha,j}\, T_{ext}(t+h-j) + s_{\beta,j}\, GHI(t+h-j) \\ & + s^{-}\, T_{int(24h)}^{-} + s^{+}\, T_{int(24h)}^{+} + s_{\mu,1}\, \overline{T}_{int(24h)} \\ & + s_{\mu,2}\, \overline{T}_{ext(24h)} + s_{\mu,3}\, \overline{GHI}_{(24h)} + s_{cc}\, H(t+h) \\ & + p_{wo}\, WO + e(t+h) \end{aligned} \quad (2)$$

where:

g	gaussian (default) link function for GAM models
$T_{int}(t+h)$	forecasted hourly internal temperature at the time step t for the forecasting horizon h (°C)
t	hourly time step (h)
h	forecasting horizon in hourly time steps ($h = 1, \ldots, 72$) (h)

[2] Non-parametric functions, where the shapes of predictor variables (i.e. relationships between dependent and independent variables) are entirely determined by the data [32] (see Fig. 2).

[3] The hour of the day in linear models (e.g. ARX) can be modelled with the use of 23 binary dummy variables (1 less than the levels of the categorical variable to avoid the dummy variable trap, which can cause a regression to fail). That is because the last category (i.e. 24th) is captured by the intercept, and is specified when the remaining 23 dummy variables are set to zero [40].

c	intercept (°C)
n	maximum lag (previous n time steps) of the input variables that are being considered in the model
i	lag count (1-5) for autoregressive inputs (i.e. previous time steps of the output variable)
j	lag count (0–5) for exogenous inputs, where count 0 is weather data at the forecasted time step
T_{int} ($t + h - i$)	observed or forecasted hourly internal air temperature at lag i before the forecasting horizon h (°C)
$s_{\Phi, i}$	smooth functions of the lagged (previous n) T_{int}
T_{int} (t + h − 24)	observed or forecasted hourly internal air temperature 24 h before the forecasting horizon h (°C)
$s_{\Phi, 24}$	smooth function of the T_{int} on the previous day at the same hour (t-24)
T_{ext} ($t + h - j$)	observed or forecasted hourly external air temperature at lag j before the forecasting horizon h (°C)
$s_{\alpha, j}$	smooth functions of the lagged (previous n) T_{ext}
GHI ($t + h - j$)	observed or forecasted Global Horizontal Irradiance at lag j before the forecasting horizon h (W/m^2)
$s_{\beta, j}$	smooth functions of the lagged (previous n) GHI
$T^{-}_{int(24h)}$	minimum internal air temperature in the past 24 h (°C)
s^{-}	smooth function of the minimum T_{int} in the past 24 h
$T^{+}_{int(24h)}$	maximum internal air temperature in the past 24 h (°C)
s^{+}	smooth function of the maximum T_{int} in the past 24 h
$\overline{T}_{int(24h)}$	mean internal air temperature in the past 24 h (°C)
$\overline{T}_{ext(24h)}$	mean external air temperature in the past 24 h (°C)
$\overline{GHI}_{(24h)}$	mean Global Horizontal Irradiance in the past 24 h (W/m^2)
$s_{\mu, 1}$, $s_{\mu, 2}$, $s_{\mu, 3}$	smooth functions of the mean values in the past 24 h of T_{int}, T_{ext} and GHI respectively
H	Hour of the day (00–23)
s_{cc}	cyclic penalized cubic regression spline smooth function of H
WO	Window Opening state (0 – closed; 1 – open)
p_{wo}	parametric coefficient of WO
e ($t + h$)	forecasting error: hourly difference between the forecasted and observed temperatures at the time step t (°C)

To constrain the complexity of the models and thus the computational time,[4] which is considerably longer for GAMs than ARX models, the maximum lag (n), of the AutoRegressive (T_{int}) and eXogenous inputs (T_{ext} and GHI) was limited. As in previous work [21,36], input variables were set to a maximum lag n of 5 previous time steps.

For one-step-ahead forecasts, the models require only the observed past internal temperatures (T_{int}) as autoregressive inputs, whilst for multi-step-ahead forecasts, the model adopts partially (when $1 < h \leq n$) or exclusively (when $h > n$) the forecasted internal temperature estimates (generated at previous time steps). Similarly, with exogenous inputs, the one-step-ahead forecasts require only the observed past weather data (T_{ext} and GHI) and the forecasted weather data for that specific time step ($t+1$). For multi-step-ahead forecasts, the model adopts the forecasted weather data partially (when $1 < h \leq n$) or exclusively (when $h >$ n).

The developed models were coded in R [37] and the GAMs were implemented using the *'Mixed GAM Computation Vehicle with Automatic Smoothness Estimation'* (*'mgcv'*) package [38,39].

3.2. Model training and validation

The accuracy of a forecasting model can only be evaluated based on how well it performs in relation to 'new' data [40], and not by comparison with the 'past' data to which it was exposed during the training period. In this study, the initial training period spans from the 13th April 2013 to the 30th June at 23:00, during which there was a marked increase in the external air temperature and the heating was turned off. The forecasting period then starts immediately after this, on the 1st July at 00:00 (initial forecasting origin). However, due to the 72-h forecasting window, it is not possible to evaluate the forecasting accuracy for the first three days, from 1st July at 00:00 to 3rd July at 23:00 for all forecasting horizons (h). The forecasting accuracy was evaluated at different forecasting horizons ($h = 1, 2, 3, 4, 5, 6, 12, 24, 36, 48, 60, 72$), using scale-dependent error metrics: *Mean Bias Error (MBE), Mean Absolute Error (MAE)* and *Root Mean Square Error (RMSE).*

Rolling origin forecasts (i.e. sliding training and forecasting windows) were performed from 1st July at 00:00 to 26th July at 23:00. However, because of the constraints imposed by using a 72-h forecasting window (as the longest forecasting horizon) a full comparison of the forecasting accuracy between the various forecasting horizons is only possible during the 19-day period from 4th July at 00:00 to 22nd July at 23:00, when complete forecasts are available for each forecasting horizon (h).

3.3. Model identification

For the identification of the optimal linear ARX model, as in the previous study [21], model selection was based on the minimisation of the *Akaike Information Criterion (AIC).* However, the consideration of additional input variables compared to the previous study [21] leads to an increase in the number of viable model combinations from 131,072 [21] to 8.4 million and 16.8 million for the ARX model and GAM respectively. This exponential increase in model combinations would render the testing of every possible combination computationally excessive. Therefore, in order to converge quickly on a near-optimal model, a *backward stepwise regression* [40] selection procedure was adopted.

For the linear ARX model, the model selection algorithm begins by including all of the considered input variables in the calculation of the AIC. The algorithm then excludes one variable at a time, re-computing the AIC after each exclusion. The excluded parameter that decreases the AIC value the most is then permanently removed, and the improved model adopted as a reference for further parameter exclusions. The selection algorithm continues removing input variables iteratively until no further decrease in the AIC is observed, whereupon the final reference model is selected. This model selection procedure defines the structure of the model and is performed only once during the initial training period.

[4] The computational time required to fit a model to the data varies considerably depending on the amount of training data and number of inputs. The fitting time (with the forecasting models for indoor temperature) might take just a fraction of a second with the linear ARX models, whereas it might take up to 2-2.5 minutes with a semi-parametric GAM model when using a single core (i.e. running the code in sequence) on an Intel i7-7700HQ CPU with 16GB of RAM.

Model identification is more challenging for GAMs, due to their more complex structures. According to Wood [41], automatic model selection procedures for complex models that consider all of the possible inputs are often unsuccessful. Since the selection procedure (described above) based on the minimisation of the AIC did not show satisfactory results, a backward stepwise regression, based on minimisation of the out-of-sample predictive accuracy (as defined by the MAE) was adopted. This approach was demonstrated by Fan and Hyndman [33] to provide good results, for semi-parametric model selection. During this selection process, only the first part of the training period of the linear ARX model (75% of the data spanning from 13 April 2013 at 00:00 to 11 June 2013 at 23:00) was used to fit the models and the remaining 19 days (25% of the data spanning from 12 June 2013 at 00:00 to 30 June 2013 at 23:00) were used to test the forecasting accuracy, as part of the backward stepwise selection process. As for the ARX models, the model selection procedure is performed only once during the initial training period.

3.4. Multi-step-ahead predictions

In 'real-world' applications any model would require forecasted weather data from one or more [42] nearby meteorological station(s) as an input. Since the uncertainty of weather forecasts increases in proportion to the length of the forecasting horizon, their reliability several days ahead (particularly in a maritime climate) is questionable [42]; as a result, forecasting overheating risks at periods well beyond the forecasting origin is likely to be unreliable. According to the UK Met Office, short-range (1–3 days ahead) weather forecasts, use data that is updated several times per day and are considered to be extremely accurate [43]. On the other hand, medium-range (3–10 days ahead) weather forecasts provide only a general synopsis on a day-to-day basis. For this reason, the developed models were constrained to forecasting indoor temperatures up to 72 h (3-days) ahead. As in the previous study [21], multi-step-ahead forecasts are performed by adopting a recursive strategy based on a rolling forecasting origin (i.e. utilising a sliding training and forecasting windows). This means that after each forecast the model's training window moves forward by one time-step (i.e. 1 h), before recalibrating the relationships of the previously selected predictors and then recalculating the subsequent forecasts. The model automatically stops forecasting when the sliding forecasting window (of 1–72 h) reaches the end of the validation period. Once rolling origin forecasts have been completed for the entire validation period, it is then possible to assess the forecasting accuracy.

3.5. Statistical significance of the forecasting accuracy

To reliably determine which model produces more accurate forecasts, it is insufficient to simply consider the forecasting accuracy. Different models will always produce different forecasts; the question is whether the differences between the predictions have statistical significance or not?

