Building and Environment 143 (2018) 390–403

Contents lists available at ScienceDirect

Building and Environment

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

Optimization-aided calibration of an urban microclimate model under uncertainty

Jiachen Mao[a,*], Yangyang Fu[b], Afshin Afshari[c], Peter R. Armstrong[c], Leslie K. Norford[a,**]

[a] Massachusetts Institute of Technology, Cambridge, MA, 02139, USA
[b] University of Colorado, Boulder, CO, 80309, USA
[c] Masdar Institute of Science and Technology (part of Khalifa University), PO Box 54224, Abu Dhabi, United Arab Emirates

ARTICLE INFO

Keywords:
Model calibration
Urban microclimate
Uncertainty
Online hyper-heuristics
Evolutionary algorithm
Simulation-based optimization

ABSTRACT

Simulation models play an important role in the design, analysis, and optimization of modern energy and environmental systems at building or urban scale. However, due to the extreme complexity of built environments and the sheer number of interacting parameters, it is difficult to obtain an accurate representation of real-world systems. Thus, model calibration and uncertainty analysis hold a particular interest, and it is necessary to evaluate to what degree simulation models are imperfect before implementing them during the decision-making process. In contrast to the extensive literature on the calibration of building performance models, little has been reported on how to automatically calibrate physics-based urban microclimate models. This paper illustrates a general methodology for automatic model calibration and applies it to an urban microclimate system. The Urban Weather Generator (UWG) is selected as the underlying simulation engine for an optimization-aided calibration based on the urban outdoor air temperature in an existing district area located in downtown Abu Dhabi (UAE) during 2017. In particular, given the time-constrained nature of engineering applications, an online hyper-heuristic evolutionary algorithm (EA) is proposed and developed in order to accelerate the calibration process. The validation results show that, in single-objective optimization, the online hyper-heuristics could robustly help EA produce quality solutions with smaller uncertainties at much less computational cost. In addition, the resulting calibrated solutions are able to capture weekly-average and hourly diurnal profiles of the urban outdoor air temperature similar to the measurements for certain periods of the year.

1. Introduction

Over the past decades, many climate projections have foreseen both global warming and sea level rise [1]. This projected climate change will potentially lead to increased food shortages, decreased fresh water supplies, and severe storm events – all of which would have a significant impact on humanity in both developing and developed regions of the world. In response to mitigating these on-going threats, the IPCC [1] urges dramatic reduction in greenhouse gas emissions and sustainable adaption of societies to a new climate context. This agenda holds a particular attention in urban areas where massive valuable assets are concentrated and more than half of the world's population resides [2]. Moreover, in some cases the anthropogenic climate change can be exacerbated by neighborhood-to-city-scale phenomena, such as the Urban Heat Island (UHI) [3].

In general, the UHI increases the peak electricity demand and likelihood of heat wave event during summer, which may cause various health problems leading to morbidity, disability, or even death [4,5]. Cities must undertake mitigation and adaptation measures to reduce the negative impacts of heat islands on the environment, the economy, and the population. However, aside from the social and economic concerns, developing effective adaptation strategies comes with a large technical challenge since an urban microclimate system comprises very complex physical relationships between many elements that may interact with each other [6]. A good understanding of the mechanism and characteristics of UHI is thus a prerequisite for decision makers to identify and adopt reliable mitigation and adaptation options, particularly during the design of new or renovated neighborhood areas.

This pressing need motivates many energy and environment research communities to expand their scope to the urban realm [7]. Great efforts have been made to incorporate the UHI effect into thermal simulations [8]. In addition, some researchers have started to examine

* Corresponding author.
** Corresponding author.
E-mail addresses: maoj@mit.edu (J. Mao), lnorford@mit.edu (L.K. Norford).

https://doi.org/10.1016/j.buildenv.2018.07.034
Received 27 April 2018; Received in revised form 20 July 2018; Accepted 21 July 2018
Available online 23 July 2018
0360-1323/ Published by Elsevier Ltd.

Nomenclature

D	training database
$F_{E,P}/F_{E,Q}$	expensive objective and constraint values in P/Q
F_S	approximated objective and constraint values in S
$\overline{m}$	mean of the absolute measured urban-rural temperature differences
m_i	measured data point i of the urban-rural temperature difference
n	number of the data points
P	parent population
Q	offspring population
S	surrogate population
s_i	simulated data point i of the urban-rural temperature difference
t	generation counter
$w_{\mathrm{CV(RMSE)}}$	weight assigned to CV(RMSE)
w_{NMBE}	weight assigned to NMBE

Abbreviation

AEC	architecture, engineering, and construction
AM	averaged model
ASHRAE	American Society of Heating, Refrigerating, and Air-conditioning Engineers
CFD	computational fluid dynamics
COP	coefficient of performance
CV(RMSE)	coefficient of variation of the root-mean-square error
EA	evolutionary algorithm
EPW	EnergyPlus Weather
ES	evolutionary strategy
FPC	fitness prediction correlation
GIGO	garbage in, garbage out
GOF	goodness-of-fit
IPCC	Intergovernmental Panel on Climate Change
LH	Latin Hypercube
MC	Monte Carlo
MOO	multi-objective optimization
NMBE	normalized mean bias error
RSM	rural station model
SA	sensitivity analysis
SHGC	solar heat gain coefficient
SVR	support vector regression
UBL	urban boundary layer
UCM	urban canopy model
UHI	Urban Heat Island
UWG	Urban Weather Generator
VDM	vertical diffusion model

the physical behavior or causal factors of urban climate change and heat island effect via mesoscale computational fluid dynamics (CFD) simulations [9], analytical and empirical algorithms [10], and physics-based urban canopy models [11]. To account for the interactions between building energy demand and urban thermal behavior, the Urban Weather Generator (UWG) was proposed and developed by Bueno et al. [12] as a physics-based simulator to quickly estimate the microclimate condition and energy consumption at the neighborhood-to-city scale. With continuous updates and validations [13–15], the UWG has been reported to be a promising urban simulation engine with exceptionally low computational requirements.

However, despite the positive progress, simulation practice to date has only penetrated a small fraction of professional communities within the AEC industry. One recognized obstacle is the discrepancy, sometimes significant, between actual and predicted values. In general, prognostic law-driven models [16] involve a suite of simplified physical relations describing the way various component disturbances (from system operation, human activity, material property, etc.) interact with each other and influence the aggregate physical behavior. Within these equations, both differential and algebraic, hundreds of parameters exist. It is common for an engineer to make ad-hoc estimates for these parameters based on limited engineering knowledge, past user experience, and an abundance of trial and error. As a result, even though many inputs seem empirically validated, the simulated output could be far from the real scenario. It is ironic that at the time when simulation is the most popular, parameters of simulation may be the least reliable, which inevitably reduces the confidence of simulated results and curtails the use of simulation models to some extent. It is hence necessary to match simulation with measurement, a process called “model calibration.”

Although some studies use calibrated models, their underlying calibration techniques are unclear. In order to dive deeper into model calibration, it is important to consider “model uncertainty” [17]. Validation of a complex-system model is notoriously difficult, especially when the purpose of the model is to look at some non-observable or unmeasured physical behavior. The reason stems from the fact that closed-loop simulations usually represent major simplifications and constraints. That is to say, “*the portion of the world captured by the model is an arbitrary ‘enclosure’ of an otherwise open, interconnected system*” [18]. Model errors are mainly caused by difficulties in capturing how exactly a system operates, due to software limitations and inaccurate parameter descriptions that cannot be completely modeled a priori. The input parameters are often calibrated manually by an expert, which may require days or weeks of work depending on model complexity. A commonly observed method tunes some specific parameters until the result meets an acceptance criteria without any uncertainty analysis.

Uncertainty quantification is often time-consuming and requires additional efforts in the overall design and/or retrofit phase of an engineering system, but can provide more robust decisions. However, not all the modeled aspects have the same level of importance and not every input parameter offers the same contribution to error propagation. As a result, uncertainty analysis is usually coupled with sensitivity analysis (SA) to measure the relative importance of various input parameters [19]. In general, SA is used to identify how the uncertainty in an output can be allocated to the uncertainties in the inputs. Once the “weak” parameters are determined, they could be set at some nominal values, thereby reducing the parameter space and increasing the calibration efficiency. The remaining influential input set is considered by a more rigorous calibration process.

Given that manually tuning the parameters can be viewed as an optimization process, it is natural to think about using computers to implement calibration in an automatic or semi-automatic way via optimization algorithms. Simulation-based optimization—wherein a simulation model is embedded in the optimization—has been increasingly applied in the building science community through mathematical and statistical methods to assist design analysis [20–22] and model calibration [23–25]. A pioneering study was conducted by Wright [26] in the 1980s, while the number of optimization-related papers has sharply increased since 2005 [21]. Many open-source tools, such as the GenOpt by Wetter [27], are now available to provide the capabilities of coupling various building performance simulations to effectively support optimization.

Generally speaking, an objective performance function is formulated to define a max/min target, while some constraint functions are employed to reduce the possibility of deviating too far from reality. Since the performance function associated with building or urban

thermal-physical behavior is usually discontinuous, non-differentiable, multi-modal, and locally-flat [28], traditional gradient-based algorithms cannot successfully search the whole parameter space. On the other hand, heuristic-based algorithms (e.g., evolutionary algorithm) have been frequently used in building or urban performance optimization, mainly owing to their abilities to obtain good solutions with some degree of efficiency and robustness. In particular, a brief literature review indicates that heuristic-based algorithms can perform reliable calibration for building energy models [29–32]. To the authors' best knowledge, there is nearly no work on performing optimization-based calibration of physics-based urban microclimate models, due partly to the expensive computational costs.

At the current stage, calibration still relies to some extent on expert judgment and engineering experience (e.g., in the selection of candidate inputs). So, we recognize that the computer is more likely to act as a supplement to optimize and accelerate the calibration by transforming manual adjustment into automatic tuning. However, application of numerical optimization in the calibration process, while abstracting the physical objective as a tractable mathematical problem, inevitably neglects some real physical details in the constraint(s). Indiscriminant use could result in mathematical match but physical mismatch, which is why some researchers would criticize such methodology. This naturally necessitates the incorporation of uncertainty analysis after calibration to test reliability.

Finally, although simulation-based optimization has been actively discussed, one practical concern is the computation time. A typical model optimization could take days or weeks of computation to find optimal or near-optimal solutions. Given the time-constrained nature of engineering applications, it is necessary to develop novel optimization methods that are able to find high-quality solutions as quickly as possible.

