# Physics-informed machine learning for metamodeling thermal comfort in non-air-conditioned buildings

**Issa Jaffal** (✉)

*Laboratoire du froid et des systèmes énergétiques et thermiques (Lafset), Cnam, HESAM Université, Paris, France*

## Abstract

There is a growing need for accurate and interpretable machine learning models of thermal comfort in buildings. Physics-informed machine learning could address this need by adding physical consistency to such models. This paper presents metamodeling of thermal comfort in non-air-conditioned buildings using physics-informed machine learning. The studied metamodel incorporated knowledge of both quasi-steady-state heat transfer and dynamic simulation results. Adaptive thermal comfort in an office located in cold and hot European climates was studied with the number of overheating hours as index. A one-at-a-time method was used to gain knowledge from dynamic simulation with TRNSYS software. This knowledge was used to filter the training data and to choose probability distributions for metamodel forms alternative to polynomial. The response of the dynamic model was positively skewed; and thus, the symmetric logistic and hyperbolic secant distributions were inappropriate and outperformed by positively skewed distributions. Incorporating physical knowledge into the metamodel was much more effective than doubling the size of the training sample. The highly flexible Kumaraswamy distribution provided the best performance with $R^2$ equal to 0.9994 for the cold climate and 0.9975 for the hot climate. Physics-informed machine learning could combine the strength of both physics and machine learning models, and could therefore support building design with flexible, accurate and interpretable metamodels.

## Keywords

adaptive thermal comfort;
overheating;
physics-informed machine learning;
metamodel;
surrogate model;
probability distributions

## Article History

Received: 05 May 2022
Revised: 25 July 2022
Accepted: 09 August 2022

## 1 Introduction

Thermal comfort in non-air-conditioned buildings is a major concern, growing with climate change. These buildings are in free-running mode, and therefore indoor temperatures are highly dependent on internal heat gains and outdoor conditions. Thermal comfort in non-air-conditioned buildings could be assessed using adaptive thermal comfort models, based on field studies (de Dear and Brager 1998; Nicol and Humphreys 2002; Li et al. 2014). These models have been adopted in energy standards, such as the European EN 16798-1 standard (CEN 2019), to define the range of acceptable indoor conditions according to outdoor temperatures.

Dynamic simulations, based on transient heat transfer models, are used to predict thermal comfort in buildings. Since building design implies studying a very large number of configurations, the computational intensiveness of dynamic simulations becomes an issue for studies such as optimization or uncertainty quantification.

Steady-state energy balance methods offer an alternative with rapid calculation times, but numerous field studies have revealed that they are unreliable for non-air-conditioned buildings, considerably underestimating the overheating risk (Fletcher et al. 2017; Lomas and Porritt 2017; Morgan et al. 2017). Therefore, rapid and reliable methods for assessing thermal comfort are needed to support the design of non-air-conditioned buildings.

Machine learning offers great opportunities to address such a need by building mathematical models to data. However, despite their success in various fields of engineering and science, conventional machine learning approaches lack interpretability and physical consistency. This has led to the emergence of the new modeling paradigm of physics-informed machine learning, also referred as knowledge-guided

E-mail: issa.jaffal@lecnam.net

machine learning, in which prior knowledge of physical principles is explicitly incorporated into models (von Rueden et al. 2019; Karniadakis et al. 2021).

Knowledge could be represented in machine learning in several ways, such as algebraic and differential equations, logic rules, and simulations results (von Rueden et al. 2019). Physics-informed machine learning is a rapidly emerging field that has already shown great promise in various areas; for instance, in solving heat transfer partial differential equations using neural networks (Zobeiry and Humfeld 2021), in climate modeling by enforcing conservation of energy in neural networks (Beucler et al., 2019), by incorporating physical principles and constraints for predicting streamflow from weather variables (Khandelwal et al. 2020).

Physics-informed machine learning models for building energy performance have been recently presented. A physics-constrained recurrent neural network (RNN) model was developed to study the thermal dynamics of buildings (Drgoňa et al. 2021). Its predictions were found to be significantly better compared to unconstrained RNN models. A physics-informed ARMAX model was developed for model predictive control (Bünning et al. 2022). The model had lower computational requirements and better accuracy compared to random forests and input convex neural networks models.

Metamodels, also referred to as surrogate models, are machine learning models that approximate simulation models. By approximating building dynamic models, metamodels could be used for building design with very low computational requirements. Polynomial regressions provide transparent metamodels, that are computationally cheaper and easier to interpret than other machine learning techniques such as artificial neuronal networks, support vector machines and kriging (Simpson et al. 2001; Li et al. 2010).

Several metamodels have been proposed to approximate dynamic models in studying thermal comfort in non-air-conditioned buildings. The number of overheating hours was metamodeled using polynomial regression, multivariate adaptive regression splines (MARS), kriging, radial basis function networks and artificial neural networks (ANN) (van Gelder et al. 2014). The same index was also metamodeled using artificial neural networks and radial basis functions (Symonds et al. 2015). Metamodels were used to study the proportion of time with overheating through support vector regression (Rackes et al. 2016), along with the proportion of time with acceptable thermal comfort using linear regression and MARS metamodels (Chen et al. 2017). Weighted temperature excess hours were also studied through linear regression metamodels (Breesch and Janssens 2010). Furthermore, degree-hours of thermal discomfort were studied using linear regression metamodels (Rossi et al. 2019).

Building optimization, which is usually a very time-intensive process, could be significantly more efficient when based on metamodels. An optimization of thermal comfort and energy consumption of a residential house, which would have not been feasible without metamodeling, was achieved with a drastic time reduction using an artificial neural network metamodel and a multiobjective genetic algorithm (NSGA-II) (Magnier and Haghighat 2010). Support vector machine metamodels provided results equivalent to those of dynamic simulation with lower computational efforts when thermal comfort and energy consumption of a building was optimized using various cost functions and optimization algorithms (Eisenhower et al. 2012).

Thus, many studies have been conducted to metamodel thermal comfort in non-air-conditioned buildings, highlighting the important role of metamodels for building design and optimization. The proposed metamodels, however, were based on generic metamodeling techniques (polynomial regression, support vector regression, artificial neural networks, etc.), without incorporating knowledge of building physics. This highlights the need to develop metamodels based on physics-informed machine learning with better physical consistency, interpretability and generalizability. Finally, the reliability of such metamodels for non-air-conditioned buildings, which have high indoor temperature variations, is still an issue.

Based on the knowledge of quasi-steady-state heat transfer in buildings, a physics-informed polynomial metamodel for assessing thermal comfort in non-air-conditioned buildings was presented by Jaffal et al. (2020). Incorporating this knowledge of heat transfer provided a flexible and transparent polynomial metamodel with low computational cost.

The accuracy of this metamodel, however, was dependent on the thermal comfort index. It had good fit with thermal comfort indices that are consistently sensitive to factors influencing thermal comfort, e.g. with maximum indoor temperature. Conversely, it was not reliable when the sensitivity was not consistent, notably with the number of overheating hours, in particular for cold climates where overheating is low. To overcome this problem, an interpolation method was developed in the same study.

To improve the reliability of the mentioned metamodel while maintaining its transparency and computational efficiency, metamodeling of thermal comfort was here based on knowledge from dynamic simulation, along with quasi-steady-state heat transfer. The considered thermal comfort index was the number of overheating hours $NH_o$, for which the metamodel form was polynomial or cumulative distribution functions (CDF) of probability distributions.

Based on the information obtained from dynamic simulations, the data used to train the metamodel were filtered and probability distributions chosen for the metamodel form. The metamodel was then trained and its accuracy with various forms studied.

## 2 Methods

### 2.1 Metamodeling

This study uses physics-informed machine learning to metamodel thermal comfort in non-air-conditioned buildings. The basic metamodel form is a polynomial that explicitly incorporates knowledge of heat transfer under quasi-steady-state conditions (Jaffal et al. 2020).

With the number of overheating hours $NH_{\text{o}}$ as a response variable, this metamodel is given by

$$NH_{\text{o}} = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n} a_{ii} x_i^2 + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_{ij} x_i x_j \tag{1}$$

where $x_i$ and $x_j$ are the metamodel features, and $a_0$, $a_i$, $a_{ii}$ and $a_{ij}$ are the coefficients to be determined from dynamic simulations. A feature $x_i$ is related to quasi-steady-state heat transfer. It is equal to a coefficient of heat transfer $H_i$ (W K$^{-1}$) for heat transfer by transmission or ventilation and to a seasonal heat quantity $Q_i$ (kWh) for internal or solar heat gains.

This polynomial metamodel was tested on an office for the cold climate of Helsinki and the hot climate of Athens with five thermal comfort indices (Jaffal et al. 2020). It was trained with a low computational cost (57 runs) using a Box-Behnken design. Good fit was obtained when an index was consistently sensitive to the influencing factors, e.g. with maximum indoor temperature ($R^2 > 0.99$). However, the metamodel was not reliable when the sensitivity was not consistent, notably with the number of overheating hours $NH_{\text{o}}$, and particularly for Helsinki with a heavy thermal mass.

To address this issue, additional physical knowledge obtained from dynamic simulation, based on dynamic heat transfer modeling, was here incorporated in the metamodel with the aim of improving the its fit, while maintaining its transparency and computational efficiency. Based on information obtained from dynamic simulations, training data were filtered and alternative metamodel forms were proposed based on a CDF of probability distributions. This avoided adding higher-order polynomial terms or using alternative machine learning models, which could be hard to interpret and decrease computational efficiency.

The response variable $NH_{\text{o}}$ is bounded between a lower (no overheating) and an upper limit (overheating throughout the occupation period). It is more convenient to work with the proportion of time with overheating $PT_{\text{o}}$, equal to the ratio between $NH_{\text{o}}$ and the total number of occupation hours $NT$, since it is bounded between 0 and 1. Therefore, the metamodel of $PT_{\text{o}}$ could have the form of a CDF of a probability distribution.

For the polynomial form, the right side of Eq. (1) can be replaced by a scalar $z$. For CDF forms, $PT_{\text{o}}$ could be expressed as

$$PT_{\text{o}} = F(z) \tag{2}$$

where $F(z)$ is a CDF of a given probability distribution. The choice of appropriate distributions for the metamodel was informed by dynamic simulations results.

The inverse of $F(z)$, $F^{-1}(z)$, is expressed as

$$F^{-1}(z) = z = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n} a_{ii} x_i^2 + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_{ij} x_i x_j \tag{3}$$

$F^{-1}(z)$ is a monotonic function called the link function. It leads to an ordinary second-order polynomial metamodel with a transformed response variable. The coefficients of this metamodel could be determined from dynamic simulations in a similar manner to those of Eq. (1).

