Estimating dynamic solar gains from on-site measured data: an ARX modelling approach

Xiang Zhang*, Dirk Saelens, Staf Roels
KU Leuven, Department of Civil Engineering, Building Physics and Sustainable Design Section, Kasteelpark Arenberg 40 Bus 2447, 3001 Heverlee, Belgium

Abstract. On site measured data in combination with statistical methods is used more and more to assess the actual performance of a building and to develop simplified models that can be used in model predictive control, fault detection and optimization of energy grids. Most of the methods are based on a simplified heat balance of the building. For this heat balance, solar gains, referring to the part of energy supplied by the sun, are a vital factor. Gauging solar gains and their time dependency in practice is, however, challenging. Most models assume the solar aperture of the building as an invariable property, although it is highly dependent on the sun position during the time of the day and year. In this study a statistical modeling method, based on in-situ measurement data, is developed to estimate the time-varying solar gains more precisely. The method integrates basis splines (B-splines) into an ARX model (Auto-regressive with eXogenous input). Verified by BES-simulation outcomes, it is demonstrated that this B-splines integrated ARX modelling approach can reflect the key information of the dynamic features of the solar gains to a large extent. Most importantly, only a limited-size of in-situ measurement data of high-frequency is required in this method, indicating its good efficiency in dynamic solar gain estimation.

Keyword. data-driven method; black-box model; dynamic solar aperture (*gA*); B-splines; autoregressive with exogenous terms (ARX) model

1. Introduction

1.1 Problem statement

Collecting and analyzing empirical data of in use buildings is becoming an important tool in assessing and optimizing our building energy use. On the one hand, data analysis on large scale datasets, also referred to as building energy epidemiology [1–3], investigates the energy demand within a population to improve our understanding of causes of variations within or between building stocks. On the other hand, on site measured data of a specific building is used in combination with statistical methods to assess e.g. the actual achieved performance of the insulation quality of the building fabric [4–7] or to construct more reliable simplified models that can be used in model predictive control [8,9], fault detection and diagnosis [10] and optimization of energy smart grids [11]. The latter methods typically are based on the indoor heat balance of the building, determined by heat gains and losses from various sources, such as heat supplied by the heating system and internal appliances, solar gains, and losses or gains through the building fabric often accompanied by air leakage. Among the gains, the part supplied by the sun through the glazed building fabric, generally referred to as solar gain, is considered very important for the overall indoor heat balance, but often hard to measure in practice. As an alternative, the BES (building energy simulation) model-based approach [12–16] is widely used to estimate the amount and dynamics of solar gains for a particular building. BES-models are physic-based or so-called white-box models. The BES-approach can offer reliable results with the adequate required information (i.e. input data), such as properties, size, and orientation of glazed building fabric and information of

* Corresponding author, KU Leuven, Department of Civil Engineering, Building Physics and Sustainable Design.
Email: jason.zhang.kul@outlook.com; jason.zhang@kuleuven.be.

shading obstacles nearby (e.g. size and location), etc. Simulation programs such as DOE-2.1E, Energy-Plus, ESP-r, HELIOS, IDA-ICE, STAR-CCM+ (CFD-based), Modelica (IDEAS) are used for this particular purpose [12,14–16]. However, the needed information (data) is often not available or, at least, hard to obtain, especially for existing buildings. A partial lack of required data could lead to certain simulation inaccuracies or even an invalid outcome. Furthermore, the BES-model involves plenty of complex differential equations and assumptions, demanding significant domain expertise and computational power, cost, and time, which sets barriers for efficient applications [17,18]. Therefore, in data-driven approaches, the solar gains are typically estimated directly from the on-site collected data of existing buildings. As systematically sorted out in the International Energy Agency (IEA) Annex 58 and 71 project [4,6], several statistical modeling tools are available, ranging from static models (e.g. simple linear regression) to dynamic ones (e.g. grey-box and auto-regression models). In those models, the total solar gain as a function of time (Sg_t) is commonly estimated by Eq. (1):

$$Sg_t = Sg_{dir_t} + Sg_{dif_t} = gA * GHI_t \quad (1)$$

with Sg_t the total solar gain and GHI_t the global horizontal irradiance (W/m²) at time *t*. The solar aperture (gA) is a 'fabricated' area generally assumed as a constant value, in m², which will be elaborated on below. For most buildings, solar gain is a significant contributor to buildings' indoor thermal dynamics [19,20] and Eq. (1) is widely used to estimate or predict the building energy performance. For example, Eq. (1) embedded in linear regression methods [21–23], ARX models (Auto-regressive with eXogenous input) [21–24], and grey-box models are used for determining building energy performance [22,25,26]. Meanwhile, Eq. (1) integrated grey-box [8,27,28] and ARX [28] models are applied in model predictive control (MPC) for optimizing building energy. In addition, Hollick et al. [19] proposed a grey-box-based Bayesian method to account for solar gains. Other relevant studies based on Eq. (1) can be found in [6,7,29–32].

However, the above-mentioned studies [6–8,19,21–24,26–32] all rely on a constant solar aperture (gA) integrated in Eq. (1), which is not in line with its dynamic nature, which is highly influenced by the sun's movement. This discrepancy could lead to considerable estimation uncertainties or prediction errors, especially in building cases with unevenly distributed windows. Let us use a simple case study to illustrate the problem. The gA-value, or solar aperture (also called the solar gain factor, solar transmittance, or effective window area, etc.), is defined as 'the equivalent area of a perfectly transparent surface that transmits the same amount of solar energy as the actual windows of the space' in a specific building [33]. Thus, without windows open to the Sun, the gA-value should theoretically be zero. Hence, for a particular east-west orientated dwelling with windows only in the front and back facade, the dynamic gA_t diagram revealing the key dynamic features would look as illustrated in Fig. 1. Assuming a typical daily global horizontal irradiance (*GHI*) plot (Fig. 2), the dynamic solar gains – real and constant gA based ones – correspond to the plot at the right side of Fig. 2. Interestingly, the constant gA assumption might offer an acceptable estimation of the total amount of daily solar gain (in J), but the dynamic details of Sg_t, in W, based on a constant gA-estimation could be quite wrong. For this specific case, the solar gain (Sg_t) peak is predicted around noon, when no solar gain exists in reality (Fig. 2). It is noteworthy that such row houses are popular in European cities. In other words, the problem unfolded by this case is not isolated.

Thus, some researchers have stated that, to get e.g. a valid model for predictive control, pre-simulated (BES-model based) solar gain dynamic details [16] or accurate measurements of dynamic solar gain [27] are strongly required. The invariant solar aperture (gA) assumption might lead to unqualified statistical models since it estimates the gA parameter with a large uncertainty and without statistical significance [34]. Therefore, it is argued that a precise

dynamic gA_t curve (i.e. a curve reflecting the dynamic gA_t value over time such as Fig. 2) can reduce the prediction uncertainties to a large extent. However, as elaborated at the beginning of this section, neither BES- models nor current constant solar aperture (gA) assumption-based data-driven methods are efficient or even feasible in obtaining elaborated dynamic solar gain information. Therefore, a more reliable technique is needed, and this study works on proposing a simple but efficient data-driven method to estimate the dynamics of Sg_t precisely (e.g., the blue curve in Fig. 2). The estimation of Sg_t with enhanced accuracy could reduce the predictive errors, such as in MPC. As stated above, compared to the existing techniques of estimating dynamic solar gain, this method is expected to estimate the key trends dynamics of Sg_t accurately but efficiently, such as limited data demanding and low expertise requirement, which is regarded as the key novelty of this study.

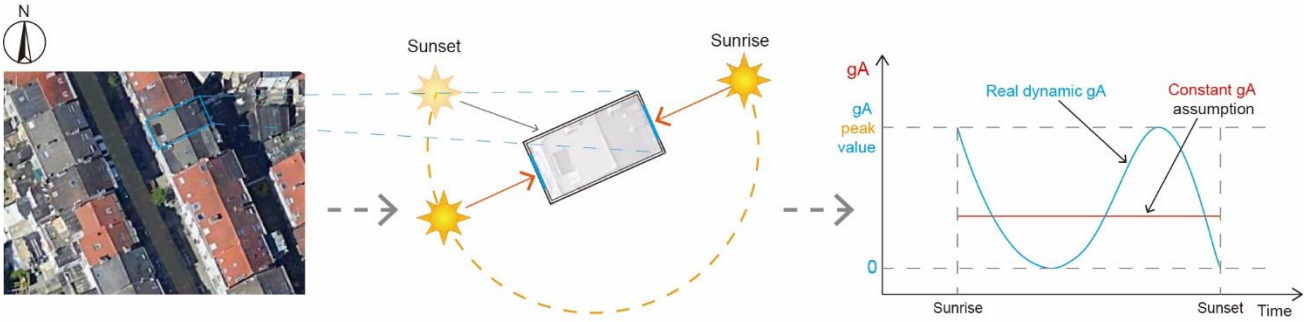


Fig. 1 The qualitative analysis on dynamic solar aperture (gA) in a particular building case.

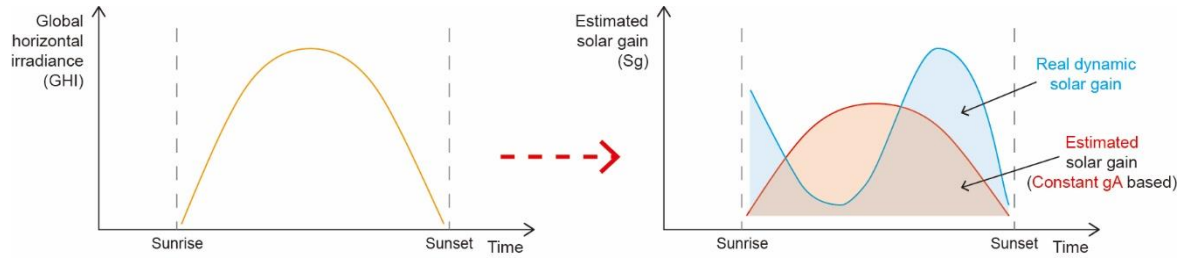


Fig. 2 The qualitative analysis on dynamic solar gain estimation gap, based on constant gA assumption.

1.2 Refining solar irradiances, gains, and apertures

The global horizontal irradiance (*GHI*) is composed by direct normal irradiance (*DNI*) and diffuse horizontal irradiance (*DHI*), formulated by Eq. (2), where θ_z specifies the zenith angle of the sun.

$$GHI_t = DNI_t * \cos\theta_z + DHI_t \tag{2}$$

$$gA_{dni_t} = \frac{Sg_{dir_t}}{DNI_t} \tag{3}$$

Thus, the total solar gain at time *t*, denoted as Sg_t in Eq. (1), similarly consist of direct (Sg_{dir_t}) and diffuse (Sg_{dif_t}) solar gains respectively, as shown in Eq. (1). In section 1.1, the importance of a good estimation of the gA_t-curve was clarified. To get the dynamic gA_t-curve for a particular building, Rasmussen et al. [33] proposed a grey-box modelling based method. On the basis of [33], Zhang et al. [35] further elaborated this method towards an estimation method for direct solar heat gains Sg_{dir_t} and to estimate the time-varying gA_{dni_t} curve. The gA_{dni_t} is a particular solar aperture customized to *DNI* (direct normal radiation) and can be understood as a tailor-made 'equivalent area', in analogy to gA, as defined in Eq. (3). It is noted that, physically, this 'equivalent area' is rotating with sun movement to ensure is it always pedicular to the solar beam (i.e., DNI). The main reason for refining gA_t (Sg_t) into gA_{dni_t}

(Sg_{dir_t}) is that the Sg_{dif_t}, supplied by diffuse horizontal irradiance (DHI_t), is hardly varying with time (sun position), and the key dynamic features of Sg_t are attributed to DNI_t [35].

The two aforesaid methods developed in [33,35] are based on in-situ measured data and grey-box modelling. In-situ measurement data is gaining increased importance in recent years [23,36–38], since on-site measurement data incorporates the complexity of reality inherently, hence, not only including the necessary input information required for BES-simulations but also some phenomena that are ignored easily and hard to be accounted for in the BES-simulations. For example, unknown shading effects, such as foliage or stains on windows, which could influence solar gains considerably, will probably escape detection. Meanwhile, the aforesaid grey-box based methods [33,35] have demonstrated their capability in revealing the key quantity and dynamic information of Sg_t (or Sg_{dir_t}) and gA_t (or gA_{dni_t}), indicating the great potential of in-situ measurement data in this respect. However, a lumped effect is still observed in the grey-box model-based methods, generally leading to an over-estimation of Sg_t (or Sg_{dir_t}), which could partially be attributed to the restriction and limitation of the simplified physical assumptions underlying grey-box modelling, making it hard to match the complexity of reality exactly [35]. Furthermore, a rather high building physics and statistical expertise is needed in grey-box modelling (e.g. expertise to propose reasonable physical knowledge based assumptions), which could set a barrier for the effective application. Hence, in this regard, ARX models might have innate advantages of model flexibility without the need of physical knowledge-based assumptions. Drawback is their physical interpretation [6], since ARX is a kind of black-box model, contrary to grey-box models. For more information on white-box (e.g. BES-model), grey-box, and black-box (e.g. ARX model) modelling, we refer to [35,39]. This study explores the ARX model-based method to estimate dynamic Sg_{dir_t} and gA_{dni_t} curves, and will compare the estimation outcomes of dynamic gA_{dni_t} of three methods (i.e. ARX, BES, and grey-box models), to understand the characteristics of the three methods in dynamic Sg_{dir_t} estimation. As mentioned, since Sg_{dir_t} and gA_{dni_t} have a higher time dependency and are more connected to the sun position than Sg_t and gA_t, this study will focus on Sg_{dir_t} and gA_{dni_t}.