This study was initiated to identify a general methodology for automatic model calibration and apply it to an urban microclimate system. Together with a companion paper [15], the proposed ideas involve various concepts and methods borrowed from allied scientific disciplines in a more mathematical and statistical point of view. The UWG is selected as the simulation engine for model calibration based on the urban outdoor air temperature in an existing district area located in downtown Abu Dhabi (UAE) during 2017. In particular, we develop and apply an online hyper-heuristic evolutionary algorithm to accelerate the calibration process. In the following content, Section 2 illustrates the calibration methodology and the concept of online hyper-heuristics. Section 3 introduces the case study for conducting the developed calibration method. Finally, results are described and discussed in Section 4, and the corresponding conclusions are given in Section 5. The proposed approach to effectively integrate field measurements and computer simulations for model calibration has the potential to open up a broad range of new engineering applications.

2. Methodology

Model calibration is commonly regarded as an inverse estimation process where selected input parameters are tuned to reconcile the outputs from simulation as closely as possible to the measurements. Generally there are three technical parts: model pre-establishment, model calibration, and model post-evaluation. In this study, the Urban Weather Generator (UWG) is selected as the simulation engine to implement the calibration, since it can be applied to different climate zones and urban configurations to yield an estimation of the UHI effect to some extent. The UWG is a stand-alone physics-based program to map a reference weather file to the microclimate conditions at a neighborhood scale via four coupled models [12]. Starting with the hourly rural meteorological data provided in the EPW file and the physical information of the target neighborhood (e.g., building, street, vegetation, etc.), the UWG outputs the hourly steady-state urban weather data at street level that can be readily used in the EPW format by standard building performance simulation software. Detailed descriptions of the newest UWG have been reported in Ref. [15]. As shown in Fig. 1, the whole process starts by establishing case-specific UWG inputs via the Excel interface facilitated by MATLAB. Then, the simulated outputs are collected along with the measured data to initiate the calibration algorithm. Once the most plausible solutions are produced, they will be further examined to evaluate the system behavior.

The core methodology of calibrating a simulation model against the real data needs to be rational, robust, and efficient. Besides, it should have the flexibility for different users with different levels of preference and target. It is recognized from previous studies [33,34] that calibrating a detailed model with numerous parameters is a highly underdetermined problem, which may yield multiple non-unique solutions. The conventional wisdom is that once a model is calibrated in a certain sense, the effect of some intended adaption measures (i.e., parameter variations) can be assessed with a high level of confidence. This idea could be misleading since the urban microclimate condition is the aggregate behavior of many components in the system. Even if the optimization algorithm exhibits an overall good calibration performance, some of the individual parameters might be inaccurately identified. Therefore, it is unlikely that any one optimal solution can be deemed the “best.” Instead, we posit that it is more reasonable to identify the most plausible solutions (that are able to perform well under some specified measure) with associated uncertainties for a fairly robust calibration method.

Within this context, the calibration framework developed in this paper generally enjoys the methodology proposed by Reddy et al. [35,36]. The basic structure, shown in Fig. 2, involves the following five steps:

(1) Prepare the baseline information of the neighborhood area as precisely as possible. This enables the UWG to simulate the microclimate condition in the target street-level area.
(2) Identify a set of candidate parameters along with their preferred

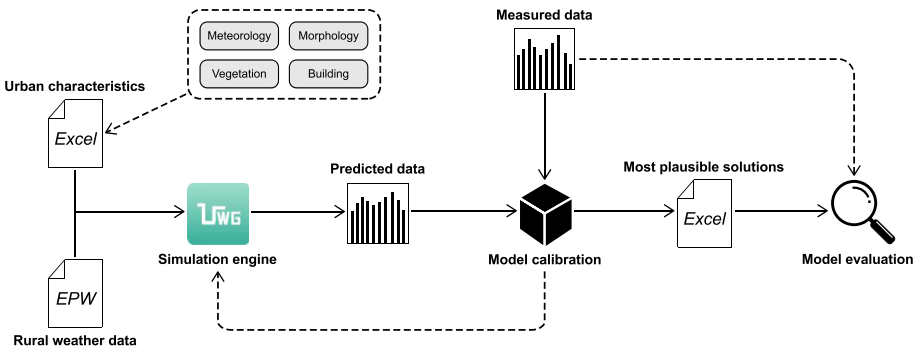


Fig. 1. General workflow of model calibration via the UWG.

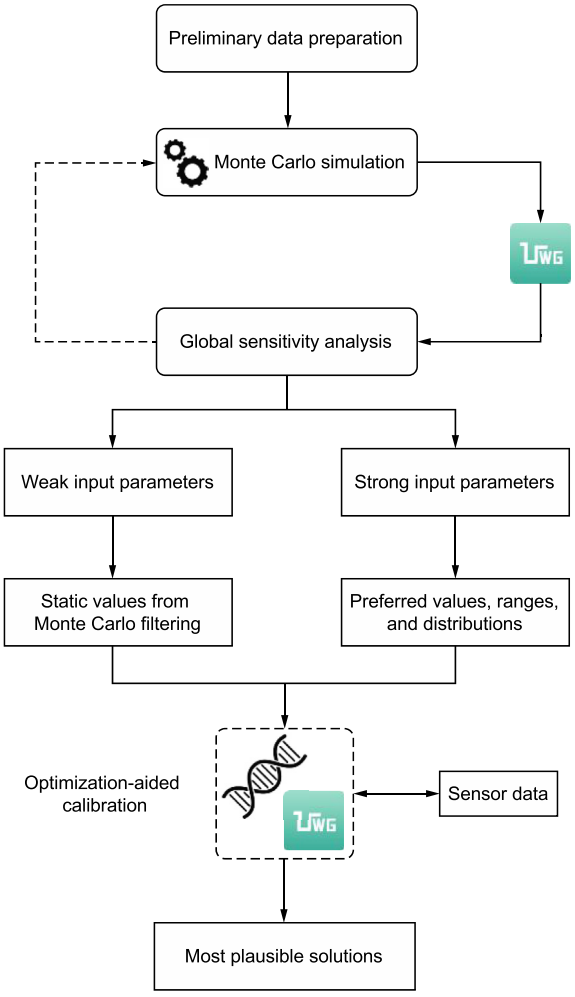


Fig. 2. Structure of an optimization-aided calibration method via the UWG. Detailed descriptions of global sensitivity analysis can be found in Ref. [15].

defaults and ranges. This provides some potential options for model calibration and adaption measures.

(3) Perform uncertainty and sensitivity analysis with different combinations of the inputs. This improves the understanding of the system by identifying the significant parameters.

(4) Apply guided search and optimization algorithm to refine the solutions automatically. This results in a set of plausible solutions that may represent the system dynamics.

(5) Evaluate the calibration effectiveness and efficiency of these solutions under uncertainty. This examines how likely the calibrated model is to yield biased system predictions.

Since our companion paper [15] has detailed the workflow of uncertainty/sensitivity analysis, this paper will further illustrate the technical guideline of search and optimization algorithm for model calibration. In particular, the Monte Carlo filtering technique and online hyper-heuristic evolutionary algorithm are proposed and developed in the present study.

2.1. Objective function

Goodness-of-fit (GOF) is an index that relates the dispersion between the measured and simulated data via a statistical model. Many researchers [32,36] have recommended and adopted the GOF as the objective for model calibration. Lower values mean that the model is more accurate and its behavior fits better with the real case. The GOF is defined as:

$$\mathrm{GOF} = \sqrt{\frac{w_{\mathrm{NMBE}}^2 \mathrm{NMBE}^2 + w_{\mathrm{CV(RMSE)}}^2 \mathrm{CV(RMSE)}^2}{w_{\mathrm{NMBE}}^2 + w_{\mathrm{CV(RMSE)}}^2}} \quad (1)$$

where

$$\mathrm{NMBE} = \frac{1}{\overline{m}} \frac{\sum_{i=1}^{n} (s_i - m_i)}{n} \quad (2)$$

$$\mathrm{CV(RMSE)} = \frac{1}{\overline{m}} \sqrt{\frac{\sum_{i=1}^{n} (s_i - m_i)^2}{n-1}} \quad (3)$$

The NMBE (normalized mean bias error) quantifies the relative error of the simulated values with respect to the measured values summed over the selected number of time steps. Positive values indicate that the model over-estimates the actual scenarios, while the under-prediction produces negative values. As a counterpart, the CV (RMSE) (coefficient of variation of the root-mean-square error) indicates the variability of the errors (i.e., uncertainty) between the simulated (s_i) and measured (m_i) urban-rural temperature differences. For numerical stability during optimization, $\overline{m}$ is set as the mean of the absolute values.

If the simulation results are plotted on an *x-y* plane with NMBE and CV(RMSE) as the axes, one would like to identify a set of trade-off optimal solutions (i.e., a Pareto set) as the "better" solution set. This method is referred to as "multi-objective optimization" or "Pareto optimization." For any given problem, the Pareto optimal set can be produced by an infinite number of Pareto points. A potential drawback of this approach is that the problem of how to select the best solution from the Pareto set is not trivial since it depends on many aspects [21].

To avoid multi-objective optimization in the present case study, we use the idea of "scalarization." Different weights are assigned to each index and the multi-objective function is then simplified as a weighted function of the criteria. Accordingly, we introduce a set of weights, w_{NMBE} and $w_{\mathrm{CV(RMSE)}}$, to impact our consolidated indices in Eqn. (1), where $w_{\mathrm{NMBE}} + w_{\mathrm{CV(RMSE)}} = 1$. Thus, the estimation of a Pareto front can be achieved by running single-objective optimizations with various weights.

In practice, energy and environment researchers would prefer the model to capture the bias (i.e., NMBE) more precisely than the variation (i.e., CV(RMSE)). This argument has been supported by the ASHRAE Guideline 14-2002 (RP-1051) [36,37], which recommends a ratio of 3:1 for w_{NMBE}: $w_{\mathrm{CV(RMSE)}}$. Therefore, in the present study, we adopt this rule and set the w_{NMBE} and $w_{\mathrm{CV(RMSE)}}$ to be 0.75 and 0.25, respectively.

2.2. Monte Carlo filtering

Before conducting a fine calibration, it is important to reduce the dimensionality of the parameter space by identifying strong and weak parameters among the set of candidate inputs. Another intent is to detect the local ranges most likely to contain the "actual" value for each parameter. This can be achieved by adopting a blind Monte Carlo (MC) search coupled with sensitivity analysis (SA). In general, MC filtering is a process of rejecting sets of model outputs that are far from reality, while the SA can produce an influential parameter set that meets some prescribed criteria. Once the strong parameters are determined, the weak parameters can be fixed at some nominal values.