For example, using the CDF of the commonly used logistic distribution, the metamodel is given by

$$PT_{\text{o}} = F(z) = \frac{1}{1+\mathrm{e}^{-z}} \tag{4}$$

The corresponding link function $F^{-1}(z)$ is the logit link function (the inverse CDF of the logistic distribution) which is given by

$$F^{-1}(z) = \operatorname{logit}(PT_{\text{o}}) = \ln\left(\frac{PT_{\text{o}}}{1-PT_{\text{o}}}\right) \tag{5}$$

To train the metamodel, values of $PT_{\text{o}}$ are first obtained from dynamic simulation planned according to predetermined sampling. The coefficients of the polynomial metamodel of Eq. (3) are then determined with logit($PT_{\text{o}}$) as response. Finally, the metamodel of the number of overheating hours $NH_{\text{o}}$ is given by the product of the trained metamodel of $PT_{\text{o}}$ (Eq. (4)) and the total number of occupation hours $NT$ as follows:

$$NH_{\text{o}} = NT \cdot PT_{\text{o}} = \frac{NT}{1+\mathrm{e}^{-z}} \tag{6}$$

### 2.2 Case study

For comparison purposes, metamodeling was performed for adaptive thermal comfort in the same office studied in Jaffal et al. (2020) (Figure 1). The studied features were

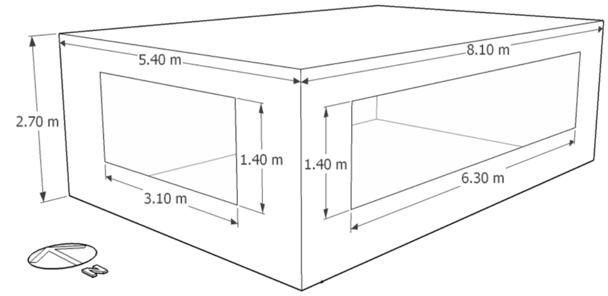


**Fig. 1** View of the office studied

related to the office facade and internal heat gains, while interior walls were assumed to be adiabatic. The office was occupied from Monday to Friday from 8:00 to 18:00 from the beginning of June until the end of September. The corresponding total number of occupation hours $NT$ was equal to 860 h.

Two typical European climates were studied: the cold climate of Helsinki, Finland, and the hot climate of Athens, Greece. Typical meteorological year (TMY2) data were used to obtain the climatic conditions. The corresponding mean outdoor air temperatures during the studied period were 14.1 °C and 25.3 °C, respectively. The number of overheating hours during the office occupation $NH_o$ was assessed with a heavy thermal mass (a concrete structure insulated from the outside) for which the metamodel of Eq. (1), trained using a Box-Behnken design, gave unreliable results for Helsinki in the mentioned study.

An adaptive model was used to assess the thermal comfort in the office as suggested in the European EN 16798-1 standard (CEN 2019). Thus, the optimal comfort temperature in the office $\theta_{comf}$ was linearly correlated to the outdoor air temperatures. Category II of thermal comfort (normal level of expectation), used for new buildings and renovations, was considered. Consequently, overheating occurred when the indoor operative temperature was higher than $\theta_{comf} + 3$ K.

Considering the office design, seven features were studied (Table 1). Each freature corresponded to a heat transfer in the office. To train the metamodel, these features were varied by varying corresponding parameters of Table 1 between lower and upper levels presented in Table 2. Coded values of these parameters (varying from −1 to 1) were used for the metamodeling. The metamodel was trained using least squares regression from dynamic simulations performed with TRNSYS software (Klein et al. 2004).

Two training samples were obtained using Latin hypercube sampling. In the first sample, LHS200, with 200 dynamic simulations performed, the range of the coded values of each parameter varied was divided into 200 intervals of length 0.01 and equal probability. To study the effect of sample size, a second sample, LHS400, was considered, with 400 dynamic simulations performed. Finally, the test sample was obtained from an additional sample of 200 runs with a random combination of the parameters of Table 2.

Prior knowledge was integrated into the metamodel, with a quasi-steady-state assumption of heat transfer in the metamodel features (Eq. (1) and Table 1), and by choosing the metamodel form and filtering the training data using information from TRNSYS dynamic simulations. Information was obtained from dynamic simulations using a one-at-a-time method in which in each parameters of Table 2 varied

**Table 1** Metamodel features and corresponding parameters varied to train the metamodel

| No. | Feature | Parameter varied |
|---|---|---|
| 1 | Heat transfer coefficient for transmission through opaque walls $H_{tr,ow}$ | Opaque wall $U$-value $U_{ow}$ (W m$^{-2}$ K$^{-1}$) |
| 2 | Heat transfer coefficient for transmission through windows $H_{tr,w}$ | Window $U$-value $U_w$ (W m$^{-2}$ K$^{-1}$) |
| 3 | Heat transfer coefficient for ventilation $H_{vent}$ | Ventilation rate $q_{v,vent}$ (m$^3$ h$^{-1}$) |
| 4 | Heat transfer coefficient for night ventilation $H_{nvent}$ | Night ventilation rate $q_{v,nvent}$ (ACH) |
| 5 | Quantity of heat due to internal gains $Q_{ig}$ | Internal heat gains during occupation $p_{ig,o}$ (W m$^{-2}$) with 0.1 $p_{ig,o}$ during inoccupation |
| 6 | Solar heat gain through the south window $Q_{so,ws}$ | Solar heat gain coefficient ($SHGC$) of the south window $SHGC_{ws}$ |
| 7 | Solar heat gain through the west window $Q_{so,ww}$ | $SHGC$ of the west window $SHGC_{ww}$ |

**Table 2** Lower and upper levels of the parameters varied for training the metamodel

| No. | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Parameter varied | $U_{ow}$ (W m$^{-2}$ K$^{-1}$) | $U_w$ (W m$^{-2}$ K$^{-1}$) | $q_{v,vent}$ (m$^3$ h$^{-1}$) | $q_{v,nvent}$ (ACH) | $p_{ig,o}$ (W m$^{-2}$) | $SHGC_{ws}$ | $SHGC_{ww}$ |
| Lower level | 0.1 | 0.7 | 100 | 0 | 15 | 0.4 | 0.4 |
| Upper level | 0.5 | 2.7 | 250 | 5 | 40 | 0.7 | 0.7 |

independently while all others were kept constant (Section 2.3).

## 2.3 Information from dynamic simulations

To obtain information from dynamic simulations, the response variable of the dynamic model, i.e. the number of overheating hours $NH_o$ given by dynamic simulation, was studied using a one-at-a-time method in which each parameter of Table 2 varied independently. The results obtained are presented in Figure 2 for Helsinki and Figure 3 for Athens, with coded values of the studied parameters (varying from −1 to 1). The central point corresponded to all the parameters at their mean values (coded values equal to zero). Eight additional simulations were conducted for each parameter varying from its lower to its upper level. Furthermore, a second-order polynomial was fitted for each parameter when at least three of the corresponding values of $NH_o$ were non-zero. Information obtained from the analysis of these results was used to support the metamodel.

For Helsinki, it is obvious that the dynamic model was not consistently sensitive to the studied parameters (Figure 2). No overheating occurred for any of the values of $U_{ow}$; when $U_w \geq -0.25$, $q_{v,vent} \geq 0$ and $q_{v,nvent} \geq -0.75$; and when $p_{ig,o} \leq 0$, $SHGC_{ws} \leq 0.5$ and $SHGC_{ww} \leq 0.75$. Therefore, the polynomial assumption of Eq. (1) could not be valid for the levels of the parameters in Table 2, even according to a single feature. This could explain why the metamodel was not reliable with $NH_o$ for the same case study using a Box-Behnken design (Jaffal et al. 2020).

Thus, to achieve good accuracy, the polynomial metamodel must be trained considering only the sensitive range of the dynamic model, i.e. when $NH_o$ is not equal to zero or the total number of occupation hours $NT$ = 860 h. Thus, the training data of samples LHS200 and LHS400 should be filtered by excluding data with $NH_o$ equal to these comfort limits. Figure 2 also reveals that, when the dynamic model was sensitive, second-order polynomials could accurately associate $NH_o$ with each studied parameter with $R^2 > 0.989$.

Furthermore, the dynamic model had low sensitivity when $NH_o$ was low. For instance, when $p_{ig,o}$ varied from 0.25 to 0.5, $\Delta NH_o/\Delta p_{ig,o}$ = 164 h; from 0.5 to 0.75, $\Delta NH_o/\Delta p_{ig,o}$ = 648 h; and from 0.75 to 1, $\Delta NH_o/\Delta p_{ig,o}$ = 1064 h. This illustrates the highly nonlinear variation of thermal conditions in a non-air-conditioned building; and thus, linear metamodels were not considered in this study. Moreover, for $q_{v,vent} = -0.5$ and $p_{ig,o} = 0.5$, when overheating was low, the values of $NH_o$ obtained were below the fitted polynomial curves. This suggested that the dynamic model had distinct behavior close to zero.

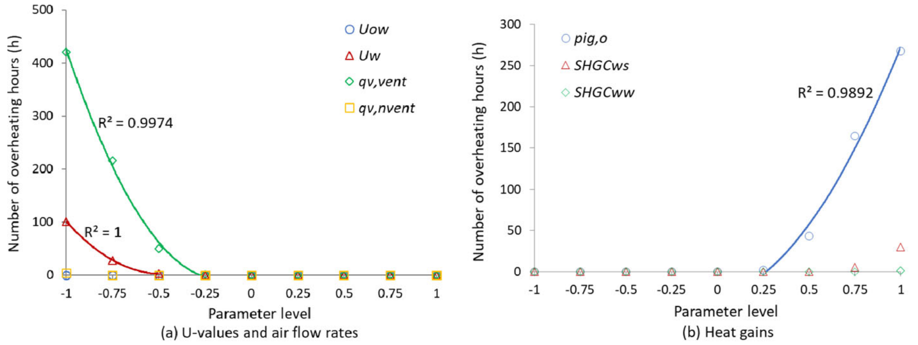


**Fig. 2** Number of overheating hours according to the parameters varied to train the metamodel for Helsinki: (a) $U$-values $U_{ow}$ and $U_w$, and airflow rates $q_{v,vent}$ and $q_{v,nvent}$; and (b) internal heat gains $p_{ig,o}$, and solar hat gain coefficients $SHGC_{ws}$ and $SHGC_{ww}$

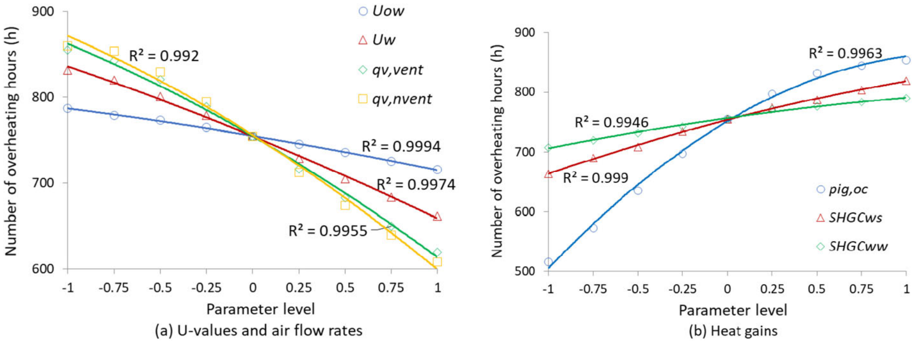


**Fig. 3** Number of overheating hours according to the parameters varied to train the metamodel for Athens: (a) $U$-values $U_{ow}$ and $U_w$, and airflow rates $q_{v,vent}$ and $q_{v,nvent}$; and (b) internal heat gains $p_{ig,o}$, and Solar Hat Gain Coefficients $SHGC_{ws}$ and $SHGC_{ww}$

For Athens, the dynamic model was consistently sensitive to the parameters studied with significant overheating observed. Overheating occurred throughout the occupation period for $q_{\mathrm{v,nvent}} = -1$ and $p_{\mathrm{ig,o}} = 1$. Thus, the training data should be filtered by excluding data with $NH_{\mathrm{o}} = NT$. The fitted curves show that $NH_{\mathrm{o}}$ could be accurately associated with each parameter using second-order polynomials with $R^2 > 0.992$. However, $NH_{\mathrm{o}}$ reached $NT$ slowly, in particular when $q_{\mathrm{v,nvent}}$ was close to −1 and $p_{\mathrm{ig,o}}$ close to 1. For instance, when $p_{\mathrm{ig,o}}$ varied from 0.25 to 0.5, $\Delta NH_{\mathrm{o}}/\Delta p_{\mathrm{ig,o}} = 140$ h; from 0.5 to 0.75, $\Delta NH_{\mathrm{o}}/\Delta p_{\mathrm{ig,o}} = 52$ h; and from 0.75 to 1, $\Delta NH_{\mathrm{o}}/\Delta p_{\mathrm{ig,o}} = 36$ h. Therefore, the dynamic model had also distinct behavior close to $NT$.