2. ARX model based *gA* estimation

2.1 A brief description of ARX modelling

Autoregressive with exogenous terms (ARX) models [40] belong to the discrete-time transfer function models, which are fundamentally linear models [41,42]. The uniqueness of the ARX model is that the past observations of independent (input) and dependent (output) variables are involved to describe the dynamics of the system [23]. In practice, the time-shifted data is represented as past observation. ARX models generally neglect any prior physical knowledge and fully base their parametric estimations on statistical correlations embedded in the data [37], which is the reason that they are classified as black-box models [43]. As illustrated in the name, AR refers to an AutoRegressive of output variable itself, while X represents eXternal inputs are modelled, indicating that ARX models express output dynamics based on both previous system output and (extra) system inputs [5]. ARX models generally show as Eq. (4-7), linking output at time *t* to a chain of past outputs and inputs [44].

$$A(q^{-1})y(t) = B(q^{-1})u(t) + e(t) \tag{4}$$

$$q^{-1}y(t) = y(t-1) \tag{5}$$

$$A(q^{-1}) = 1 + a_1q^{-1} + a_2q^{-2} + \cdots + a_nq^{-n} \tag{6}$$

$$B(q^{-1}) = b_0 + b_1q^{-1} + b_2q^{-2} + \cdots + b_nq^{-n} \tag{7}$$

Where $y(t)$, $u(t)$, $e(t)$ and, q^{-1} represent output, input variable, equation error, and backward shift operator. $A(q^{-1})$ and $B(q^{-1})$ are polynomials in q^{-1}. Note that Eq. (4-7) is the simplest form of the ARX model with one external input, and it can be extended to incorporate more input data, such as the case shown in Eq. (8). Meanwhile, the order (n) in Eq. (6-7) signifies how many data points from the past (previous observations) for input and output are modeled [29]. As stated, ARX is a specific type of black-box model, indicating the coefficients estimated by ARX models generally don't have direct physical meaning or interpretation. Nevertheless, Jiménez et al. [45] and Naveros et al. [46] have demonstrated the link between R-C (resistance-capacitance) networks and ARX models [43]. Related key information is sorted out in IEA EBC Annex 58 project [6].

2.2 Linking ARX model and RC networks

Treating a building as one thermal zone with only heating input and solar gains as heat sources to the indoor environment (i.e. indoor air mass), the dynamics of indoor temperature (T_i) at time *t* can be simplified and represented by an ARX model [45], where T_e, Ph, GHI and e are exterior temperature (°C), heating input (W), global horizontal irradiance (W/m^2), and residuals. By introducing $A(q^{-1})$, $B_1(q^{-1})$, $B_2(q^{-1})$, and $B_3(q^{-1})$, this particular ARX model can be written as Eq. (8) [6,45]. Note that $A(q^{-1})$ and $B_1(q^{-1})$ are shown in Eq. (6-7). $B_2(q^{-1})$, and $B_3(q^{-1})$ is an analog to $B_1(q^{-1})$ with similar model structure.

$$A(q^{-1})T_{i,t} = B_1(q^{-1})T_{e,t} + B_2(q^{-1})Ph_t + B_3(q^{-1})GHI_t + e_t \quad (8)$$

$$B_2(1)Ph_t = A(1)T_{i,t} - B_1(1)T_{e,t} - B_3(1)GHI_t - e_t \quad (9)$$

$$Ph = HLC * (Ti - Te) - gA * GHI + e \quad (10)$$

$$HLC = \frac{A(1)}{B_2(1)} = \frac{B_1(1)}{B_2(1)} \quad (11)$$

$$gA = \frac{B_3(1)}{B_2(1)} \quad (12)$$

Additionally, if we assume the heat loss coefficient (*HLC*) and solar aperture (*gA*) as steady properties of the building, we can estimate them in a steady-state form by putting $q^{-1} = 1$. Thus, Eq. (8) can be rewritten as Eq. (9), which has to meet the indoor heat balance [47] shown as Eq. (10). The overall heat loss coefficient (*HLC*; W/K), is an indicator quantifying both the performance of thermal insulation and airtightness of the building envelope [38]. It is defined as the heat flow rate through the building envelope per unit temperature difference (W/K) between indoor and outdoor [48]. The reader is referred to [44] for more details. Thus, based on Eq. (9-10), the heat loss coefficient and constant solar aperture can be calculated as Eq. (11-12) [45].

However, Eq. (8) will also incorporate solar gains via opaque building envelope parts (marked as Sg_{en_t} in Fig. 3) into $B_3(q^{-1}) * GHI_t$ (so-called 'lumped effect' in [35]), which might weaken the physical interpretation of the estimated gA via Eq. (12). The physical process of solar radiation affecting the indoor temperature dynamics is relatively complicated, attributed to different types of irradiances (e.g. direct or diffuse) and influencing paths (i.e. via glazed or opaque envelope). Elaborated details can, amongst others, be found in [49,50]. Among the various types of solar energy gains, the three key parts are Sg_{dir_t}, Sg_{dif_t}, and Sg_{en_t}, visualized in Fig. 3. Sg_{dir_t} and Sg_{dif_t} are two types of solar gains entering only via glazing, contributed by direct normal irradiance (*DNI*) and diffuse horizontal irradiance (*DHI*) respectively. On the contrast, Sg_{en_t} represents solar gains solely through the opaque building envelope, attributed to global horizontal irradiance (*GHI*), which embeds both *DNI* and *DHI* naturally, as shown in Eq. (2). For elaborated details of the three main types of solar gains, the readers refer to [35]. As explained in section 1, Sg_{dir_t} and gA_{dni_t} are the key research interest of this study. Hence, Eq. (9) is further elaborated into Eq. (13), to reduce the potential 'lumped effect' and control the model complexity at the same time.

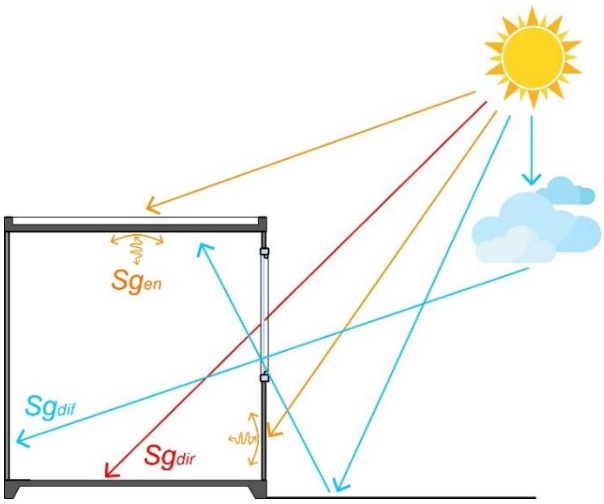


Fig. 3 Three key types of solar gains.

Thus, theoretically, $B_4(q^{-1}) * GHI_t$ comprises both Sg_{en_t} and Sg_{dif_t} in Fig. 3. And, assuming $A_{en} * GHI$ jointly represent both Sg_{en_t} and Sg_{dif_t}, Eq. (10) is rewritten as Eq. (14). Note, the mixed physical meaning of A_{en} in this equation, which accounts for solar gains from *DHI* through both glazed and opaque building fabric and solar gains from *DNI* penetrating the opaque envelope only. Hence, similar to Eq. (12), a constant gA_{dni} can be estimated via Eq. (15), based on Eq. (13-14). Another proof of *HLC* and constant gA_{dni} estimation is clarified in Appendix A. $B_4(q^{-1})$, and $B_5(q^{-1})$ is an analog to $B_1(q^{-1})$ with similar model configuration. Thus, the three main types of solar gain- Sg_{dir_t}, $Sg_{dif_t}$and Sg_{en_t}- are considered in the following ARX models, which will be explored in section 5.1.

$$B_2(1)Ph_t = A(1)T_{i,t} - B_1(1)T_{e,t} - B_4(1)DNI_t - B_5(1)GHI_t - e_t \quad (13)$$

$$Ph = HLC * (Ti - Te) - gA_{dni} * DNI - A_{en} * GHI + e \quad (14)$$

$$gA_{dni} = \frac{B_4(1)}{B_2(1)} \quad (15)$$

Section 5 will investigate whether Eq. (15) is also valid when estimating a time-varying gA_{dni_t}, by comparing its estimation with the outcomes offered by the Building Energy Simulation (BES) and grey-box models, studied in [35]. Thus, this study proposes two main research questions: 1) Is the ARX modelling technique-based Eq. (15) able to estimate gA_{dni_t} reliably? If so, 2) what are the performance and main characteristics of the three techniques (ARX-, BES-, grey-box models) in revealing gA_{dni_t} dynamics?

2.3 B-splines integrated ARX models

In general, there are several available solutions to model non-linear relationships, including polynomial regression, spline regression, and generalized additive model (GAM). Polynomial regression is a straightforward approach, which models dependent (output) variable as an nth degree polynomial in independent (input) variable. Generally, a higher degree of the polynomial is required for richer dynamics of the non-linear relationship, but a high degree polynomial is commonly unwanted. To overcome this drawback, spline regression divides the dataset into several portions firstly and fits a low degree polynomial on each portion subsequently. Meanwhile, those pieces of polynomials are restricted to be continuous, as well as the first and second derivative at the connection points (called knot), to offer an integral smooth curve (spline) for the full dataset. However, the fitting outcome might be strongly influenced by the selected number of portions and locations of connection points. Thus, tailored to the research aim in this study, there is a more effective way of modeling non-linear relationships by treating it as a sum of weighted pre-constructed basis splines (B-splines) [33], and the flexible 'smoother' in GAM is essentially based on this

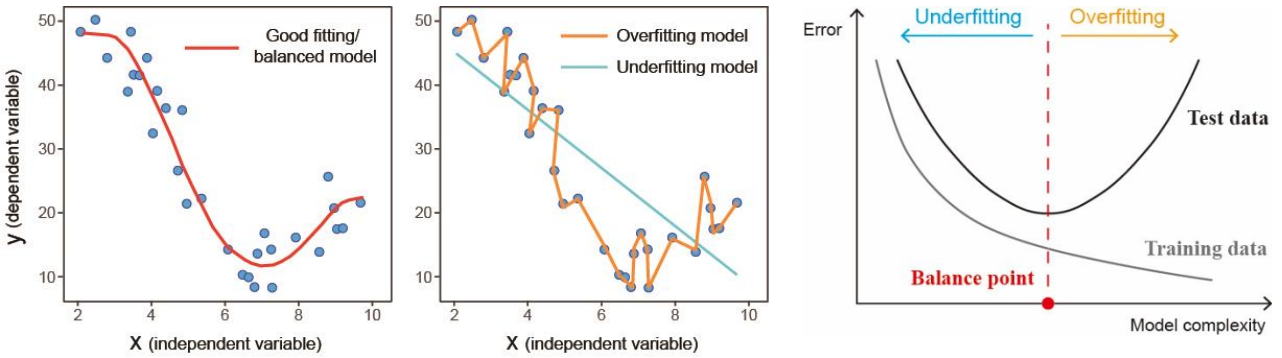
technique [51]. The advantage of this technique is mainly attributed to that a linear combination of B-spline of a given degree can equivalently replace any spline function of the specific degree, and only one unique combination for each spline function exists [52]. The compact definition of B-splines and elaborated practical information can be found in respectively [53] and [33,35]. A general formula of B-splines is written as Eq. (16) and examples are visualized in Fig. 16. The key benefit of B-splines based modeling of non-linear relationships includes 1) the 'arbitrary' parts of polynomial and spline regression (e.g. degree and number of portions) are pre-set and 'fixed' in the B-splines pre-construction stage; 2) transferring the complex non-linear modeling in polynomial and spline regressions to a simple linear regression modeling, as shown in Eq. (16) where $Bs_{i,m}(t)$ are pre-constructed B-splines with order (m) and number (q), and ϕ_i is the unknown regression coefficient (i.e. weighting or scaling factors) for the ith B-spline.

$$S(t) = \sum_{i=1}^{q} \phi_i Bs_{i,m}(t) \tag{16}$$

Thus, $S(t)$ is a sum of scaled q B-splines in total, with scaling factors ϕ_i for the ith basis spline $Bs_{i,m}(t)$. By varying the order (m) and number (q) of basis splines, $S(t)$ can represent any possible spline, describing any non-linear relationship between t and $S(t)$. The main property of the B-spline is that the sum of the entire B-spline series equals one at any value of t for the full range of interest [54]. In addition, m^{th} order of spline indicates it is composed of piecewise polynomial functions of degree *m-1*, and 'typically linear (m = 2), quadratic (m = 3), or cubic (m = 4) splines are used' [33]. Due to the merits of B-splines, they are used in this study to model the non-linear relationship between dynamic gA_{dni_t} versus time t. Alternatively stated, gA_{dni_t} will be represented by a sum of several weighted B-splines, as $S(t)$ in Eq. (16). Hence, the distribution of B-splines has to follow the physical interpretation of gA_{dni_t}, indicating that no B-splines construction is needed when no solar radiation is present - after sunset and before sunrise. To follow this criterion, in this study, the B-splines are customized according to the presence hours of DNI_t. Fig. 15 and Fig. 16 serve as a reference, and details will be elaborated in section 4.3.

2.4 Model selection and validation

As explained in section 2.1, the ARX model can incorporate more extra input data and historical observations to (endlessly) extend the model. In general, for a particular dataset (so-called training data), a better data fitting will be observed when the complexity of the ARX model increases, statistically represented as an increased loglikelihood-value (*LogL*). However, the risk of 'over-fitting' grows with increased model complexity. Any measurement dataset of a system contains both the key dynamic feature of the system , shown as the red curve in Fig. 4 (a) (i.e., the research interest), and local noise (e.g. measurement error). The local noise varies in different datasets measured in an identical system. When a model over-fits this dataset, the better model fitting performance is contributed by matching the local noise of a particular dataset instead of the main system characteristic of interest, as visualized by the orange curve in Fig. 4 (a),Thus, Fig. 4 (b) shows that an over-fitted model will enhance data-fitting in the training data by over-fitting local noise, to minimize the error (i.e., the discrepancy between model estimation and measurement), but will poorly extrapolate to new test datasets [55]. Hence, the optimal model is the one with its complexity located at the 'balance point'. A stepwise regression is generally used to identify the optimal ARX model, which mainly consists of two ways - forward selection and backward elimination [56]. In regression models (e.g. ARX model), the backward elimination method, which starts from a model with all *k* candidate regressors (i.e. inputs) of current and historic data, is easier in application [56,57]. The potential input without statistical significance in the regression model can be eliminated in the updated model with the rest *k-1* regressors, until no further regressor can be removed (i.e. the remained inputs all show statistical significance) [58].



(a) Good/over-/under-fitting performance of statistical models (b) Balance point of model complexity

Fig. 4 Identification of optimal model complexity.

In addition, the model updating also requires a considerable statistical improvement on data-fitting. However, to identify the suitable number of B-splines, the forward selection is more convenient [33,35], since, generally, only partial coefficients designated to B-splines are estimated with statistical significance. In other words, only a part of the B-splines plays a vital role in dynamic gA_{dni_t} curve construction, but it is necessary to keep also the part of B-splines without statistical significance to approach toward an 'accurate' estimation of the dynamic gA_{dni_t}-curve. The general application of forward selection is expanding the model progressively by adding one component enhancing the statistical fitting the most until no model expansion is needed. For details, the reader is referred to [33,35] and the application will be elaborated in Section 5.1. In this study, a considerable improvement is defined as a significant increase in likelihood-value (*LogL*) and remarkable decrease in *AIC* (Akaike information criterion) or *BIC* (Bayesian information criterion) values [59]. This strategy is applied to maintain the identification or prediction capacity of the model but prevent potential overfitting [60]. Note that, the process of B-spline integration (i.e. from the constant gA_{dni}-based model to B-splines integrated ones) will increase the number of parameters sharply but is crucial for the dynamic gA_{dni} curve identification. Thus, in the process of B-splines integration, the pre-set model selection rule might not be applied strictly (for instance a decrease in *BIC*-value is not always feasible). More discussion will be conducted in section 5.1. Moreover, the selected optimal model will be further tested to check whether it meets the assumption of 'white noise residuals', which indicates the ARX model has captured the key dynamics of the system. To test this assumption, auto-correlation function (ACF) and the cumulated periodogram (CP) techniques are used. The details are elaborated on in section 4.1.