According to the *Diebold-Mariano* (*DM*) *test* [44,45] if the null hypothesis (H_0), that both forecasts have equal accuracy, is rejected (e.g. if the p-value ≤ 0.10), then the alternative hypothesis (H_1) of different accuracy can be accepted at the 90% confidence level. The DM test is specific to a given forecasting horizon h (which is a required input for the test) and is based on comparing the forecasting errors of the two competing models, which are known as the loss differentials (d). The loss differential function can be either based on absolute (Eq. (3); as adopted in this study) or squared errors and must be covariance stationary for the test to be valid.

$$d_t = |e_{1,t}| - |e_{2,t}| \tag{3}$$

where:

d_t	loss differential at the hourly time step t
$e_{1,t}$	forecasting error of the first model at the time step t
$e_{2,t}$	forecasting error of the second model at the time step t

Harvey et al. [46] proposed a modification of the DM test to address limitations associated with small sample sizes and heavy-tailed distributions, a problem which becomes increasingly severe as h increases. The modified DM test (Eq. (4)) differs from the original in two ways: firstly, it multiplies the original statistics by a correction factor (k)(Eq. (5)), which depends on the sample size (N) and forecasting horizon (h); and secondly, it compares the statistics with critical values from a *Student t-test* distribution with (N-1) degrees of freedom, rather than with the standard normal distribution (i.e. the critical values and p-values of the test depend on the sample size N and will tend towards the values of the standard normal distribution when N is large). According to Harvey et al., the modified DM test (Eq. (4)) "constitutes the best available approach to assessing the significance of observed differences between the performance of two forecasts" [45, p.291].

$$DM_{1,2} = \frac{\bar{d}_{1,2}}{\hat{\sigma}_{\bar{d}_{1,2}}} k \tag{4}$$

where:

$DM_{1,2}$	modified Diebold-Mariano test for two competing forecasts (1 and 2)
$\bar{d}_{1,2}$	mean loss differential of a sample with size N
$\hat{\sigma}_{\bar{d}_{1,2}}$	estimate of the standard deviation of $\bar{d}_{12}$
k	correction factor for the modified Diebold-Mariano test

$$k = \sqrt{\frac{N + 1 - 2h + N^{-1} h(h-1)}{N}} \tag{5}$$

where:

k	correction factor for the modified Diebold-Mariano test
N	sample size
h	forecasting horizon on which the forecasting errors of the two competing models have been calculated

In R [37], the *modified DM test* is available in the *'Forecasting Functions for Time Series and Linear Models'* package, known as *'forecast'* [47]. In this study, to evaluate the significance of the different forecasting accuracies at different forecasting horizons), the modified DM tests were carried out by considering the absolute loss function (Eq. (3)) and by testing three different alternative hypotheses (H_1, H_2, H_3) at the 90% confidence level (wherein a 95% CI is considered excessively restrictive in order to identify a statistical difference several time steps ahead). The three alternative hypotheses test whether model 1 is significantly more accurate than model 2 and vice versa, based on whether the competing model has a higher (H_1, one-sided test), lower (H_2, one-sided test) or different accuracy (H_3, two-sided test).

3.6. Forecasting window opening states

The main aim of including an auxiliary model to predict the WO state (as explained in Section 1.3) is to improve the overall forecasting accuracy. Prior to determining this, it is necessary to consider how accurately the WO state can be predicted in residential settings at an hourly time step; the majority of previous studies in the literature have adopted 5 or 10-minutely predictive time steps [24–26].

Based on findings from the literature, a *logistic* univariate GAM with multiple predictors was developed (Eq. (6)). GAM is essentially an extension of the *Generalized Linear Model (GLM)* approach

which is considerably more flexible because the relationships between the independent and dependent variables are not assumed to be linear. In addition, the use of GAMs avoids the pitfalls of dealing with higher order polynomial terms to model non-linear relationships in linear models where it is not necessary to know, *a priori*, the type of function which best describes the relationship [32]. Here the relationships s_1, s_2 and s_3 of the internal temperature (T_{int}), external temperature (T_{ext}) and Global Horizontal solar Irradiance (GHI) respectively, are represented by smooth functions that can assume non-linear relationships (i.e. the probability of a state to vary across the range of the input variables). Since the WO model was auxiliary to the main system model, the internal temperature being forecasted by the main model cannot be used as an input to the WO model at the same time step. Therefore, the indoor temperature at the previous hourly time step was used instead. As the time of day (H) is known to be influential in relation to WO [23], this parameter was included in the GAM. Lastly, because the interaction with the windows might also depend on the day of the week (D)(e.g. working individuals might be absent during weekdays) D was also included as an input to the model.

The general equation of the auxiliary WO state model can be written in the form shown in Eq. (6).

$$\ell(\mathrm{WO}(t)) = \mathrm{c} + s_1\,T_{int}(t-1) + s_2\,T_{ext}(t) + s_3\,\mathrm{GHI}(t) + s_{cc}\mathrm{H}(t) + \mathrm{pD(t)} \quad (6)$$

where:

ℓ	logistic (logit) link function for binary output with the GAM model
WO(t)	predicted Window Opening state at the time step t (0 – closed; 1 – open)
c	intercept
T_{int} (t-1)	internal temperature at the previous time step (t-1) (°C)
T_{ext} (t)	external temperature at the time step t (°C)
GHI (t)	Global Horizontal Irradiance at the time step t (W/m^2)
s_1, s_2, s_3	smooth functions of the predictor variables T_{int}, T_{ext} and GHI respectively
H	Hour of the day (00–23)
s_{cc}	cyclic penalized cubic regression spline smooth function of H
D	Day of the week (Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday)
p	parametric coefficient of D

Because the data (Fig. 1) showed an intensification of window opening behaviour from 26 June to 17 July with changes in occupant behaviour between the peak (17–20 July), end (21–23 July) and after (24–28 July) the heatwave, it was decided to extend the validation period of the *logistic* GAM model to the whole month of July 2013 (from 1st July at 00:00 to 30th July at 23:00).

3.7. Discrimination criteria of the auxiliary logistic model of the window opening state

A cut-off threshold of 0.5 was adopted to classify the predicted hourly values of the windows into the two possible states: window closed (0 – if WO $\leq$ 0.5); and window open (1 – if WO $>$ 0.5). In order to validate the ability of a *logistic* model, a confusion matrix was used to compare modelled outcomes with the observed Positive (P) and Negative (N) states using four classification categories: True Positive (TP), correctly predicted open, False Positive (FP), predicted open but actually closed, True Negative (TN), correctly predicted as closed and False Negative (FN), predicted closed but actually open. These parameters enable the calculation of: sensitivity or True Positive Rate (TPR = TP / P); specificity or True Negative Rate (TNR = TN / N); fall-out or False Positive Rate (FPR = 1 - TNR); miss rate or False Negative Rate (FNR = 1 - TPR); and the proportion of correct predictions, ACCuracy (ACC = (TP + TN) / (P + N)) [26]. Models with a strong predictive value are characterized by TPRs higher than FPRs [24].

4. Results

4.1. Model identification

In order to automatically select near-optimal models, backward stepwise regressions, based on the minimisation of the AIC and MAE were adopted for ARX and GAM models respectively. During the model identification process, a number of the inputs (including T_{int}, T_{ext}, and/or GHI) were discarded from both the GAM and ARX models at some of the previous time steps (Table 1). The internal temperature that was recorded at the same time on the previous day (T_{int} (t-24)), as well as the minimum and maximum internal temperature in the past 24 h ($T^{-}_{int(24h)}$ and $T^{+}_{int(24h)}$), and the mean GHI in the past 24 h ($\overline{GHI}_{(24h)}$) were selected in 3 out of the 4 models. Conversely, terms describing the mean internal and external temperatures in the past 24 h ($\overline{T}_{int(24h)}$ and $\overline{T}_{ext(24h)}$) were never selected.

Although the hour of the day (H) was included in GAM models as a non-linear smooth function it was omitted by the selection algorithm for dwelling B. In order to evaluate the effect of including the WO state variable into the forecasting model for indoor temperatures, two model variants were created for both the ARX and GAM model (one with and one without the WO variable, see Table 1).

Examining the fitting of the GAM provides a useful means of understanding how optimal relationships are attributed to the various variables (Fig. 2). It is evident from this analysis that the autoregressed variables of T_{int} assume the most dominant weights, and the nearer they are temporally located to the value that is being forecasted, the higher their weighting. Moreover, the final result is the sum of positive and negative effects, which in the ARX models is always linear, whereas in the semi-parametric GAM models might be non-linear.

The extremes on the y-axes of the plots (Fig. 2) indicate that, with the exception of the GHI, the exogenous inputs have considerably lower weights than the autoregressed variables and they, therefore, act as a tuning effect on the predicted dependent variable. It should be noted that since the exogenous variables (i.e. T_{ext}, GHI) are not normalised their ranges are different to one another and to that of the autoregressed variables (T_{int} (t-i)). For this reason, their absolute influence on the dependent variable (T_{int} (t)) cannot be directly compared via the parameter weightings.

For both dwellings, when the WO state is equal to 1 (i.e. window open), the relationships are negative which indicates a reduction in the predicted temperature. Nevertheless, the WO coefficients (for WO state = 1) are low in absolute terms, with −0.03 and −0.05 applied to dwellings A and B respectively.

4.2. Indoor temperature forecasts without the window opening state

Forecasts with the GAMs produced considerably lower MBEs than those from the ARX models for forecasting horizons up to 24 h (Table 2). At longer forecasting horizons, however ($24 < h \leq 48$), the improvement in the MBE becomes smaller and once $h > 48$ the MBE with the GAMs is worse than for the ARX models. The MAE and RMSE provide a similar perspective, suggesting that: GAMs are capable of producing more accurate forecasts for $h \leq 6$ h; whilst for $h = 12$ h, the forecasting accuracy of the two models is very similar; but when $h \geq 24$ h, ARX models are much better. Analyses using the modified Diebold-Mariano (DM) test confirmed that the improved forecasting accuracy of GAMs was statistically significant at