The results also show that $NH_{\mathrm{o}}$ was monotonic, increasing according to $U$-values and airflow rates and decreasing with heat gains. Moreover, $NH_{\mathrm{o}}$ increased more rapidly at low levels (in Helsinki) than when approaching $NT$ (in Athens). This suggests that the dynamic model response could be positively skewed. Thus, symmetric probability distributions may not be appropriate for the metamodel form; and positively skewed distributions may be more fitting. This is not conclusive, however, because a building's thermal behavior differs from one climate to another. Several probability distributions that may be appropriate are presented in Section 2.5.

## 2.4 Training data filtering

Based on the information obtained from dynamic simulations, the training data of both the LHS200 and LHS400 samples were filtered to consider only the sensitive range of the dynamic model in the training, i.e. by excluding data with response $NH_{\mathrm{o}}$ equal to the comfort limits ($NH_{\mathrm{o}} = 0$ or $NH_{\mathrm{o}} = NT = 860$ h). Moreover, to consider the distinct behavior of the dynamic model close to these comfort limits, additional training data filtering was conducted by excluding data with $NH_{\mathrm{o}}$ close to these latter.

Thus, several data were used to train the metamodel, in each a data point was considered when $NH_{\mathrm{o}}$, given by dynamic simulation, was in the following range:

$$\frac{NT(100\% - DP)}{2} < NH_{\mathrm{o}} < \frac{NT(100\% + DP)}{2} \tag{7}$$

where $NT$ is the total number of occupation hours ($NT = 860$ h) and $DP$ is the domain percentage considered for the training data (%).

For training the polynomial metamodel, six domain percentages $DP$ were studied for each climate and each sample, ranging from 90% to 100%, in increments of 2%. These are presented in Table 3 with the corresponding domains of $NH_{\mathrm{o}}$ considered for the training. For simplicity, training of the metamodel with probability distribution forms was conducted with only three values of $DP$: 90%, 95% and 100%.

**Table 3** Domain percentages $DP$ and corresponding domains of $NH_{\mathrm{o}}$ considered for the training of the polynomial metamodel

| Domain percentage $DP$ | Training domain (h) |
|---|---|
| 90% | ]43.0, 817.0[ |
| 92% | ]34.4, 825.6[ |
| 94% | ]25.8, 834.2[ |
| 96% | ]17.2, 842.8[ |
| 98% | ]8.6, 851.4[ |
| 100% | ]0, 860[ |

## 2.5 Probability distributions studied

An abundance of probability distributions is available in the literature. In particular, a probability distribution could be a member of the exponential family (normal, logistic, gamma, etc.). The choice of distributions studied was here informed by dynamic simulation results with the aim of making the metamodel response consistent with that of the dynamic model. To preserve metamodel simplicity and computational efficiency, simple and well-known distributions were selected.

Both symmetric and asymmetric distributions were studied and the premise that the response variable $NH_{\mathrm{o}}$ is positively skewed was tested. The CDFs of the studied distributions are presented with their link functions in Table 4, and CDF curves are illustrated in Figure 4.

The symmetric logistic distribution (Eq. (4) and Figure 4(a)) plays an important role in the statistical literature and it is the most common in regression when

**Table 4** Cumulative distribution functions (CDF) of the probability distributions studied with their corresponding link functions (inverse CDF)

| Distribution | CDF | Link function |
|---|---|---|
| Logistic | $PT_{\mathrm{o}} = \frac{1}{1+\mathrm{e}^{-z}}$ | $z = \mathrm{logit}(PT_{\mathrm{o}}) = \ln\left(\frac{PT_{\mathrm{o}}}{1-PT_{\mathrm{o}}}\right)$ |
| Hyperbolic secant | $PT_{\mathrm{o}} = \frac{2}{\pi}\arctan\left(\mathrm{e}^{\frac{\pi}{2}z}\right)$ | $z = \frac{2}{\pi}\ln\left(\tan\left(\frac{2}{\pi}PT_{\mathrm{o}}\right)\right)$ |
| Gumbel | $PT_{\mathrm{o}} = \mathrm{e}^{-\mathrm{e}^{-z}}$ | $z = -\ln(-\ln(PT_{\mathrm{o}}))$ |
| Complementary Gumbel | $PT_{\mathrm{o}} = 1-\mathrm{e}^{-\mathrm{e}^{z}}$ | $z = \ln(-\ln(1-PT_{\mathrm{o}}))$ |
| Skew logistic | $PT_{\mathrm{o}} = \frac{1}{(1+\mathrm{e}^{-z})^{\alpha}}$ | $z = \ln\left(\frac{PT_{\mathrm{o}}^{1/\alpha}}{1-PT_{\mathrm{o}}^{1/\alpha}}\right)$ |
| Fréchet | $PT_{\mathrm{o}} = \mathrm{e}^{-z^{-\alpha}}$ | $z = -\ln(-PT_{\mathrm{o}}^{1/\alpha})$ |
| Kumaraswamy | $PT_{\mathrm{o}} = 1-(1-z^{\alpha})^{\beta}$ | $z = \left(1-(1-PT_{\mathrm{o}})^{\frac{1}{\beta}}\right)^{\frac{1}{\alpha}}$ |

the response is bounded between lower and upper limits. Its CDF is comparable to that of the standard normal distribution, but it is generally preferable because of its simplicity and more readily interpretable results. It has been extensively used when studying building energy performance, notably thermal comfort such as in Haldi and Robinson (2008), and in Takasu et al. (2017).

The hyperbolic secant distribution (Figure 4(a)) is an alternative to the logistic distribution, although it is much less known. It is also symmetric, but it is more leptokurtic, i.e. it has heavier tails and a higher peak at the mean. Its CDF is given by

$$F(z)=\frac{2}{\pi}\arctan\left(e^{\frac{\pi}{2}z}\right) \tag{8}$$

Considering that the response $NH_o$ of the dynamic model appeared to be positively skewed, logistic and hyperbolic secant distributions may not guarantee good fit, and positively skewed distributions would provide better fit. This was first tested with the metamodel having forms of asymmetric Gumbel distribution together with its complementary (Figure 4(a)). The Gumbel distribution is a special case of the generalized extreme value distribution. Its CDF is expressed as follows:

$$F(z)=e^{-e^{-z}} \tag{9}$$

and that of its complementary as follows:

$$F'(z)=1-e^{-e^{z}} \tag{10}$$

The Gumbel distribution is positively skewed. It has light lower tail and upper tail similar to that of the logistic distribution. Its complementary is consequently negatively skewed with a light upper tail. The Gumbel distribution is widely used in various areas of engineering, notably to study earthquakes, flood frequency, rainfall and wind speed (Kotz and Nadarajah 2000).

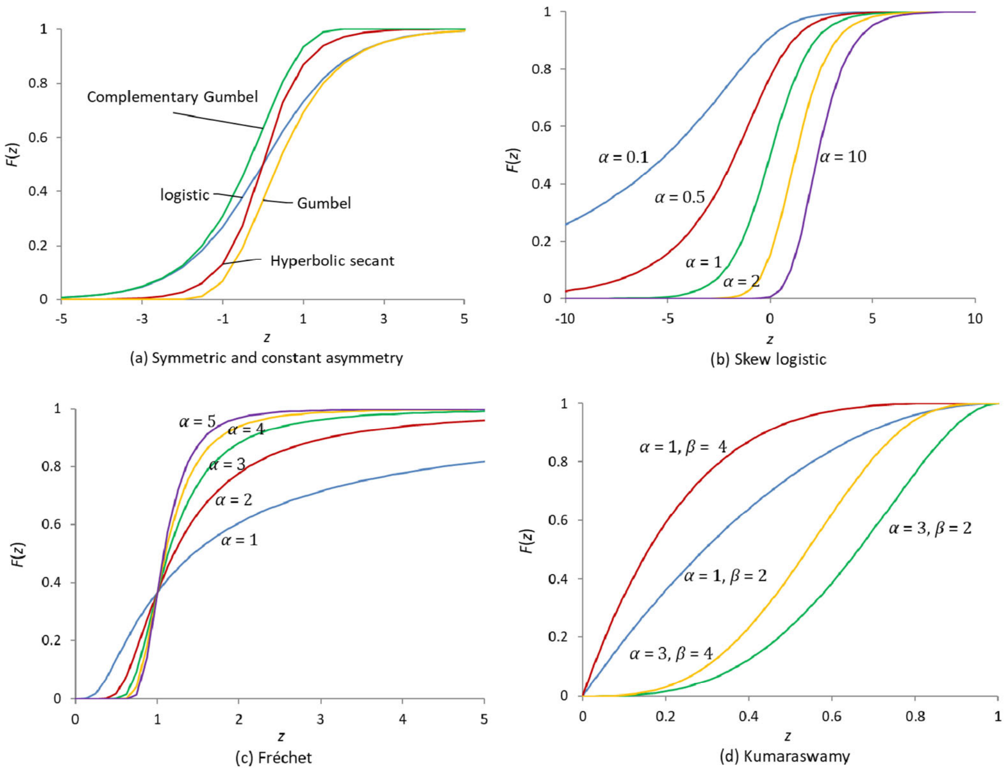
However, the Gumbel distribution and its complementary have constant positive and negative skewness, respectively, and metamodeling would be better with flexible distributions. A simple flexible distribution is the skew logistic (generalized logistic type I) which is a generalization of the logistic distribution (Nagler 1994). Its CDF is given by

$$F(z)=\frac{1}{\left(1+e^{-z}\right)^{\alpha}} \tag{11}$$

where $\alpha > 0$ is the shape (skewness) parameter.

This distribution can have different shapes with negative skewness ($\alpha < 1$) or positive skewness ($\alpha > 1$); and corresponds to the standard logistic distribution for $\alpha = 1$ (Figure 4(b)).

The Fréchet distribution (Fréchet 1928) was also considered. Like the Gumbel distribution, it is a special



(a) Symmetric and constant asymmetry

(b) Skew logistic

(c) Fréchet

(d) Kumaraswamy

**Fig. 4** Cumulative distribution functions (CDF) of the probability distributions: (a) symmetric (logistic and hyperbolic secant) and constant asymmetry (Gumbel and its complementary), (b) skew logistic, (c) Fréchet, and (d) Kumaraswamy

case of the generalized extreme value distribution. It is also positively skewed, but it is characterized by a heavy upper tail (Figure 4(c)). Its CDF is given by

$$F(x) = \mathrm{e}^{-z^{-\alpha}}, \quad z > 0 \tag{12}$$

where $\alpha > 0$ is the shape parameter.

The Fréchet distribution has many applications including earthquakes, rainfall and wind speed (Kotz and Nadarajah 2000; Harlow 2002).

Finally, the two-parameter Kumaraswamy distribution was studied (Kumaraswamy 1980). This is a double-bounded distribution defined on the finite interval [0, 1]. Its CDF is expressed as

$$F(z) = 1 - \left(1 - z^{\alpha}\right)^{\beta}, \quad 0 < z < 1 \tag{13}$$

where $\alpha > 0$ and $\beta > 0$ are the parameters that govern the shape of the CDF as exemplified in Figure 4(d).