3. B-splines integrated grey-box modeling

As stated in section 1, Zhang et al. [35] have proposed a B-spline integrated grey-box modelling technique to estimate a time-varying gA_{dni_t}-curve, and compared it with BES-simulation results. In short, a grey-box model is a combination of white- and black-box models, serving as an intermediate strategy [61]. In grey-box modelling, prior knowledge on the building and its physical behavior is used to construct system equations, while statistical estimation techniques (e.g. maximum likelihood) are used to estimate the physical parameters. Lumped RC (resistance-capacitance) models are generally applied when modelling the thermal behavior of buildings, commonly visualized as an electric circuit analogy [6,62]. For example, Fig. 5 visualized a second-order grey-box RC model [63], with corresponding system equations - Eq. (17-18) - and an additional observation equation, Eq (19). In Eq. (17-19) and Fig. 5, T_i, T_e, and T_a represent the indoor, envelope, and ambient temperature, R_{ie} and R_{ea} indicate the thermal resistances between the indoor air mass to the inner envelope and building fabric to the ambient air respectively, C_i and C_e stand for interior mass and envelope heat capacities and Ph and Sg describe the heat fluxes from heaters and the sun (i.e. solar gain), while $\{\omega_i,t\}$, and $\{\omega_e,t\}$ are independent standard Wiener processes, and $\sigma_{w,i}^2$ and $\sigma_{w,e}^2$ mean the incremental variances of the Wiener processes [64]. The reader is referred to [64] for more details.

$$dT_i = \frac{1}{R_{ie}C_i}(T_e - T_i)dt + \frac{1}{C_i}(Sg + Ph)dt + \sigma_{w,i}d\omega_i \quad (17)$$

$$dT_e = \frac{1}{R_{ie}C_e}(T_i - T_e)dt + \frac{1}{R_{ea}C_e}(T_a - T_e)dt + \sigma_{w,e}d\omega_e \quad (18)$$

$$Y_k = T_{ik} + e_k \quad (19)$$

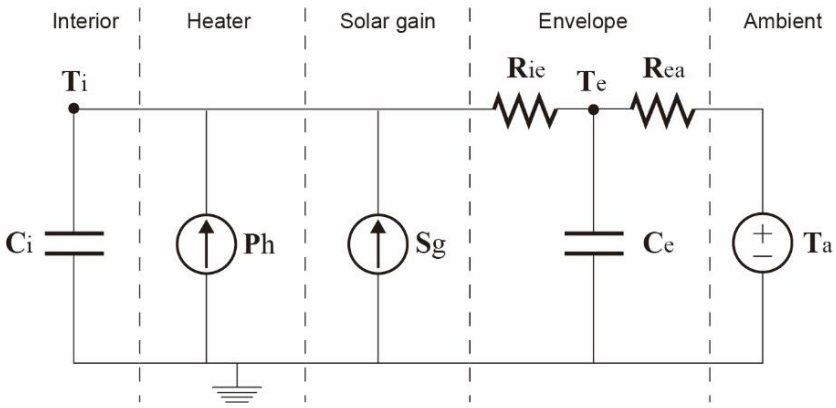


Fig. 5 A two-states model (*TiTe* model).

Note that, the *TiTe* model illustrated above is a simplified model, which didn't split the solar gain *Sg* into two parts (i.e. Sg_{dir} and Sg_{dif}). More works have been done in [35] to elaborate *Sg* as a sum of Sg_{dir} and Sg_{dif}. For the integration of B-splines into the grey-box model in [35], in brief, Eq. (3) is transferred to Eq. (20), where gA_{dni_t} is replaced by a sum of scaled B-splines $S(t)$, as stated in Eq. (16).

$$Sg_{dir_t} = gA_{dni_t}DNI_t = \sum_{i=1}^{q} \phi_i Bs_{i,m}(t)\, DNI_t \quad (20)$$

4. Study case

4.1 Description of the test building

This study is a part of the Ph.D. research of the first author, aiming to develop, evaluate and optimize on-site data-based dynamic solar gain estimation techniques. It is also part of the IEA EBC Annex-71 project: 'Building energy performance assessment based on in-situ measurements' [65]. The basic strategy of the Ph.D. research is exploring solar gain estimation techniques step-by-step alongside an increased complexity of cases: from reduced size to full-scale building cases and from simplified heating signal to real occupants' heating profiles. In this study, a 'simplified building' with a 'simplified heating signal' is initially explored.



Fig. 6 The image of the portable site office in scenario 1.

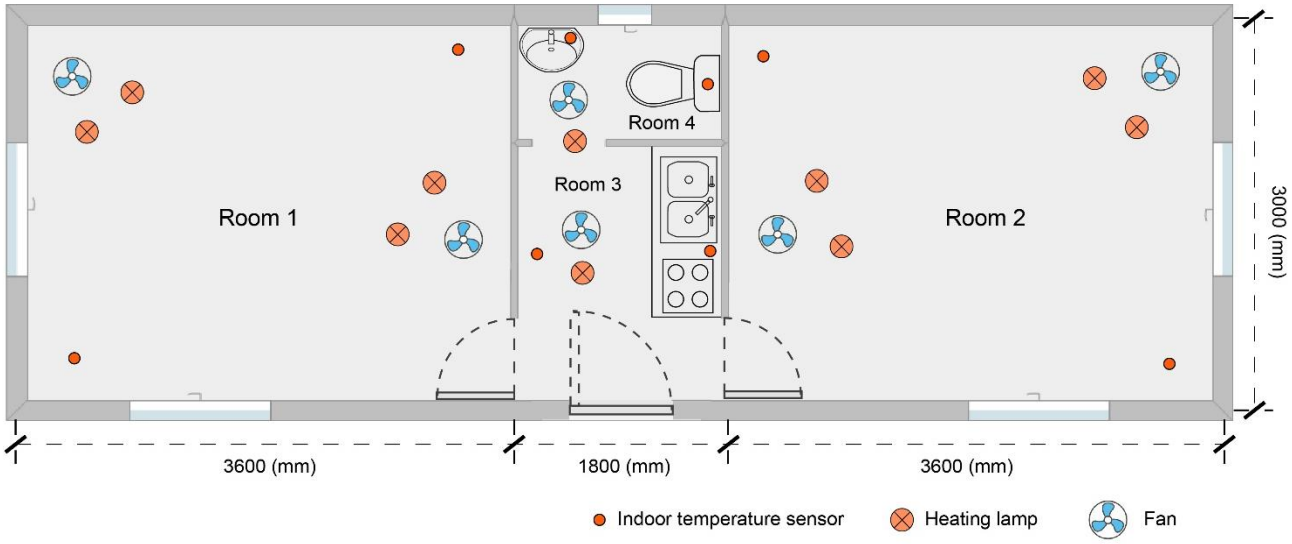


Fig. 7 The layout of the portable site office with PRBS scenario 1.



Fig. 8 the parking lot where the test building is located.



Fig. 9 The site layout around the test building.

A portable site office (Fig. 6) is treated as a 'simplified building' to reduce the complexity of actual building cases [5]. For example, the building is raised 30 cm above the ground to avoid heat transfer between test building and the ground, simplifying the boundary conditions out of interest (Fig. 6). In addition, the size of the test building is much smaller than common buildings. The overall dimensions are 9*3*3m^3, accommodating four rooms, two main test rooms, a tiny kitchen, and a toilet (Fig. 7). The distribution of five windows in the building is illustrated in Fig. 7. Except for the small window of room 4 (toilet) with a size of 0.5*0.6m^2 (0.33*0.4m^2 for the glazing part), the other four windows have the same size of 1.42*1.08m^2 with a glazed area of 0.98*0.72m^2. As shown in Fig. 8, the test building is located at a parking lot (50.86N, 4.68E) of KU Leuven, Belgium, with the front facade with entrance door facing south-east (SE140°) (Fig. 9).

4.2 Two datasets with pseudo random binary sequence (PRBS) heating signal

Regarding the 'simplified heating signal', a PRBS (Pseudo-random binary sequence) heating signal was selected and applied to the test building. The key feature of a PRBS sequence [66] is its white-noise property of auto-correlation and its uncorrelation with other signals [7,67]. Hence, when applying a PRBS as heating signal, no correlation between heating input and other input data (e.g. solar radiation or outdoor temperature) will be observed. The heating system in the building is visualized in Fig. 10. In addition, two scenarios are explored, as shown in Fig. 6 and Fig. 11, under two seasons respectively (i.e. winter and summer). The main purpose of designing two scenarios is to test the performance of the proposed technique in revealing time-varying solar gains under different scenarios of the window distribution and typical seasons. The layouts of the two scenarios with the respective distribution of the heating system are visualized in Fig. 7 and Fig. 12. Scenario 1 (Fig. 7), leaving all windows uncovered is measured in winter, and scenario 2 with two southern windows shaded (Fig. 11) is monitored under the summer weather. In the latter case, the peak of the PRBS signal is reduced from 3000 to 1800W (from winter to summer) by decreasing the number of ten to six lamps (see Fig. 7 and Fig. 12), to avoid over-heating. Subsequently, two datasets (marked as DS1 and DS2) are measured and constructed, as summarized in Table 1.

Table 1 Two datasets measured under different scenarios with PRBS signal

Dataset (code)	Scenarios	Measurement period	Frequency
Scenario 1 in January dataset (DS1)	All windows uncovered, shown in Fig. 6 and Fig. 7.	From 17 January (19:00) to 24 January 2020 (9:55); in winter	60 mins (hourly)
Scenario 2 in August dataset (DS2)	Two windows covered shown in Fig. 11 and Fig. 12.	From 6 August (00:00) to 12 August 2020 (23:55); in summer	30 mins

Fig. 13 and Fig. 14 visualize the datasets. It is acknowledged that both DS1 and DS2 stay a distance from reality, from indoor thermal comfort perspective, since high temperatures are observed in both two datasets, especially in the summer case (DS2). For the details, the heating input (W), indoor temperature (°C) are measured inside the building, and outdoor air dry-bulb temperature (°C), wind speed (m/s), direct normal (W/m^2), global horizontal (W/m^2), and diffuse horizontal irradiance (W/m^2) are monitored at a nearby weather station [68,69]. Note that the original sampling time of aforesaid in-situ data is 1 min (or 5 mins). Original in-situ data is mathematically averaged to 60 (or 30 mins) for the DS1 and DS2 dataset respectively (Table 1). It is expected that the scenario 2 dataset is more dynamic, due to the added shading devices, thus, a lower frequency (30 mins) is applied. In addition, as shown in Fig. 7 and Fig. 12, the indoor temperature is measured by eight air temperature sensors. A representative temperature is calculated via averaging firstly, as a pre-processing before the 60 (or 30 mins) re-sampling.



Fig. 10 The heating and sensor system in the test building.



Fig. 11 The test building under scenario 2.

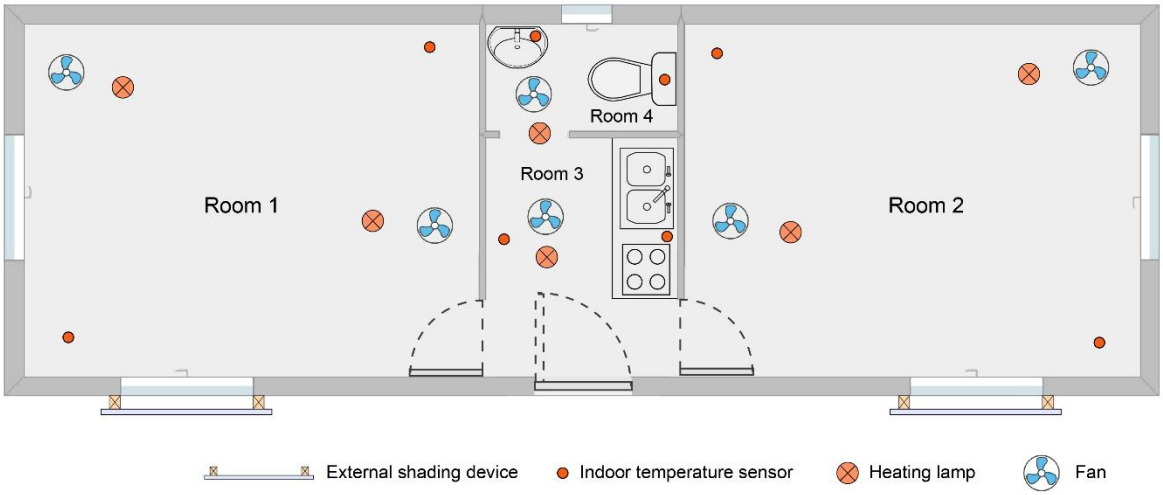


Fig. 12 The PRBS scenario 2 of the test building.