We have illustrated the details of regression-based SA through MC sampling in a companion paper [15]. In particular, the Latin Hypercube (LH) sampling strategy is applied in this study due to its efficient stratification properties [38]. After an LHMC batch run and a regression-based analysis are completed, the strong parameters (with an average of the absolute standardized regression coefficients large than 2.5% over 24 h in one day [15]) can be identified and the GOF indices can be computed for each trial. Those input vectors that result in unfavorable GOF values will be ruled out. On the other hand, the

information contained in the "good" input vectors could be used to set values for the weak parameters, which will then be removed from further consideration in the calibration process. In particular, we use the average values of these promising input vectors to fix the weak parameters.

Once the MC filtering has identified a set of strong parameters, we can use the information advantageously to further fine-tune these influential inputs via heuristic-based search methods, such as the evolutionary algorithm.

2.3. Online hyper-heuristic evolutionary algorithm

Evolutionary algorithms (EAs) have been widely used in a variety of complex real-world applications. However, EAs need to perform a large number of fitness (or objective) function evaluations to obtain optimal or near-optimal solutions. For engineering problems, each fitness function is evaluated via physics-based simulation, which often makes the whole process computationally expensive. Hyper-heuristics represent a class of methods that could address this barrier and reduce computational cost.

A hyper-heuristic search method seeks to automate, often by incorporation of statistical or machine learning techniques, the process of handling several simpler heuristics to efficiently solve computational search problems. The overall goal is to reduce the number of numerical simulations along a search path at the algorithm level. One popular paradigm is to construct a so-called surrogate or meta-model that can approximate the behavior of the original fitness function in the optimization process. The idea of hyper-heuristics can be traced back to the 1960s [39], while managing approximate models in optimization via EAs has received continuous attention since the 1990s [40].

Based on how surrogate models are used during the evolution, the hyper-heuristics can be further classified as *offline* and *online*. Offline hyper-heuristics utilize a surrogate model that is trained in advance, separately from the evolution process. This needs to be pre-validated and the surrogate must be updated to support a new case study. In online hyper-heuristics, a surrogate model is continuously retrained during the evolution and thus can make use of the most recent data. The self-updating mechanism without any pre-simulated database enables the online method to flexibly allow for many "plug-and-play" applications. Therefore, online hyper-heuristics have been very popular in engineering optimization problems [41].

Although many researchers have utilized offline hyper-heuristics to calibrate building energy models [42,43], little is known to the authors about the use of online hyper-heuristics. Very recently, some studies have been published to develop the online hyper-heuristic EAs for simulation-based multi-objective optimization in building design [44,45]. Therefore, we opt to embrace the methodology proposed by Brownlee and Wright [44] and develop an online hyper-heuristic EA to calibrate an urban microclimate model. There are many possible ways to implement the general idea of online hyper-heuristics. The aim of this paper is not to explore this whole space but simply to illustrate that one fairly straightforward application works well and that online surrogate model helps. As an illustrative example, the evolutionary strategy (ES) [46] in the MATLAB environment and the support vector regression (SVR) [47] from LIBSVM [48] are selected for the present study due to their wide applications. Table 1 summarizes the associated hyper-parameters. Although the convergence behavior of EA could be significantly affected by the hyper-parameters [61], we do not intend to explore this topic here and merely adopt some common settings.

Fig. 3 depicts the procedure of an individual-based model management strategy as described in Ref. [49]. In each generation, a local surrogate (SVR) model is trained if necessary. The role of SVR is to approximately evaluate the offspring in each generation and identify the most promising individuals among them, which will then be evaluated via expensive simulation. The latter are, of course, recorded and can be exploited as training patterns in the forthcoming generations. It is important to note that the surrogate model is *not* used directly to examine the convergence behavior.

The pseudo-code outline of the proposed online hyper-heuristic EA is shown in Fig. 4. The blue part (Steps 5-14) depicts where and how a local surrogate model is built within the EA. At generation t, a surrogate (SVR) model needs to be updated if its accuracy is not acceptable. The SVR model is trained using the database D that consists of previous populations P and their expensive evaluations $F_{E,P}$ (Steps 5-10). Then, crossover and mutation operators are used to generate a surrogate population S (with size $|S| > |P|$) (Step 11). The SVR approximates the objective values in S (Step 12), and the ES ranks S (Step 13) with the highest ranking solutions inserted into the offspring population Q (with size $|Q| = |P|$) (Step 14). Q are then evaluated using the expensive UWG simulations, which will be appended into the database D if they are not in D before. Finally, the offspring Q are assigned to be the parent P in the next generation.

The whole process continues until the evolution meets a stop criterion, e.g., the search exceeds the maximum number of generations or the maximum number of expensive simulations. In addition, we use a fitness prediction correlation (FPC) to measure the surrogate model accuracy. Usually the FPC is the Spearman's rank correlation, between the exact and approximated values, for a solution set. The threshold FPC for retraining is set as 0.7 [44] for the present case study.

3. Case study

The case study was conducted in District E3, which is representative of the large city blocks in downtown Abu Dhabi (see Fig. 5). The climate in Abu Dhabi is characterized by a very hot summer from July to September, a mild winter from December to February, and insufficient rainfall throughout the year. The total land area of District E3 is about 193,351 m^2. Altogether 70 buildings are considered in the model, including 59 residential buildings, five office buildings, three hotels, one school, one mosque, and one hospital. As shown in Fig. 5(b), there is a perimeter of high-rise buildings on the outer borders surrounding a number of medium- and low-rise buildings in the inner part.

The parameters used for establishing the baseline model can be categorized into four groups: meteorological factors, urban characteristics, vegetation variables, and building systems. Descriptions of the

Table 1
Hyper-parameters of the evolutionary algorithm and support vector regression used in the model calibration.

Hyper-parameter	Setting
Evolutionary algorithm	
Algorithm type	Evolutionary strategy
Objective function	Goodness-of-fit
Parent population size $\|P\|$	120
Offspring population size $\|Q\|$	120
Surrogate population size $\|S\|$	360
Maximum number of generations	60
Maximum number of expensive evaluations	2040
Crossover algorithm	Laplace
Crossover probability	0.8
Mutation algorithm	Power
Mutation probability	0.005
Proportion of elitism	0.2
Support vector regression	
Regression type	ε-SV regression
Kernel function	Gaussian radial basis function
Optimization method	Grid search
Cross-validation	5 folds

Note: The hyper-parameters are set based on prevailing engineering practices, previous related studies [40,41,44,49,60], and trial and error. In particular, the surrogate population size is set to be three times of the parent population size [44], while the parent population size is set to be 10 times of the number of design parameters for single-objective optimization [60]. The settings for the crossover and mutation operator follow the suggestions in Ref. [60].

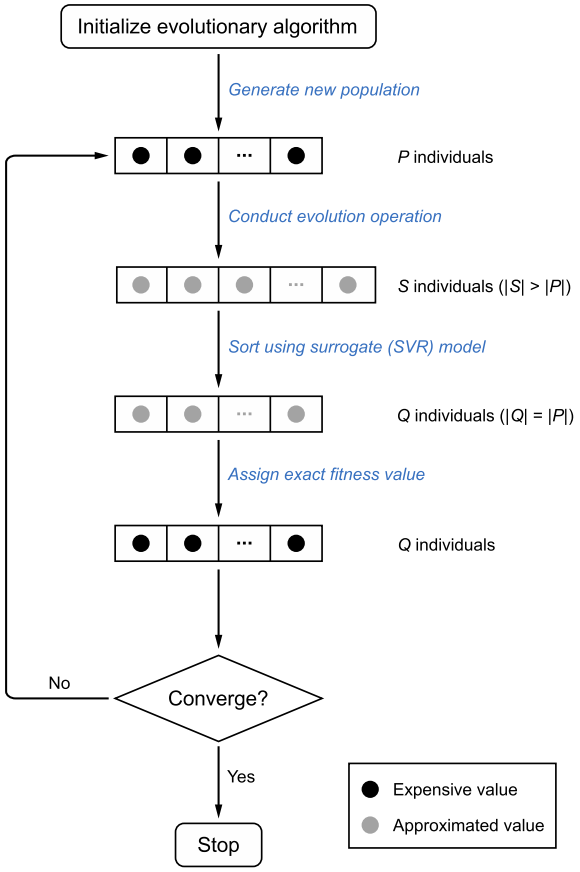


Fig. 3. General flowchart of an online hyper-heuristic EA (mainly from Ref. [49]).

Pseudo-code Online hyper-heuristic evolutionary algorithm

```
initialize(P_t) at t = 1                          ▷ Initialize new population
F_{E,P_t} = obj-and-const-calc(P_t)               ▷ Expensive objective and constraint in P
D = cache(P_t, F_{E,P_t})                         ▷ Store training data P and F into D
while not terminated do
    for all objective and constraint do
        compute FPC in P_t                        ▷ Surrogate model performance
        if FPC < 0.7 then
            update-model(D)                       ▷ Retrain model within the current generation
        end if
    end for
    S = make-new-pop(P_t)                         ▷ Create surrogate population S (with size |S| > |P|)
    F_S = obj-and-const-approx(S)                 ▷ Approximated objective and constraint in S
    S = rank(F_S, S)                              ▷ Rank individuals based on fitness values
    Q_t = extract(S)                              ▷ Pre-select Q individuals (with size |Q| = |P|)
    F_{E,Q_t} = obj-and-const-calc(Q_t)           ▷ Expensive objective and constraint in Q
    D = D ∪ (Q_t, F_{E,Q_t})                      ▷ Add expensive evaluations into D
    P_{t+1} = Q_t                                 ▷ Assign Q as the next generation
    t = t + 1                                     ▷ Increase generation counter
end while
```

Fig. 4. Pseudo-code of an online hyper-heuristic EA for single-objective optimization.

corresponding values can be found in Ref. [15]. A summary of the key input parameters in the baseline model is shown in Table 2. In order to reflect the overall impact of building systems on the urban microclimate in a specific neighborhood area, we use an averaged model (AM) to include one generic building type based on the weighted average values of key building design parameters. The AM has been validated against the detailed model in Ref. [15]. In addition, based on our previous studies [12–15], we have selected 30 candidate input parameters with associated uncertainties for the case study (see Table 3).