This distribution was created by Kumaraswamy (1980) for applications in hydrology. It is very flexible; depending on the values and $\alpha$ and $\beta$, its probability density function can be unimodal, uniantimodal, increasing, decreasing or constant, in the same way as the beta distribution (Jones 2009), but it is considerably simpler. The Kumaraswamy distribution has received considerable interest in hydrology. Despite its numerous advantages however, it has been little explored in the literature, as pointed out by Jones (2009).

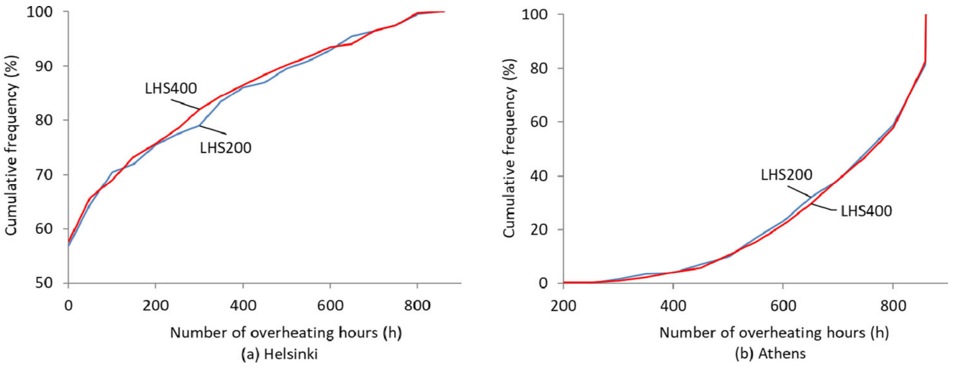
## 3 Results and discussion

### 3.1 Dynamic simulation results

Two Latin hypercube samples, LHS200 and LHS400, were used to train the metamodel. The cumulative frequencies of the number of overheating hours $NH_o$, as given by the results of the corresponding dynamic simulations, are illustrated in Figure 5. The simulation numbers used to train the polynomial metamodel versus the domain percentage *DP* are illustrated in Figure 6.

For Helsinki, with both samples, more than the half the data points had $NH_o$ outside the domains considered; and were thus excluded from the training data. For instance, with the LHS200 sample, 64% of the 200 data points had $NH_o \leq 43$ h and were thus excluded for $DP = 90\%$; and 57% of them had $NH_o = 0$ and were excluded for $DP = 100\%$. The maximum value of $NH_o$ in the same sample was 846 h with only one data point excluded for $DP \leq 98\%$. Thus, out of 200 simulations, training was performed with a minimum of 71 for $DP = 90\%$ and a maximum of 86 for $DP = 100\%$.

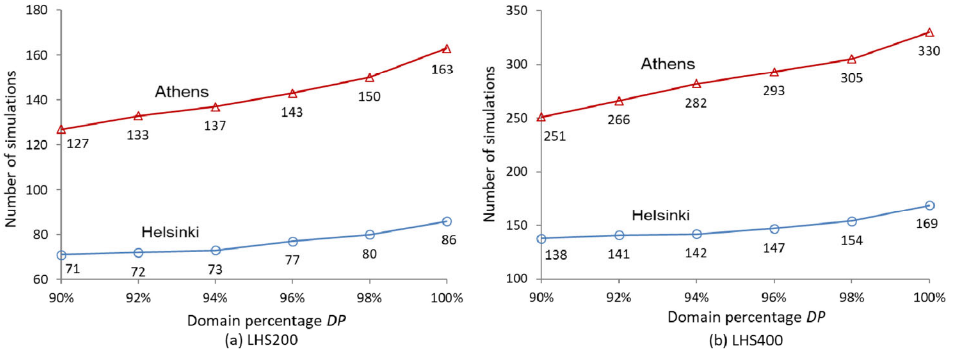
For Athens, $NH_o$ was higher than 277 h for the LHS200



(a) Helsinki

(b) Athens

**Fig. 5** Cumulative frequencies of the number of overheating hours $NH_o$ as given by the results of LHS200 an LHS400 samples: (a) Helsinki and (b) Athens



(a) LHS200

(b) LHS400

**Fig. 6** Number of simulations used to train the polynomial metamodel for Helsinki and Athens: (a) LHS200 sample and (b) LHS400 sample

sample, and higher than 200 h for the LHS400 sample; and thus none of the data points were excluded for having $NH_o$ below the domain considered for training. Conversely, many data points were excluded for having $NH_o$ above this domain. For instance, with the LHS200 sample, 37% of the 200 dynamic simulations had $NH_o \geq 817$ h and were thus excluded for $DP$ = 90%; and 19% of them had $NH_o = NT =$ 860 h and were excluded for $DP$ = 100%. The corresponding numbers of dynamic simulations included in the training were, respectively, 127 and 163. Finally, comparable percentages of data points excluded were obtained of the LHS400 sample for both climates.

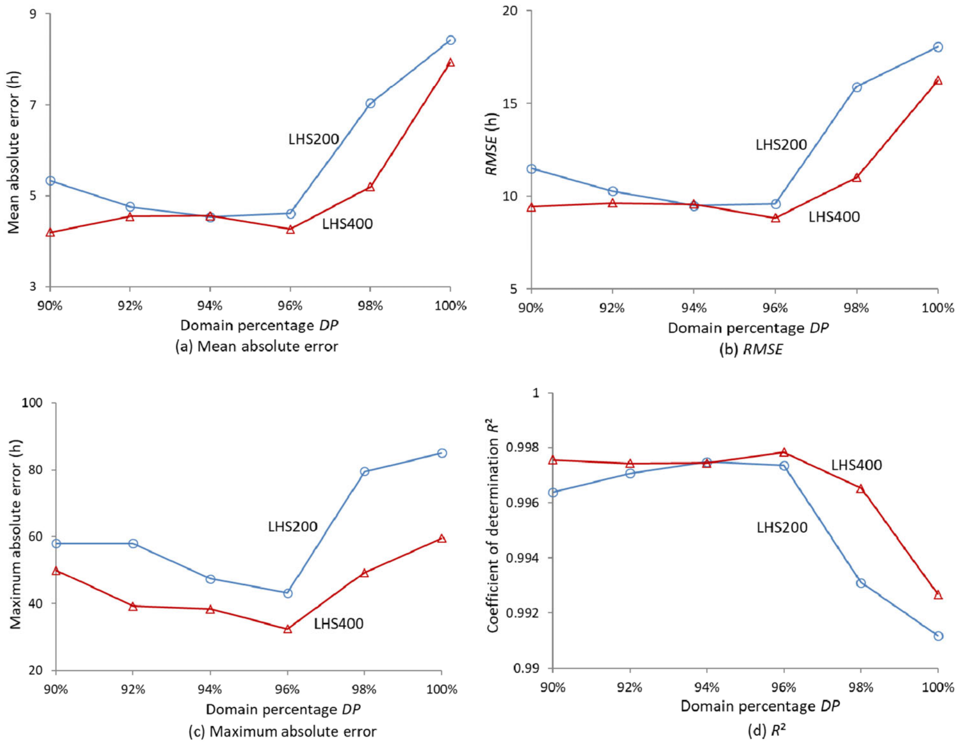
For the metamodel forms based on probability distributions, the metamodel was trained with the same data as the polynomial metamodel (LHS200 and LHS400 samples), without performing any additional dynamic simulation. Three domain percentages $DP$ were studied, 90%, 95% and 100%. For $DP$ = 95%, training with the LHS200 sample was performed with 76 data points for Helsinki, and 140 for Athens; and training with the LHS400 sample with, respectively, 143 and 290 data points.

Figure 6(a) suggests that, for Helsinki, both samples had wide ranges of $NH_o$: $0 \leq NH_o \leq 846$ h for the LHS200 sample, and $0 \leq NH_o \leq 825$ h for the LHS400 sample. Thus, all the studied probability distributions may be suitable for this cold climate. However, for Athens, with 277 h $\leq NH_o \leq$ 860 h for the LHS200 sample, and 200 h $\leq NH_o \leq$ 860 h for the LHS400 sample, the results did not cover a large part of the domain of $NH_o$ (the lowest 32% for the LHS200 sample and 23% for LHS400 sample). Thus, probability distributions with left tails were not considered for this hot climate, and only the very flexible Kumaraswamy distribution was studied.

## 3.2 Polynomial form

The results of polynomial metamodel of Eq. (1) were compared with those of dynamic simulation using a test sample of 200 runs. The corresponding mean and maximum absolute errors, $RMSE$ and coefficient of determination $R^2$ according to the domain percentage $DP$ are illustrated in Figure 7 for Helsinki and Figure 8 for Athens.

The results revealed that the domain percentage $DP$ had a more significant impact on accuracy than the sample size. For Helsinki with $DP$ = 100%, increasing the number of simulations performed from 200 to 400 simulations decreased the mean absolute error, $RMSE$ and maximum absolute error by 6%, 10% and 30% respectively, and increased $R^2$ from 0.9912 to 0.9927. Reducing the $DP$ from 100% to 96% with the LHS200 sample, reduced these errors by 45%, 47% and 49%, respectively, and increased $R^2$ to 0.9974, although the size of the corresponding training sample was also reduced from 86 to 77. A comparable improvement of accuracy was obtained for the LHS400 sample, which gave the best accuracy for Helsinki with $DP$ = 96%.



(a) Mean absolute error

(b) *RMSE*

(c) Maximum absolute error

(d) $R^2$

**Fig. 7** Errors and coefficients of determination $R^2$ of the polynomial metamodel for Helsinki according to the domain percentage: (a) mean absolute error, (b) $RMSE$, (c) maximum absolute errors and (d) coefficient of determination $R^2$

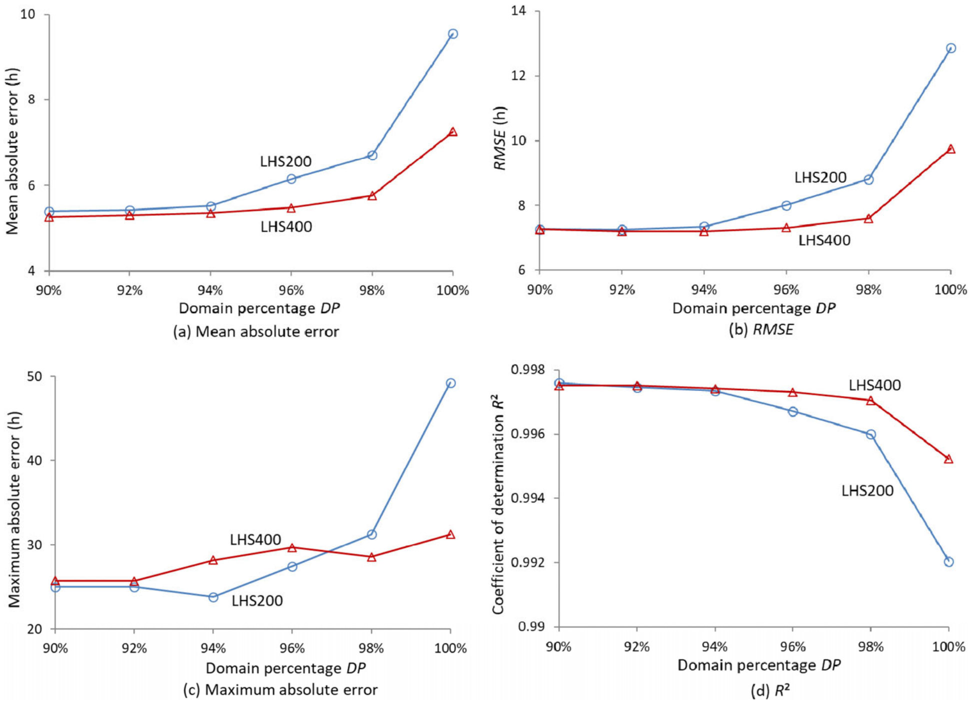


**Fig. 8** Errors and coefficients of determination $R^2$ of the polynomial metamodel for Athens according to the domain percentage: (a) mean absolute error, (b) *RMSE*, (c) maximum absolute errors and (d) coefficient of determination $R^2$

A dramatic improvement of accuracy was obtained for the same climate due to filtering the data. For the LHS200 sample, when training with *DP* = 96% instead of using all the 200 results, the mean absolute errors, *RMSE* and maximum absolute errors were reduced by 88%, 84% and 78%, respectively, and $R^2$ increased from 0.9074 to 0.9974. For the LHS400 sample, when training with *DP* = 96% instead of using all the 400 results, the corresponding error reductions, were 88%, 84% and 86%, respectively, and $R^2$ increased from 0.9219 to 0.9978.