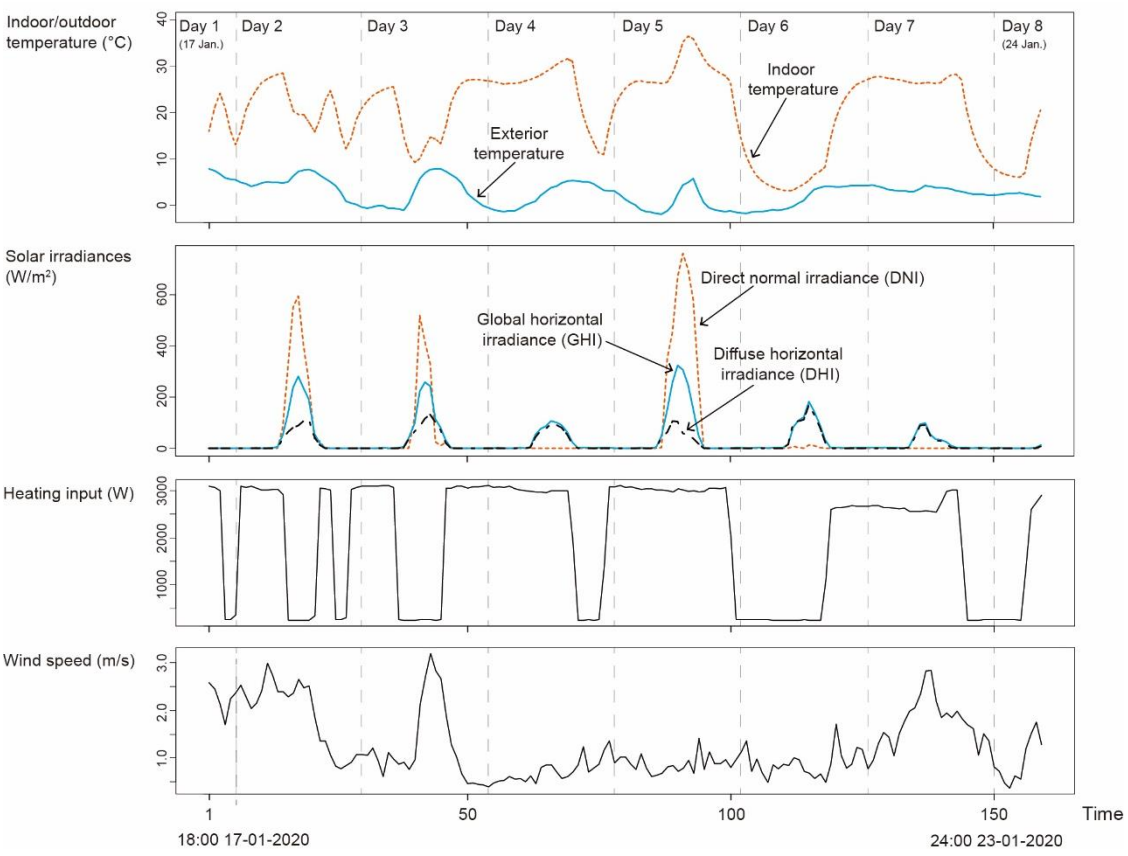


Fig. 13 The plots of DS1 dataset.

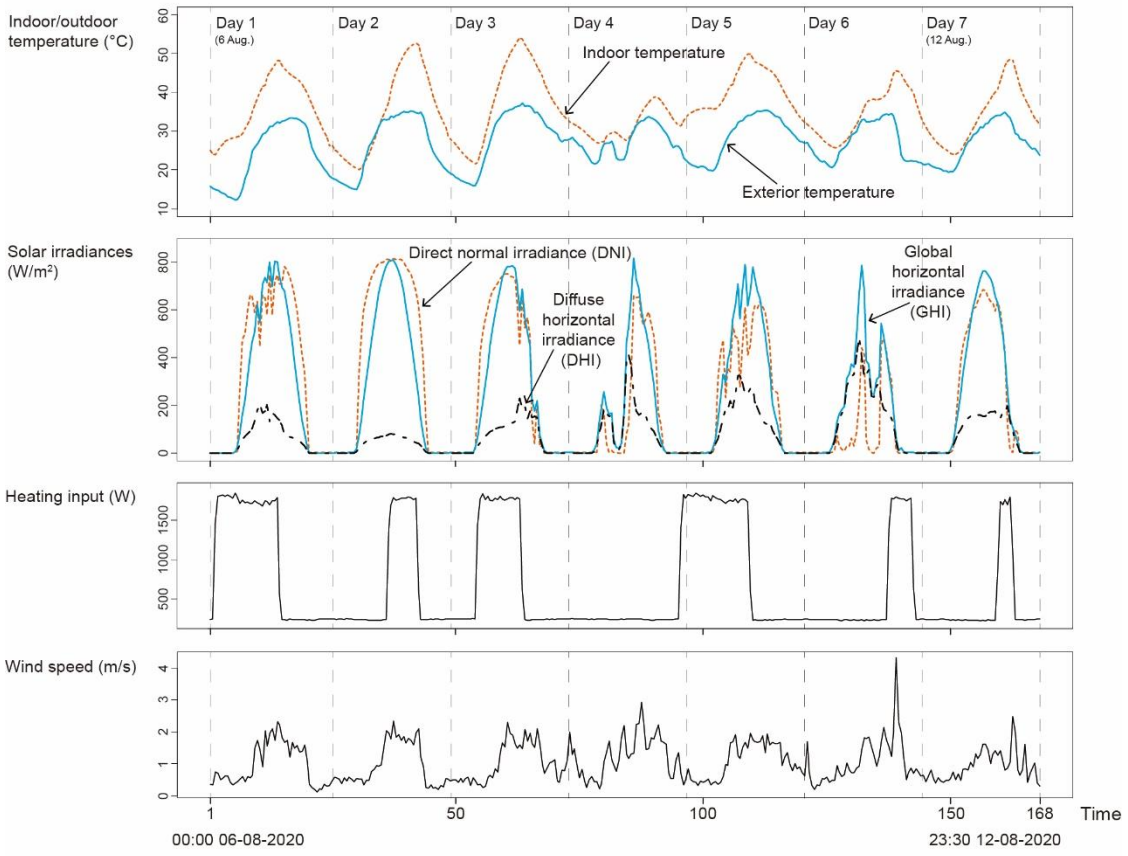


Fig. 14 The plots of DS2 dataset.

4.3 Pre-constructed customized B-splines for the datasets

As stated in section 2.3, the pre-construction of basis splines follows the occurrence of direct normal irradiance. At the location of the test building (50.86N, 4.68E), the appearance of direct normal irradiance (DNI_t) on 17-24 January (DS1) and 6-12 August 2020 (DS2) is basically from 8:00 to 17:00 and from 6:30 to 20:00 respectively, as visualized in Fig. 15. The appearance of DNI_t generally refers to the daily period when DNI_t value is above zero. Thus, the pre-constructed B-splines for DS1 and DS2 are shown in Fig. 16, using five basis splines as an example. In the next section, the customized B-splines data serves as additional input data, to support the exploration of ARX models with time-varying solar aperture. In this study, the data pre-processing and ARX model construction are conducted in the R-studio platform supported by the R 4.1.0 version. The R-core package 'Splines' (version 3.6.3) [70] is used to generate basis splines. In addition, the 'tidyverse' package [71] is also required to enable the 'nls' function, which enables the customization of the upper and lower boundaries of coefficient estimation in regression.

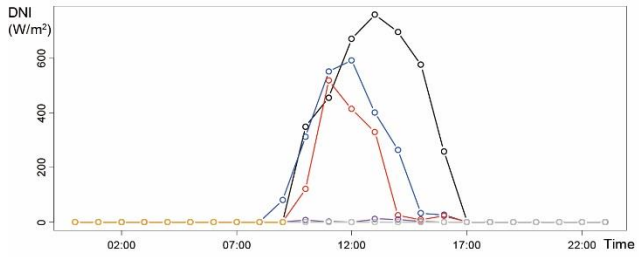


Fig. 15 Daily DNI_t plots of DS1 (up) and DS2 (down) datasets.

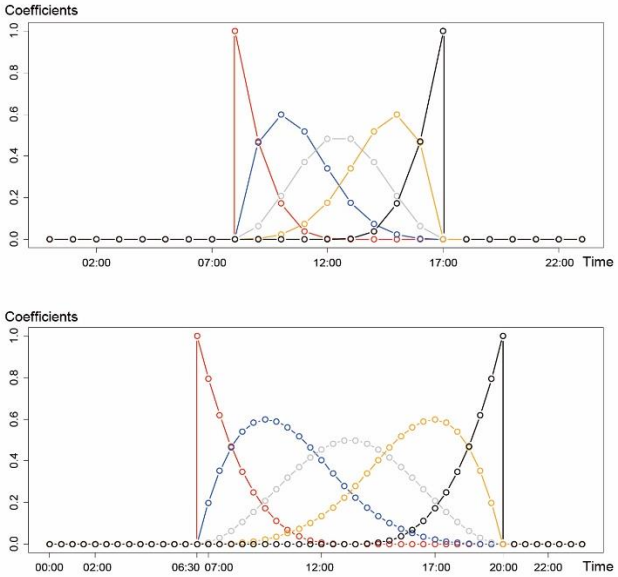


Fig. 16 Customized daily B-splines for DS1 (up) and DS2 (down) datasets, taking five B-splines as an example.

As elaborated in section 2.1, in general, the ARX model (so-called standard ARX model) is designed to give 'full freedom' when estimating coefficients, varying from positive to negative infinity. The 'full freedom' of the standard ARX model can support the optimal data-fitting, but with a certain risk of converging to a physically unreasonable but statistically optimal solution, such as a model with the best data-fitting performance but determining a negative solar aperture. Therefore, in this study, a value of zero is set as lower boundary for the DNI_t related coefficients (e.g. 'DNI ',' DNI_l1 ', '$bs1$', '$bs2$', etc. in Fig. 17 and Fig. 18), to ensure the physical interpretation of the estimated gA_{dni_t} outcomes. For the coefficient of the other parameters considered - $T_{i,t}$, $T_{e,t}$, Ph_t, and GHI_t (e.g. Ti_l1, Te, Ph, and GHI_l1 etc. in Fig. 17-18), no boundaries are set in the subsequently proposed ARX models, to stay closer to the standard ARX model, reaching their full potential in data fitting and offering meaningful gA_{dni_t} estimation at the same time. In this regard, the ARX models explored in Fig. 17-18, are not pure black-box models, strictly speaking, since the physical knowledge of non-negative gA_{dni_t} is incorporated.

5. Results

5.1 Model selections for two datasets

A two-stages selection strategy is used in this study to identify the optimal B-splines integrated ARX model to each of the two datasets. To be specific, as visualized in Fig. 17 and Fig. 18, the (backward elimination based) first step aims to identify the 'optimal' ARX model, under constant gA_{dni} assumption, for each dataset. Thus, two ARX models, assuming gA_{dni} steady, are offered for DS1 and DS2 respectively in step 1, which are ARX_DS1 (Fig. 17) and ARX_DS2 (Fig. 18) models. In step 2, based on ARX_DS1 and ARX_DS2, the constant gA_{dni} values integrated within them are further elaborated to dynamic curves by the B-splines fitting (section 2.3). As clarified in section 2.4, the forward selection is selected in step 2 to identify the suitable number(s) of B-splines needed for ARX_DS1 and ARX_DS2, by increasing the number of integrated B-splines step-by-step. The (forward selection based) second step points out that five B-splines is most suitable for both ARX_DS1 and ARX_DS2 models, leading to ARX_DS1_S5 and ARX_DS2_S5 respectively as most optimal models (Fig. 17-18). Note that, both step 1 and 2 follows the pre-set model selection strategy, elaborated in section 2.4, choosing the model with the highest *LogL* and lowest *AIC* and *BIC*-values among candidates in each stage. In Fig. 17 and Fig. 18, the key information of the model identification process is visualized by a color code. In the square brackets, parameters with statistical significance are marked as black. In contrast, the green color indicates the parameter is statistically insignificant, with a specific case that the orange color is marked for DNI_t related parameters with statistical insignificance but non-zero estimation on the coefficients. The 'bs3' in ARX_DS1_S5 and 'DNI_l1' in ARX_DS2 models are examples of orange code. On the right hand of the square brackets, red and blue indicate increasing in *LogL*-value and reduction in *AIC* (or *BIC*) respectively, and vice versa (marked as black).

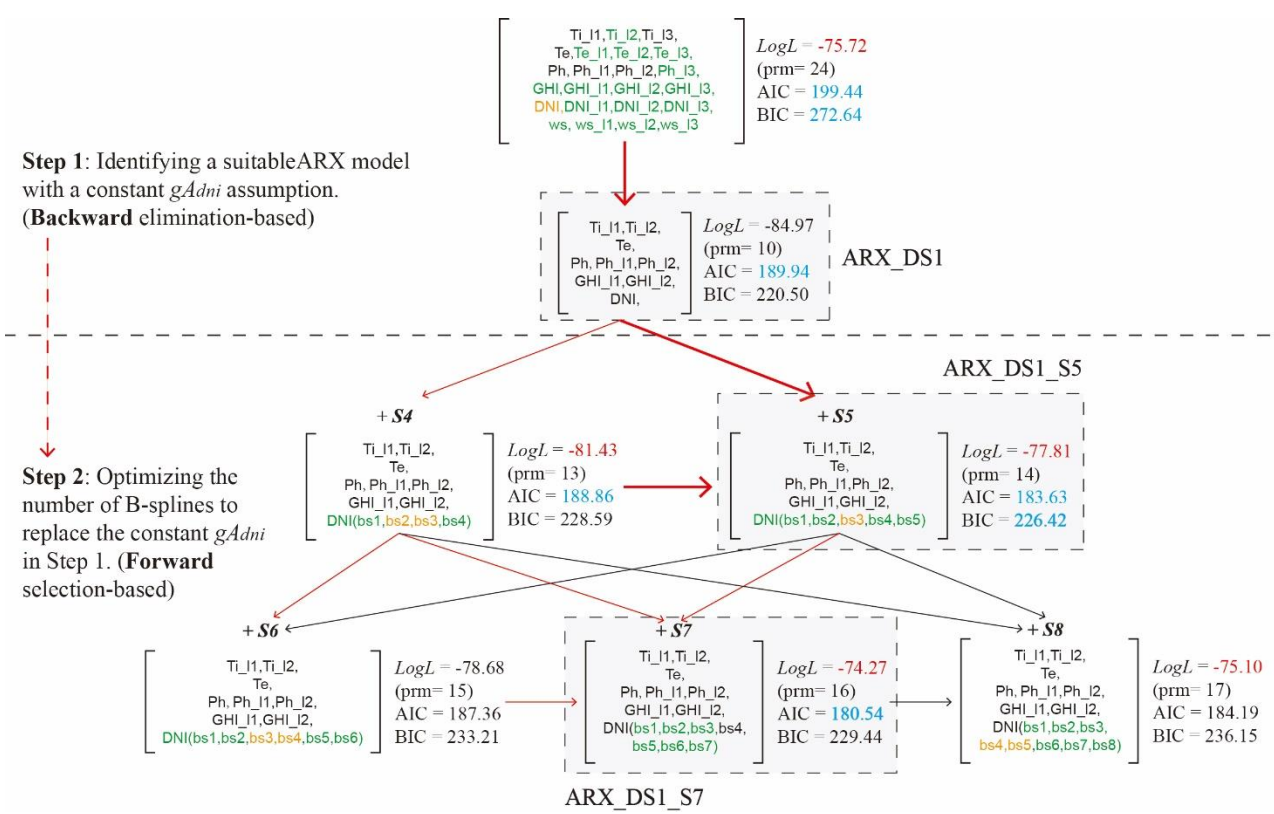


Fig. 17 ARX model selection of DS1 dataset.