In this study, we use the urban-rural outdoor air temperature difference to calibrate the UWG. A good fit between the predicted and measured temperature trajectories is one indication that the model has captured the thermal dynamics well. Other calibration studies based on energy use instead of air temperature may increase the solution uncertainty, since various combinations of the system properties can easily produce similar consumptions. We also find that some researchers have used air temperature to analyze or calibrate building thermal simulations [31,32,51,52].

The rural reference weather data was collected from the Masdar Institute Field Station (24.436 N, 54.612E), an isolated laboratory building in Masdar City. The station was built near the Abu Dhabi International Airport, 28 km from the downtown area (see Fig. 5(a)). The measured rural data during calendar year 2017 is used for running the UWG. Contemporaneous air temperature observations from the sensor arrays attached to six lamp poles at different sites in District E3 are used for conducting the developed calibration approaches. Each array includes temperature measurements (at 3, 4, 5.5, 7, and 8.5 m), relative humidity measurements (at 3 m), and wind measurements (at 6 m, only available in four of the six stations). The locations of these sensors are depicted as the red-cross points in Fig. 5(c). The temperature observations at 8.5 m are selected for the subsequent analysis, as they present consistent performances.

4. Result and discussion

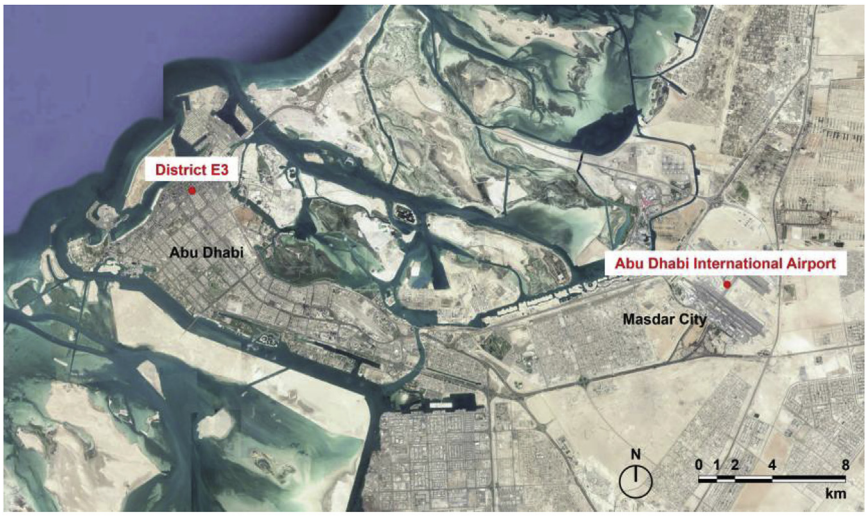
First, similar to our previous study [15], an LHMC sample matrix of size $N = 1000$ was generated via the UWG for summer (August 1–7) and another for winter (February 1–7) in 2017. For each sampled trial, the predicted weekly-average diurnal profiles of the urban outdoor air temperature were saved to enable sensitivity analysis. Once the weak parameters were identified using the same criteria specified in Ref. [15], their values were set based on the information contained in those input vectors with the lowest GOFs. The strong parameters, on the other hand, were considered to be tuned in the subsequent EA-based optimization. The objective function used to guide the EA is the GOF based on weekly-average diurnal profiles of the urban-rural outdoor air temperature difference in February 1–7 and August 1–7, 2017. That is, we are calibrating the strong parameters only based on the measurements in one summer week and one winter week.

In order to evaluate the associated uncertainty as predicted by the calibrated solutions, we ran 10 independent trials for both the traditional and online hyper-heuristic EAs. All the trials used the same stop criteria: the evolution ends when the number of expensive evaluations exceeds the maximum 2040 or the evolution lasts more than 60 generations. The goal is to find the algorithm setup returning the lowest GOF for the same number of evaluations within fixed computation budgets (i.e., better quality solution given the same effort); this goal precluded use of a convergence criterion. Also provided were comparisons of the computational effort required to obtain the same quality of solutions. An internal cache of previously-evaluated solutions is maintained in the online hyper-heuristic EA for each run, to avoid re-evaluating the same solutions. The solutions drawn from the cache do not count toward the limit of 2040. Finally, the resulting solution sets from the 10 trials of the online hyper-heuristic EA were evaluated using the measurements.

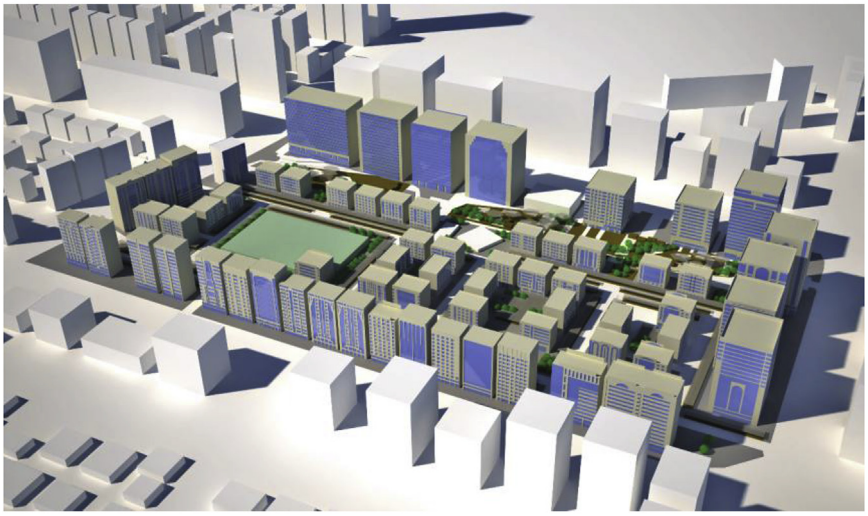
4.1. Performance of the Monte Carlo filtering

Using the regression-based analysis via Monte Carlo sampling, we identified 12 significant parameters for summer and winter in 2017 (see Table 4). The results are very consistent with those in our previous study [15], when we performed identical sensitivity analysis (SA) on the same UWG model for summer and winter in 2016. Thus, the UWG is a fairly robust simulator to approximate the thermal behavior of an urban microclimate system with a stable level of sensitivity to its inputs. On the other hand, the sensitivity difference between 2016 and 2017, although quite minor, indicates that the rural weather condition may have some impact on the UWG model. Future studies can consider this issue as more data becomes available.

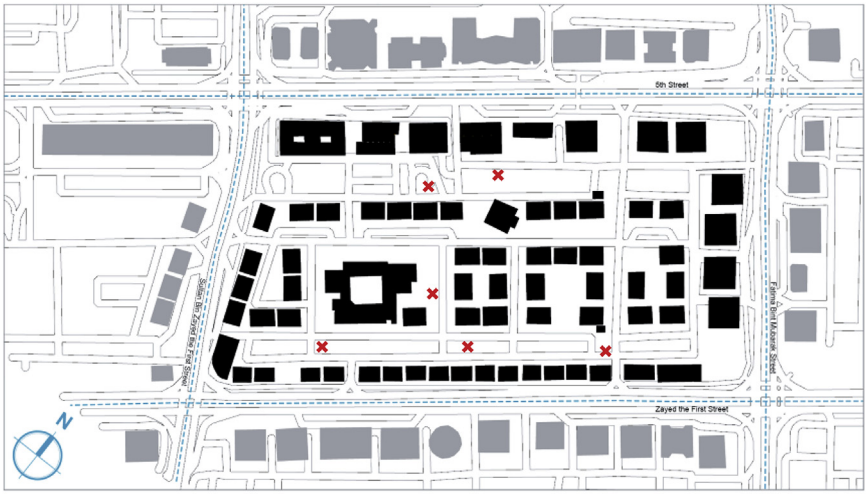
Given the results of SA, the weak parameters were then removed



(a) Satellite plan



(b) 3D model



(c) Site view

Fig. 5. District E3 in downtown Abu Dhabi (UAE).

Table 2
Inputs of the UWG used in the baseline model with field data from District E3 in Abu Dhabi (mainly from Ref. [15]).

Parameter	Setting
General information	
Location	Abu Dhabi
Latitude	24.490°
Longitude	54.366°
Simulation time-step	300 s
Weather data time-step	3600 s
Meteorological factors	
Daytime urban boundary layer height	700 m
Nighttime urban boundary layer height	80 m
Reference height of the VDM	150 m
RSM temperature reference height	10 m
RSM wind reference height	10 m
Circulation coefficient	1.2
UCM-UBL exchange coefficient	0.3
Heat flux threshold for daytime conditions	200 $W m^{-2}$
Heat flux threshold for nighttime conditions	50 $W m^{-2}$
Minimum wind velocity	0.1 $m s^{-1}$
Rural average obstacle height	0.1 m
Urban characteristics	
Average building height	35 m
Fraction of waste heat into canyon	0.3
Building density	0.24
Vertical-to-horizontal ratio	2.2
Urban area characteristic length	1000 m
Road albedo	0.165
Pavement thickness	1.25 m
Traffic sensible anthropogenic heat (peak)	19.6 $W m^{-2}$
Traffic latent anthropogenic heat (peak)	2.0 $W m^{-2}$
Vegetation variables	
Urban vegetation coverage	0.01
Urban tree coverage	0.01
Start month of vegetation participation	January
End month of vegetation participation	December
Vegetation albedo	0.25
Latent fraction of grass	0.6
Latent fraction of tree	0.7
Rural vegetation coverage	0.01
Building systems	
Glazing ratio	0.48
Wall U-value	2.50 $W m^{-2} K^{-1}$
Roof U-value	0.70 $W m^{-2} K^{-1}$
Window U-value	3.25 $W m^{-2} K^{-1}$
Window SHGC	0.58
Infiltration rate	0.60 ACH
Lighting load density	10 $W m^{-2}$
Equipment load density	13 $W m^{-2}$
Occupancy density	19 m^2 $person^{-1}$
Indoor air temperature set point	22 °C
Chiller COP	2.5

Note.
(a) Detailed physical definition of the parameters can be found in Refs. [12,50]. In particular, Ref. [12] illustrates the parameters in the rural station model and urban canopy-building energy model, while Ref. [50] illustrates the parameters in the vertical diffusion model and urban boundary layer model.
(b) For the averaged model [15], the wall-, window-, and roof-related parameters are averaged based on the wall, window, and roof area, respectively. The internal heat gains are averaged based on the floor area. The infiltration level is averaged based on the building volume.

from the subsequent calibration process. Their static values were determined, somewhat arbitrarily, using the average of the 20 input vectors with the lowest GOFs. In addition, many strong parameters may have associated higher uncertainties [15] and perhaps be more likely in need of tweaking, which can be done through EA-based optimization. The ranges of these to-be-tuned parameters were assigned by considering the uncertainties based on local building design/energy codes, prevailing engineering practices, and our previous investigations [12–15]. Table 4 summarizes the current settings to initiate the traditional and online hyper-heuristic EAs for the following calibration tests. Each strong parameter is bounded via a uniform distribution in order to evaluate the mathematical optimality in terms of calibration performance.

