Accepted manuscript. To cite this article:
Ardeshir Mahdavi & Farhang Tahmasebi (2016): On the quality evaluation of behavioural models for building performance applications, Journal of Building Performance Simulation, DOI: 10.1080/19401493.2016.1230148

On the quality evaluation of behavioural models for building performance applications

Ardeshir Mahdavi, Farhang Tahmasebi

ABSTRACT

Building performance assessment applications require multiple categories of input information. These include, aside from building construction and systems and external conditions, representations of inhabitants. It has been suggested that the representation of people as passive and static entities is unlikely to yield reliable building performance assessment and building operation planning. Rather, adequate representations of building inhabitants should account for dynamics of inhabitants' presence in buildings and their control-oriented actions (e.g., interactions with buildings indoor environmental control devices and systems). To address these requirements, many recent model development efforts have explored the potential of advanced mathematical formalisms. However, the resulting occupancy-related behavioural models have rarely gone through a rigorous evaluation process. The present contribution is indeed motivated primarily by the lack of explicit procedures and guidelines for the evaluation of proposed user-related behavioural models. Specifically, we formulate a number of conditions that are necessary for systematic and dependable quality assessment of buildings' inhabitants. Toward this end, we discuss both general model evaluation requirements and specific circumstances pertaining to behavioural models of building inhabitants. Thereby, using specific instances of such models, we intend to identify the requirements of a rigorous quality assurance process with regard to behavioural models in building performance assessment applications.

1 MOTIVATION AND BACKGROUND

Building performance simulation models typically require input information on context (climate), building geometry, construction, systems, and internal processes. Whereas the specification methods regarding physical building components and properties (pertaining, for example, to buildings' fabric and construction) in building performance simulation are fairly well established, representations of inhabitants (presence, movement, behaviour, perception, and evaluation) are frequently rudimentary. Specifically, simplistic representations of people as passive and static entities have been suggested to diminish the reliability of building performance assessment and building operation planning processes (e.g. D'Oca et al. 2014; Liang et al. 2016). Rather, adequate representations of building inhabitants need to address not only building inhabitants' passive presence, but the multi-dimensional scope and the dynamic nature of their actions (e.g., interactions with buildings indoor environmental control devices and systems). A further, related phenomenon that needs to be considered in any model development activity is the inhabitants' behavioural diversity (inter-individual differences amongst attitudes, preferences, and habits) (Mahdavi & Tahmasebi, 2015; O'Brien & Gunay, 2016).

Conventional representations of buildings' inhabitants in performance simulation models mostly consist of fixed schedules (so-called diversity profiles) and rule-based action models. As such, these kinds of representations do not realistically reflect the inherent temporal fluctuations of occupancy-related processes and events (e.g., entering, leaving, and moving in buildings, operation of devices such as windows, blinds, luminaires, manipulation of control set-points, equipment usage). There has been thus recently a considerable number of efforts – for instance, by the professionals in the building performance simulation community – to develop more sophisticated dynamic models of people's presence and actions in buildings in terms of stochastic algorithms (see for example, Parys et al. 2011; Feng et al. 2015) and agent-based representations (see for example, Langevin et al. 2015; Liao et al. 2010).

A significant number of such efforts have focused on the potential of probabilistic methods and associated formalisms. Thereby, a stated objective has been to replace fix schedules and rule-based actions models in performance simulation with high-resolution probabilistic models. A number of such models have been and are being incorporated in building performance simulation applications. However, this process has not been immune to a number of unwarranted claims, misconceptions, and fallacies (Mahdavi 2011, 2015; Mahdavi & Tahmasebi, 2016). Models have been at times prematurely promoted as valid and reliable, despite wanting empirical evidence and despite ignorance regarding the down-stream deployment scenarios. The inclusion of sophisticated and realistic behavioural models in building performance assessment application is of course desirable as such. However, it must be done in a careful and systematic manner, lest confusion and poor decision making result due to uncritical implementation and application of all kinds of insufficiently tested behavioural models.

Given this background, the present contribution is primarily motivated by the lack of general procedures and guidelines for the evaluation of proposed user-related behavioural models. To encourage a deeper discourse in this area, we specifically formulate a number of conditions that are necessary for systematic and dependable enrichment of building performance assessment applications with behavioural representations of buildings' inhabitants. Toward this end, we discuss both general model evaluation requirements as well as specific circumstances pertaining to models of building inhabitants. Moreover, we present, as an illustrative case study, a potentially paradigmatic model evaluation process using a comparison of a number of recently proposed behavioural models. Thereby, our main objective is to promote a rigorous process toward quality assurance while considering and integrating behavioural representations in building performance assessment tools and processes.