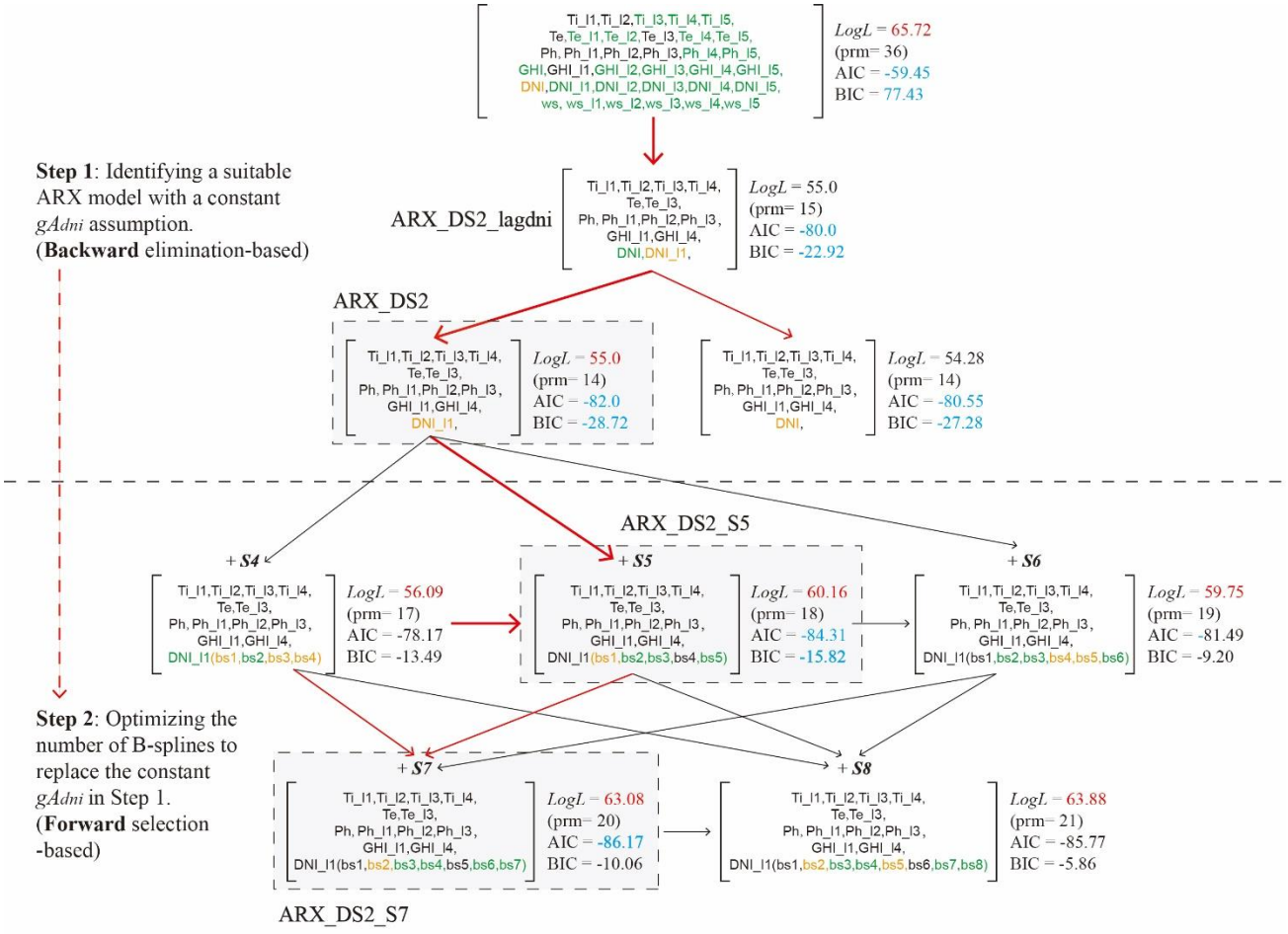


Fig. 18 ARX model selection of DS2 dataset.

Note that, DS1 is in 60 min and DS2 in 30 mins frequency (Table 1). Thus, the lags, denoted as '$_\ l_n$' in Fig. 17 and Fig. 18 refer to historical data in n times of 60 mins (DS1) or 30 mins (DS2). For example, '$T_{i_}\ l_2$' in ARX_DS1 and '$T_{i_}\ l_4$' both represent the indoor temperature observations two hours ago. Thus, in step 1, the starting models of DS1 and DS2, in order 3 (three hourly lags considered) and order 5 (five 30 mins lags incorporated) respectively, indicates the modeling of a maximum of three (or 2.5) hours of historical data. Incorporating additional lags in ARX_DS1 and ARX_DS2 leads to no statistical convergences, and the selected two models in step 1 also show that only the previous data within two hours of solar radiation (e.g. 'GHI_l2' in ARX_DS1 and 'GHI_l4' in ARX_DS2), heating inputs (e.g. 'Ph_l2' in ARX_DS1 and 'Ph_l4' in ARX_DS2), etc, are statistically vital to the modelling of the present indoor temperature ($T_{i,t}$). Considering that the portable site office is a lightweight and reduced-size building, the aforesaid modelling of 2 hours delay seems physically reasonable. Besides, interestingly, in the ARX_DS2_lagdni model (Fig. 18), the estimated coefficient of the current time step (*DNI*) is zero with no statistical significance. Thus, ARX_DS2 only models 'DNI_l1' and neglects '*DNI*'. This indicates that the direct normal radiation from 30 mins ago, instead of the current direct normal radiation, determines the present indoor temperature ($T_{i,t}$), which also seems reasonable. Physically, the solar radiation (e.g. DNI_t) heats the internal surface of the building initially by radiation, and the heated internal surface warms the indoor air (T_i) by convection subsequently. Given that the test building is a lightweight building, the 'delay' phenomenon might be relatively short, Thus, the wider time interval of the sampling (i.e. 60 mins) in DS1 could be the reason that the 'delay' effect is not found in the ARX_DS1 model and it could also be the reason for a better performance of the residuals of ARX_DS2_S5 in the auto-correlation function (ACF) and the cumulated periodogram (CP), etc. (see Appendix B). To sum up, as stated, ARX_DS1_S5 and ARX_DS2_S5 are

chosen as optimal B-splines integrated ARX models for DS1 and DS2 datasets respectively, because of their low *AIC* and lowest *BIC* among candidate models (Fig. 17 and Fig. 18). In addition, it is noteworthy that seven B-splines integrated ARX models (ARX_DS1_S7 in Fig. 17 and ARX_DS2_S7 in Fig. 18) have the smallest *AIC*-value among 'peers'. In addition, the residuals of the chosen models of interest have been checked (see Appendix B) and can be treated as 'white noise'.

5.2 Comparison of estimated gA_{dni_t} curves from different approaches

As mentioned in section 3, based on the same datasets (Table 1), the two gA_{dni_t}-curves have been estimated based on the B-splined integrated grey-box model, and compared to corresponding simulation outcomes from the traditional BES-model. To understand the performance of the B-splines integrated ARX modeling technique proposed in this study, Fig. 19 and Fig. 20 compare the estimated dynamic gA_{dni_t}-curves of the selected B-spline integrated ARX-models (i.e. ARX_DS1_S5 and ARX_DS2_S5) for both datasets with the corresponding outcomes from grey-box and BES-model approaches, studied in [35]. Note that the BES-simulation in [35] is not a full model, but only considers the effective glazed envelope of the building, without modelling the opaque walls, roof, etc. Considering only the dynamic Sg_{dir_t} information is of interest in [35], there is no need for additional model calibration, since the simulation reliability of the platform (Modelica with IDEAS library) has been verified in several approaches [72]. Hence, the simulated Sg_{dir_t} in [35] is regarded as trustable and comparable to the ARX modelling outcomes in this study.

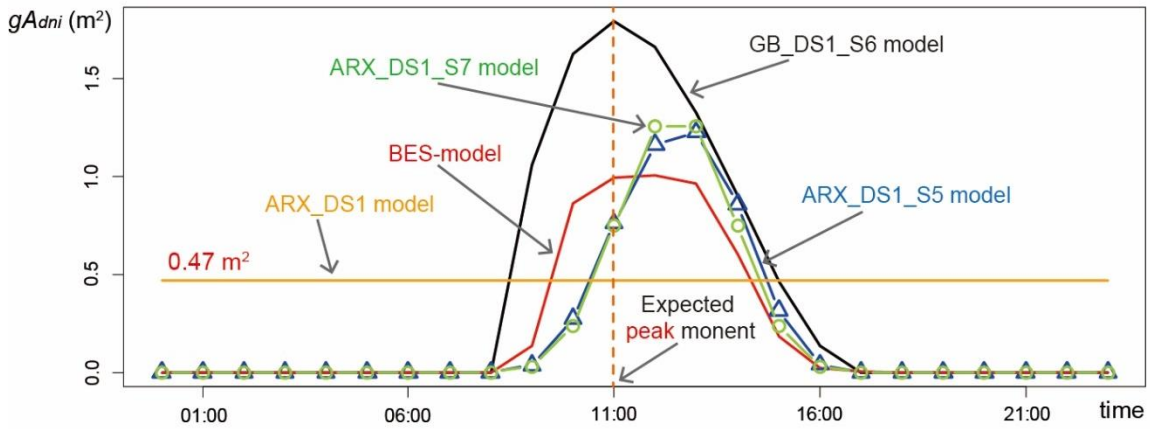


Fig. 19 Estimation outcomes of dynamic $gA_{dni_ds1_t}$ of DS1 dataset under different approaches.

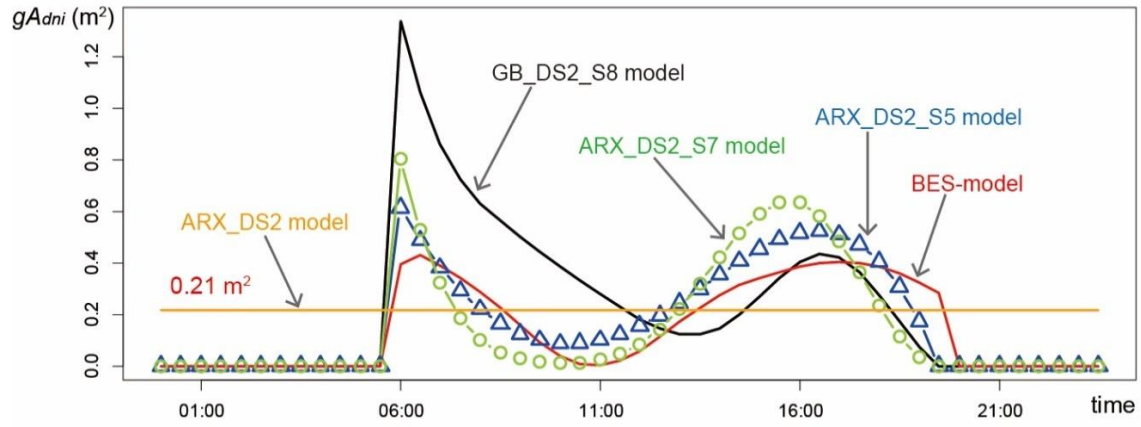


Fig. 20 Estimation outcomes of dynamic $gA_{dni_ds2_t}$ of DS2 dataset under different approaches.

The qualitative characteristics of the building with and without shading are firstly analyzed and visualized in Fig. 21. Based on the qualitative analysis from Fig. 21, it is clear that all of the three approaches (BES-, ARX-, and grey-box models) can reveal the key qualitative time-dependent features of $gA_{dni_ds1_t}$ and $gA_{dni_ds2_t}$ for DS1 and DS2 respectively. To be specific, one peak is expected for the first dataset (DS1) and two peaks for the second (DS2).

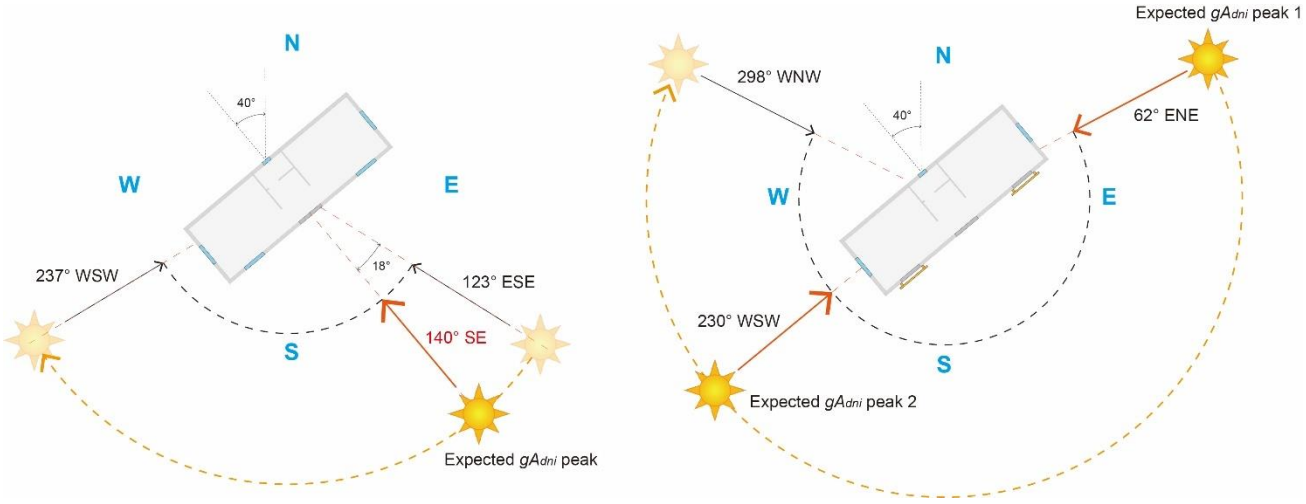


Fig. 21 Expected peak(s) of daily dynamic $gA_{dni_ds1_t}$ and $gA_{dni_ds2_t}$ for January DS1 (left) and August DS2 (right).

For the quantitative aspect, taking the BES-simulation as the benchmark, a general over-estimation of the dynamic gA_{dni_t} is observed in the grey-box model-based approach. This could be attributable to a so-called 'lumped effect' [35]. In brief, the 'lumped effect' refers to the limitation of data-driven approaches, to distinguish different physical phenomena which are statistically correlated. It leads to the estimated value of the particular physical variable that might lose a clear physical interpretation, with partial gain (or loss) from (to) another physical parameter. For example, the grey-box model might incorporate a part of the solar gains via the opaque envelope (Sg_{en_t}) into the direct solar gains via the glazed parts of the windows (Sg_{dir_t}) (Fig. 3). Note that, only the glazing parts of the window system are constructed in the BES-model, to ensure a clear physical interpretation of its simulation output (i.e. Sg_{dir_t}). The reader is referred to [35] for more details.