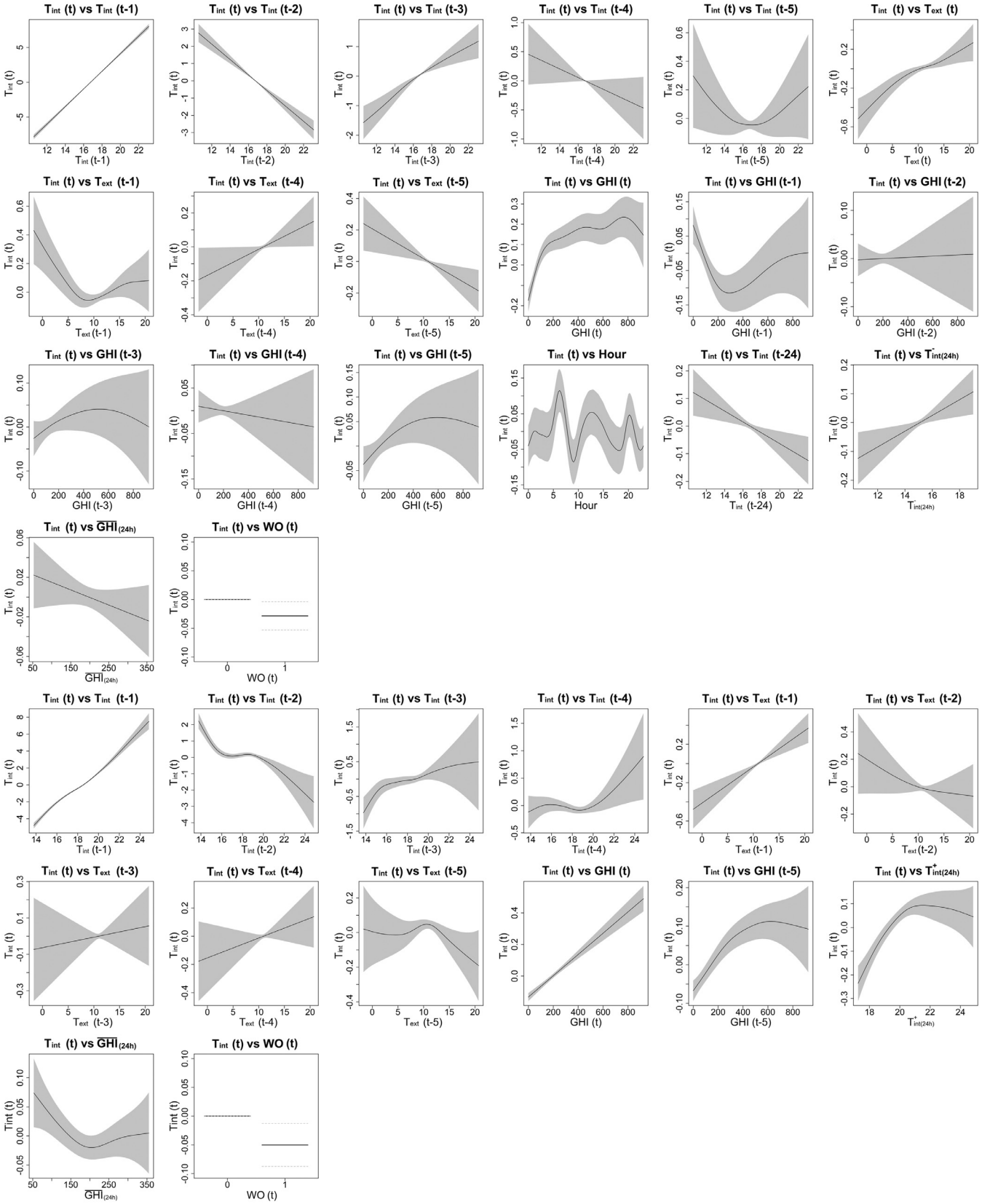

Fig. 2. Relationships of the selected GAM models with WO, for dwelling A (upper 4 rows) and B (bottom 3 rows); the grey bands / dashed lines indicate the 95% confidence intervals of the assigned relationships.

Table 1
Selected predictor variables for dwellings A and B, for ARX and GAM models, with and without (w/o) WO.

Predictor variables	Dwelling A				Dwelling B			
	ARX		GAM		ARX		GAM	
	w/o WO	with WO	w/o WO	with WO	w/o WO	with WO	w/o WO	with WO
WO	$\times_m$	p_m	$\times_m$	p_m	$\times_m$	p_m	$\times_m$	p_m
T_{int} (t-1)	p		s		p		s	
T_{int} (t-2)	p		s		p		s	
T_{int} (t-3)	p		s		p		s	
T_{int} (t-4)	p		s		×		s	
T_{int} (t-5)	×		s		p		×	
T_{ext} (t)	p		s		p		×	
T_{ext} (t-1)	p		s		p		s	
T_{ext} (t-2)	p		s		×		s	
T_{ext} (t-3)	×		×		×		s	
T_{ext} (t-4)	p		×		p		s	
T_{ext} (t-5)	p		s		×		s	
GHI (t)	p		s		p		s	
GHI (t-1)	p		s		×		×	
GHI (t-2)	×		s		×		×	
GHI (t-3)	p		s		p		×	
GHI (t-4)	×		s		p		×	
GHI (t-5)	p		s		×		s	
T_{int} (t-24)	p		s		p		×	
$\bar{T}^{-}_{int(24\ h)}$	p		s		p		×	
$\bar{T}^{+}_{int(24\ h)}$	p		×		p		s	
$\bar{T}_{int(24\ h)}$	×		×		×		×	
$\bar{T}_{ext(24\ h)}$	×		×		×		×	
$\overline{GHI}_{(24\ h)}$	p		s		×		s	
Hour (H)	n/a		s_{cc}		n/a		×	

Legend: $\times_m$ = manually excluded WO variable; p_m = manually included parametric WO variable; × = discarded predictor variable; p = selected parametric variable; s = selected smooth variable; s_{cc} = selected variable as cyclic penalized cubic regression spline smooth; n/a = hour variable (H) is not applicable in the ARX model.

Table 2
Forecasting accuracy of GAM vs. ARX models in two dwellings during the 2013 heatwave, without (w/o) the WO state, including the modified Diebold-Mariano comparison tests (DM test).

Forecasting horizon h (hours)	Dwelling A							Dwelling B						
	ARX (w/o-WO)			GAM (w/o-WO)			DM test	ARX (w/o-WO)			GAM (w/o-WO)			DM test
	MBE (°C)	MAE (°C)	RMSE (°C)	MBE (°C)	MAE (°C)	RMSE (°C)		MBE (°C)	MAE (°C)	RMSE (°C)	MBE (°C)	MAE (°C)	RMSE (°C)	
1	−0.02	**0.13**	0.21	−0.01	**0.13**	0.21	√	−0.05	**0.12**	0.15	−0.01	**0.10**	0.13	√
2	−0.04	**0.25**	0.36	−0.01	**0.24**	0.35	√	−0.10	**0.21**	0.27	−0.03	**0.18**	0.24	√
3	−0.06	**0.35**	0.48	−0.02	**0.33**	0.45	√	−0.14	**0.28**	0.35	−0.04	**0.24**	0.32	√
4	−0.08	**0.44**	0.58	−0.03	**0.41**	0.54	e/a	−0.18	**0.33**	0.42	−0.04	**0.29**	0.39	√
5	−0.10	**0.50**	0.66	−0.04	**0.48**	0.61	e/a	−0.21	**0.37**	0.47	−0.04	**0.33**	0.45	√
6	−0.12	**0.57**	0.73	−0.05	**0.54**	0.68	e/a	−0.25	**0.41**	0.52	−0.05	**0.37**	0.50	√
12	−0.20	**0.81**	0.99	−0.10	**0.78**	0.98	e/a	−0.40	**0.59**	0.70	−0.06	**0.53**	0.78	e/a
24	−0.27	**0.92**	1.13	−0.14	**0.98**	1.25	e/a	−0.56	**0.79**	0.91	0.08	**0.99**	2.49	n/a
36	−0.31	**0.92**	1.13	−0.22	**1.03**	1.30	e/a	−0.65	**0.88**	1.02	1.00	**2.23**	10.11	n/a
48	−0.34	**0.93**	1.13	−0.29	**1.06**	1.32	e/a	−0.71	**0.94**	1.08	4.23	**5.76**	40.57	n/a
60	−0.35	**0.94**	1.14	−0.39	**1.11**	1.41	e/a	−0.76	**0.98**	1.12	15.76	**17.51**	166.8	n/a
72	−0.36	**0.95**	1.14	−0.47	**1.21**	1.54	e/a	−0.80	**1.01**	1.14	57.37	**59.33**	697.1	n/a

Legend: √ = the GAM model has significantly better accuracy at the 90% probability level; e/a = equal accuracy / no difference; n/a = test not applicable because the assumption of covariance stationarity of the loss differential function is violated.

the 90% probability level, but only up to $h = 3$ h for dwellings A and $h = 6$ h for dwelling B.

Whereas there is a comparable forecasting accuracy between the GAM and ARX models for $h = 12$ h (Fig. 3), for dwelling B, a localised disruption in the GAM forecast occurs on the 7th of July (Fig. 4). This is because when forecasting temperatures close to or above the maximum temperatures experienced during the training period some of the predictor variables contain estimates of the relationships which encompass a broad confidence interval (Fig. 2). Therefore, until the model has been exposed to such hot conditions the out of range values predicted by these terms remain highly uncertain. The recursive strategy used by GAMs for multi-step-ahead forecasts means that such errors compound exponentially. Thus, whilst the local over-prediction (seen in Fig. 4 on 7 July), is not unduly pronounced at short forecasting horizons ($h \leq 6$) it degenerates quickly as the forecasting horizon (h) increases (Table 2). This local disruption is evident in the MBE, MAE and RMSE for $h \geq 24$ h (Table 2), being most pronounced in the RMSE metric, which is highly sensitive to outliers. For the ARX model, the errors are much smaller (Table 2) thereby avoiding the local disruptions that were observed with the GAM (Fig. 5 cf. Fig. 4) by allowing only linear relationships using the same regression coefficients throughout the whole range of temperatures.