2 ABOUT MODEL VALIDATION

A central trust of scientific activity is the development of models that are used to describe phenomena and predict events. Given the persistence and historical evolution of model development activity across a variety of scientific disciplines (see, for example, Hulley et al. 2013, Oleckno & Anderson 2002), one would expect that there would be no need to revisit the question of model validation in the rather narrow context of the occupancy-related behavioural models. However, at least a brief treatment would be in order, given the aforementioned shortcomings in the building inhabitants model development domain. Note that a considerable number of such shortcomings can be shown to be the consequence of the following three circumstances:

- Firstly, systematic occupancy-related studies in the context of the built environment belong to a relatively young field of inquiry. The strength of research standards in a domain typically results from expected utility and a critical mass of projects and researchers. As compared to many other areas of scientific inquiry (such as medical sciences or information technology), research pertaining to inhabitants' behaviour in building is rather underdeveloped. A closer instance for comparison purposes would be research on human comfort in general and thermal comfort in particular. The latter has a longer tradition and is arguably better established. But even in the thermal comfort domain many open research questions and challenges persist (Schweiker & Wagner, 2016; Shipworth et al. 2016).

- Secondly, a perilous problem for both model development and model evaluation lies in the rather limited availability of large-scale observational data. Consequently, the demographic basis of the majority of proposed behavioural models is often very small. The coverage and representativeness of behavioural models of buildings' inhabitants depends on the availability and fidelity of observational data. As such data is still hard to come by, models are often developed and disseminated without sufficient empirical

backing. This circumstance has also affected the aforementioned thermal comfort research, albeit to a lesser degree.

- Thirdly, behavioural models require – in principle – the concurrent consideration of multiple parameters of physical, physiological, psychological, and socio-cultural nature. To conduct field or controlled studies addressing this complex pattern of potential causal factors is indeed anything but trivial. The multifariousness of potential influencing and contributing factors to behaviour actions creates as it were a kind of background "noise". Against this background, it is often difficult to discern the typically low-strength "signal" of causal factors hypothesised to be behind behavioural manifestations.

Obviously a number of the above-mentioned challenges in behavioural model development and evaluation cannot be met in the short run. Collection of vast amount of reliable observational data in the course of field studies is laborious, time-consuming, and costly. Likewise, conducting experimental behavioural studies is exceedingly difficult and the corresponding results cannot be readily generalised. These observations, however, do not absolve the invested community from trying to do better, and an indispensable precondition for doing better is a self-critical assessment of the past efforts in model development and application. Specifically, avoiding certain unnecessary but regrettably common mistakes and fallacies would help to further the behavioural modelling discourse in a more reasoned manner (Mahdavi 2015). Specifically:

- One should not confuse simulation (computational, typically dynamic representation of a system's behaviour) with prediction;
- One should neither assume nor claim that the mismatch between simulation-based predictions and observations of energy use (the so-called performance gap) is necessarily, or automatically, or exclusively due to behavioural factors. Long-term accurate predic-

tions of building performance indicators are difficult (if not impossible) to make due to an extensive list of uncertainties, pertaining not only to internal (occupancy-related) processes, but also to building fabric, building systems, and especially boundary conditions (i.e., weather conditions);

- One should not use (or at least be careful while using) the expression "deterministic", which has a weighty philosophical baggage, while meaning to refer to fixed diversity profiles (e.g., assumed fixed schedules of occupants' presence) and rule-based behavioural models;
- One should not claim building performance simulation results would be necessarily more "accurate" if we simply replace occupancy-related diversity profiles and rule-based assumptions with more detailed probabilistic ones (see for example Tahmasebi & Mahdavi 2015, 2016);
- One should not claim that "people behave randomly". There is nothing illegitimate about constructing black-box models of inhabitants' control actions toward generation of realistic patterns. But this does not point to the absence of a motivational (and potentially causally effective) field shaped by physiological, psychological, and social factors.
- One should not confuse code-based benchmarking with energy use prediction. Specifically, we should not assume that a specific modelling approach or technique can be appropriately applied to all kinds of use cases (see Gaetani et al. 2016, Mahdavi & Tahmasebi 2016).
- One should properly and meticulously document the model development and evaluation procedures (research design, empirical basis, hypotheses and assumed causal factors, limitations, etc.), such that others could independently retrace, comprehend, and reappraise them;