From this viewpoint, the ARX-model generally suffers less from this 'lumped effect' (i.e. over-estimation phenomenon). Although the over-estimation might not be fully eliminated by the ARX-model approach, it performs generally better than the grey-box approach for both datasets. This can be attributed to the fact that the ARX model has enhanced flexibility in model construction than the grey-box model, since the latter might force the model to simplified physical assumptions. Furthermore, it shows some additional potentials. For instance, in scenario 2, when the sun faces the main façade of the building with its windows covered, the $gA_{dni_ds2_t}$ should gradually reduce to zero (Fig. 21), since no glazed area is exposed to the DNI_t for a while. However, the two dynamic $gA_{dni_ds2_t}$curves estimated by both ARX_DS2_S5 and the grey-box model (GB_DS2_S8) never reach the bottom (i.e. $gA_{dni_ds2_t}$ equals to zero) with values always above 0.1 m^2. Thus, the above-mentioned 'lumped effect' exists in both ARX_DS2_S5 and GB_DS2_S8 models. However, if we select the optimal model via the lowest *AIC* instead of the smallest *BIC* (i.e. the pre-set model selection criteria in section 5.1), the ARX_DS2_S7 model will be identified for scenario 2 (Fig. 18). The dynamic $gA_{dni_ds2_t}$ curve estimated by the ARX_DS2_S7 model shows the potential to reach the zero-value estimation of $gA_{dni_ds2_t}$ (Fig. 20) for the particular period when the sun faces the two covered windows (Fig. 21). Thus, it could be interesting to conduct more case studies in future research to explore potentially 'best' model selection guidelines for the dynamic gA_{dni_t} estimation. Generally, BIC is more sensitive to the number

of parameters involved in the statistical models, with a stricter penalty than AIC when the model tends to become more complicated. For a normal statistical model with all involved parameters statistically significant, BIC works well [73]. However, in the B-splines integrated models, it might show difficulties, since, in general, only partial basis splines play a vital role in gA_{dni_t}-determination and show statistical significances in those models (see Table 2). However, to apply Eq. (16), it is not allowed to remove the statistically insignificant ones, leading to limited improvement on data-fitting but a considerable increase in the number of parameters.

Table 2 Partial estimated coefficients of ARX_DS2_S7 model

Parameters	Estimate of coefficients	p-value	Statistical significance
Basis spline 1 (bs1)	4.685e-04	0.04376	*
bs2	4.659e-05	0.84445	--
bs3	0.0	1.0	--
bs4	0.0	1.0	--
bs5	6.247e-04	0.01565	*
bs6	0.0	1.0	--
bs7	0.0	1.0	--

For example, the estimated coefficients of B-splines for the ARX_DS2_S7 model are summarized in Table 2, where only 'bs1' and 'bs5' (the first and fifth basis splines) are statistically significant. The other five B-splines are statistically insignificant with 'zero' estimate for four out of five. Hence, there is a possibility that ARX_DS1_S7 is treated as somewhat 'over-punished' in the BIC performance, leading to the selection of the ARX_DS1_S5 as the most suitable model for DS2. More studies might be required to get a deeper understanding. But overall, treating the outcome of BES-model simulation as a benchmark, it seems that the ARX-based technique has a better performance than the grey-box model. However, the ARX-based technique might have a risk of missing some precise details. For example, as illustrated in Fig. 21, for the DS1 dataset, a peak closer to the sunrise (around 11 a.m.) is expected for $gA_{dni_ds1_t}$ which is revealed precisely by both grey-box and BES-models approaches. However, the peak is estimated two hours later (around 1 pm) by the ARX model-based technique (e.g. ARX_DS1_S5 and ARX_DS1_S7 models). For the grey-box model selection, the reader is referred to [35] for elaborated details. In brief, GB_DS1_S6 and GB_DS2_S8 models are identified as the 'optimal' B-splines integrated grey-box models for DS1 and DS2 respectively.

6. Discussion and conclusion

Although the BES-simulation approach also has its limitation (as elaborated in [35]), it still serves as a reference for the grey-box and ARX models based time-varying solar aperture estimations. In general, based on the current case study, the ARX-based approach proposed in this study offers a closer similarity of gA_{dni_t} with BES-model simulation outcomes, compared to the grey-box model-based technique [35]. This could be attributed to the fact that ARX models do not start from building physics-based assumptions. Even a very complicated grey-box models still largely simplifies the reality. From this perspective, the black-box approach of the ARX model, liberated from the pre-set assumptions, might be able to mimic the actual behavior closer than the grey-box models. The interesting phenomenon shown in the ARX_DS2 model selection could serve as a piece of evidence (section 5.1). As mentioned, instead of the current direct normal irradiance (DNI_t) the first lag (DNI_{t-1}) results in an enhanced statistical fitting. Therefore, the coefficient of first lag DNI_{t-1} is replaced by the sum of weighted B-splines to obtain the time-varying solar aperture. The potential physical interpretation has been illustrated in section 5.1. Therefore, this example might show a better capability of ARX to deal with actual physical phenomena.

However, as discussed in section 5.2, the black box approach in the ARX model could increase the risk of offering an optimal model in statistics but physically unreasonable. As stated in section 5.1, for the first scenario (DS1), a two-hour delay of the solar aperture peak is detected in the ARX model approach. Besides, an additional potential weakness of the ARX model approach is its sensitivity to the process of model reduction (backward elimination) or extension (forward selection). Specifically, removing (or adding) different components (in each step) could impact the statistical significance of existing parameters, and could finally lead to slightly different models in model selection. This is also the reason that the backward elimination is recommended for the common ARX model identification, since it can incorporate many parameters at the starting point, with a lower risk of losing the optimal model. The two stages strategy applied in this study is ideally to be merged into one-step. However, in the ideal one-step model identification, each coefficient related to the direct solar gains (e.g. 'DNI', 'DNI_l1', 'DNI_l2', and 'DNI_l3' in the start model of Fig. 17) should be replaced individually by the sum of a certain number of weighted B-splines in the initial stage (i.e. the first model of the backward elimination). This is almost infeasible in practice, since this single-step strategy will result in an exponentially boosting of the number of parameters involved, resulting in no convergence in the data fitting. Hence, from this point of view, the (feasible) two steps strategy used in this study might have a potential risk of losing the theoretically 'optimal' model. Nevertheless, the above (potential) limitations didn't prevent the good performance of the ARX model to estimate the time-varying solar aperture in both scenarios of this study.

To sum up, to respond to the first research question proposed in section 2.2, under the assumption of dynamic gA_{dni_t} supported by the weighted B-splines fitting, the application of Eq. (15) can still contribute to a meaningful estimation outcome in practice. For the second research question, in brief, all of the three aforesaid estimation approaches (BES, ARX, and grey-box models) can reveal the key qualitative characteristics of dynamic daily gA_{dni_t} under different window distribution scenarios very well (Fig. 19 and Fig. 20). From the quantitative perspective, treating BES-simulation outcome as reference, the B-splines-based ARX model seems to offer a better agreement, compared to the B-splines integrated grey-box model. Considering the potential risks, more cases studies are required to test the robustness of the B-splines integrated ARX model technique in characterization of dynamic solar gains.

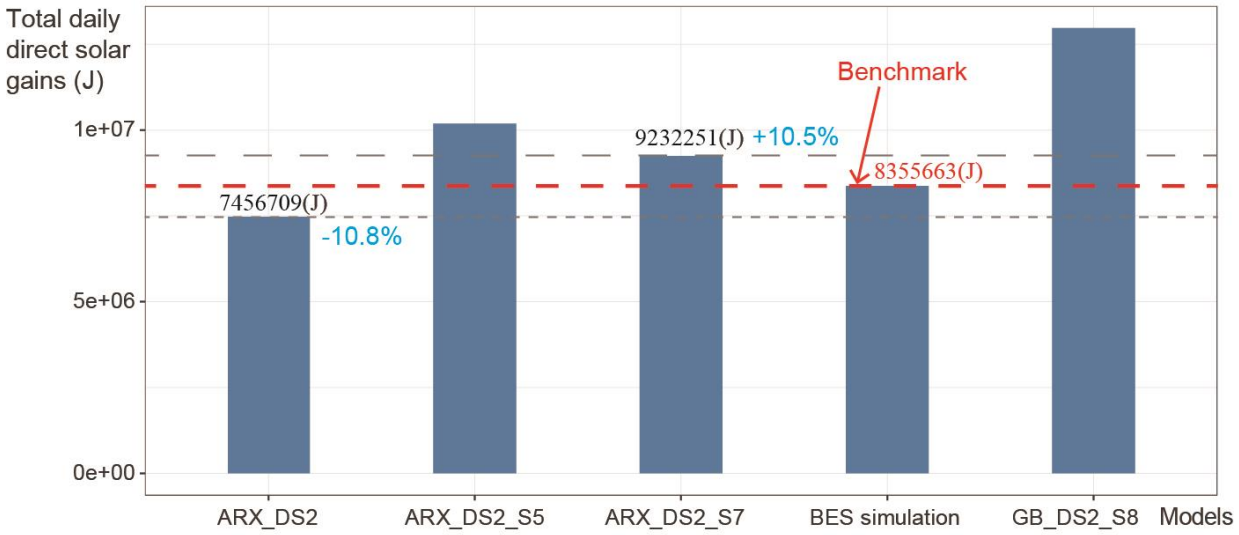


Fig. 22 Estimation outcomes of the total daily Sg_{dir_t} in Day 2 of DS2 dataset (7 August) from different models.

Finally, this study further demonstrated that the constant (direct) solar aperture values (i.e. gA_{dni} or gA) assumptions widely used in ARX models, might offer an acceptable approximate estimation in the total daily amount of solar

gains, but will generally lead to inaccurate estimations at a certain moment in time, especially in cases with unevenly distributed windows. To illustrate this, Fig. 22 compares the total amount of daily Sg_{dir_t}, marked as $Sg_{dir,daily}$, on 7 August 2020 estimated by different models. To be specific, $Sg_{dir,daily}$ indicates the total direct solar gain (in J) penetrating glazed envelop in one day, which is formulated as Eq. (21). where *t* expresses the number of chronological observations with a total of 48 observations for a particular day (in 30 mins frequency), and DNI_t data of this specific day is used in Eq. (21). The blue (or red) zone areas in Fig. 2 could serve as a reference to visualize the definition of $Sg_{dir,daily}$.

$$Sg_{dir,daily} = \sum_{t=1}^{48} Sg_{dir_t} = \sum_{t=1}^{48} gA_{dni_t} \times DNI_t \quad (21)$$

Both ARX_DS2 (with constant gA_{dni} assumption) and ARX_DS2_S7 (with spline-based time-varying solar aperture) predict a very similar amount of daily total solar gains as the BES-model simulated reference value. However, when we zoom in on the dynamic solar gains as predicted by the above-mentioned models (Fig. 23), the ARX_DS2 model estimates, unlike reality, the highest gains when the sun is facing the two blocked windows. Therefore, in this particular case with uneven distribution of windows (i.e., PSO scenario 2), it is reasonable to expect that, in data-driven models, elaborated solar gain estimations could enhance the performance of model predictive control (MPC). In future work, the comparison of MPC performances between data-driven models, based on constant or dynamic solar gain coefficient (*gA*) assumptions, could be an interesting topic. Meanwhile, a recent study by Bünning et al. [74] shows elaborating dynamic solar gain details might not be necessary for a building with glazed building fabrics (almost) equitably allocated in four orientations. Thus, the necessity of taking the dynamic solar aperture (gA) assumption is also of interest for future studies. Moreover, as stated, this study is based on the PRBS heating signal, leading its temperature might be unrealistic regarding the thermal comfort (e.g., extreme indoor temperature above 40 degrees Celsius is noticed in DS2; Fig. 14). Thus, more case studies on buildings with more realistic conditions are of interest to investigate the 'robust' of this technique proposed, in the future study.

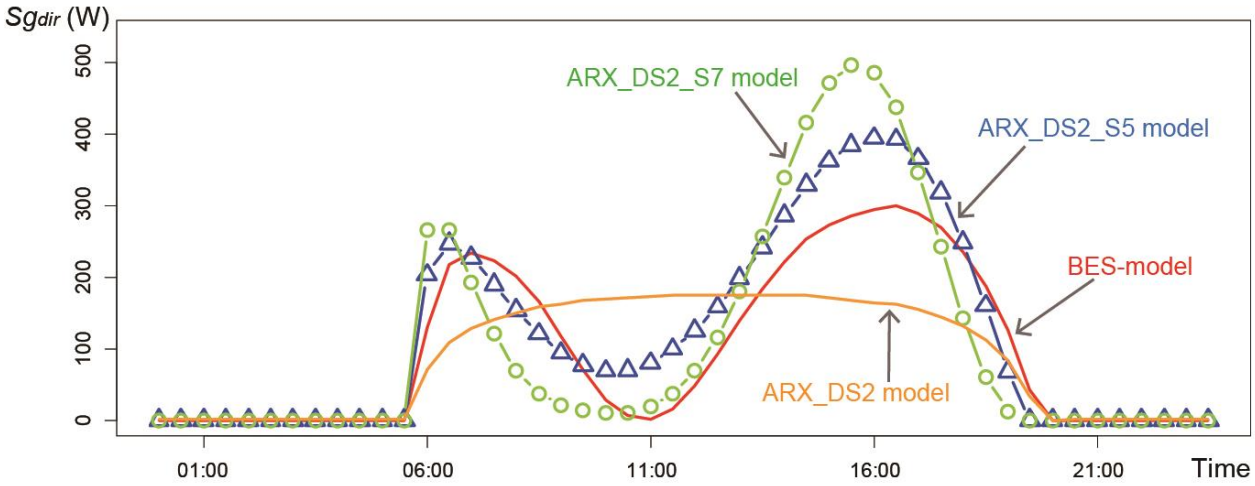


Fig. 23 Estimation outcomes of dynamic solar gains (Sg_{dir_t}) during the second day of the DS2 dataset (7 August) under different modelling approaches.

Acknowledgment

This research was funded by the Research Foundation Flanders (FWO), application number G0D2519N, and by KU Leuven, grant C24/18/040. This support is gratefully acknowledged.