Furthermore, the present method performed much better than the interpolation method used for the same case study (Jaffal et al. 2020), while being simpler. With the LHS400 sample and *DP* = 96%, the *RMSE* was reduced by 55% and the maximum absolute error by 46%, while $R^2$ increased from 0.9892 to 0.9978.

For Athens with *DP* = 100%, increasing the number of simulations conducted from 200 to 400 decreased mean absolute error, *RMSE* and maximum absolute error by 24%, 24% and 37%, respectively, and increased $R^2$ from 0.9920 to 0.9952. Reducing *DP* from 100% to 92% with the LHS200 sample reduced these errors by 43%, 44% and 49%, respectively, and increased $R^2$ to 0.9975. The best accuracy was with domain percentages of 90% and 92%, between which the variation of the errors and $R^2$ was not significant; and for which sample size had low impact.

Filtering the data also had a dramatic impact on accuracy for Athens. For the LHS200 sample, when training with *DP* = 92% instead of using all 200 results, the mean absolute errors, *RMSE* and maximum absolute errors were reduced by 67%, 65% and 66%, respectively, and $R^2$ increased from 0.9803 to 0.9975. For the LHS400 sample, the corresponding error reductions were 62%, 59% and 56%, respectively, while $R^2$ increased from 0.9853 to 0.9975.

## 3.3 Logistic and hyperbolic secant forms

The metamodel with probability distribution forms was trained for Helsinki using the same dynamic simulation results obtained for the polynomial metamodel, considering three domain percentages *DP*: 90%, 95% and 100%. The results are presented for the LHS400 sample (400 runs) that gave more accurate results than the LHS200 sample (200 runs). For comparison, the results of the LHS200 sample are presented in each section for configurations considered to be the best, with the $R^2$ coefficient used for the choice. For the logistic and hyperbolic secant forms (Eq. (4) and Eq. (8), respectively), errors and $R^2$ coefficients obtained with the LHS400 sample are presented in Table 5.

The logistic distribution showed better accuracy than the more leptokurtic hyperbolic secant distribution. Its best fit was for a domain percentage *DP* = 90% with $R^2$ = 0.9960, but the difference with the fit for *DP* = 95% ($R^2$ = 0.9953) was low as compared to the difference with that for *DP* = 95% and *DP* = 100% ($R^2$ = 0.9778).

The accuracy obtained with these symmetric probability

**Table 5** Errors and coefficients of determination $R^2$ of the metamodel with the logistic and hyperbolic secant forms trained with the LHS400 sample

| Metamodel form | Logistic | | | Hyperbolic secant | | |
|---|---|---|---|---|---|---|
| Domain percentage *DP* | 90% | 95% | 100% | 90% | 95% | 100% |
| Mean absolute error (h) | 7.5 | 7.4 | 13.7 | 9.2 | 9.1 | 16.9 |
| *RMSE* (h) | 12.4 | 13.2 | 30.9 | 14.8 | 15.9 | 38.7 |
| Maximum absolute error (h) | 54.8 | 59.3 | 140.4 | 60.0 | 74.6 | 175.2 |
| $R^2$ | 0.9960 | 0.9953 | 0.9778 | 0.9945 | 0.9934 | 0.9667 |

distributions, however, was no better than that achieved with the polynomial form, and as such they were not appropriate alternatives. The mentioned best accuracy of these forms had mean absolute error, *RMSE* and maximum absolute error higher by, respectively, 75%, 40% and 69% than those of the polynomial form with *DP* = 96% ($R^2$ = 0.9978).

With the LHS200 sample, the best fit was obtained for the logistic form with *DP* = 90%, with mean absolute error, *RMSE* and maximum absolute error equal to 9.1 h, 16.5 h and 74.7 h, respectively, and $R^2$ = 0.9929. Thus, although the accuracy improvement between the LHS200 and LHS400 was significant, it was again less effective than considering, for the LHS400 sample, a domain percentage of *DP* = 95% instead of *DP* = 100%.

### 3.4 Gumbel and complementary Gumbel forms

The positively skewed Gumbel distribution (Eq. (9)) provided very good results in contrast to its complementary (Eq. (10)) which is negatively skewed and was inappropriate (Table 6). This is in accordance with the results of Figures 2 and 3 in which the dynamic model response appeared to be positively skewed.

The domain percentages also had a significant impact on accuracy. For the Gumbel form, with *DP* = 95%, mean absolute error, *RMSE* and maximum absolute error were reduced by 13%, 21% and 16%, respectively, compared to those with *DP* = 100%. However, between the results for *DP* = 90% and *DP* = 95%, the difference was significant only in terms of maximum absolute error (38.7 h and 33.0 h, respectively).

The metamodel with the Gumbel form and *DP* = 95% had lower mean absolute error and *RMSE* than those of with the polynomial form and *DP* = 96%, by 18% and 24%, respectively. Its maximum absolute error, however, was not lower (33.0 h for the former and 32.5 h for the latter).

Moreover, the Gumbel form showed dramatically better accuracy when compared to the logistic form. The mean absolute error, *RMSE* and maximum absolute error of the former with *DP* = 95%, were lower by more than 40% than those of the latter with *DP* = 90%. The best accuracy of the LHS200 sample was with the Gumbel form and *DP* = 90%, with a mean absolute error, *RMSE* and maximum absolute error equal to 4.1 h, 9.1 h and 45.8 h, respectively, and $R^2$ = 0.9977. Thus, the error reduction using the LHS400 sample was significant with mean absolute error, *RMSE* and maximum absolute error reduced by 15%, 26% and 28%, respectively, with *DP* = 95%.

The asymmetry in the dynamic model results, however, could be higher or lower than that provided by the Gumbel distribution. Flexible asymmetric distributions, trained with the same samples, were studied and are presented below.

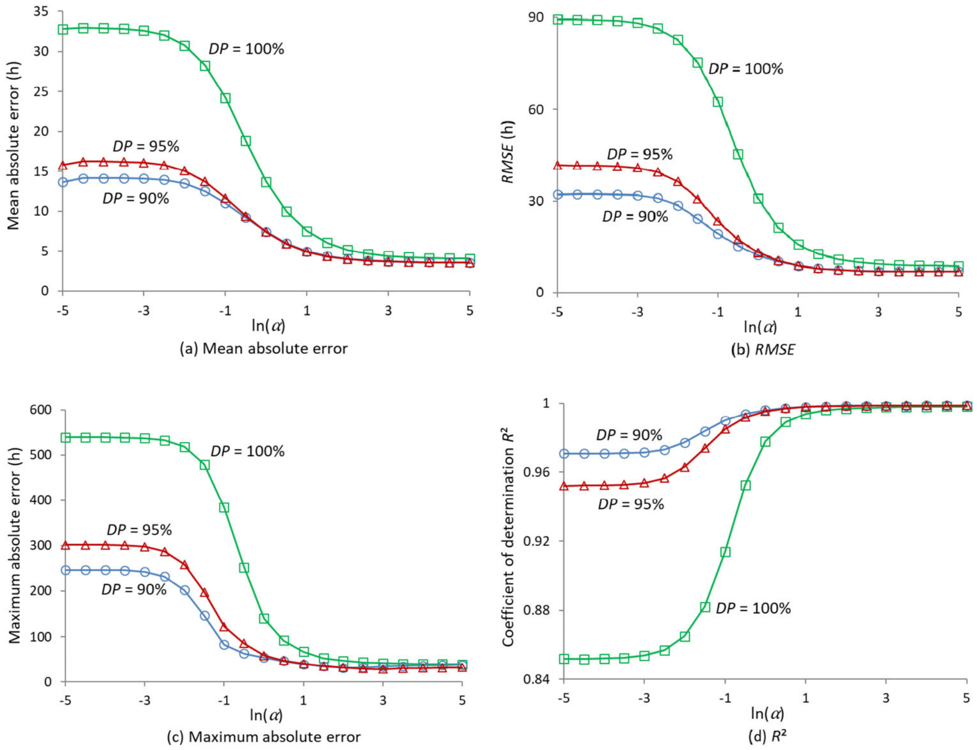
### 3.5 Skew logistic form

The first flexible distribution studied for the metamodel form was the skew logistic (Eq. (11)). This distribution could be negatively skewed ($0 < \alpha < 1$) or positively skewed ($\alpha > 1$). As expected, only positively skewed shapes were appropriate (Figure 9).

The best fit was with *DP* = 95% and $\ln(\alpha) = 5$ with a mean absolute error, *RMSE* and maximum absolute error

**Table 6** Errors and coefficients of determination $R^2$ of the metamodel with Gumbel and Complementary Gumbel forms trained with the LHS400 sample

| Metamodel form | Gumbel | | | Complementary Gumbel | | |
|---|---|---|---|---|---|---|
| Domain percentage *DP* | 90% | 95% | 100% | 90% | 95% | 100% |
| Mean absolute error (h) | 3.5 | 3.5 | 4.0 | 9.4 | 10.0 | 20.6 |
| *RMSE* (h) | 6.9 | 6.8 | 8.5 | 17.0 | 20.3 | 53.0 |
| Maximum absolute error (h) | 38.7 | 33.0 | 39.4 | 79.9 | 113.3 | 324.8 |
| $R^2$ | 0.9987 | 0.9988 | 0.9981 | 0.9922 | 0.9885 | 0.9354 |



(a) Mean absolute error

(b) *RMSE*

(c) Maximum absolute error

(d) $R^2$

**Fig. 9** Errors and coefficient of determination $R^2$ for three domain percentages *DP* according to the logarithm of the shape parameter $\alpha$ for the metamodel with the skew logistic form trained with the LHS400 sample: (a) mean absolute error, (b) *RMSE*, (c) maximum absolute error and (d) coefficient of determination $R^2$

equal to 3.5 h, 6.8 h and 32.4 h, respectively, and $R^2 = 0.9987$. Nevertheless, values of these indicators did not vary significantly after $\ln(\alpha) = 3$, i.e. $\alpha = 20$; except for the maximum absolute error, which increased with $DP = 95\%$ from 28.7 h with $\ln(\alpha) = 3$ to 32.4 h with $\ln(\alpha) = 5$.