Following the first warm period, the non-linear relationships in the GAMs are recalculated and as a result, the error in subsequent forecasts of impending high indoor temperatures are greatly

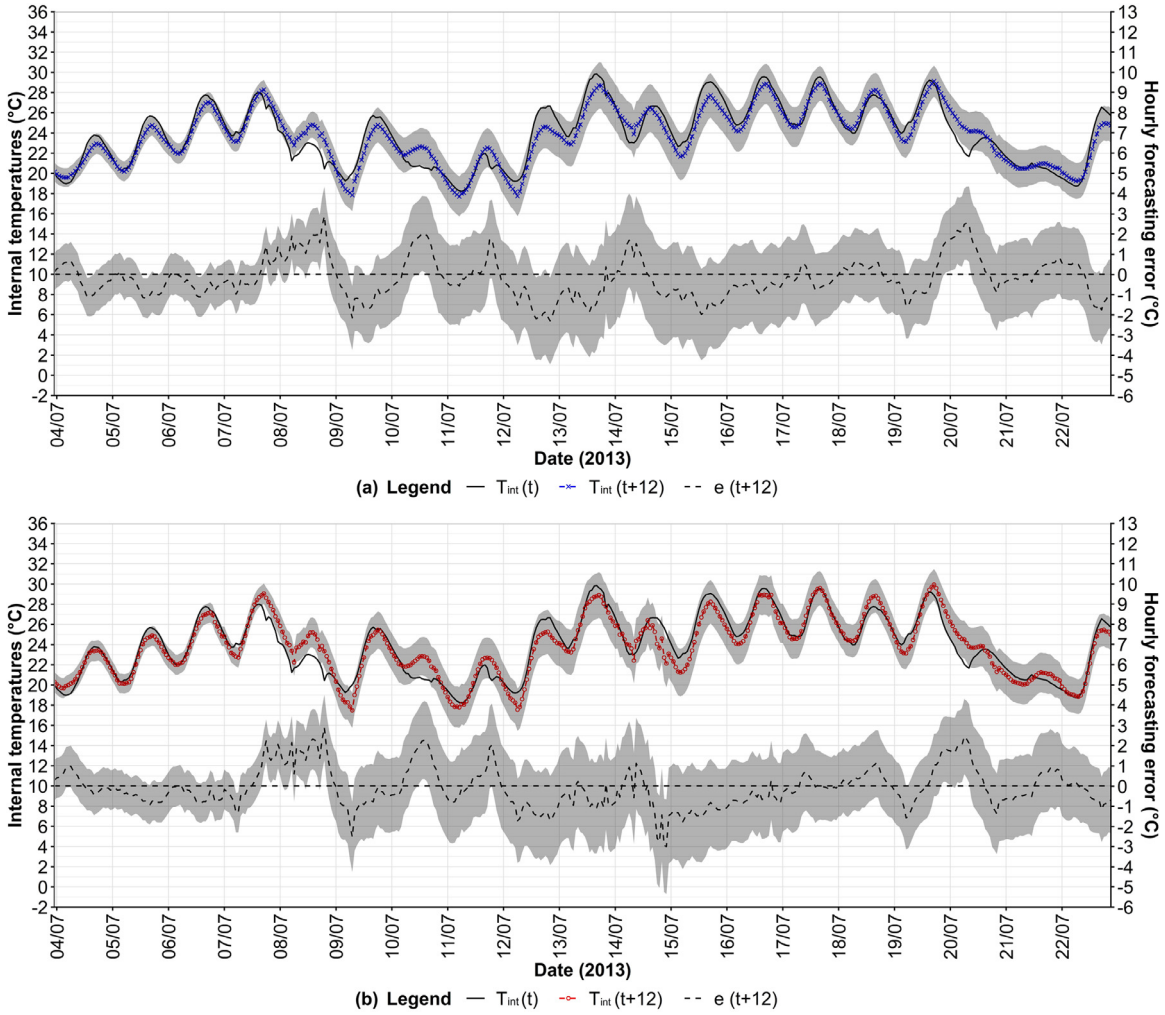

Fig. 3. Dwelling A: observed, T_{int} (t), and predicted, T_{int} $(t+h)$, hourly internal temperatures with hourly forecasting error, e$(t+h)$, and the 95% predictive intervals (grey bands) for the 12 h forecasting horizon (h), with ARX model (a) and GAM (b).

reduced (Fig. 4). However, in terms of reliability in a 'real-world' application, it is concerning that a non-linear model might fail temporarily when rapidly approaching a considerably warmer period for the first time.

4.3. Relationships in the logistic GAM for the prediction of the window opening state

For the *logistic* GAM the relationships (Fig. 6, y-axis) of the independent variables are expressed as logit functions (i.e. log-odds or logarithm of the odds; logit$(p)=\ln[p/(1\text{-}p)]$). These values can be converted to the probability of a window being open as follows (p = odds / (1 + odds); odds = exp[logit(p)]): -4 = 1.8%; -3 = 4.8%, -2 = 11.9%; -1 = 26.9%; 0 = 50%, 1 = 73.1%; 2 = 88.1%; 3 = 95.3; 4 = 98.2; 5 = 99.3%; 6 = 99.8%.

The probability of the windows being open increases considerably at higher internal temperatures (T_{int}) but decreases at higher external air temperatures (T_{ext}). Whereas GHI has almost no effect on the WO state for dwelling A, the probability of opening the windows increases linearly with GHI for dwelling B. Similarly, the influence of the hour of the day (H) shows a considerably different effect for the two dwellings. For dwelling A, the probability of the windows being open remains close to 50% most of the time but is slightly higher in the late morning and at midday. Whilst, for dwelling B, the probability of the windows being open is highest ($p \approx 85\%$) during the early morning and lowest during the evening ($p \approx 10\%$). Even though the day of the week (D) has less influence on the WO, there is a small amount of variability during the week. For example, in dwelling A, there is a lower chance of the windows being open on Sundays compared to the rest of the week. Whilst for dwelling B, there is a higher probability of windows being opened on the weekends and also on Tuesdays (Fig. 6).

4.4. Forecasting the window opening state using logistic GAMs

During summer 2013, the occupants of both dwellings opened the windows for longer periods of time as the temperatures rose (Fig. 1), until eventually leaving them continuously open in dwelling A from 26 June to 17 July. Whereas during the training period of the logistic model (13 April – 30 June) the windows were open ($t_{open,tr}$) 48.8% and 25.8% of the time for dwellings A and B respectively, during the validation period of the logistic model (whole of July 2013) the window opening time ($t_{open,val}$) increased to 77.4% and 53.9% of the time for dwellings A and B respectively (Table 3). As a result, dwelling A recorded a slightly unbalanced testing period with 576 Positives (P), i.e. hours when the window was open, and 168 Negatives (N), i.e. hours when the window was closed, whilst dwelling B is considerably more balanced with 401 P and 343 N.

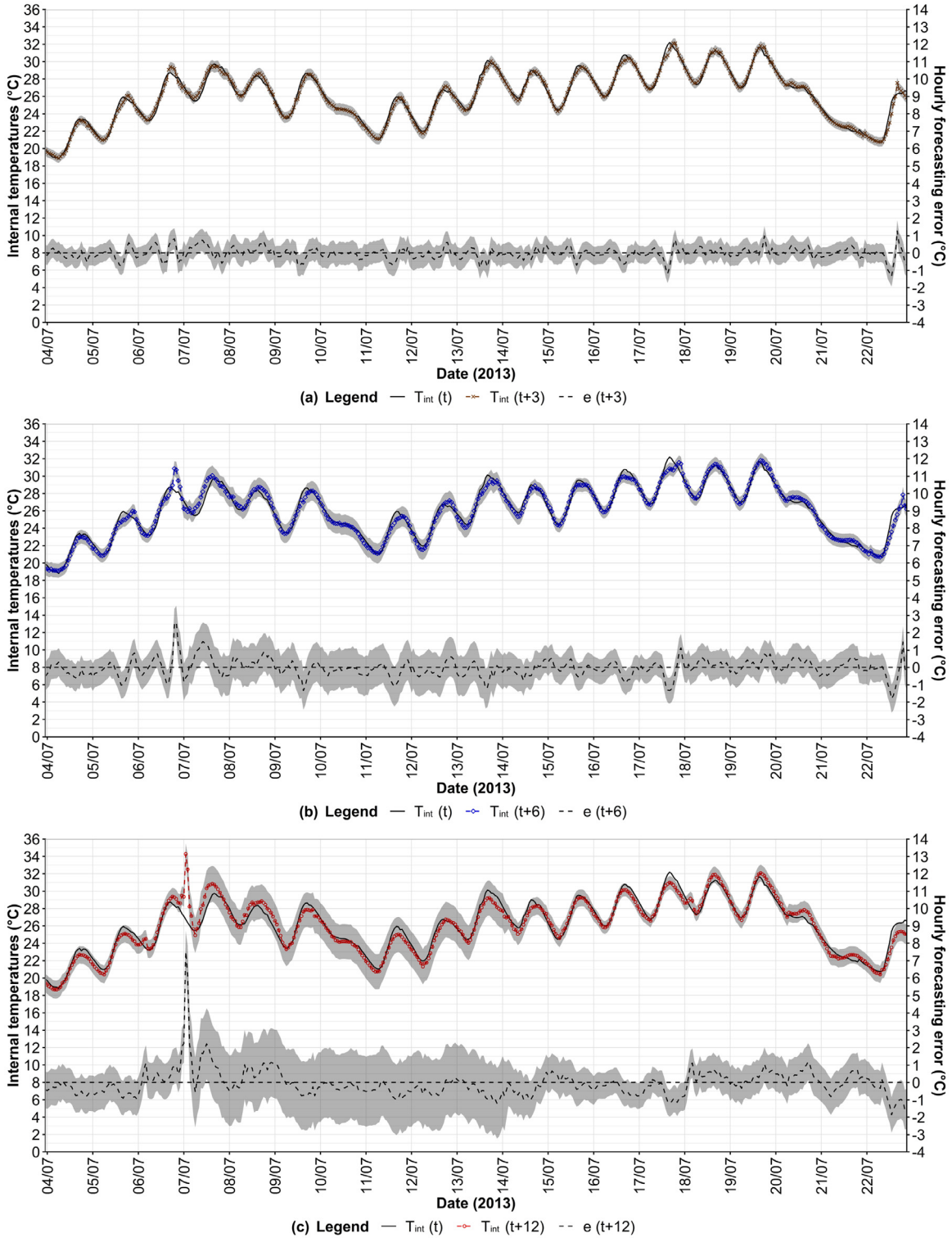

Fig. 4. Dwelling B: observed, T_{int} (t), and predicted, T_{int} $(t+h)$, hourly internal temperatures with hourly forecasting error, e$(t+h)$, and the 95% predictive intervals (grey bands) for 3 h (a), 6 h (b) and 12 h (c) forecasting horizons (h), with GAM.

Table 3
Percentage-time windows are open during training ($t_{open,tr}$) and validation ($t_{open,val}$) periods and discrimination of the logistic GAMs for the hourly window opening state for dwellings A and B.