- One should not claim an occupancy model is "validated" without (or with just a "quick and dirty") comparison of calculations and using only a limited set of observations, specifically, one should not conflate data sets for model development and model evaluation. Testing a model based on the same data set, which was used for its development, is unsound methodologically and hence entirely unconvincing;
- One should not extrapolate from a single limited behavioural study to all kinds of populations, building types, locations, and climates. Specifically, it is hard to see why black-box models – devoid of first explicit principles based causal explanations – should be generally applicable;
- One should safe-guard against bias in model evaluation. As such, internal evaluation by model developers does not provide conclusive evidence for a model's general reliability. While not easy to conduct, external evaluation procedures, double blind studies, and round robin tests are undoubtedly in a better position to convincingly support the evaluation of a model's credibility;
- One should be extremely careful while incorporating insufficiently documented and rudimentarily tested behavioural models in simulation tools lest potential users are misled into assuming such models necessarily capture "reality".

In the next section of this paper, we address some of these considerations based on a specific illustrative case study of behavioural models. The material for this case study is taken from a previously published paper of the authors that explored the reliability of various models pertaining to inhabitants' operation of windows for natural ventilation in buildings (Tahmasebi & Mahdavi, 2016). In the present context, the results are not so much of interest in the original narrow sense of model comparison. Rather, we use this case study here paradigmatically to elaborate on a number of central model evaluation issues. Note that the case study itself has a number of key limitations (small set of reference empirical data from only one loca-

tion, small number of models considered, etc.). We could of course argue, Popperian style, that strictly speaking, models cannot be "verified", even with large amount of affirmative evidence. A single counter-example, on the other hand, suffices to "falsify" a model. This is, however, not the point we are making here. In the domain under discussion (assessment of inhabitants' behavioural models), it would be perhaps unwise to set unrealistically high standards regarding models' predictive performance. Consequently, the treatment of this case study's material does not attempt here to definitively evaluate the selected models. For such an objective, neither the original empirical basis upon which those models were developed, nor the empirical basis we used to examine their performance are large enough. Consequently, the case study has a different purpose: The structure and embedded procedure of this external evaluation exercise of a number of window operation models provides a useful context to specifically address a number of the aforementioned model evaluation challenges.

3 CASE STUDY: EXTERNAL EVALUATION OF WINDOW OPERATION MODELS

3.1 *Introductory remarks*

As already mentioned, the following treatment of external model evaluation issues uses material from a case study from one of our previous publications (Tahmasebi & Mahdavi, 2016). Specific details concerning the model comparison process related to this case study may be found in the aforementioned reference. Our focus in the present context and the respective use of the case study is, however, the critical discussion of a number of typical challenges in behavioural model evaluation. Toward this end, we first provide a description of the evaluation case study, followed by an extended discussion of respective results and their general implications.

3.2 *Selected window operation models for the external evaluation study*

As a case in point, the following external evaluation study specifically addresses the performance of window operation models. We studied three existing stochastic and three simple non-stochastic models. The stochastic models (referred here as A, B, and C) are derived based of occupant behaviour at office buildings and are widely referenced in the building performance simulation community. They are all Markov chain based logistic regression models that estimate the probability of window opening and closing actions based on the previous window state and a number of occupancy-related and environmental independent variables. To our knowledge, at least two of these models are implemented within well-known building performance simulation tools, namely model A in ESP-r (2016) and model C in IDA ICE (2016), despite the rather limited underlying empirical basis and despite the lack of conclusive evidence for their conclusive general validity and applicability.

The non-stochastic models (referred as D, E, and F) are defined based on simple rules according to the common practice in use of building performance simulation tools without integration of stochastic models – models D and F are, for example, integrated in EnergyPlus (2016).

In our study, we also included additional variations of models A and C (denoted as A* and C*), as the original models did not capture a key behavioural feature in the building under study where the inhabitants are requested not to leave the windows open when they leave the office due to storm damage risk. In addition, we considered two benchmark pseudo-models (denoted as G and H), whose purpose is to put the performance of the selected models into perspective. For the sake of clarity, a brief description of the aforementioned models is provided below:

- Model A, developed by Rijal et al. (2007), estimates the probability of opening and closing windows based on outdoor and operative temperature, when operative temperature is outside a dead-band (Comfort temperature ± 2°C). This model is derived

based on data obtained from 15 office buildings in UK between March 1996 and September 1997.