With $\ln(\alpha) = 5$, the domain percentage $DP = 95\%$ gave mean absolute error, *RMSE* and maximum absolute error values lower by respectively 13%, 22% and 19% than those obtained with $DP = 100\%$. The difference between the results for $DP = 90\%$ and $DP = 95\%$ was significant only in terms of maximum absolute error (38.1 h for $DP = 90\%$ and 32.4 h for $DP = 95\%$). The best fit of the skew logistic form gave results similar to those of the Gumbel form (Section 3.4), the only noticeable difference being a slightly lower maximum absolute error (32.4 h for the former and 33 h for the latter).

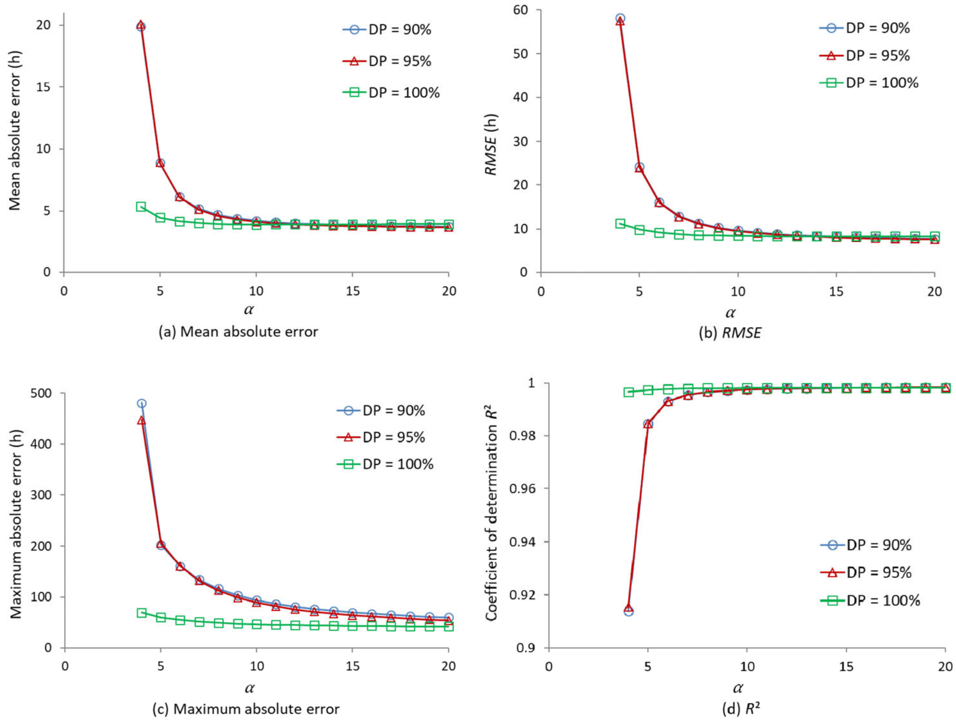
With the LHS200 sample, the best obtained accuracy was for $DP = 90\%$ and $\ln(\alpha) = 5$, with mean absolute error, *RMSE* and maximum absolute error equal to 4.2 h, 9.1 h and 46.0 h, respectively, and $R^2 = 0.9977$. Thus, these errors were reduced by 15%, 26% and 30%, respectively, using the LHS400 sample.

Finally, it should be noted that, when studying building energy performance, where simplifications are very common, low sensitivity to shape parameters, such as that obtained for $\alpha > 20$, is an important advantage. In particular, in building standards, it is convenient that such parameters do not vary significantly from one building to another.

## 3.6 Fréchet form

The second flexible distribution studied was the positively skewed Fréchet distribution, which has a heavy upper tail (Eq. (12)). Good accuracy was obtained for high values of the shape parameter $\alpha$ (Figure 10); i.e. when the upper tail of this distribution was not too heavy to approximate the number of overheating hours $NH_o$ (Figure 2). For each domain percentage *DP*, the best fits were obtained for $\alpha$ at its highest value ($\alpha = 20$). The best one was for $DP = 95\%$ with mean absolute error, *RMSE* and maximum absolute error equal to 3.7 h, 7.6 h and 55 h, respectively, and $R^2$ equal to 0.9984. A lower maximum absolute error (42.9 h), however, was obtained for $DP = 100\%$ with the same value of $\alpha$. The difference between fits with $DP = 90\%$ and $DP = 95\%$ was only significant in terms of maximum absolute error, with a lowest value of 60.8 h for the former. For the LHS200 sample, the best accuracy was for $DP = 90\%$ and $\alpha = 20$, with mean absolute error, *RMSE* and maximum absolute error equal to 4.2 h, 9.4 h and 44.9 h, respectively, and $R^2 = 0.9976$.

Finally, the Fréchet form was not a better alternative to the Gumbel and skew logistic forms since it offered



(a) Mean absolute error

(b) *RMSE*

(c) Maximum absolute error

(d) $R^2$

**Fig. 10** Errors and coefficient of determination $R^2$ for three domain percentage *DP* according to the shape parameter $\alpha$ for the metamodel with the Fréchet distribution trained with the LHS400 sample: (a) mean absolute error, (b) *RMSE*, (c) maximum absolute error and (d) coefficient of determination $R^2$

comparable mean absolute error and *RMSE*, but a higher maximum absolute error.

## 3.7 Kumaraswamy form

The last form considered was that of the Kumaraswamy distribution, which has two shape parameters $\alpha$ and $\beta$ (Eq. (13)). Due to its high flexibility, this form was studied for Helsinki and also for Athens. The best accuracy was obtained with domain percentages *DP* of 95% For Helsinki and 90% for Athens. The corresponding errors and coefficient of determination $R^2$ according to the shape parameters are presented in Figure 11 for Helsinki and Figure 12 for Athens, with $\alpha$ equal to 1.5, 2, 2.5 and 3.

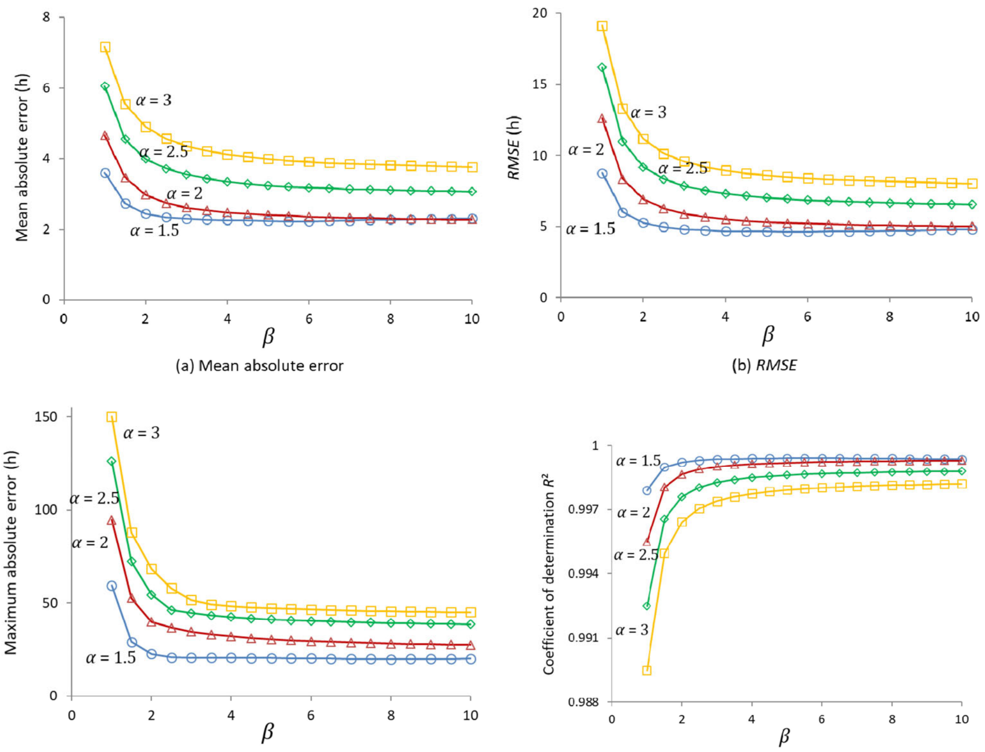
The results reveal that this distribution had excellent fits for both climates when values of $\alpha$ and $\beta$ were appropriate. Moreover, when the value of one parameter ($\alpha$ in Helsinki and $\beta$ in Athens) was appropriate, the sensitivity to the second was low; e.g. for $\alpha$ = 1.5 in Helsinki, $\beta$ had low impact on accuracy when it was higher than 3; and for $\beta$ = 1.5 in Athens, $\alpha$ had limited impact except for maximum absolute error.

The best fits were obtained for $1.5 < \alpha < 2$ and $\beta > 5$ in Helsinki; and for $1 < \alpha < 2$ and $1.5 < \beta < 2.5$ in Athens. For each domain percentage *DP* studied, Table 7 presents the errors and $R^2$ for fits which were nearly the best: with $\alpha$ = 1.75 and $\beta$ = 9 for Helsinki, and with $\alpha$ = 1.5 and $\beta$ = 1.5 for Athens. For each climate, the difference between fits with *DP* = 90% (best results in Athens) and *DP* = 95% (best results in Helsinki) was only significant in terms of maximum absolute error.

The Kumaraswamy form outperformed all the studied forms. For Helsinki, Table 8 presents the error reduction of its best fit (*DP* = 95%, $\alpha$ = 1.75, $\beta$ = 9) as compared to those of the polynomial (*DP* = 96%), logistic (*DP* = 90%) and Gumbel (*DP* = 95%) forms. It should be noted that the skew logistic form had results similar to those of the Gumbel form, while the Fréchet form had comparable mean absolute error and *RMSE* and higher maximum absolute error than the latter.

For Athens, the Kumaraswamy form with *DP* = 90%, gave slightly better results than the polynomial form with *DP* = 92%, with mean absolute error *RMSE* and maximum absolute errors lower by 3%, 3%, and 8%, respectively. This low difference was due to the fact that the values of $NH_o$, as given by dynamic simulation for the LHS400 sample, varied between 200 h and *NT* = 860 h (Figure 7); thus a polynomial could have trained the results without the need to approximate the behavior of the dynamic model for *NT* close to zero. Therefore, the best fit of the Kumaraswamy form was achieved with low values of the shape parameters ($\alpha$ = 1.5 and $\beta$ = 1.5) close to ($\alpha$ = $\beta$ = 1) for which this form coincides with a second-order polynomial form.