4.2. Performance of the online hyper-heuristics

The overall goal of our calibration method is to minimize the GOF, stated by Eqn. (1), via parameter tuning. The EA is executed to search the design space of the 12 significant parameters identified by the SA. In order to compare the performances of the traditional and online hyper-heuristic EAs, we applied both EAs to the same calibration problem for 10 independent trials. The two EAs have the same settings for crossover operators, mutation operators, elitist strategies, etc.

Fig. 6 shows the convergence behavior of the traditional and online hyper-heuristic EAs. For 2040 simulation runs, the online hyper-heuristic EA has, on average, converged to a slightly better near-optimal value than the traditional EA. That is, given the same computation budget, the online hyper-heuristic EA is able to find a slightly better objective value, on average, than the traditional EA for the present case study. In addition, the average number of expensive evaluations (i.e., the UWG simulations) required for the online hyper-heuristic EA to reach a near-optimal value is only around 1000, while that for the traditional EA roughly exceeds 2000. However, it is important to note that Fig. 6 only depicts the number of computationally expensive evaluations (i.e., simulation runs). The online hyper-heuristic EA should require a larger number of objective function evaluations than the traditional EA, because it needs to evaluate an additional surrogate population in each generation. The computation time for the function evaluations in the surrogate is negligible compared with that for the expensive evaluations via the UWG, even though the number of additional function evaluations in the surrogate is large. One could also compute the total number of function evaluations, including both expensive and surrogate evaluations, to see the difference (an option for future work), but what really matters in practice is the number of expensive evaluations during a simulation-based optimization. Therefore, the online hyper-heuristic EA seems to be about *twice* as fast, on average, as the traditional EA for the same level of accuracy. In this case study, where a single simulation costs about 2 min on a single thread, using the online hyper-heuristic EA can approximately save at least one day, on average, for one calibration trial if no parallel computation is applied.

In addition, the optimum uncertainties of the two EAs (i.e., the shaded area inside the dashed lines in Fig. 6) are quite different. Over these 10 trials, the online hyper-heuristic EA has a smaller uncertainty band, especially after ~1000 expensive evaluations, which means that the results of one single trial of the online hyper-heuristic EA are more reliable and robust. In contrast, the traditional EA yields a much wider uncertainty band, which somewhat compromises the best fitness value identified in only one calibration trial. In single-objective optimization, a well-performed EA can produce most solutions that are expected to be clustered around the global optimum. Some others could be clustered around local optima and some outlying individuals may exist as well. Outliers are generated due to mutation, which intends to prevent the solution from being trapped in local optima. Searching solutions mostly in a specific parameter space would act in favor of the online surrogate model, whose predictive capability is progressively increased as more objective functions are evaluated. Therefore, the online hyper-heuristics can fairly produce more confident solutions when computation time is limited.

The solutions obtained from both the traditional and online hyper-heuristic EAs are summarized in Table 5. We found great consistency between the two algorithms over 10 trials, leading us to conclude that the EA is able to identify the sub-ranges of most strong parameters in our current settings. Besides, the solution uncertainties from the online hyper-heuristic EA are, in general, smaller than those from the traditional EA. This reinforces our argument that, in single-objective

Table 3
Uncertainty of the model parameters for District E3 in Abu Dhabi (mainly from Ref. [15]).

Group	No	Parameter	Unit	Distribution	Uncertainty
Meteorological factors	A1	Daytime urban boundary layer height	m	Uniform	500–1000
	A2	Nighttime urban boundary layer height	m	Uniform	50–100
	A3	Reference height of the VDM	m	Uniform	100–200
	A4	Circulation coefficient	–	Uniform	0.8–1.2
	A5	UCM-UBL exchange coefficient	–	Uniform	0.1–0.9
	A6	Heat flux threshold for daytime conditions	$W\ m^{-2}$	Uniform	150–250
	A7	Heat flux threshold for nighttime conditions	$W\ m^{-2}$	Uniform	40–60
Urban characteristics	B1	Average building height	m	Normal	35 ± 5
	B2	Fraction of waste heat into canyon	–	Uniform	0.1–0.9
	B3	Building density	–	Normal	0.25 ± 0.10
	B4	Vertical-to-horizontal ratio	–	Normal	2.2 ± 0.5
	B5	Urban area characteristic length	m	Uniform	800–1200
	B6	Road albedo	–	Normal	0.165 ± 0.080
	B7	Traffic sensible anthropogenic heat (peak)	$W\ m^{-2}$	Normal	20 ± 5
Vegetation variables	C1	Urban grass coverage	–	Uniform	0–0.1
	C2	Urban tree coverage	–	Uniform	0–0.1
	C3	Vegetation albedo	–	Normal	0.25 ± 0.05
	C4	Latent fraction of grass	–	Uniform	0.45–0.75
	C5	Latent fraction of tree	–	Uniform	0.5–0.9
	C6	Rural vegetation coverage	–	Uniform	0–0.1
Building systems	D1	Glazing ratio	–	Normal	0.5 ± 0.15
	D2	Wall U-value	$W\ m^{-2}\ K^{-1}$	Normal	2.5 ± 1
	D3	Window U-value	$W\ m^{-2}\ K^{-1}$	Normal	3.25 ± 1
	D4	Window SHGC	–	Normal	0.60 ± 0.15
	D5	Infiltration rate	ACH	Uniform	0.1–0.7
	D6	Chiller COP	–	Uniform	2–4
	D7	Indoor air temperature set point	°C	Uniform	20–24
	D8	Equipment load density	$W\ m^{-2}$	Normal	13 ± 3
	D9	Lighting load density	$W\ m^{-2}$	Normal	10 ± 3
	D10	Occupancy density	$m^2\ person^{-1}$	Uniform	15–25

Note.
(a) The parameter associated with aleatory uncertainty is assumed to have a normal distribution, which is suitable for measured physical properties. The parameter associated with epistemic uncertainty is approximated with a uniform distribution, which represents that all the values are equally likely to happen. Detailed descriptions can be found in Ref. [15].
(b) For the parameter assumed to have a normal distribution, the uncertainty is represented as ($\mu \pm 3\delta$), where μ is the mean and δ is the standard deviation of the distribution.
(c) The parameter uncertainty is mainly assigned based on the data taken from local building design/energy codes provided by the Abu Dhabi Municipality (via personal contact), prevailing engineering practices, and previous related investigations [12–15].
(d) Detailed physical definition of the parameters can be found in Refs. [12,50]. In particular, Ref. [12] illustrates the parameters in the rural station model and urban canopy-building energy model, while Ref. [50] illustrates the parameters in the vertical diffusion model and urban boundary layer model.

optimization, the online surrogate model can help EA produce the solutions that are robustly closer to the global optimum with much less computation time.

One interesting observation in Table 5 is that neither algorithm is able to identify an appropriate sub-range for some parameters (e.g., D5), since the corresponding standard deviation is quite comparable to the average value. On the other hand, the optimal solution of other parameters (e.g., A5) can be determined almost surely. Under a modern optimization lens, there are two possible reasons. First, if the objective function is not quite as sensitive to the parameter (compared to other parameters), its solution could be easily varied during a purely mathematical search and the resulting uncertainty would become large. Second, even if the parameter is quite influential, an algorithm may still fail to reach a single optimal or near-optimal region, since the objective function with respect to this parameter could be highly non-convex, multi-modal, and/or locally-flat. Thus, the observed uncertainties from purely mathematical search in this study would naturally come up with an appealing motivation to investigate the convexity of the target parameter space in urban-scale simulation settings. Future studies could be considered along this direction.

Whether the proposed algorithm's performance can be considered "good" from the perspective of why the calibration was considered in the first place is addressed in the next subsection.

4.3. Performance of the calibrated model

Now we analyze in more detail the behavior of the solutions obtained from the online hyper-heuristic EA. Given the measured data at hand, four periods in 2017 are considered for validation of the calibrated UWG models, i.e., January 15–21, February 8–14, July 15–21, and August 8–14. Based on the assumption that the measurement uncertainties impose a limit to the model accuracy that one could hope to achieve, we compared the values measured by the sensors, the values predicted by the optimization-calibrated models, and the values predicted by the manually-calibrated baseline model (detailed in Table 2). The final results are shown in Fig. 7 and Table 6, where the accuracy is assessed in terms of how close the predicted values are to the measured values and whether the uncertainty bands are narrow enough to be of practical use while bounding the measurements.

Overall, the calibrated solutions (black curve) produce weekly-average diurnal profiles of the urban outdoor air temperature similar to the baseline solution (red curve), as shown in Fig. 7. Sometimes the calibrated models represent the observed behavior better than the baseline model (e.g., January). This is very encouraging since the baseline model has been manually refined and calibrated for over one year [15] via exhaustive investigations of the local construction documents, regular on-site visits, and detailed discussions with experienced engineers and building management personnel. In addition to the performance guarantee from the calibrated models, one great advantage of the methodology proposed in this paper lies in the time consumed to

Table 4
List of the strong and weak parameters from the Monte Carlo filtering and global sensitivity analysis.

Strong parameter	Unit	Range
Nighttime urban boundary layer height (A2)	m	[50, 100]
Reference height of the VDM (A3)	m	[100, 200]
UCM-UBL exchange coefficient (A5)	–	[0.1, 0.9]
Average building height (B1)	m	[30, 40]
Fraction of waste heat into canyon (B2)	–	[0.1, 0.9]
Building density (B3)	–	[0.15, 0.35]
Urban area characteristic length (B5)	m	[800, 1200]
Infiltration rate (D5)	ACH	[0.1, 0.7]
Chiller COP (D6)	–	[2, 4]
Indoor air temperature set point (D7)	°C	[20, 24]
Equipment load density (D8)	$W\ m^{-2}$	[10, 16]
Occupancy density (D10)	$m^2\ person^{-1}$	[15, 25]
Weak parameter	**Unit**	**Value**
Daytime urban boundary layer height (A1)	m	753.31
Circulation coefficient (A4)	–	0.98
Heat flux threshold for daytime conditions (A6)	$W\ m^{-2}$	199.04
Heat flux threshold for nighttime conditions (A7)	$W\ m^{-2}$	50.59
Vertical-to-horizontal ratio (B4)	–	2.19
Road albedo (B6)	–	0.16
Traffic sensible anthropogenic heat (peak) (B7)	$W\ m^{-2}$	20.35
Urban grass coverage (C1)	–	0.04
Urban tree coverage (C2)	–	0.04
Vegetation albedo (C3)	–	0.25
Latent fraction of grass (C4)	–	0.61
Latent fraction of tree (C5)	–	0.70
Rural vegetation coverage (C6)	–	0.04
Glazing ratio (D1)	–	0.50
Wall U-value (D2)	$W\ m^{-2}\ K^{-1}$	2.52
Window U-value (D3)	$W\ m^{-2}\ K^{-1}$	3.18
Window SHGC (D4)	–	0.60
Lighting load density (D9)	$W\ m^{-2}$	9.90

Note.
(a) The values of the weak parameters are set using the average of the top 20 promising input vectors in terms of the GOF.
(b) The identified strong parameters along with the associated ranges will be adjusted through an EA-based calibration process.