Dwelling	$t_{open,tr}$ (%)	$t_{open,val}$ (%)	P	TP	FN	N	TN	FP	TPR (%)	FNR (%)	TNR (%)	FPR (%)	ACC (%)
A	48.8	77.4	**576**	542	34	**168**	19	149	**94.1**	5.9	11.3	**88.7**	**75.4**
B	25.8	53.9	**401**	274	127	**343**	202	141	**68.3**	31.7	58.9	**41.1**	**64.0**

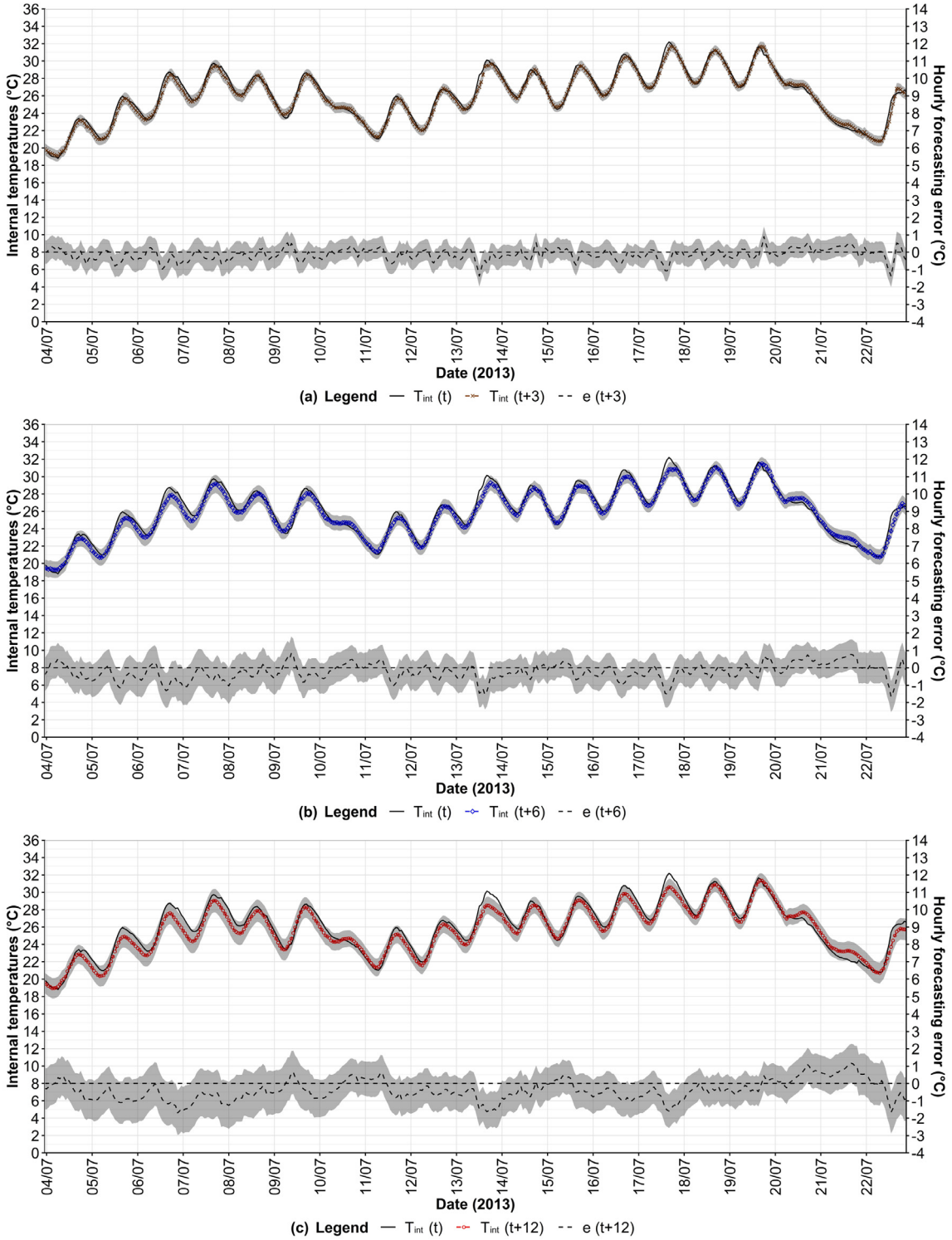

Fig. 5. Dwelling B: observed, T_{int} (t), and predicted, T_{int} $(t+h)$, hourly internal temperatures with hourly forecasting error, e$(t+h)$, and the 95% predictive intervals (grey bands) for 3 h (a), 6 h (b) and 12 h (c) forecasting horizons (h), with ARX model.

In dwelling A, although the occupants behaved atypically, leaving the window open for almost the entire heatwave (Fig. 1), the model managed to capture this general tendency (Fig. 7), but it produced FNs on cooler evenings and nights (e.g. 1–2, 2–3, 4, 8–9 and 11 July). As might be expected, when similar ranges of indoor and outdoor temperatures were experienced (Fig. 7), the model predicted the same outcome (i.e. window open). However, on certain days (e.g. 17–18, 19 and 24–27 July), the windows were apparently closed for non-temperature related reasons that are unaccounted for by the model, which led to FPs. In dwelling B, the occupants tended to operate the windows on a daily basis (Fig. 7), with the only exception being on the 17th of July. The model was able to replicate the daily opening pattern (Fig. 7) but generated FPs on colder days (e.g. 3–4 July). However, the model, cannot predict a change in the occupants' behaviour (i.e. when leaving the windows closed during the day and opening them only in the evening, instead of leaving them open night and day) at the peak of the heatwave (i.e. 17–19 July), which

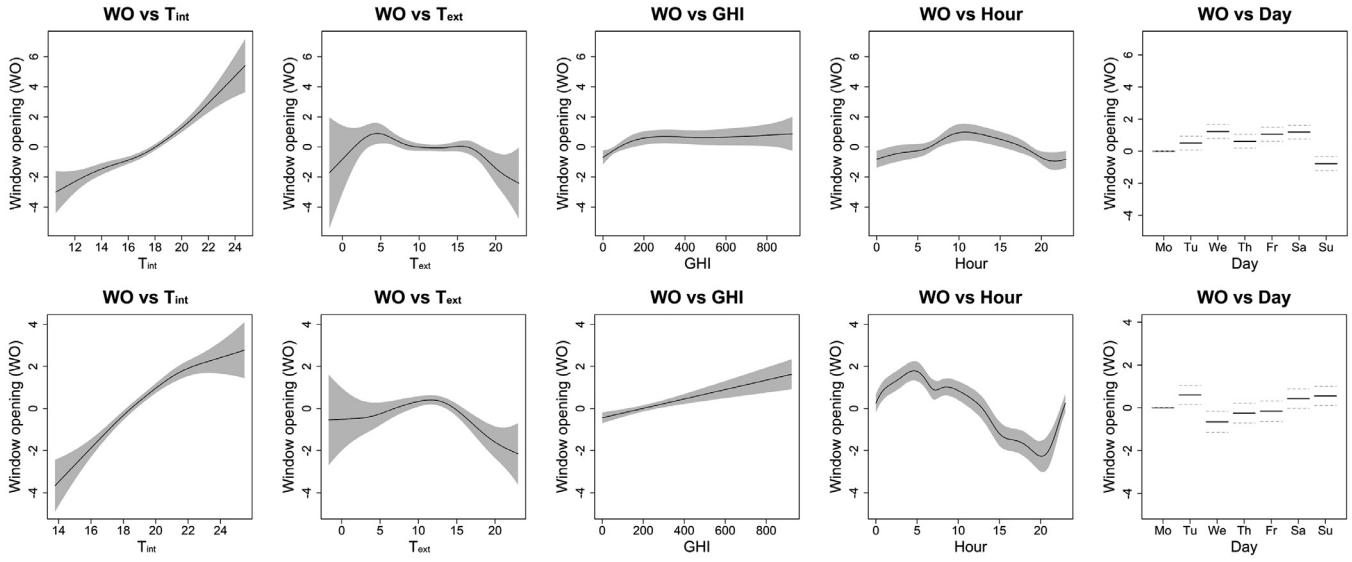

Fig. 6. Relationships of the logistic GAMs for the prediction of the hourly Window Opening (WO) for dwellings A (upper row) and B (bottom row); the grey bands / dashed lines indicate the 95% confidence intervals of the assigned relationships.