With the LHS200 sample, the best accuracy for Helsinki



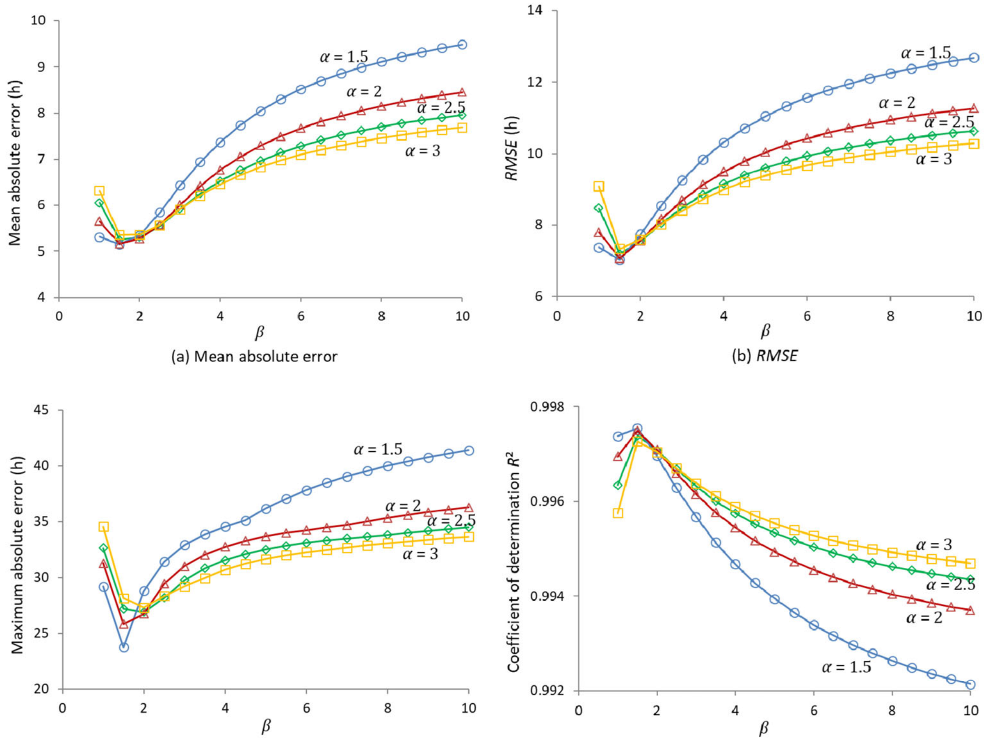
(a) Mean absolute error

(b) *RMSE*

(c) Maximum absolute error

(d) $R^2$

**Fig. 11** Errors and coefficient of determination $R^2$ of the metamodel with the Kumaraswamy form for Helsinki according to the shape parameters $\alpha$ and $\beta$: (a) mean absolute error, (b) *RMSE*, (c) maximum absolute error and (d) coefficient of determination $R^2$. The results were obtained with the LHS400 sample and a domain percentage *DP* of 95%



(a) Mean absolute error

(b) *RMSE*

(c) Maximum absolute error

(d) $R^2$

**Fig. 12** Errors and coefficient of determination $R^2$ of the metamodel with the Kumaraswamy form for Athens according to the shape parameters $\alpha$ and $\beta$: (a) mean absolute error, (b) *RMSE*, (c) maximum absolute error and (d) coefficient of determination $R^2$. The results were obtained with the LHS400 sample and a domain percentage *DP* of 90%

**Table 7** Errors and $R^2$ for the metamodel with the Kumaraswamy form trained with the LHS400 sample for shape parameters $\alpha$ and $\beta$ giving nearly the best fits

| Location | Helsinki ($\alpha$ = 1.75, $\beta$ = 9) | | | Athens ($\alpha$ = 1.5, $\beta$ = 1.5) | | |
|---|---|---|---|---|---|---|
| Domain percentage *DP* | 90% | 95% | 100% | 90% | 95% | 100% |
| Mean absolute error (h) | 2.1 | 2.1 | 2.7 | 5.1 | 5.3 | 6.6 |
| *RMSE* (h) | 4.5 | 4.5 | 6.0 | 7.0 | 7.2 | 9.0 |
| Maximum absolute error (h) | 20.6 | 19.0 | 28.8 | 23.7 | 28.1 | 32.2 |
| $R^2$ | 0.9994 | 0.9994 | 0.9990 | 0.9975 | 0.9974 | 0.9959 |

**Table 8** Error reduction for Helsinki with the Kumaraswamy form (*DP* = 95%, $\alpha$ = 1.75, $\beta$ = 9) as compared to the polynomial (*DP* = 96%), logistic (*DP* = 90%) and Gumbel (*DP* = 95%) forms

| Metamodel form | Polynomial | Logistic | Gumbel |
|---|---|---|---|
| Mean absolute error reduction | −52% | −72% | −41% |
| *RMSE* reduction | −49% | −64% | −34% |
| Maximum absolute error reduction | −42% | −65% | −42% |

was for *DP* = 95%, $\alpha$ = 1.5 and $\beta$ = 5; with mean absolute error, *RMSE* and maximum absolute error equal to 2.4 h, 5.6 h and 23.1 h, respectively, and $R^2$ = 0.9992. The errors were thus between 15% and 20% lower with the LHS400 sample. For Athens, the best accuracy was for *DP* = 95%, $\alpha$ = 1.75 and $\beta$ = 1.5; with mean absolute error, *RMSE* and maximum absolute error equal to 5.2 h, 7.4 h and 22.7 h, respectively, and $R^2$ = 0.9976. Thus, the sample did not have a major effect for this hot climate (slightly lower mean absolute error and *RMSE,* and slightly higher maximum absolute error with the LDH400 sample). This was also the case for the polynomial form with suitable domain percentages (Figure 8).

To illustrate the accuracy of the metamodel with the Kumaraswamy form, Figure 13 presents, for Helsinki (*DP* = 95%, $\alpha$ = 1.75, $\beta$ = 9) and Athens (*DP* = 90%, $\alpha$ = 1.5, $\beta$ = 1.5), a comparison between its results and those of the TRNSYS dynamic simulation for the same testing sample of 200 simulations (corresponding errors and $R^2$ presented in Table 7).

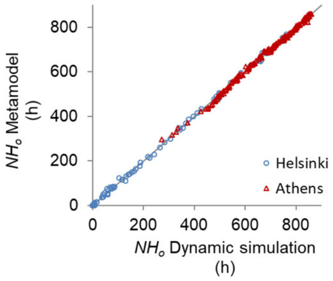


**Fig. 13** Number of overheating hours $NH_o$ as given by dynamic simulation and the metamodel with Kumaraswamy form trained with the LHS400 sample for Helsinki (*DP* = 95%, $\alpha$ = 1.75, $\beta$ = 9) and Athens (*DP* = 90%, $\alpha$ = 1.5, $\beta$ = 1.5)

After determining the coefficients and shape parameters of the metamodel, the number of overheating hours $NH_o$ could be explicitly expressed as a function of the features of Table 1; and as such thermal comfort could be assessed very rapidly with a small volume of data required. For instance, the metamodel with Kumaraswamy form that offered for Helsinki the results in Figure 13, gives $NH_o$ according to the coded values of the features as follows:

$$\begin{cases} NH_o = 1-\left(1-z^{1.75}\right)^9 & 0<z<1 \\ NH_o = 0 & z \le 0 \\ NH_o = 860 & z \ge 1 \\ R^2 = 0.9994 & \end{cases} \tag{14}$$

with

$$\begin{aligned} z = {} & -8.84\times10^{-2} - 6.01\times10^{-2}H_{\mathrm{tr,ow}} - 1.78\times10^{-1}H_{\mathrm{tr,w}} - 2.84\times10^{-1}H_{\mathrm{vent}} \\ & - 8.38\times10^{-2}H_{\mathrm{nvent}} + 3.14\times10^{-1}Q_{\mathrm{ig}} + 1.39\times10^{-1}Q_{\mathrm{so,ws}} + 7.45\times10^{-2}Q_{\mathrm{so,ww}} \\ & - 6.32\times10^{-4}H_{\mathrm{tr,ow}}^2 + 2.66\times10^{-3}H_{\mathrm{tr,w}}^2 + 2.76\times10^{-2}H_{\mathrm{vent}}^2 \\ & + 1.47\times10^{-3}H_{\mathrm{nvent}}^2 - 6.14\times10^{-2}Q_{\mathrm{ig}}^2 - 1.45\times10^{-2}Q_{\mathrm{so,ws}}^2 - 3.47\times10^{-3}Q_{\mathrm{so,ww}}^2 \\ & + 1.35\times10^{-3}H_{\mathrm{tr,ow}}H_{\mathrm{tr,w}} + 7.60\times10^{-3}H_{\mathrm{tr,ow}}H_{\mathrm{vent}} - 7.39\times10^{-3}H_{\mathrm{tr,ow}}H_{\mathrm{nvent}} \\ & + 1.56\times10^{-2}H_{\mathrm{tr,ow}}Q_{\mathrm{ig}} + 7.61\times10^{-3}H_{\mathrm{tr,ow}}Q_{\mathrm{so,ws}} + 3.15\times10^{-3}H_{\mathrm{tr,ow}}Q_{\mathrm{so,ww}} \\ & + 2.19\times10^{-2}H_{\mathrm{tr,w}}H_{\mathrm{vent}} - 1.76\times10^{-2}H_{\mathrm{tr,w}}H_{\mathrm{nvent}} + 4.72\times10^{-2}H_{\mathrm{tr,w}}Q_{\mathrm{ig}} \\ & + 2.12\times10^{-2}H_{\mathrm{tr,w}}Q_{\mathrm{so,ws}} + 1.23\times10^{-2}H_{\mathrm{tr,w}}Q_{\mathrm{so,ww}} - 2.92\times10^{-2}H_{\mathrm{vent}}H_{\mathrm{nvent}} \\ & + 6.71\times10^{-2}H_{\mathrm{vent}}Q_{\mathrm{ig}} + 2.91\times10^{-2}H_{\mathrm{vent}}Q_{\mathrm{so,ws}} + 1.59\times10^{-2}H_{\mathrm{vent}}Q_{\mathrm{so,ww}} \\ & + 4.64\times10^{-2}H_{\mathrm{nvent}}Q_{\mathrm{ig}} + 1.80\times10^{-2}H_{\mathrm{nvent}}Q_{\mathrm{so,ws}} + 8.69\times10^{-3}H_{\mathrm{nvent}}Q_{\mathrm{so,ww}} \\ & - 5.52\times10^{-2}Q_{\mathrm{ig}}Q_{\mathrm{so,ws}} - 3.18\times10^{-2}Q_{\mathrm{ig}}Q_{\mathrm{so,ww}} - 1.24\times10^{-2}Q_{\mathrm{so,ws}}Q_{\mathrm{so,ww}} \end{aligned}$$

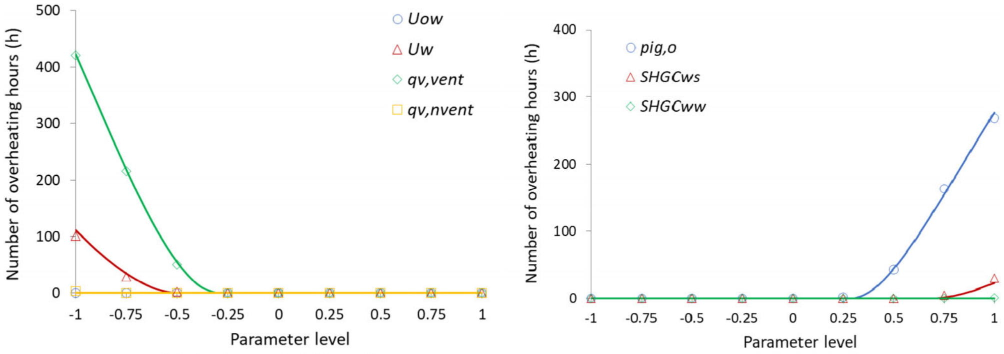
Besides offering accurate predictions and very rapid calculation, such machine learning models are easy to use by non-expert users with only small building data requirements. Unlike many machine learning models such as, neural networks and support vector machines, they are easily interpretable, contributing to an improved understanding of design problems, with insight into the relationships between building design and indoor thermal comfort.

The second order polynomial of Eq. (14) reveals a strong association between the quantity of heat due to internal gains $Q_{ig}$ and the number of overheating hours $NH_o$ with a corresponding coefficient of $3.14\times10^{-1}$. Moreover, the coefficients of all the heat transfer coefficients $H_{tr,ow}$, $H_{tr,w}$, $H_{vent}$ and $H_{nvent}$ were negative, suggesting that transmission and ventilation heat transfer improved thermal comfort. The shape parameters of the Kumaraswamy distribution ($\alpha = 1.75$, $\beta = 9$) give a knowledge about the positive skewness of the dynamic model response (Figures 2 and 14).