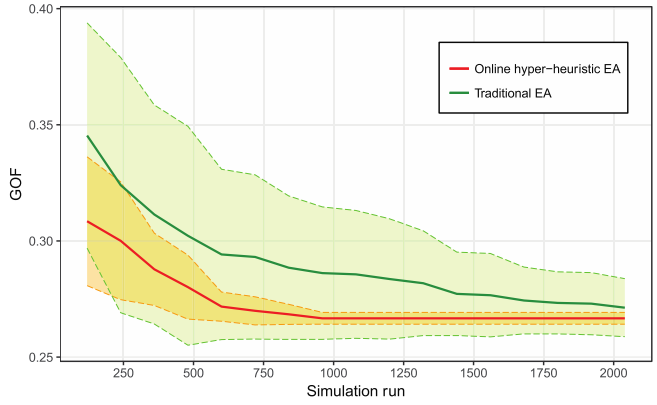


Fig. 6. Convergence behavior of the traditional and online hyper-heuristic EAs in terms of the UWG simulation runs (i.e., expensive evaluations). The solid line represents the average of the best fitness values over the 10 trials for the two algorithms, respectively. The shaded area inside the dashed lines represents one standard deviation of the best fitness values over the 10 trials for the two algorithms, respectively.

obtain them. For this case study, a total of only ~1000 simulations (which take at most two days without parallel computing) on average have resulted in an improved urban microclimate model. So, it seems highly cost-effective and convenient to use the developed algorithm in the process of model calibration.

A further consideration of the hourly data is shown in Fig. 8, where four days in each validation period are depicted. In general, the calibrated solutions are able to capture most trends as well as peaks and valleys of the measurements. This is impressive because in this study we calibrated the model parameters for a whole year based on just two weekly-average diurnal profiles of the urban-rural outdoor air temperature difference. More data with higher spatial and temporal resolution are being collected and it is reasonable to expect that, if these data were used to calibrate the model, the resulting solutions could achieve higher accuracies with lower uncertainties. Whether the improved performance is commensurate with the extra resources and efforts required to perform accurate measurement and conduct optimization-aided calibration on urban-scale models is not clear and should be investigated in the future.

Finally, one particular aspect to notice in Fig. 8 is the unexpected behavior during July 17–18, where large discrepancies between predicted and measured data are evident. A good fit to the weekly-average diurnal profile (i.e., the one with low GOF shown in Fig. 7) may not necessarily predict the hourly diurnal profile more accurately. One possible reason is that, there may be some short-term physical activities (e.g., anonymous wind patterns) during that time which can largely affect the corresponding urban microclimate condition, while such physical activities have not been adequately modeled in the current UWG or have not been reflected in the given rural weather data. This strong bias needs to be revisited in our future studies and may call into question a conclusion from a previous UWG assessment that the location of the reference station has minimal impact on the estimate of temperatures at district level because the urban-canopy energy balance is weakly influenced by advection in the urban boundary layer [13]. Nevertheless, in most cases, the differences are quite acceptable given the state of the art of urban microclimate modeling [9–15].

4.4. Discussion and future work

For simulation-based optimization problems, hyper-heuristics have a catalytic effect in obtaining solutions, but have been overshadowed by the success of parallel computing [53]. Indeed, some researchers have recently tried to improve the run-time of optimization algorithms via surrogate models [54–56]. However, most studies focused on offline hyper-heuristics where a surrogate or meta-model is trained in advance. This would require additional effort to build a database for a specific case study and the algorithm cannot fully guarantee equivalent quality solutions if the true simulation engine is discarded. Online hyper-heuristics, on the other hand, have a self-updating mechanism without any pre-simulated database to produce well-verified solutions [44,45]. Therefore, we advocate the use of online hyper-heuristics among the relevant research communities that are interested in simulation-based optimization.

The present study only considered a model calibration based on the urban outdoor air temperature via single-objective optimization. Energy use or other urban microclimate conditions (e.g., air humidity) at multi-layer level could be further incorporated into the calibration process if the corresponding data were available at sufficiently high frequency. Given hourly or sub-hourly data of the outdoor microclimate and/or building energy use in an urban neighborhood, one could conduct simultaneous model calibration via multi-objective optimization (MOO) [57]. The performance of the online hyper-heuristic EA could be further evaluated in this MOO setting. Another direction for future studies would be to develop automatic physics-based calibration methods [58] or Bayesian calibration methods [62] in order to further improve the results from a purely heuristic-based search. A joint mathematics- and physics-based calibration approach that is able to effectively integrate the actual measurements and computer simulations has the potential to improve the way building and urban systems are designed and operated.

At a more fundamental and philosophical level, any comparison between the predicted and measured performance can be viewed in

Table 5
Calibration results of the selected parameters over the 10 trials of EA-based optimization.

Parameter	Unit	Result from traditional EA	Result from online hyper-heuristic EA
Nighttime urban boundary layer height (A2)	m	79.51 ± 13.92	89.57 ± 9.98
Reference height of the VDM (A3)	m	171.96 ± 29.09	195.09 ± 11.86
UCM-UBL exchange coefficient (A5)	–	0.10 ± 0.00	0.10 ± 0.00
Average building height (B1)	m	34.48 ± 2.42	35.68 ± 1.16
Fraction of waste heat into canyon (B2)	–	0.14 ± 0.03	0.16 ± 0.01
Building density (B3)	–	0.33 ± 0.02	0.33 ± 0.01
Urban area characteristic length (B5)	m	1116.43 ± 108.50	1176.47 ± 29.97
Infiltration rate (D5)	ACH	0.24 ± 0.13	0.15 ± 0.09
Chiller COP (D6)	–	3.25 ± 0.87	3.86 ± 0.05
Indoor air temperature set point (D7)	°C	21.22 ± 1.50	20.65 ± 1.05
Equipment load density (D8)	$W\ m^{-2}$	14.19 ± 1.09	14.45 ± 0.43
Occupancy density (D10)	$m^2\ person^{-1}$	20.13 ± 3.09	22.43 ± 0.93

Note: The results are presented as "average ± one standard deviation" from the solutions over the 10 trials for the two algorithms, respectively.

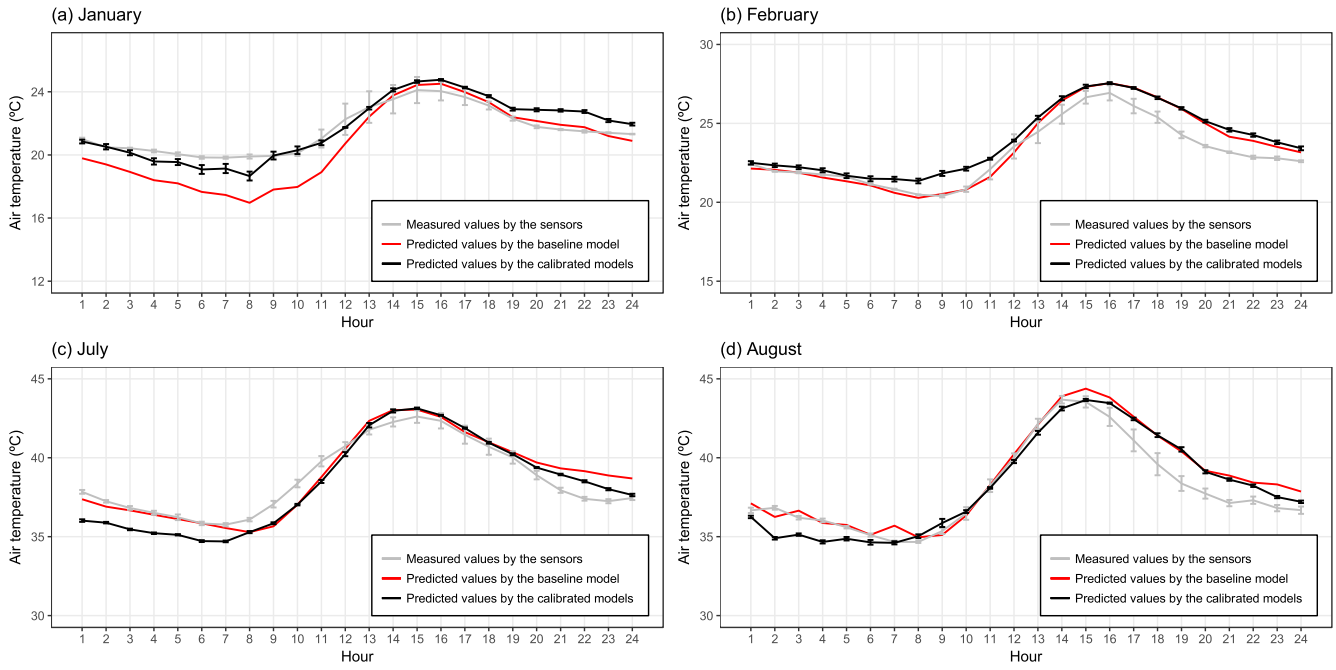


Fig. 7. Weekly-average diurnal profiles of the urban outdoor air temperature: (a) Data between January 15 and 21, 2017; (b) Data between February 8 and 14, 2017; (c) Data between July 15 and 21, 2017; (d) Data between August 8 and 14, 2017. The error bar represents one standard deviation of the measured values by the sensors and the predicted values by the 10 calibrated models, respectively. The baseline model has been manually calibrated via detailed investigations [15].

Table 6
Performance of the baseline and calibrated UWG models based on two weekly-average diurnal profiles of the urban-rural outdoor air temperature difference.