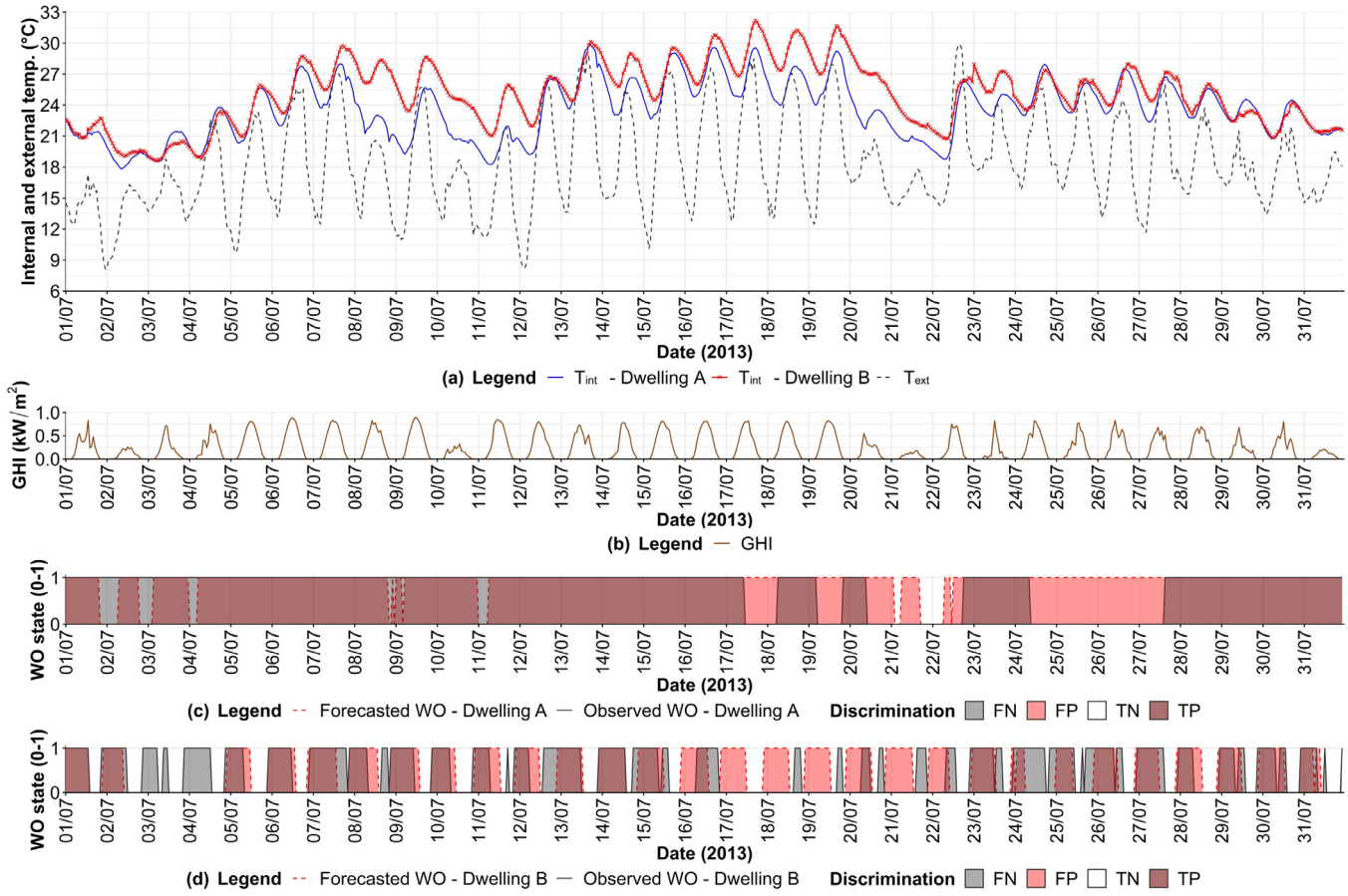

Fig. 7. (a) Hourly averages of the observed internal temperatures (T_{int}) in dwelling A and B, and external air temperatures (T_{ext}); (b) Global Horizontal solar Irradiance (GHI); (c,d) observed Window Opening (grey shading) state and forecasted Window Opening (red shading) with the logistic GAM for dwellings A and B.

led to a large number of FPs until the end of the heatwave (i.e. 22 July). Overall, in both cases, the logistic models performed with adequate predictive discrimination achieving high TPRs (94.1% and 68.3% for dwellings A and B respectively), and with lower FPRs (88.7% and 41.1% for dwellings A and B respectively) and an adequate ACC (75.4% and 64.0% for dwellings A and B respectively).

4.5. Indoor temperature forecasts incorporating the window opening state

Adding the actual monitored WO state as a parametric input produced very similar results to the models without the WO input (Table 4). In fact, for dwelling A, the modified DM comparison tests suggested that models with and without the window

Table 4
Forecasting accuracy of GAM vs. ARX models in two dwellings during the 2013 heatwave, with the WO state, including the modified Diebold-Mariano comparison tests (DM test) vs. the results without the WO state (Table 2).

Forecasting horizon h (hours)	Dwelling A						Dwelling B					
	ARX (with WO)			GAM (with WO)			ARX (with WO)			GAM (with WO)		
	MAE (°C)	RMSE (°C)	DM test	MAE (°C)	RMSE (°C)	DM test	MAE (°C)	RMSE (°C)	DM test	MAE (°C)	RMSE (°C)	DM test
1	**0.13**	0.20	e/a	**0.13**	0.21	e/a	**0.13**	0.21	w/o	**0.13**	0.21	w/o
2	**0.25**	0.36	e/a	**0.24**	0.35	e/a	**0.25**	0.36	w/o	**0.24**	0.34	w/o
3	**0.35**	0.48	e/a	**0.33**	0.46	e/a	**0.35**	0.48	w/o	**0.32**	0.44	w/o
4	**0.43**	0.58	e/a	**0.41**	0.55	e/a	**0.43**	0.58	w/o	**0.39**	0.52	w/o
5	**0.50**	0.66	e/a	**0.48**	0.62	e/a	**0.50**	0.66	w/o	**0.45**	0.59	w/o
6	**0.57**	0.73	e/a	**0.54**	0.68	e/a	**0.56**	0.73	w/o	**0.51**	0.64	w/o
12	**0.81**	1.00	e/a	**0.79**	0.98	e/a	**0.81**	1.00	w/o	**0.73**	0.90	w/o
24	**0.94**	1.14	e/a	**1.01**	1.27	e/a	**0.91**	1.13	e/a	**0.86**	1.09	n/a
36	**0.95**	1.14	e/a	**1.10**	1.34	e/a	**0.92**	1.13	e/a	**0.86**	1.10	n/a
48	**0.95**	1.14	e/a	**1.13**	1.37	e/a	**0.93**	1.13	e/a	**0.90**	1.13	n/a
60	**0.96**	1.15	e/a	**1.17**	1.46	e/a	**0.94**	1.14	e/a	**0.95**	1.21	n/a
72	**0.96**	1.15	e/a	**1.26**	1.57	e/a	**0.95**	1.14	e/a	**1.01**	1.27	n/a

Legend: e/a = models with and without the WO input have an equal accuracy at the 90% probability level; w/o = the model without the WO input has a significantly better accuracy at the 90% probability level; n/a = test not applicable because the assumption of covariance stationarity of the loss differential function is violated.

opening parameter had statistically equal accuracy. On the other hand, for dwelling B, with the addition of the WO state the forecasting accuracy was significantly worse at the 90% probability level when $h \leq 12$ h, for both ARX model and GAM. Whilst for dwelling B, the addition of the WO state resulted in a forecast which avoided the previously observed local disruptions (Fig. 4); however, their absence should not be attributed to the addition of the WO state as an input, but rather to the slightly different structure of the model. Depending on the identified model structure, local disruptions might still appear due to the general instability of GAMs when forecasting outside of the range of the predictor variables upon which the models were trained.

5. Discussion

The results demonstrate that the inclusion of substantially more input variables to the ARX models than in the authors' previous study [21] did not improve their accuracy at shorter forecasting horizons. For example, the 6 h forecasts produced MAEs of 0.57 and 0.41 °C for dwellings A and B respectively (Table 2) compared to MAEs of 0.21, 0.51 and 0.55 °C in [21]. Over longer forecasting horizons, such as 72 h, ARX models produced an MAE of 0.95 and 1.01 °C (Table 2), for dwellings A and B respectively which is higher than the MAEs of 0.49, 0.63 and 0.69 °C recorded in the previous study [21]. However, the lower forecasting accuracy, reported here, should not be attributed to poorer model performance but rather to the extended period over which it was evaluated. In the previous study [21], the forecasting accuracy was computed for only one week of data where the day of, and the day after the two-day heatwave produced the largest forecasting errors. The intensive and long-lasting nature of the 2013 heatwave used in this study enabled errors to be computed over a 19-day period, during which there were several pronounced drops in the outdoor and indoor temperatures. The mean zonal indoor temperatures were also approximately 6.5 °C (dwelling A) and 7.3 °C (dwelling B) above the corresponding indoor temperatures during the initial training period. Considering these forecasting challenges, the ARX model can be considered to have performed well and with good generalisation ability.

In the absence of previous results from the literature, the forecasting accuracy of the semi-parametric GAMs can be best assessed by comparison with the forecasts of the linear ARX models. The GAMs produced statistically better forecasts than the ARX models (at the 90% level) for horizons up to 6 h ahead (with MAEs of 0.54 and 0.37 °C for dwellings A and B respectively at 6 h cf. 0.57 and 0.41 °C with the ARX models, Table 2). For forecasting horizons beyond 12 h, the GAMs were not significantly better than the ARX models.

The findings of this study concur with the established forecasting literature in a number of important aspects. Firstly, research by Taieb [48] and Teräsvirta, Van Dijk and Medeiros [49] shows that in cases where the time series is only weakly non-linear, or if there is only a rare occurrence of non-linear features (Fig. 2), the use of more complex non-linear models is not justified since simpler linear models already provide a good approximation, especially for short time series and long forecasting horizons. Secondly, Ferracuti et al. [30] have demonstrated that recursive linear ARX models are more accurate than NARX models for long-term indoor temperature predictions in air-conditioned buildings. This concurs with the study by Teräsvirta, Van Dijk and Medeiros [49], where the researchers found that autoregressive single hidden layer feedforward neural networks (without Bayesian regularisation) were not capable of improving upon linear autoregressive models, especially at longer forecasting horizons. It is known that with ANNs there is a risk of explosive models (i.e. models where error gradients grow exponentially) causing models to become unstable, with implausible forecasts at long forecasting horizons [49]. Similarly, with GAMs there is a risk of instability at long forecasting horizons when predicting outside the range upon which the dependent variables were trained. Here it was shown that GAMs were vulnerable to disruptions when rapidly approaching a considerably warmer period for the first time, rendering them highly uncertain, and difficult to control at longer forecasting horizons.

Whilst this study has intentionally focused on testing the models on shorter time series data (i.e. without data from previous years and heatwaves), training the models on historical data from past heatwaves could potentially obviate this issue. However, any changes to the building fabric or occupancy in the interim would invalidate the previously established relationships embedded in a historically trained model. In addition, this approach is predicated on the assumption that suitable historical data exists. Moreover, this is a problematic assumption in the case of overheating forecasting since climate change projections suggest a continued upward trend in global summertime temperatures and with an increased frequency of extreme heat events [4,5]. Considering these factors collectively, the use of GAMs (in this context) should be constrained to shorter forecasting horizons, especially when automatic model selection procedures are adopted. In contrast, linear

ARX models appear to be a more reliable[5] choice. When computational time is considered, ARX models are also favoured due to their minimal fitting times. In contrast, GAMs require much longer fitting times and high-dimensional settings are more difficult to handle (i.e. for each predictor variable a function has to be estimated instead of a single slope parameter in the linear model) [50]. Nonetheless, GAMs can be safely fitted even when the nature of the underlying structure is unclear or is mostly linear [50]. Conversely, when forecasting at shorter horizons, and when the computational time is less relevant, the potentially higher forecasting accuracy of GAMs might be advantageous.