Appendix A

Developing an ARX model for the RC model (thermal network), instead of treating it as a pure black-box, contributes to maintaining the physical information of parameters [46]. This study proposes an alternative way of connecting ARX- and RC- models to enhance the physical interpretation of estimation outcomes of the ARX model. Let's start from a simple first-order RC-model with a system equation written in Eq (A.1).

$$dT_i = T_{i,t+1} - T_{i,t} = \left(\frac{1}{RC}T_{e,t} - \frac{1}{RC}T_{i,t} + \frac{1}{C}Ph_t + \frac{1}{C}gA_{dni_t}DNI_t + \frac{1}{C}A_{en_t}GHI_t\right)dt + \sigma\omega_t \quad \text{(A.1)}$$

Meanwhile, Eq. (A.1) can be rewritten as Eq. (A. 2) and Eq. (A. 3-4) are generated via Eq. (A. 2) by varying t to $t-1$, $t-2$, ect.

$$T_{i,t+1} = T_{i,t} + \left(-\frac{1}{RC}\boldsymbol{T_{i,t}} + \frac{1}{RC}T_{e,t} + \frac{1}{C}Ph_t + \frac{1}{C}gA_{dni_t}DNI_t + \frac{1}{C}A_{en_t}GHI_t\right)dt + \sigma\omega_t \quad \text{(A.2)}$$

$$T_{i,t} = T_{i,t-1} + \left(-\frac{1}{RC}\boldsymbol{T_{i,t-1}} + \frac{1}{RC}T_{e,t-1} + \frac{1}{C}Ph_{t-1} + \frac{1}{C}gA_{dni_{t-1}}DNI_{t-1} + \frac{1}{C}A_{en_{t-1}}GHI_{t-1}\right)dt + \sigma\omega_{t-1} \quad \text{(A.3)}$$

$$T_{i,t-1} = T_{i,t-2} + \left(-\frac{1}{RC}T_{i,t-2} + \frac{1}{RC}T_{e,t-2} + \frac{1}{C}Ph_{t-2} + \frac{1}{C}gA_{dni_{t-2}}DNI_{t-2} + \frac{1}{C}A_{en_{t-2}}GHI_{t-2}\right)dt + \sigma\omega_{t-2} \quad \text{(A.4)}$$

By replacing the bold part of Eq. (A.2) by Eq (A.3) and repeating a similar campaign for Eq. (A.3) via integration of Eq. (A.4). Then, the expression of $T_{i,t+1}$ can be extended into an endless form consisted by T_i , T_e , Ph, DNI and GHI at time *t* and past observations (e.g. at *t*-1 and *t*-2), which is a similar ARX model form in Eq. (A.5).

$$\begin{aligned}
T_{i,t+1} = {} & dtT_{i,t} - \frac{1}{RC}(dt)^2T_{i,t-1} + \frac{1}{R^2C^2}(dt)^3T_{i,t-2} - \frac{1}{R^3C^3}(dt)^3T_{i,t-2} + \dots \\
& + \frac{1}{RC}dtT_{e,t} - \frac{1}{R^2C^2}(dt)^2T_{e,t-1} + \frac{1}{R^3C^3}(dt)^3T_{e,t-2} + \dots \\
& + \frac{1}{C}dtPh_t - \frac{1}{RC^2}(dt)^2Ph_{t-1} + \frac{1}{R^2C^3}(dt)^3Ph_{t-2} + \dots \\
& + \frac{1}{C}dtA_{en_t}GHI_t - \frac{1}{RC^2}(dt)^2A_{en_{t-1}}GHI_{t-1} + \frac{1}{R^2C^3}(dt)^3A_{en_{t-2}}GHI_{t-2} + \dots \\
& + \frac{1}{C}dtgA_{dni_t}DNI_t - \frac{1}{RC^2}(dt)^2gA_{dni_{t-1}}DNI_{t-1} + \frac{1}{R^2C^3}(dt)^3gA_{dni_{t-2}}DNI_{t-2} + \dots \\
& + \sigma\omega_t + \sigma\omega_{t-1}\,dt + \sigma\omega_{t-2}(dt)^2 + \dots
\end{aligned} \quad \text{(A.5)}$$

It is not hard to find that the expanded version of Eq. (A.2), Eq. (A.5), can be regarded as an ARX model, with the possibility of endless extension by integrating more historical observations of both system input and output. Hence, the $B_1(1)$, $B_2(1)$ and $B_4(1)$ of Eq. (A.5), referring Eq. (13-15), are:

$$B_1(1) = \frac{1}{RC}dt - \frac{1}{R^2C^2}(dt)^2 + \frac{1}{R^3C^3}(dt)^3 - \frac{1}{R^4C^4}(dt)^4 + \cdots \quad \text{(A.6)}$$

$$B_2(1) = \frac{1}{C}dt - \frac{1}{RC^2}(dt)^2 + \frac{1}{R^2C^3}(dt)^3 - \frac{1}{R^3C^4}(dt)^4 + \cdots \quad \text{(A.7)}$$

$$B_4(1) = \frac{1}{C}dtgA_{dni_t} - \frac{1}{RC^2}(dt)^2\,gA_{dni_{t-1}} + \frac{1}{R^2C^3}(dt)^3gA_{dni_{t-2}} - \frac{1}{R^3C^4}(dt)^4gA_{dni_{t-3}} + \cdots \quad \text{(A.8)}$$

In case gA_{dni} is a constant value, with $gA_{dni_t} = gA_{dni_{t-1}} = gA_{dni_{t-2}} = gA_{dni_{t-3}} = \cdots$ in Eq. (A.8), applying Eq. (15) leads to the constant-value of gA_{dni}, as shown in Eq. (A.9). Similarly, Eq. (11) is rewritten as Eq. (A. 10) to gain the *HLC*-determination.

$$\frac{B_4(1)}{B_2(1)} = gA_{dni}\left(\frac{\frac{1}{C}dt - \frac{1}{RC^2}(dt)^2 + \frac{1}{R^2C^3}(dt)^3 - \frac{1}{R^3C^4}(dt)^4 + \cdots}{\frac{1}{C}dt - \frac{1}{RC^2}(dt)^2 + \frac{1}{R^2C^3}(dt)^3 - \frac{1}{R^3C^4}(dt)^4 + \cdots}\right) = gA_{dni} \tag{A.9}$$

$$\frac{B_1(1)}{B_2(1)} = \frac{\frac{1}{RC}dt - \frac{1}{R^2C^2}(dt)^2 + \frac{1}{R^3C^3}(dt)^3 - \frac{1}{R^4C^4}(dt)^4 + \cdots}{\frac{1}{C}dt - \frac{1}{RC^2}(dt)^2 + \frac{1}{R^2C^3}(dt)^3 - \frac{1}{R^3C^4}(dt)^4 + \cdots} = \frac{1}{R} = HLC \tag{A.10}$$

Appendix B

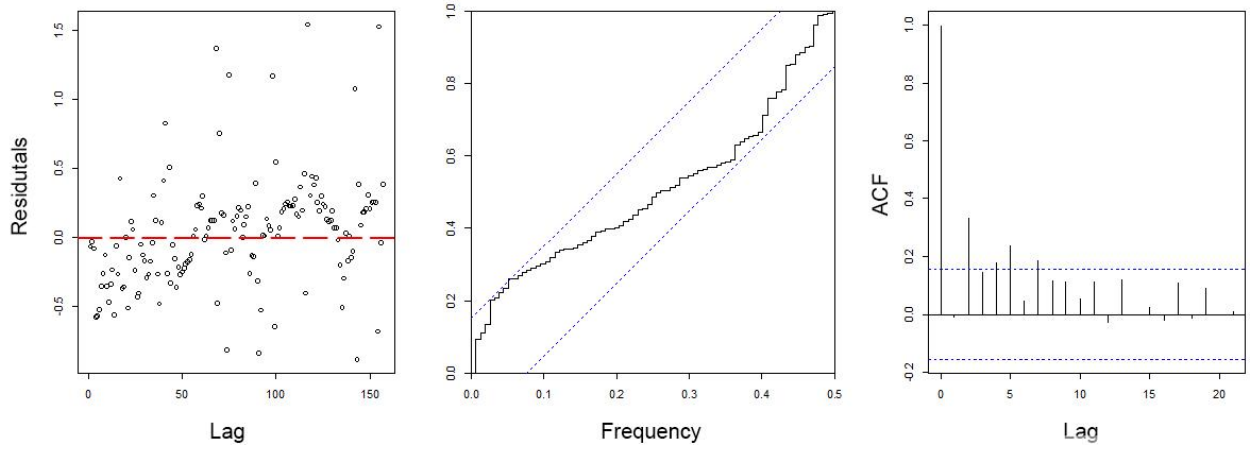


Fig. 24 The plots of residuals versus time, the auto-correlation function (ACF) and the cumulated periodogram (CP) of the residuals for the ARX_DS1_S5 model of DS1 dataset.

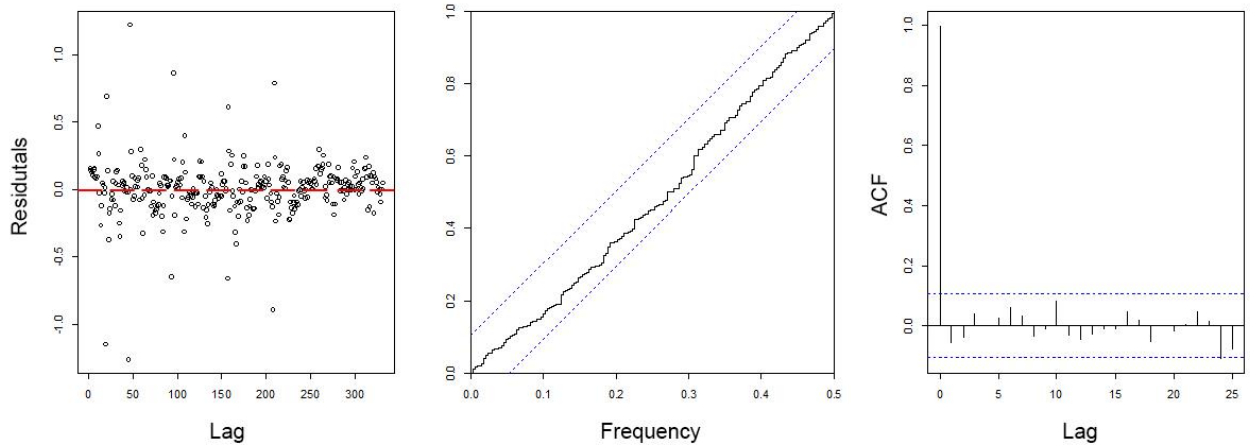


Fig. 25 The plots of residuals versus time, the auto-correlation function (ACF) and the cumulated periodogram (CP) of the residuals for the ARX_DS2_S5 model of DS2 dataset.

Nomenclature

Abbreviations

ARX	auto-regressive with eXogenous input model
GB	grey-box model
GAM	generalized addictive model
RC	resistance-capacitance
BES	building energy simulation
MPC	model predictive control

PRBS	pseudo random binary sequence
LogL	loglikelihood-value
AIC	Akaike information criterion
BIC	Bayesian information criterion
ACF	auto-correlation function
CP	cumulated periodogram
HLC	heat loss coefficient

Mathematical notation

$y(t)$	output variable of ARX model
$u(t)$	input variable of ARX model
$e(t)$	equation error
q^{-1}	backward shift operator
$Bs_{i,m}$	the ith B-spline of order m
ϕ_i	scaling factor of the ith B-spline
S	spline function
ω	Wiener process
σ_w	standard deviation of Wiener process
e	measurement error

Physical parameters

gA	solar aperture [m^2]
gA_{dni}	solar aperture tailored to direct normal radiation [m^2]
GHI	global horizontal irradiance [W/m^2]
DNI	direct normal irradiance [W/m^2]
DHI	diffuse horizontal irradiance [W/m^2]
Sg	total solar gain [W]
Sg_{dir}	direct solar gain [W]
Sg_{dif}	diffuse solar gain [W]
Sg_{en}	heat transfer of solar energy through opaque envelope [W]
P_h	heating input [W]
θ_z	zenith angle of the sun [°]
t	time [s]
T	temperature [°C]
Y_k	measured temperature [°C]
C	heat capacity [J/K]
R	thermal resistance [K/W]
prm	number of parameters [-]
ws	wind speed [m/s]

Subscripts

t	time
q	number of (B-splines)
k	discrete time step
a	ambient air
i	internal air
e	building envelope
ie	interaction between i and e
ea	interaction between e and a
dir	direct
dif	diffuse
dni	(tailored to) direct normal irradiance
en	(via) building envelope
ds1	dataset 1 of PSO
ds2	dataset 2 of PSO
_l1	one lag in the specific sampling time

_l2 two lags in the specific sampling time

Reference

[1] Hamilton I, Summerfield A, Oreszczyn T, Ruyssevelt P. Using epidemiological methods in energy and buildings research to achieve carbon emission targets. Energy Build 2017;154:188–97. https://doi.org/https://doi.org/10.1016/j.enbuild.2017.08.079.

[2] Hamilton IG, Summerfield AJ, Lowe R, Ruyssevelt P, Elwell CA, Oreszczyn T. Energy epidemiology: a new approach to end-use energy demand research. Build Res Inf 2013;41:482–97. https://doi.org/10.1080/09613218.2013.798142.

[3] IEA-EBC ANNEX-70. Building Energy Epidemiology: Analysis of Real Building Energy Use at Scale. 2022. https://doi.org/https://www.iea-ebc.org/projects/project?AnnexID=70.

[4] Bauwens G, Ritosa K, Roels S, Deltour J, Masy G, Senave M, et al. EBC Annex 71: Building Energy Performance Assessment Based on In-situ Measurements:Physical Parameter Identification. Leuven: 2021.

[5] Zhang X, Ritosa K, Saelens D, Roels S. Comparing statistical modeling techniques for heat loss coefficient estimation using in-situ data. J Phys Conf Ser 2021;2069:12101. https://doi.org/10.1088/1742-6596/2069/1/012101.

[6] Madsen H, Bacher P, Bauwens G, Deconinck AH, Reynders G, Roels S, et al. IEA EBC Annex 58: Report of Subtask 3, part 2: Thermal performance characterisation using time series data – statistical guidelines. 2016.

[7] Bauwens G. In Situ Testing of a Building's Overall Heat Loss Coefficient - Embedding Quasi-stationary and Dynamic Tests in a Building Physical and Statistical Framework. KU Leuven, 2015.

[8] De Coninck R, Magnusson F, Åkesson J, Helsen L. Toolbox for development and validation of grey-box building models for forecasting and control. J Build Perform Simul 2016;9:288–303. https://doi.org/10.1080/19401493.2015.1046933.

[9] Killian M, Mayer B, Kozek M. Cooperative fuzzy model predictive control for heating and cooling of buildings. Energy Build 2016;112:130–40. https://doi.org/https://doi.org/10.1016/j.enbuild.2015.12.017.

[10] Andriamamonjy A, Klein R, Saelens D. Automated grey box model implementation using BIM and Modelica. Energy Build 2019;188–189:209–25. https://doi.org/https://doi.org/10.1016/j.enbuild.2019.01.046.