Scenario	January 15–21			February 8–14			July 15–21			August 8–14		
Evaluation metric	NMBE	CV(RMSE)	GOF	NMBE	CV(RMSE)	GOF	NMBE	CV(RMSE)	GOF	NMBE	CV(RMSE)	GOF
Measured values *v.s.* predicted values (baseline model)	−0.27	0.43	0.29	0.36	0.69	**0.40**	0.18	1.06	**0.38**	0.57	0.87	0.61
Measured values *v.s.* predicted values (calibrated model)	0.05	0.21	**0.08**	0.77	0.90	0.78	−0.41	1.22	0.55	0.17	0.89	**0.33**

The bold type highlights which model performs better in terms of the GOF for each scenario.
Note: For the measured values by the sensors and predicted values by the calibrated models, we use the average to evaluate the performance. The baseline model has been manually calibrated via detailed investigations [15].

terms of not only how well the simulation agrees with the measurement, but also *whether the simulation program is good enough for its intended purposes* [36]. The question of interest is, then, how to determine a "good enough" solution? In particular, there should be a broad consensus among the scientific and engineering communities when it comes to the specification of the accuracy bound at different scales for the calibration to be deemed satisfactory. Scientists prefer to describe such bound as *uncertainty*, which is one key idea we want to deliver in this paper. The input uncertainties indicate the difficulties in capturing the inherent physical properties (e.g., model parameter values) during a specific simulation setting. It is assumed that, if the model is calibrated within the prescribed criteria, it seems closer to the physical reality as the input parameters (i.e., the "knobs") are tuned properly.

However, just because all the selected knobs yield a desired output, we cannot guarantee that each knob is tuned correctly [36]. Most simulation models have many degrees of freedom and, with judicious fiddling, can be easily manipulated to produce any desired behavior with both plausible model structures and parameter values. More often

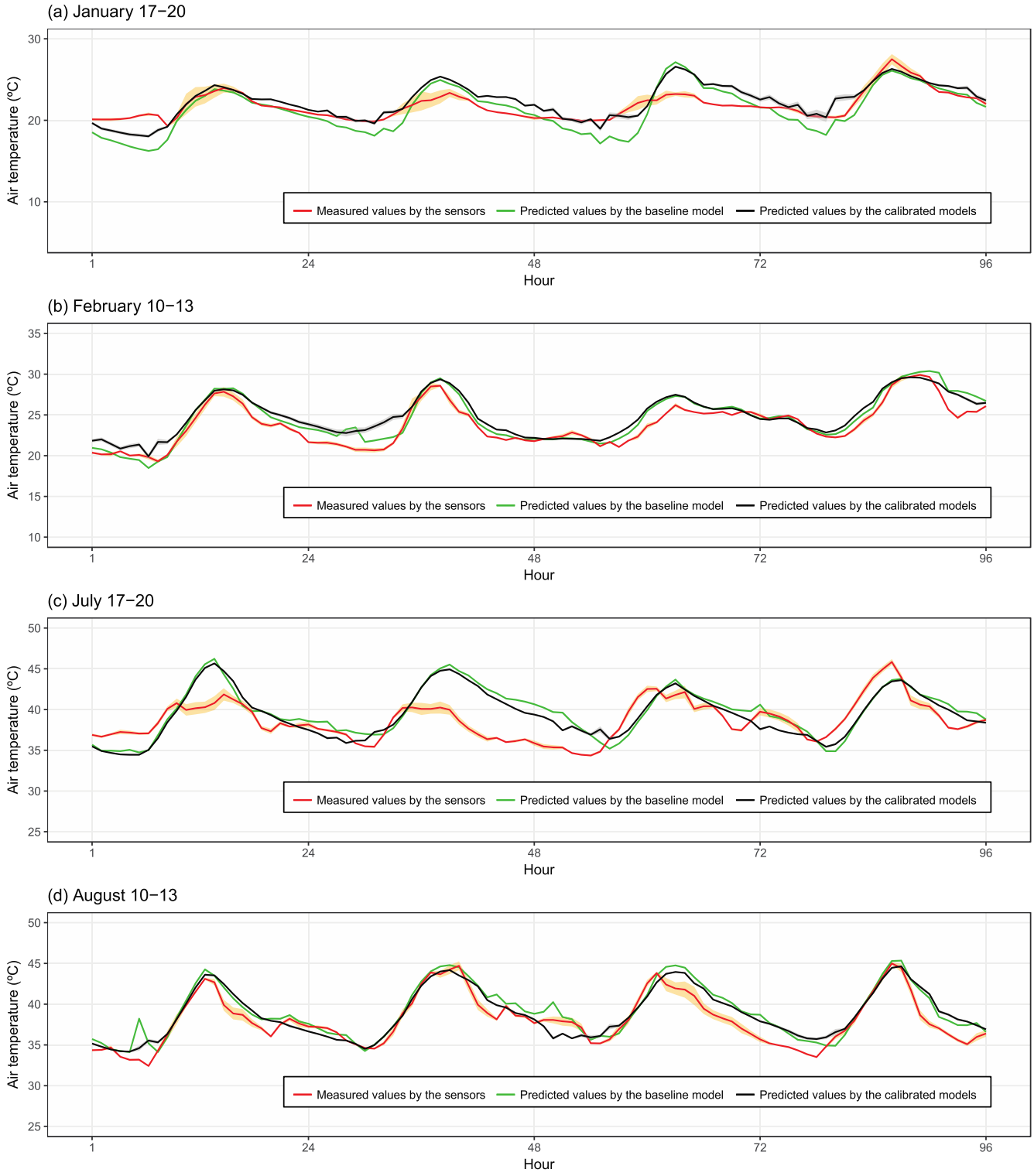


Fig. 8. Hourly diurnal profiles of the urban outdoor air temperature. The solid line represents the average of the measured values by the sensors and the predicted values by the 10 calibrated models, respectively. The shaded area represents one standard deviation of the measured values by the sensors and the predicted values by the 10 calibrated models, respectively. The baseline model has been manually calibrated via detailed investigations [15].

than not, this calibration process can be seen as GIGO (garbage in, garbage out) [16]. The reason may lie in the fact that, during an optimization process, we usually obtain one specific parameter combination leading to a "local" optimum within the search space. Thus, the output (or solution) uncertainties, which demonstrate the consistency of a calibration process, should also be considered with great care after calibration. Earlier, Kaplan et al. [59] looked at this issue and concluded that one can never hope to identify the parameters correctly, in part because we do not know what is correct. Given this unanswerable (and maybe hopeless) situation, limiting ourselves to one plausible solution would be quite misleading. It is hence much more reasonable to incorporate both the input and output uncertainties when evaluating and using a calibrated model.

5. Conclusion

The extreme complexity of a building or urban system leads to difficulties in estimating the benefits and drawbacks of present and future adaptation strategies to climate change and energy concern. The design, analysis, and optimization of modern building and urban systems may benefit significantly from the implementation of energy and environmental simulation tools at different scales. However, in many cases, studies have revealed large discrepancies between modeled and measured values, which somewhat undermines our confidence in the practical value of these simulation tools. A well-calibrated model is hence one of the key bases for practitioners to perform simulation-based analysis.

This paper illustrates a general methodology for automatic model calibration and applies it to an urban microclimate system simulated by the Urban Weather Generator (UWG). The UWG model is calibrated using heuristic-based optimization based on the urban outdoor air temperature in an existing district area located in downtown Abu Dhabi (UAE). In consideration of the time-constrained nature of engineering applications, an online hyper-heuristic evolutionary algorithm (EA) is proposed and developed to accelerate the calibration process. To our best knowledge, this is the first time that an automatic calibration is systematically performed on physics-based urban microclimate models.

We started by recognizing that calibration remains an indeterminate or over-parametrized problem which could yield non-unique solutions. Hence, it is more reasonable to identify a set of most plausible solutions and to incorporate uncertainty when evaluating and using a calibrated model. Together with a companion paper [15], we performed global sensitivity analysis, Monte Carlo filtering, and optimization-aided calibration on the UWG model using the measurements in 2017. Validation of the proposed calibration method was more of an empirical nature.

Based on 30 candidate inputs from the meteorological factors, urban characteristics, vegetation variables, and building systems, the regression-based analysis is able to identify 12 strong parameters for summer and winter in 2017. This is highly consistent with the results we obtained for the same case study in 2016 [15], indicating that the UWG is a fairly robust simulator to approximate the urban microclimate condition with a stable level of sensitivity to its inputs. These 12 parameters are then considered into an optimization-aided calibration process based on the measurements in one summer week and one winter week during 2017.

The proposed online hyper-heuristic EA is roughly twice as fast, on average over 10 trials for the present case study, as the traditional EA to achieve the same objective. In addition, both algorithms are able to identify the sub-ranges of most strong parameters in our current settings, while the solutions from the online hyper-heuristic EA have generally smaller uncertainties. In single-objective optimization, searching solutions mostly in a specific parameter space would act in favor of the online hyper-heuristics, which can thus help EA produce the solutions that are robustly closer to the global optimum with much less computation time.

The solutions obtained from the online hyper-heuristic EA can produce weekly-average diurnal profiles of the urban outdoor air temperature similar to the manually-calibrated baseline solution, which has been extensively investigated for over one year. This is encouraging because a total of only $\sim$1000 simulations (which take at most two days without parallel computing for the present study) could result in an improved urban microclimate model. In addition, the calibrated solutions are able to capture most trends as well as peaks and valleys of the measured data on an hourly basis for certain periods of the year. Despite some yet unexplained behaviors, the calibrated models generally perform well. Therefore, the automatic calibration method proposed in this study is expected to improve the model performance to some extent, both effectively and efficiently.

6. Declaration of interest

None.

Acknowledgment

This work was funded under the Cooperative Agreement between the Masdar Institute of Science and Technology, Abu Dhabi, UAE and the Massachusetts Institute of Technology, Cambridge, MA, USA, Reference Number 02/MI/MIT/CP/11/07633/GEN/G/00. The authors gratefully acknowledge the Abu Dhabi Municipality for the administrative and technical support, as well as the anonymous reviewers for their fruitful comments.