The forecasting accuracies presented in this study are in line with previous studies involving the prediction of internal temperatures; although most previous research has focused on offices with mechanical cooling and with higher data resolutions. Mustafaraj et al. [28] observed MAEs of 0.27–0.38 °C for an ARX model predicting 1.5 h ahead; cf. MAEs of 0.25 and 0.21 °C for dwellings A and B at $h = 2$ (Table 2). Forecasts by Mustafaraj et al. using a NARX model [28] were considerably better, with MAEs of 0.23–0.27 °C at 2 h ahead, which is very close to the MAEs achieved with the GAMs for $h = 2$, 0.24 °C and 0.18 °C for dwellings A and B ($h = 2$, Table 2). Ferracuti et al. [30] produced 3 h summertime temperature forecasts with RMSEs of 0.33 °C and 0.36 °C for ARX and NARX models respectively; which are close to the values of 0.48 °C and 0.35 °C for the ARX model, and 0.45 °C and 0.32 °C for the GAM, for dwellings A and B respectively ($h = 3$, Table 2). However, these results must be viewed in relation to the validation data used to test the models. Notably, the forecasts performed here took place in free-running dwellings with considerably higher indoor temperature variability than that observed in the studies by Mustafaraj et al. [28] and Ferracuti et al. [30].

Considering the stochastic and embedded nature of residential window operation, the newly developed *logistic* GAM performed with good discrimination ability. For both dwellings, the TPR was encouragingly high, 94.1% and 68.3% for dwellings A and B respectively, but this was achieved at the expense of a high FPR, 88.7% and 41.1% for dwellings A and B respectively (Table 3). The high FPR for dwelling A may be partially attributable to the considerably unbalanced testing period (i.e. substantially more P than N). Overall, the TPRs were higher than the corresponding FPRs and the models showed an adequate ACCs, 75.4% and 64.0% for dwellings A and B respectively (Table 4). In cases where the TPR is low or when the discrimination is poor, relying on an auxiliary stochastic model to supply the WO state to the main model is unlikely to be reliable and could potentially decrease the forecasting accuracy of the main indoor temperature model. In addition, at longer forecasting horizons, the discrimination ability of the model would be hampered by the additional uncertainty in the estimated indoor temperatures of the main forecasting model.

Integration of the actual, measured window opening states into the GAMs, rather than relying on the auxiliary logistic models, showed variable results (Table 4 cf. Table 2). At best, the models incorporating the known window state produced forecasts of equal accuracy (in dwelling A) but conversely (in dwelling B) the inclusion of the WO state significantly reduced the accuracy of the model. There are two reasons why the additional information supplied to the forecasting models is incapable of improving the predictions. Firstly, the coefficients that were attributed to the WO state in the forecasting models were relatively small in this study compared to other predictor variables (Fig. 2) and according to Binder and Tutz [50] in developing GAMs, it is advisable to include only those predictor variables that are truly influential. Secondly, the actual cooling effect provided by an open window cannot be reduced to a constant value, as it is considered by the predictive models. In reality, the actual effect of the WO on the indoor temperatures depends on the temperature difference between the indoor and outdoor environments, which is at a maximum overnight but can be small or even negative during the central hours of the day. This is especially true during heatwaves, when indoor temperatures may even exceed the outdoor temperatures during the late afternoon and evening. Lastly, the operation of windows will directly affect indoor temperatures, with its 'effect' partially embedded in the indoor temperatures that are incorporated as model inputs (i.e. the autoregressive terms) and which are seen to have the highest influence on the model predictions (Fig. 2). Therefore, it can be concluded that even with exact knowledge of the WO state, its inclusion into the main forecasting model for indoor temperatures is not (in isolation) capable of improving the forecasting accuracy. As a consequence, the WO state should not be included in the forecasting model due to its low influence on the dependent variable that could, at times, also negatively affect the overall predictive accuracy.

6. Conclusions

The ability of linear ARX models and semi-parametric GAMs to forecast indoor temperatures over the intense and long-lasting UK heatwave of 2013 was investigated using hourly data from two bedrooms, in two houses, located near to the town of Loughborough in the UK Midlands. A backward stepwise regression based on minimisation of the AIC (for ARX models) and MAE (for GAMs) was adopted for the model selection process. Recursive multi-step-ahead forecasts were produced by both the models using a rolling forecasting origin for the entire duration of the heatwave. Forecasts were made for time horizons of 1–6, 12, 24, 36, 48, 60 and 72 h ahead, including the 95% prediction intervals, in order to provide a credible interval for the forecasted temperatures. The accuracy of the predictions was evaluated using the MBE as a measure of the bias, and MAE and RMSE to assess out-of-sample accuracy. Modified DM tests were adopted to assess whether differences in the accuracies of the GAMs and ARX models, and the inclusion of the actual window opening state, were significant at the 90% probability level.

Comparisons between the ARX models and GAMs showed that although the GAMs were capable of slightly improved forecasting accuracy, the improvements were only statistically significant up to 3–6 h ahead. For longer forecasting horizons, ARX models provided an accuracy that was either equal to, or greater than the GAMs, with an MAE (up to 72 h ahead) that was typically below 1 °C for the entire heatwave. Considering the potential uncertainty associated with the non-linear GAMs relationships when exposed to higher temperature ranges for the first time, the subsequent risk of instability at longer forecasting horizons, higher computational time requirements, lower accuracy at longer forecasting horizons and marginal improvement of the predictive accuracy at shorter horizons; the adoption of such models appears unjustified for forecasting elevated internal temperatures in free-running buildings.

Logistic GAMs were shown to be capable of adequately predicting whether or not a window was open in situations where the windows were operated with a discernible frequency. The TPR was consistently higher than the FPR, with an adequate ACC, of 64.0–75.4%. However, the *logistic* window opening models could not account for the sudden unpredictable changes in occupant behaviour occurring during extreme events. In situations where occupants

[5] Linear models assume a linear relationship between the independent and dependent variables, which remains constant throughout the whole range of the predictor variables. In this application, the linear model structure prevents the generation of local disruptions that might otherwise affect semi-parametric models (containing initially highly uncertain non-linear relationships) when first approaching a hotter period (which is the case for models which have not been trained on historical data from past heatwaves).

might open the windows for reasons unaccounted for by the model (or not open them at all), the reliability of these models to provide accurate predictions is questionable and should, therefore, be considered on a case-by-case basis.

In relation to the prediction of indoor temperatures, forecasts based upon exact knowledge of the window states did not improve the forecasting accuracy of either the ARX or GAM models and in some cases had a negative effect on the forecasting accuracies.

Overall this work suggests that more complex non-linear models do not necessarily produce better forecasts and are not well indicated for predictions at long forecasting horizons. Particular attention should be given to the use of GAMs when there is a likelihood of predicting out-of-range which could render the model unstable. By definition, there will always be limited data at the lower and upper ranges of the independent variables, which engenders increasing uncertainty when forecasting beyond the ranges for which the models were originally trained, with errors that are likely to amplify at longer forecasting horizons.

Future work will involve longitudinal testing of the prototyped forecasting models using larger datasets to quantify the reliability of predictions for different room, dwelling and household configurations across a wide range of geo-social contexts.

Declarations of interest

None

Acknowledgements

This research was made possible by the Engineering and Physical Sciences Research Council (EPSRC) support for the *'London-Loughborough (LoLo) Centre for Doctoral Training in Energy Demand'* (grant EP/L01517X/1). Monitored data, indispensable to this study, was made available by the open access LEEDR project home energy dataset [18], which was funded by the EPSRC: '*LEEDR: Low Effort Energy Demand Reduction*', (grant EP/I000267/1).

References

[1] National House Building Council (NHBC) Foundation, Overheating in new homes: a review of the evidence, 2012. ISBN: 978-1-84806-306-8. [Online]. [Accessed 27 March 2017] Available at: http://www.zerocarbonhub.org/sites/default/files/resources/reports/Overheating_in_New_Homes-A_review_of_the_evidence_NF46.pdf.

[2] Zero Carbon Hub (ZCH), Overheating in homes: the big picture, 2016. [Online]. [Accessed 8 February 2017] Available at: http://www.zerocarbonhub.org/sites/default/files/resources/reports/ZCH-OverheatingInHomes-TheBigPicture-01.1.pdf.

[3] K.J. Lomas, S.M. Porritt, Overheating in buildings: lessons from research, Build. Res. Inf. 45 (1-2) (2017) 1–18, doi:10.1080/09613218.2017.1256136.

[4] G.A. Meehl, C. Tebaldi, More Intense, More Frequent, and Longer Lasting Heat Waves in the 21st Century, Science 305 (2004) 994–997, doi:10.1126/science.1098704.

[5] Intergovernmental Panel on Climate Change (IPCC), Climate Change 2014: synthesis report, 2014. ISBN: 978-92-9169-143-2. [Online]. [Accessed 12 April 2018]. Available at: https://www.ipcc.ch/report/ar5/syr/.

[6] A. De Bono, G. Giuliani, S. Kluster, P. Peduzzi, Impacts of summer 2003 heat wave in Europe, United Nations Environment Programme, Environ. Alert Bull. 2 (2004) 1–4, doi:10.1017/S0147547903000218. [Online]. [Accessed 21 March 2017]. Available at: https://archive-ouverte.unige.ch/unige:32255.

[7] P.A. Stott, D.A. Stone, M.R. Allen, Human contribution to the European heatwave of 2003, Nature 432 (2004) 610–614 doi:, doi:10.1038/nature03089.

[8] R.S. McLeod, M. Swainson, Chronic overheating in low carbon urban developments in a temperate climate, Renew Sust. Energ Rev. 74 (2017) 201–220, doi:10.1016/j.rser.2016.09.106.

[9] S.M. Porritt, P.C. Cropper, L. Shao, C.I. Goodier, Ranking of interventions to reduce dwelling overheating during heat waves, Energy Build 55 (2012) 16–27, doi:10.1016/j.enbuild.2012.01.043.

[10] A. Mavrogianni, A. Pathan, E. Oikonomou, P. Biddulph, M. Davies, A. Mavrogianni, A. Pathan, E. Oikonomou, P. Biddulph, Inhabitant actions and summer overheating risk in London dwellings, Build. Res. Inf. 45 (1-2) (2016) 119–142, doi:10.1080/09613218.2016.1208431.

[11] P. Symonds, J. Taylor, A. Mavrogianni, M. Davies, C. Shrubsole, I. Hamilton, Z. Chalabi, Overheating in English dwellings: comparing modelled and monitored large-scale datasets, Build. Res. Inf. 45 (1-2) (2017) 195–208, doi:10.1080/09613218.2016.1224675.