- Model A*, a variation of Model A, always returns a closing action upon each occupant's last departure.
- Model B, developed by Yun and Steemers (2008), is derived based on summer data (from 13 June to 15 September 2006) obtained from a naturally ventilated office building in UK without night time ventilation. It estimates the probability of opening windows upon first arrival and the probability of window opening and closing actions within intermediate occupancy interval (i.e., after first arrival and before last departure) based on indoor temperature.
- Model C, developed by Haldi and Robinson (2009), estimates the probability of opening and closing actions at arrival times (first and intermediate ones), intermediate occupancy intervals, and the departure times (intermediate and last ones) based on a number of occupancy-related and environmental independent variables (see Tahmasebi & Mahdavi 2016, for the list of independent variables, and the original and adjusted estimates of the coefficients used in this study). This model has been developed based on data obtained from 14 south-facing cellular offices in a building located in the suburb of Lausanne, Switzerland for a period covering December 19th, 2001 to November 15th, 2008.
- Model C*, a variation of Model C, always returns a closing action upon each occupant's last departure.
- Model D, a non-stochastic model, operates as follows: windows are opened if indoor temperature is greater than outdoor temperature and indoor temperature is greater than 26 °C. Otherwise the windows are closed.

- Model E, a non-stochastic model, can be specified as follows: Windows are opened if indoor temperature is higher than outdoor temperature and also higher than 26°C. Windows are closed if the indoor temperature is less than 22°C.
- Model F, a non-stochastic model, operates as follows: windows are opened if the operative temperature is greater than the comfort temperature calculated from the EN15251 adaptive comfort model (2007). Following the definition of comfort temperature for free-running period in EN15251, the windows can be opened only if weighted running average of the previous 7 daily average outdoor air temperatures is above 10°C and below 30°C.
- Model G, a benchmark pseudo-model, "predicts" windows are always open.
- Model H, a benchmark pseudo-model, "predicts" windows are always closed.

In case of the stochastic window operation models, to conduct the evaluation in a comprehensive manner, we used both original and adjusted coefficients of the logit functions. Whereas the original coefficients are published by model developers, the adjusted coefficients are obtained from re-fitting the models to a separate set of data obtained from the building under study in the calibration period. We specify the models with original coefficients with a subscript "O" and the ones with calibrated coefficients with a subscript "C". As mentioned before, the latter option (adjusting model coefficients based on observations in actual buildings) has no relevance to model deployment scenarios pertaining to building design support, but may be of some interest in operation scenarios of existing buildings.

The above described process of model selection and specification of the external evaluation study already highlights some of the typical challenges in the external validation studies of behavioural models. Aside from not having gone through a prior external validation study, most published models are limited even in the scope of the underlying internal validation: The published models are often derived based on limited data – typically from a single building – rendering those as non-representative in statistical terms (population, climate,

building typology, etc.). Moreover, even for this limited base, models' documentations typically leave many questions open or include questionable assumptions (for instance, the assumption that inhabitants' degree of freedom in operating windows is independent of facility management issues in a typical office building). Likewise, hidden assumptions pertaining, for example, to the assumed one-to-one relationship between an inhabitant and a window, make it difficult for the user to judge if and to which extent socially relevant interaction patterns between inhabitants and the related implications for the window operation are captured in the model.

3.3 *Empirical data for model calibration and evaluation*

An office area at TU Wien (Vienna, Austria) including an open space with multiple workstations and a single-occupancy closed office acted as the data source for external model assessment. We specifically focused on seven workstations, at which each occupant has access to one manually operable casement window. The occupants' presence, state of windows and a number of indoor environment variables (including air temperature, humidity, and CO_2 concentration) are monitored on a continuous basis. Outdoor environmental parameters (including air temperature and precipitation) are also continuously monitored via building's weather station. For the present study, we used 15-minute interval data from a calendar year (referred to as calibration period) to calibrate the coefficients of stochastic window operation models. As such, this option is only of interest, if the model deployment scenario involves already existing buildings (e.g., model use for optimisation of building operation). A separate set of data obtained from another calendar year (referred to as validation period) was used to evaluate the predictive performance of the models.

Note that, in this paradigmatic scenario, efforts were made to satisfy a number of generic model evaluation requirements formulated in the first section of this paper. These included, for example, collection of long-term high-resolution data, a rather rigorous data quality check,

and obviously separate data sets for calibration of model coefficients and model comparison. However, a central problem remains: Data available for model evaluation was in this case only from one building and for a relatively small number of inhabitants. This circumstance may remain, at least for some time, unavoidable (large repositories of observational data from different locations and building types are, while highly desirable, not available). This underlines the importance of candid and detailed model documentation, as alluded to in the introduction of the paper.