To further illustrate the accuracy of the metamodel, its results with the Kumaraswamy form were confronted with those of TRNSYS dynamic simulation obtained with the one-at-a-time method, as illustrated in Figure 14 for Helsinki and Figure 15 for Athens.

The symbols in these figures represent the same results as in Figures 2 and 3 used to obtain information from dynamic simulation. The solid lines were obtained with the Kumaraswamy form ($DP$ = 95%, $\alpha = 1.75$, $\beta = 9$ for Helsinki; and $DP$ = 90%, $\alpha = 1.5$, $\beta = 1.5$ for Athens). Each line represents the results of 20 metamodel runs obtained by varying the corresponding parameter in Table 2, in addition to a run for the central point (all coded values equal to zero). This comparison suggests that the metamodel with the Kumaraswamy form could accurately approximate the dynamic model. In particular, it was able to consider the distinct behavior of $NH_o$ for low overheating in Helsinki and high overheating in Athens.

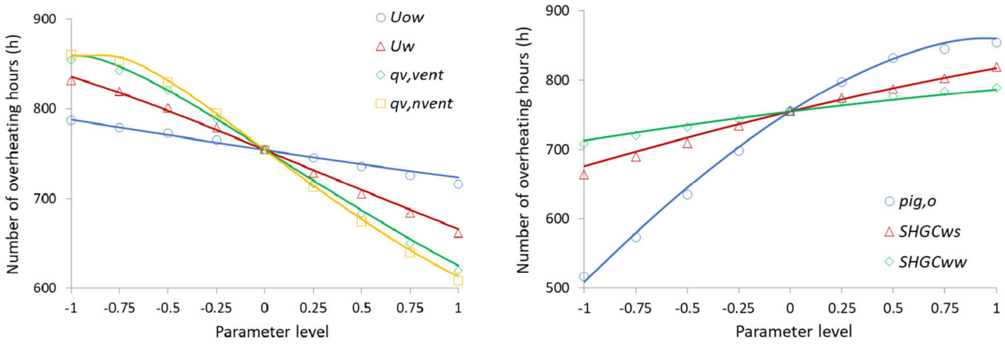
Finally, the fit of the metamodel could be further improved in future studies by using various available statistical techniques: optimized Latin Hypercube sampling, weighted least squares regression, etc. Better fit may be also achieved with alternative probability distributions. To preserve the computational efficiency and interpretability of the metamodel, however, it is important to consider simple probability distributions such as those studied here.



(a) U-values and air flow rates

(b) Heat gains

**Fig. 14** Number of overheating hours $NH_o$ for Helsinki as given by the metamodel with the Kumaraswamy form (solid lines) and by TRNSYS dynamic simulation using a one-at-a-time method (symbols): (a) $U$-values $U_{ow}$ and $U_w$, and airflow rates $q_{v,vent}$ and $q_{v,nvent}$; and (b) internal heat gains $p_{ig,o}$, and solar heat gain coefficients $SHGC_{ws}$ and $SHGC_{ww}$



(a) U-values and air flow rates

(b) Heat gains

**Fig. 15** Number of overheating hours $NH_o$ for Athens as given by the metamodel with the Kumaraswamy form (solid lines) and by TRNSYS dynamic simulation using a one-at-a-time method (symbols): (a) $U$-values $U_{ow}$ and $U_w$, and airflow rates $q_{v,vent}$ and $q_{v,nvent}$; and (b) internal heat gains $p_{ig,o}$, and solar heat gain coefficients $SHGC_{ws}$ and $SHGC_{ww}$

## 4 Conclusions

Physics-informed machine learning, using prior knowledge of quasi-steady-state heat transfer and dynamic simulation results, offered great advantages in metamodeling thermal comfort in non-air-conditioned buildings: a flexible and transparent metamodel, accurate predictions, better physical consistency and generalizability, and low computational cost.

The metamodel in the form of symmetric probability distributions failed to give reliable results for studying thermal comfort, and generally, symmetry is not common in the energy performance of a building. Thus, for metamodeling building energy performance, flexible probability distributions would provide better accuracy than the widely used logistic distribution using the same training samples, and therefore without additional computational effort.

In this study, the Kumaraswamy distribution offered the best performance, with an excellent trade-off between simplicity, flexibility and accuracy. This distribution should play an important role in metamodeling thermal comfort, and generally, in studying building energy performance.

Physics-informed machine leaning of thermal comfort in non-air-conditioned buildings is a challenging issue: the dynamic model response is highly non-linear and asymmetric, thermal comfort could be expressed by a large number of indices of different nature, occupant behavior has a significant impact on indoor conditions, etc. The metamodeling presented in this study could be extended to various thermal comfort indices with knowledge related to building physics, dynamic simulation results, occupant behavior and thermal comfort perception.

Future machine learning models for building energy analysis, should attempt to combine the strength of both physics and machine learning models. This could be achieved, for instance, by incorporating knowledge of thermodynamics, heat and mass transfer together with expert knowledge. Moreover, physical consistency should become a prime criterion for choosing a machine learning model for building energy analysis. A related issue is the incorporation of energy regulation requirements into machine learning algorithms, together with physical knowledge.

The advantages offered by physics-based machine learning, the diversity of physical phenomena involved in a building, and the major role of expert knowledge, suggest that physics-based machine learning could soon become a mainstream approach in building energy modeling. Even the incorporation of basic physical knowledge could be more effective than any advanced machine learning technique.

## Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

## References

Beucler T, Rasp S, Pritchard M, et al. (2019). Achieving conservation of energy in neural network emulators for climate modeling. arXiv:1906.06622

Breesch H, Janssens A (2010). Performance evaluation of passive cooling in office buildings based on uncertainty and sensitivity analysis. *Solar Energy*, 84: 1453–1467.

Bünning F, Huber B, Schalbetter A, et al. (2022). Physics-informed linear regression is competitive with two Machine Learning methods in residential building MPC. *Applied Energy*, 310: 118491.

CEN (2019). EN 16798-1: Energy performance of buildings – ventilation for buildings – Part 1: Indoor environmental input parameters for design and assessment of energy performance of buildings addressing indoor air quality, thermal environment, lighting and acoustics. Brussels: European Committee for Standardization.

Chen X, Yang H, Sun K (2017). Developing a meta-model for sensitivity analyses and prediction of building performance for passively designed high-rise residential buildings. *Applied Energy*, 194: 422–439.

de Dear R, Brager G (1998). Developing an adaptive model of thermal comfort and preference. *ASHRAE Transactions*, 104(1): 145–167.

Drgoňa J, Tuor AR, Chandan V, et al. (2021). Physics-constrained deep learning of multi-zone building thermal dynamics. *Energy and Buildings*, 243: 110992.

Eisenhower B, O'Neill Z, Narayanan S, et al. (2012). A methodology for meta-model based optimization in building energy models. *Energy and Buildings*, 47: 292–301.

Fletcher MJ, Johnston DK, Glew DW, et al. (2017). An empirical evaluation of temporal overheating in an assisted living Passivhaus dwelling in the UK. *Building and Environment*, 121: 106–118.

Fréchet M (1928). Sur la loi de probabilité de l'écart maximum. In Annales de la societe Polonaise de Mathematique. (in French)

Haldi F, Robinson D (2008). On the behaviour and adaptation of office occupants. *Building and Environment*, 43: 2163–2177.

Harlow DG (2002). Applications of the Fréchet distribution function. *International Journal of Materials and Product Technology*, 17: 482–495.

Jaffal I, Inard C, Ghaddar N, et al. (2020). A metamodel for long-term thermal comfort in non-air-conditioned buildings. *Architectural Engineering and Design Management*, 16: 441–472.

Jones MC (2009). Kumaraswamy's distribution: A beta-type distribution with some tractability advantages. *Statistical Methodology*, 6: 70–81.

Karniadakis GE, Kevrekidis IG, Lu, Perdikaris P, et al. (2021). Physics-informed machine learning. *Nature Reviews Physics*, 3: 422–440.

Khandelwal A, Xu S, Li X, et al. (2020). Physics guided machine learning methods for hydrology. arXiv:2012.02854.

Klein SA, Beckman WA, Mitchell JW, et al. (2004). TRNSYS 16 – A TRaNsient SYstem Simulation program, user manual. Madison, WI, USA: Solar Energy Laboratory, University of Wisconsin–Madison.

Kotz S, Nadarajah S (2000). Extreme Value Distributions: Theory and Applications. Singapore: World Scientific.

Kumaraswamy P (1980). A generalized probability density function for double-bounded random processes. *Journal of Hydrology*, 46: 79–88.

Li YF, Ng SH, Xie M, et al. (2010). A systematic comparison of metamodeling techniques for simulation optimization in Decision Support Systems. *Applied Soft Computing*, 10: 1257–1273.

Li B, Yao R, Wang Q, et al. (2014). An introduction to the Chinese Evaluation Standard for the indoor thermal environment. *Energy and Buildings*, 82: 27–36.

Lomas KJ, Porritt SM (2017). Overheating in buildings: Lessons from research. *Building Research and Information*, 45: 1–18.

Magnier L, Haghighat F (2010). Multiobjective optimization of building design using TRNSYS simulations, genetic algorithm, and Artificial Neural Network. *Building and Environment*, 45: 739–746.

Morgan C, Foster JA, Poston A, et al. (2017). Overheating in Scotland: Contributing factors in occupied homes. *Building Research and Information*, 45: 143–156.

Nagler J (1994). Scobit: an alternative estimator to logit and probit. *American Journal of Political Science*, 38: 230–255.

Nicol JF, Humphreys MA (2002). Adaptive thermal comfort and sustainable thermal standards for buildings. *Energy and Buildings*, 34: 563–572.

Rackes A, Melo AP, Lamberts R (2016). Naturally comfortable and sustainable: Informed design guidance and performance labeling for passive commercial buildings in hot climates. *Applied Energy*, 174: 256–274.

Rossi MM, Oliveira Favretto AP, Grassi C, et al. (2019). Metamodels to assess the thermal performance of naturally ventilated, low-cost houses in Brazil. *Energy and Buildings*, 204: 109457.

Simpson TW, Poplinski JD, Koch PN, et al. (2001). Metamodels for computer-based engineering design: survey and recommendations. *Engineering With Computers*, 17: 129–150.

Symonds P, Taylor J, Chalabi Z, et al. (2015). Performance of neural networks vs. Radial basis functions when forming a metamodel for residential buildings. *International Journal of Civil and Environmental Engineering*, 9: 1594–1598.

Takasu M, Ooka R, Rijal HB, et al. (2017). Study on adaptive thermal comfort in Japanese offices under various operation modes. *Building and Environment*, 118: 273–288.

van Gelder L, Das P, Janssen H, et al. (2014). Comparative study of metamodelling techniques in building energy simulation: Guidelines for practitioners. *Simulation Modelling Practice and Theory*, 49: 245–257.

von Rueden L, Mayer S, Garcke J, et al. (2019). Informed machine learning–towards a taxonomy of explicit integration of knowledge into machine learning. *Learning*, 18: 19–20.

Zobeiry N, Humfeld KD (2021). A physics-informed machine learning approach for solving heat transfer equation in advanced manufacturing and engineering applications. *Engineering Applications of Artificial Intelligence*, 101: 104232.