[11] Reynders G, Nuytten T, Saelens D. Robustness of reduced-order models for prediction and simulation of the thermal behavior of dwellings. Proc. BS2013 13th Conf. Int. Build. Perform. Simul. Assoc. Chambéry, Fr., 2013.

[12] Cho K, Cho D. Solar Heat Gain Coefficient Analysis of a Slim-Type Double Skin Window System: Using an Experimental and a Simulation Method. Energies 2018;11. https://doi.org/10.3390/en11010115.

[13] Wright JL, Barnaby CS, Collins MR, Kotey NA. Solar Gain through Windows with Shading Devices: Simulation Versus Measurement 2009.

[14] Loutzenhiser PG, Manz H, Moosberger S, Maxwell GM. An empirical validation of window solar gain models and the associated interactions. Int J Therm Sci 2009;48:85–95. https://doi.org/https://doi.org/10.1016/j.ijthermalsci.2008.01.011.

[15] Manz H, Loutzenhiser P, Frank T, Strachan PA, Bundi R, Maxwell G. Series of experiments for empirical validation of solar gain modeling in building energy simulation codes—Experimental setup, test cell characterization, specifications and uncertainty analysis. Build Environ 2006;41:1784–97. https://doi.org/https://doi.org/10.1016/j.buildenv.2005.07.020.

[16] Picard D, Drgoňa J, Kvasnica M, Helsen L. Impact of the controller model complexity on model predictive control performance for buildings. Energy Build 2017;152:739–51. https://doi.org/https://doi.org/10.1016/j.enbuild.2017.07.027.

[17] Malkawi A, Augenbroe G. Advanced building simulation. New York,YN: Spon Press; 2004.

[18] Han JM, Choi ES, Malkawi A. CoolVox: Advanced 3D convolutional neural network models for predicting solar

radiation on building facades. Build Simul 2021. https://doi.org/10.1007/s12273-021-0837-0.

[19] Hollick FP, Gori V, Elwell CA. Thermal performance of occupied homes: A dynamic grey-box method accounting for solar gains. Energy Build 2020;208:109669. https://doi.org/https://doi.org/10.1016/j.enbuild.2019.109669.

[20] Oliveti G, Arcuri N, Bruno R, De Simone M. An accurate calculation model of solar heat gain through glazed surfaces. Energy Build 2011;43:269–74. https://doi.org/https://doi.org/10.1016/j.enbuild.2010.11.009.

[21] Senave M, Roels S, Verbeke S, Lambie E, Saelens D. Sensitivity of characterizing the heat loss coefficient through on-board monitoring: A case study analysis. Energies 2019. https://doi.org/10.3390/en12173322.

[22] Jiménez MJ, Madsen H. Models for describing the thermal characteristics of building components. Build Environ 2008;43:152–62. https://doi.org/https://doi.org/10.1016/j.buildenv.2006.10.029.

[23] Senave M, Roels S, Reynders G, Verbeke S, Saelens D. Assessment of data analysis methods to identify the heat loss coefficient from on-board monitoring data. Energy Build 2020;209:109706. https://doi.org/https://doi.org/10.1016/j.enbuild.2019.109706.

[24] Senave M, Reynders G, Bacher P, Roels S, Verbeke S, Saelens D. Towards the characterization of the heat loss coefficient via on-board monitoring: Physical interpretation of ARX model coefficients. Energy Build 2019;195:180–94. https://doi.org/https://doi.org/10.1016/j.enbuild.2019.05.001.

[25] Olazo-Gómez Y, Herrada H, Castaño S, Arce J, Xamán JP, Jiménez MJ. Data-Based RC Dynamic Modelling to Assessing the In-Situ Thermal Performance of Buildings. Analysis of Several Key Aspects in a Simplified Reference Case toward the Application at On-Board Monitoring Level. Energies 2020;13. https://doi.org/10.3390/en13184800.

[26] Díaz-Hernández HP, Torres-Hernández PR, Aguilar-Castro KM, Macias-Melo E V, Jiménez MJ. Data-Based RC Dynamic Modelling Incorporating Physical Criteria to Obtain the HLC of In-Use Buildings: Application to a Case Study. Energies 2020;13. https://doi.org/10.3390/en13020313.

[27] Sourbron M, Verhelst C, Helsen L. Building models for model predictive control of office buildings with concrete core activation. J Build Perform Simul 2013;6:175–98. https://doi.org/10.1080/19401493.2012.680497.

[28] Erfani Beyzaee A, Yu X, Kull TM, Bacher P, Jafarinejad T, Roels S, et al. Analysis of the impact of predictive models on the quality of the model predictive control for an experimental building. Proc Build Simul 2021 2021.

[29] Lambie E. Evaluation of the Impact of Energy Renovation Measures: Experimental Study of Residential Buildings. Eval van Impact van Energ Renov Exp Stud van Resid Gebouw 2021.

[30] Senave M. Characterization of the Heat Loss Coefficient of Residential Buildings Based on In-Use Monitoring Data. Karakterisering van Warmteverliescoëfficiënt van Resid Gebouw Op Basis van Metingen Tijdens Gebr 2019.

[31] Stamp SF. Assessing uncertainty in co-heating tests: Calibrating a whole building steady state heat loss measurement method 2016.

[32] Hollick FP. Developing a dynamic method to assess the thermal performance of occupied dwellings 2020.

[33] Rasmussen C, Frölke L, Bacher P, Madsen H, Rode C. Semi-parametric modelling of sun position dependent solar gain using B-splines in grey-box models. Sol Energy 2020;195:249–58. https://doi.org/10.1016/J.SOLENER.2019.11.023.

[34] Thilker CA, Bacher P, Bergsteinsson HG, Junker RG, Cali D, Madsen H. Non-linear grey-box modelling for heat dynamics of buildings. Energy Build 2021;252:111457. https://doi.org/https://doi.org/10.1016/j.enbuild.2021.111457.

[35] Zhang X, Rasmussen C, Saelens D, Roels S. Time-dependent solar aperture estimation of a building: comparing grey-box and white-box approaches. Renew Sustain Energy Rev 2022;161:112337. https://doi.org/https://doi.org/10.1016/j.rser.2022.112337.

[36] Lambie E, Saelens D. Identification of the Building Envelope Performance of a Residential Building: A Case Study. Energies 2020;13. https://doi.org/10.3390/en13102469.

[37] Senave M, Reynders G, Sodagar B, Verbeke S, Saelens D. Mapping the pitfalls in the characterisation of the heat loss coefficient from on-board monitoring data using ARX models. Energy Build 2019;197:214–28. https://doi.org/https://doi.org/10.1016/j.enbuild.2019.05.047.

[38] Senave M, Roels S, Verbeke S, Saelens D. Analysis of the influence of the definition of the interior dwelling temperature on the characterization of the heat loss coefficient via on-board monitoring. Energy Build 2020;215:109860. https://doi.org/https://doi.org/10.1016/j.enbuild.2020.109860.

[39] Li Y, O'Neill Z, Zhang L, Chen J, Im P, DeGraw J. Grey-box modeling and application for building energy simulations - A critical review. Renew Sustain Energy Rev 2021;146:111174. https://doi.org/https://doi.org/10.1016/j.rser.2021.111174.

[40] Ljung L. System identification : theory for the user. 2nd ed. Englewood Cliffs (N.J.) : Prentice-Hall; 1999.

[41] Ferracuti F, Fonti A, Ciabattoni L, Pizzuti S, Arteconi A, Helsen L, et al. Data-driven models for short-term thermal behaviour prediction in real buildings. Appl Energy 2017;204:1375–87. https://doi.org/https://doi.org/10.1016/j.apenergy.2017.05.015.

[42] Niu J, Tian Z, Lu Y, Zhao H. Flexible dispatch of a building energy system using building thermal storage and battery energy storage. Appl Energy 2019;243:274–87. https://doi.org/https://doi.org/10.1016/j.apenergy.2019.03.187.

[43] François A, Ibos L, Feuillet V, Meulemans J. Estimation of the thermal resistance of a building wall with inverse techniques based on rapid active in situ measurements and white-box or ARX black-box models. Energy Build 2020;226:110346. https://doi.org/https://doi.org/10.1016/j.enbuild.2020.110346.

[44] Roels S, Bacher P, Bauwens G, Castaño S, Jiménez MJ, Madsen H. On site characterisation of the overall heat loss coefficient: Comparison of different assessment methods by a blind validation exercise on a round robin test box. Energy Build 2017. https://doi.org/10.1016/j.enbuild.2017.08.006.

[45] Jiménez MJ, Madsen H, Andersen KK. Identification of the main thermal characteristics of building components using MATLAB. Build Environ 2008;43:170–80. https://doi.org/https://doi.org/10.1016/j.buildenv.2006.10.030.

[46] Naveros I, Ghiaus C, Ruíz DP, Castaño S. Physical parameters identification of walls using ARX models obtained by deduction. Energy Build 2015;108:317–29. https://doi.org/https://doi.org/10.1016/j.enbuild.2015.09.021.

[47] Bauwens G, Roels S. Co-heating test: A state-of-the-art. Energy Build 2014;82:163–72. https://doi.org/https://doi.org/10.1016/j.enbuild.2014.04.039.

[48] ISO. ISO 52016-1:2017. Energy Perform Build Needs Heat Cool Intern Temp Sensib Latent Heat Load—Part 1 2017.

[49] Kontoleon KJ. Glazing solar heat gain analysis and optimization at varying orientations and placements in aspect of distributed radiation at the interior surfaces. Appl Energy 2015;144:152–64. https://doi.org/https://doi.org/10.1016/j.apenergy.2015.01.087.

[50] Hu M, Xiao F, Wang L. Investigation of demand response potentials of residential air conditioners in smart grids using grey-box room thermal model. Appl Energy 2017;207:324–35. https://doi.org/https://doi.org/10.1016/j.apenergy.2017.05.099.

[51] Ross N. Generalized Additive Models in R n.d. https://noamross.github.io/gams-in-r-course/chapter1.

[52] Prautzsch H, Boehm W, Paluszny M. B-spline representation BT - Bézier and B-Spline Techniques. In: Prautzsch H, Boehm W, Paluszny M, editors., Berlin, Heidelberg: Springer Berlin Heidelberg; 2002, p. 59–75. https://doi.org/10.1007/978-3-662-04919-8_5.

[53] Schoenberg IJ. Contributions to the problem of approximation of equidistant data by analytic functions. Part B. On the problem of osculatory interpolation. A second class of analytic approximation formulae. Q Appl Math 1946;4:112–41. https://doi.org/10.1090/qam/16705.

[54] Christensen O. B-splines BT - Functions, Spaces, and Expansions: Mathematical Tools in Physics and Engineering.

In: Christensen O, editor., Boston: Birkhäuser Boston; 2010, p. 203–14. https://doi.org/10.1007/978-0-8176-4980-7_10.

[55] Rouchier S. Building energy statistical modelling. 2021.

[56] Kialashaki A, Reisel JR. Modeling of the energy demand of the residential sector in the United States using regression models and artificial neural networks. Appl Energy 2013;108:271–80. https://doi.org/https://doi.org/10.1016/j.apenergy.2013.03.034.

[57] Fan C, Xiao F, Wang S. Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. Appl Energy 2014;127:1–10. https://doi.org/https://doi.org/10.1016/j.apenergy.2014.04.016.

[58] Montgomery DC. Applied Statistics and Probability for Engineers 2003.

[59] Claeskens G, Hjort NL. Model selection and model averaging. Cambridge Books 2008.

[60] Ahmed AI, McLeod RS, Gustin M. Forecasting underheating in dwellings to detect excess winter mortality risks using time series models. Appl Energy 2021;286:116517. https://doi.org/https://doi.org/10.1016/j.apenergy.2021.116517.

[61] Yu X, Georges L, Imsland L. Data pre-processing and optimization techniques for stochastic and deterministic low-order grey-box models of residential buildings. Energy Build 2021;236:110775. https://doi.org/https://doi.org/10.1016/j.enbuild.2021.110775.

[62] Crabb JA, Murdoch N, Penman JM. A simplified thermal response model. Build Serv Eng Res Technol 1987;8:13–9. https://doi.org/10.1177/014362448700800104.

[63] Shamsi MH, Ali U, Mangina E, O'Donnell J. A framework for uncertainty quantification in building heat demand simulations using reduced-order grey-box energy models. Appl Energy 2020;275:115141. https://doi.org/https://doi.org/10.1016/j.apenergy.2020.115141.

[64] Bacher P, Madsen H. Identifying suitable models for the heat dynamics of buildings. Energy Build 2011. https://doi.org/10.1016/j.enbuild.2011.02.005.

[65] Deb C, Gelder L V, Spiekman M, Pandraud G, Jack R, Fitton R. Measuring the heat transfer coefficient (HTC) in buildings: A stakeholder's survey. Renew Sustain Energy Rev 2021;144:111008. https://doi.org/https://doi.org/10.1016/j.rser.2021.111008.

[66] Godfrey KR. Correlation methods. Automatica 1980;16:527–34. https://doi.org/https://doi.org/10.1016/0005-1098(80)90076-X.

[67] Madsen H, Holst J. Estimation of continuous-time models for the heat dynamics of a building. Energy Build 1995;22:67–79. https://doi.org/https://doi.org/10.1016/0378-7788(94)00904-X.

[68] Roels S, Deurinck M. The effect of a reflective underlay on the global thermal behaviour of pitched roofs. Build Environ 2011;46:134–43. https://doi.org/https://doi.org/10.1016/j.buildenv.2010.07.005.

[69] Saelens D, Roels S, Hens H. The inlet temperature as a boundary condition for multiple-skin facade modelling. Energy Build 2004;36:825–35. https://doi.org/https://doi.org/10.1016/j.enbuild.2004.01.005.

[70] R Core Team. R: A language and environment for statistical computing 2020. http://www.r-project.org/.

[71] Wickham H, Grolemund G. R for data science: import, tidy, transform, visualize, and model data. " O'Reilly Media, Inc."; 2016.

[72] Jorissen F, Reynders G, Baetens R, Picard D, Saelens D, Helsen L. Implementation and verification of the IDEAS building energy simulation library. J Build Perform Simul 2018;11:669–88. https://doi.org/10.1080/19401493.2018.1428361.

[73] Bhat HS, Kumar N. On the derivation of the bayesian information criterion. Sch Nat Sci Univ Calif 2010;99.

[74] Bünning F, Heer P, Smith RS, Lygeros J. Improved day ahead heating demand forecasting by online correction methods. Energy Build 2020;211:109821. https://doi.org/https://doi.org/10.1016/j.enbuild.2020.109821.

View publication stats