References

[1] T.F. Stocker, D. Qin, G.K. Plattner, et al., Climate Change 2013: the Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, 2014.
[2] United Nations, Department of Economic and Social Affairs, Population Division, World Population Prospects: the 2015 Revision, Key Findings and Advance Tables, (2015).
[3] T.R. Oke, City size and the urban heat island, Atmos. Environ. 7 (8) (1967) 769–779 (1973).
[4] D.P. Johnson, J.S. Wilson, The socio-spatial dynamics of extreme urban heat events: the case of heat-related deaths in Philadelphia, Appl. Geogr. 29 (3) (2009) 419–434.
[5] K. Gabriel, W.R. Endlicher, Urban and rural mortality rates during heat waves in Berlin and Brandenburg, Germany, Environ. Pollut. 159 (8) (2011) 2044–2050.
[6] V. Masson, C. Marchadier, L. Adolphe, et al., Adapting cities to climate change: a systemic modelling approach, Urban Clim. 10 (2014) 407–429.
[7] C.F. Reinhart, C.C. Davila, Urban building energy modeling – a review of a nascent field, Build. Environ. 97 (2016) 196–202.
[8] D.B. Crawley, Estimating the impacts of climate change and urbanization on building performance, J. Build. Perform. Simulat. 1 (2) (2008) 91–115.
[9] X.-X. Li, T.Y. Koh, D. Entekhabi, et al., A multi-resolution ensemble study of a tropical urban environment and its interactions with the background regional atmosphere, J. Geophys. Res.: Atmosphere 118 (17) (2013) 9804–9818.
[10] M. Ignatius, N.H. Wong, S.K. Jusuf, Urban microclimate analysis with consideration of local ambient temperature, external heat gain, urban ventilation, and outdoor thermal comfort in the tropics, Sustain. Cities Soc. 19 (2015) 121–135.
[11] S.-H. Lee, S.-U. Park, A vegetated urban canopy model for meteorological and environmental modelling, Boundary-layer Meteorol. 126 (1) (2008) 73–102.
[12] B. Bueno, L.K. Norford, J. Hidalgo, et al., The urban weather generator, J. Build. Perform. Simulat. 6 (4) (2013) 269–281.
[13] B. Bueno, M. Roth, L.K. Norford, et al., Computationally efficient prediction of canopy level urban air temperature at the neighborhood scale, Urban Clim. 9 (2014) 35–53.
[14] A. Nakano, Urban Weather Generator User Interface Development: towards a Usable Tool for Integrating Urban Heat Island Effect within Urban Design Process, SM thesis Massachusetts Institute of Technology, 2015.
[15] J. Mao, J.H. Yang, A. Afshari, et al., Global sensitivity analysis of an urban microclimate system under uncertainty: design and case study, Build. Environ. 124 (2017) 153–170.
[16] A. Saltelli, M. Ratto, T. Andres, et al., Global Sensitivity Analysis: the Primer, John Wiley & Sons, 2008.
[17] C.J. Hopfe, J.L.M. Hensen, Uncertainty analysis in building performance simulation for design support, Energy Build. 43 (10) (2011) 2798–2805.
[18] R. Rosen, Life Itself: a Comprehensive Inquiry into the Nature, Origin, and Fabrication of Life, Columbia University Press, 1991.
[19] W. Tian, A review of sensitivity analysis methods in building energy analysis, Renew. Sustain. Energy Rev. 20 (2013) 411–419.
[20] R. Evins, A review of computational optimisation methods applied to sustainable building design, Renew. Sustain. Energy Rev. 22 (2013) 230–245.
[21] A.-T. Nguyen, S. Reiter, P. Rigo, A review on simulation-based optimization methods applied to building performance analysis, Appl. Energy 113 (2014) 1043–1058.
[22] V. Machairas, A. Tsangrassoulis, K. Axarli, Algorithms for optimization of building design: a review, Renew. Sustain. Energy Rev. 31 (2014) 101–112.
[23] T.A. Reddy, Literature review on calibration of building energy simulation programs: uses, problems, procedures, uncertainty, and tools, Build. Eng. 112 (1) (2006) 226–240.
[24] D. Coakley, P. Raftery, M. Keane, A review of methods to match building energy simulation models to measured data, Renew. Sustain. Energy Rev. 37 (2014) 123–141.
[25] E. Fabrizio, V. Monetti, Methodologies and advancements in the calibration of building energy models, Energies 8 (4) (2015) 2548–2574.
[26] J.A. Wright, The Optimised Design of HVAC Systems, Loughborough University of Technology, 1986 PhD thesis.
[27] M. Wetter, Generic Optimization Program User Manual Version 3.0.0, Lawrence Berkeley National Laboratory, 2009.
[28] M. Wetter, E. Polak, A convergent optimization method using pattern search

algorithms with adaptive precision simulation, Build. Serv. Eng. Technol. 25 (4) (2004) 327–338.
[29] J.J. Robertson, B.J. Polly, J.M. Collis, Reduced-order modeling and simulated annealing optimization for efficient residential building utility bill calibration, Appl. Energy 148 (2015) 169–177.
[30] T. Yang, Y. Pan, J. Mao, et al., An automated optimization method for calibrating building energy simulation models with measured data: orientation and a case study, Appl. Energy 179 (2016) 1220–1231.
[31] G.R. Ruiz, C.F. Bandera, T.G.-A. Temes, et al., Genetic algorithm for building envelope calibration, Appl. Energy 168 (2016) 691–705.
[32] G.R. Ruiz, C.F. Bandera, Analysis of uncertainty indices used for building envelope calibration, Appl. Energy 185 (2017) 82–94.
[33] A. Saltelli, Sensitivity analysis for importance assessment, Risk Anal. 22 (3) (2002) 579–590.
[34] R. Oliva, Model calibration as a testing strategy for system dynamics models, Eur. J. Oper. Res. 151 (3) (2003) 552–568.
[35] J. Sun, T.A. Reddy, Calibration of building energy simulation programs using the analytic optimization approach (RP-1051), HVAC R Res. 12 (1) (2006) 177–196.
[36] T.A. Reddy, I. Maor, C. Panjapornpon, Calibrating detailed building energy simulation programs with measured data—Part I: general methodology (RP-1051), HVAC R Res. 13 (2) (2007) 221–241.
[37] ASHRAE, Guideline 14-2002, Measurement of Energy and Demand Savings, American Society of Heating, Refrigerating, and Air-conditioning Engineers, Atlanta, Georgia, 2002.
[38] J.C. Helton, J.D. Johnson, C.J. Sallaberry, et al., Survey of sampling-based methods for uncertainty and sensitivity analysis, Reliab. Eng. Syst. Saf. 91 (10) (2006) 1175–1209.
[39] B. Dunham, D. Fridshal, R. Fridshal, et al., Design by natural selection, Synthese 15 (1) (1963) 254–259.
[40] Y. Jin, A comprehensive survey of fitness approximation in evolutionary computation, Soft Comput. 9 (1) (2005) 3–12.
[41] Y. Tenne, C.-K. Goh, Computational Intelligence in Expensive Optimization Problems 2 Springer Science & Business Media, 2010.
[42] Z. O'Neill, B. Eisenhower, Leveraging the analysis of parametric uncertainty for building energy model calibration, Build. Simulat. 6 (2013) 365–377.
[43] M. Manfren, N. Aste, R. Moshksar, Calibration and uncertainty analysis for computer models–a meta-model based approach for integrated building energy simulation, Appl. Energy 103 (2013) 627–641.
[44] A.E.I. Brownlee, J.A. Wright, Constrained, mixed-integer and multi-objective optimisation of building designs by NSGA-II with fitness approximation, Appl. Soft Comput. 33 (2015) 114–126.
[45] W. Xu, A. Chong, O.T. Karaguzel, et al., Improving evolutionary algorithm performance for integer type multi-objective building system design optimization, Energy Build. 127 (2016) 714–729.
[46] N. Hansen, A. Ostermeier, Adapting arbitrary normal mutation distributions in evolution strategies: the covariance matrix adaptation, Proceedings of IEEE International Conference on Evolutionary Computation, 1996, pp. 312–317.
[47] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[48] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. (TIST) 2 (3) (2011) 27.
[49] Y. Jin, Surrogate-assisted evolutionary computation: recent advances and future challenges, Swarm Evol. Comput. 1 (2) (2011) 61–70.
[50] B. Bueno, J. Hidalgo, G. Pigeon, et al., Calculation of air temperatures above the urban canopy layer from measurements at a rural operational weather station, J. Appl. Meteorol. Climatol. 52 (2) (2013) 472–483.
[51] C. Spitz, L. Mora, E. Wurtz, et al., Practical application of uncertainty analysis and sensitivity analysis on an experimental house, Energy Build. 55 (2012) 459–470.
[52] C. Cornaro, D. Sapori, F. Bucci, et al., Thermal performance analysis of an emergency shelter using dynamic building simulation, Energy Build. 88 (2015) 122–134.
[53] Y. Zhang, Parallel EnergyPlus and the development of a parametric analysis tool, Proceedings of 11th International Building Performance Association (IBPSA) Conference, Glasgow, UK, 2009, pp. 1382–1388.
[54] B. Eisenhower, Z. O'Neill, S. Narayanan, et al., A methodology for meta-model based optimization in building energy models, Energy Build. 47 (2012) 292–301.
[55] F. Boithias, M El Mankibi, Pierre Michel, Genetic algorithms based optimization of artificial neural network architecture for buildings' indoor discomfort and energy consumption prediction, Build. Simulat. 5 (2) (2012) 95–106.
[56] V. Congradac, F. Kulic, Recognition of the importance of using artificial neural networks and genetic algorithms to optimize chiller operation, Energy Build. 47 (2012) 651–658.
[57] L.G.R. Santos, A. Afshari, L.K. Norford, et al., Evaluating approaches for district-wide energy model calibration considering the Urban Heat Island effect, Appl. Energy 215 (2018) 31–40.
[58] K. Sun, T. Hong, S.C. Taylor-Lange, et al., A pattern-based automated approach to building energy model calibration, Appl. Energy 165 (2016) 214–224.
[59] M.B. Kaplan, J. McFerran, J. Jansen, et al., Reconciliation of a DOE2. 1C model with monitored end-use data for a small office building, Build. Eng. 96 (1) (1990) 981–993.
[60] K. Deep, K.P. Singh, M.L. Kansal, et al., A real coded genetic algorithm for solving integer and mixed integer optimization problems, Appl. Math. Comput. 212 (2) (2009) 505–518.
[61] A. Alajmi, J. Wright, Selecting the most efficient genetic algorithm sets in solving unconstrained building optimization problem, Int. J. Sustain. Built. Environ. 3 (1) (2014) 18–26.
[62] Y. Heo, R. Choudhary, G.A. Augenbroe, Calibration of building energy models for retrofit analysis under uncertainty, Energy Build. 47 (2012) 550–560.