[12] R. McLeod, C. Hopfe, A. Kwan, An investigation into future performance and overheating risks in Passivhaus dwellings, Build. Environ. 70 (2013) 189–209, doi:10.1016/j.buildenv.2013.08.024.

[13] P. De Wilde, The gap between predicted and measured energy performance of buildings: a framework for investigation, Autom. Constr. 41 (2014) 40–49, doi:10.1016/j.autcon.2014.02.009.

[14] E. Mantesi, C.J. Hopfe, M.J. Cook, J. Glass, P. Strachan, The modelling gap: quantifying the discrepancy in the representation of thermal mass in building simulation, Build. Environ. 131 (2017) 74–98, doi:10.1016/j.buildenv.2017.12.017.

[15] K.J. Lomas, T. Kane, Summertime temperatures and thermal comfort in UK homes, Build. Res. Inf. 41 (3) (2013) 259–280, doi:10.1080/09613218.2013.757886.

[16] A. Beizaee, K.J. Lomas, S.K. Firth, National survey of summertime temperatures and overheating risk in English homes, Build. Environ. 65 (2013) 1–17, doi:10.1016/j.buildenv.2013.03.011.

[17] S. Firth, T. Kane, V. Dimitriou, T. Hassan, F. Fouchal, M. Coleman, L. Webb, REFIT Smart Home dataset, 2016, doi:10.17028/rd.lboro.2070091. [Online]. [Accessed 11 September 2017]. Available at: https://figshare.com/articles/REFIT_Smart_Home_dataset/2070091.

[18] R. Buswell, L. Webb, P. Cosar-Jorda, D. Marini, S. Brownlee, M. Thomson, S.-H. Yang, R. Kalawsky, LEEDR project home energy dataset, 2018, doi:10.17028/rd.lboro.6176450. [Online]. [Accessed 19 July 2018]. Available at: https://figshare.com/articles/LEEDR_project_home_energy_dataset/6176450.

[19] A. Foucquier, S. Robert, F. Suard, L. Stéphan, A. Jay, State of the art in building modelling and energy performances prediction: a review, Renew. Sustain. Energy Rev. 23 (2013) 272–288, doi:10.1016/j.rser.2013.03.004.

[20] F. Amara, K. Agbossou, A. Cardenas, Y. Dubé, S. Kelouwani, Comparison and Simulation of Building Thermal Models for Effective Energy Management, Smart Grid Renew. Energy 6 (2015) 95–112, doi:10.4236/sgre.2015.64009.

[21] M. Gustin, R.S. McLeod, K.J. Lomas, Forecasting indoor temperatures during heatwaves using time series models, Build. Environ. 143 (2018) 727–739, doi:10.1016/j.buildenv.2018.07.045.

[22] National Health Service (NHS), Heatwave: how to cope in hot weather, 2016. [Online]. [Accessed 3 December 2018]. Available at: https://www.nhs.uk/live-well/healthy-body/heatwave-how-tocope-in-hot-weather/.

[23] G.Y. Yun, K. Steemers, Time-dependent occupant behaviour models of window control in summer, Build. Environ. 43 (2008) 1471–1482, doi:10.1016/j.buildenv.2007.08.001.

[24] M. Schweiker, F. Haldi, M. Shukuya, D. Robinson, Verification of stochastic models of window opening behaviour for residential buildings, J. Build. Perform. Simul. 5 (1) (2012) 55–74, doi:10.1080/19401493.2011.567422.

[25] V. Fabi, R.K. Andersen, S. Corgnati, Verification of stochastic behavioural models of occupants' interactions with windows in residential buildings, Build. Environ. 94 (2015) 371–383, doi:10.1016/j.buildenv.2015.08.016.

[26] F. Haldi, D. Robinson, Interactions with window openings by office occupants, Build. Environ. 44 (2009) 2378–2395, doi:10.1016/j.buildenv.2009.03.025.

[27] A. Mechaqrane, M. Zouak, A comparison of linear and neural network ARX models applied to a prediction of the indoor temperature of a building, Neural Comput. Appl. 13 (2004) 32–37, doi:10.1007/s00521-004-0401-8.

[28] G. Mustafaraj, G. Lowry, J. Chen, Prediction of room temperature and relative humidity by autoregressive linear and nonlinear neural network models for an open office, Energy Build 43 (2011) 1452–1460, doi:10.1016/j.enbuild.2011.02.007.

[29] B. Thomas, M. Soleimani-Mohseni, Artificial neural network models for indoor temperature prediction: investigations in two buildings, Neural Comput. Appl. 16 (2007) 81–89, doi:10.1007/s00521-006-0047-9.

[30] F. Ferracuti, A. Fonti, L. Ciabattoni, S. Pizzuti, A. Arteconi, L. Helsen, G. Comodi, Data-driven models for short-term thermal behaviour prediction in real buildings, Appl. Energy 204 (2017) 1375–1387, doi:10.1016/j.apenergy.2017.05.015.

[31] Q. Zhang, S.-C. Zhu, Visual Interpretability for Deep Learning: a Survey, Front. Inf. Technol. Electron. Eng. 19 (1) (2018) 27–39, doi:10.1631/FITEE.1700808.

[32] K. Larsen, GAM: the Predictive Modeling Silver Bullet, MultiThreaded (2015). [Online]. [Accessed 17 June 2018]. Available at: http://multithreaded.stitchfix.com/blog/2015/07/30/gam/.

[33] S. Fan, R.J. Hyndman, Short-term load forecasting based on a semi-parametric additive model, IEEE Trans. Power Syst. 27 (1) (2012) 134–141, doi:10.1109/TPWRS.2011.2162082.

[34] Met Office, July 2013 heat wave, 2013. [Online]. [Accessed 3 December 2018]. Available at: https://www.metoffice.gov.uk/learning/learn-about-the-weather/weather-phenomena/casestudies/heat-wave-july2013.

[35] Met Office, UK and regional series, 2019. [Online]. [Accessed 3 December 2018]. Available at: https://www.metoffice.gov.uk/climate/uk/summaries/datasets#yearOrdered.

[36] R.J. Hyndman, S. Fan, Density Forecasting for Long-Term Peak Electricity Demand, IEEE Trans. Power Syst. 25 (2010) 1142–1153, doi:10.1109/TPWRS.2009.2036017.

[37] R Core Team, R: A language and environment for statistical computing, 2017. [Online]. [Accessed 7 May 2017]. Available at: https://www.r-project.org/.

[38] S.N. Wood, Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models, J. R. Stat. Soc. 73 (1) (2011) 3–36, doi:10.1111/j.1467-9868.2010.00749.x.

[39] S.N. Wood, 'mgcv' package for R (version 1.8-28): Mixed GAM Computation Vehicle with Automatic Smoothness Estimation, 2019. [Online]. [Accessed 14 May 2018] Available at: https://cran.rproject.org/web/packages/mgcv/mgcv.pdf.

[40] R.J. Hyndman, G. Athanasopoulos, Forecasting: principles and practice, OTexts, 2nd ed., 2018 ISBN: 978-0987507112.
[41] S.N. Wood, 'mgcv' package: Generalized Additive Model Selection, 2018. [Online]. [Accessed 20 June 2018]. Available at: https://stat.ethz.ch/R-manual/Rdevel/library/mgcv/html/gam.selection.html.
[42] M. Gustin, A. Oraiopoulos, R.S. Mcleod, K.J. Lomas, A new empirical model incorporating spatial interpolation of meteorological data for the prediction of overheating risks in UK dwellings, in: Proceedings of PLEA 2017: the 33rd Passive and Low Energy Architecture International Conference, 3, 2017, pp. 3786–3793. Edinburgh (UK), Available at: https://plea2017.net/
[43] Met Office, 10 day weather forecast, 2016. [Online]. [Accessed 27 November 2017]. Available at: https://www.metoffice.gov.uk/guide/weather/10-day-forecast.
[44] F.X. Diebold, R.S. Mariano, Comparing predictive accuracy, J. Bus. Econ. Stat. 20 (1) (2002) 143–144, doi:10.1198/073500102753410444.
[45] F.X. Diebold, F.X. Diebold, Comparing predictive accuracy, twenty years later: a personal perspective on the use and abuse of Diebold-Mariano tests, J. Bus. Econ. Stat. 33 (1) (2015) 1, doi:10.1080/07350015.2014.983236.
[46] D. Harvey, S. Leybourne, P. Newbold, Testing the equality of prediction mean squared errors, Int. J. Forecast 13 (1997) 281–291, doi:10.1016/S0169-2070(96)00719-4.
[47] R.J. Hyndman, G. Athanasopoulos, C. Bergmeir, G. Caceres, L. Chhay, M. O'Hara-Wild, F. Petropoulos, S. Razbash, E. Wang, F. Yasmeen, 'forecast' package for R (version 8.5): Forecasting functions for time series and linear models, 2019. [Online]. [Accessed 3 October 2017]. Available at: http://pkg.robjhyndman.com/forecast.
[48] S.B. Taieb, Machine Learning Strategies for multi-step-ahead Time Series Forecasting, *PhD Thesis*, Université Libre de Bruxelles, Belgium, 2014. [Online]. [Accessed 18 February 2019]. Available at: https://souhaib-bentaieb.com/pdf/2014_phd.pdf.
[49] T. Teräsvirta, D. Van Dijk, M.C. Medeiros, Linear models, smooth transition autoregressions, and neural networks for forecasting macroeconomic time series: a re-examination, Int. J. Forecast. 21 (2005) 755–774, doi:10.1016/j.ijforecast.2005.04.010.
[50] H. Binder, G. Tutz, A comparison of methods for the fitting of generalized additive models, Stat. Comput. 18 (1) (2008) 87–99, doi:10.1007/s11222-007-9040-0.
[51] Met Office, Heatwave, 2018. [Online]. [Accessed 16 July 2018]. Available at: https://www.metoffice.gov.uk/learning/temperature/heatwave.
[52] World Meteorological Organization (WMO), Guidelines on the definition and monitoring of extreme weather and climate events, 2016. [Online]. [Accessed 16 April 2018]. Available at: https://www.wmo.int/pages/prog/wcp/ccl/opace/opace2/documents/DraftversionoftheGuidelinesontheDefinitionandMonitoringofExtremeWeatherandClimateEvents.pdf.