3.4 *Calibrated simulation model of the office area*

The previous studies on evaluation of stochastic window operation models (Schweiker et al. 2012, Fabi et al. 2015) did not address models' feedback. This circumstance represents a special problem in behavioural model validation, as the impact of behavioural models' output (for instance window states) on the models' input (for instance indoor temperature) is ignored. It is of course logically impossible to obtain empirical data matching every possible sequence of actions predicted by behavioural models. Hence, we need to emulate building's response to behavioural impulses virtually, i.e., via calibrated simulation. Therefore, we suggest the use of a calibrated simulation model as a platform for evaluation of behavioural models whose output (e.g., window states) influences models' input (e.g., indoor temperature). This necessitates a model that can reliably represent the building's behaviour.

For the purposes of the present case study, we first subjected the building model to an optimisation based calibration to adjust the fixed parameters governing the multi zone air flow simulations (for the details of the calibration procedure, see Tahmasebi & Mahdavi 2012). Secondly, we incorporated the monitored data pertaining to occupancy, plug loads, use of lights, and operation of heating system into the calibrated building model as a set of full-year data streams in terms of 15-minute intervals. This data set was obtained in the validation period. The resulting model, when fed with actual window operation data as the benchmark

model, predicts the hourly indoor temperatures in validation year with a Normalized Mean Bias Error of 2.8% and a Coefficient of Variation of Root-Mean-Square Error of 4.8%.

The described building simulation model served as a platform, into which the selected window operation models were integrated, such that in each variation of the building model, the occupants' interactions with windows are represented using one of the selected window models. For each occupant in the building, individual occupancy data and zone-level indoor environmental factors are provided for the window operation model. That is, at each simulation time-step, the window model is executed separately for each occupant. We also built a benchmark model, which contained the actual operation of windows based on the monitored data obtained in the validation period.

As using calibrated building performance simulation for evaluation of occupant behaviour models necessitates the deployment of real-year – preferably on-site – weather data, the building model was exposed to the outdoor environmental conditions in the validation period. This was accomplished by generating a weather data file from the on-site weather station measurements. The measured dataset included outdoor air temperature, air humidity, atmospheric pressure, global horizontal radiation, diffuse radiation, wind speed, and wind direction.

3.5 *Evaluation scenarios for window operation predictions*

We took two approaches to evaluate window operation models in view of their potential in predicting the occupants' interaction with windows:

i. Use of a set of monitored data pertaining to indoor and outdoor environment as well as occupants' presence and interaction with windows. In this case, which is typical for almost all past model evaluation efforts (see, for example, Fabi et al. 2015; Schweiker et al. 2012; Haldi and Robinson, 2009), the impact of window operation models' output on indoor environmental input is neglected.

ii. Use of a calibrated building performance model populated with the same set of moni-

tored data. Here, the calibrated building model simulates the impact of window operation models' output on indoor environmental input.

In both approaches, we evaluated the performance of window operation models to predict inhabitants' interactions with windows for a one-year-long validation period, whereby the models are fed with monitored occupancy-related and outdoor environmental data from the same period according to their independent variables. The required indoor environmental factors, however, are provided from different sources. That is, in the first approach from the measurements in the same period, and in the second approach from the building simulation output.

3.6 *Evaluation statistics*

One of the fundamental challenges of evaluation procedures pertaining to behavioural models of building inhabitants pertains to the paucity of systematically classified model performance metrics. The pertinent professional community has arguably not converged toward a systematic and expressive set of statistics for behavioural models' predictive performance. Some of the responsible factors for this negligence were already alluded to in the introductory sections of this paper. Given the variety of domains and application scenarios of behavioural models, the definition of a definitive set of evaluation statistics is indeed unlikely to be a trivial undertaking.

Whereas an ultimate ontology of fit-for-purpose metrics for behavioural model evaluation cannot be provided here (and may be even ultimately unattainable), a potentially important first attempt can be made. Behavioural models typically aim at predictions of "states" and "events" (or "actions"). In this taxonomy (see Mahdavi 2011), events can be system-related (e.g., switching lights on/off) or occupancy-related (e.g., entering into – or leaving – a space). States can refer to systems (e.g., position of shades/windows), indoor environment (e.g., temperature, illuminance), outdoor environment (e.g., solar radiation), and inhabitants' presence (i.e., present versus absent).

The central step in model evaluation is of course the comparison of predicted and monitored events and states (see Figure 1). We suggest that, from the large number of indicators, which have been used in previous – predominantly internal – evaluation studies of inhabitants' behavioural models (as well as in studies in relatively close fields such as thermal comfort), two broad categories can be inferred: The indicators addressing aggregate aspects of models' predictions, and the indicators addressing the interval-by-interval congruence between predictions and measurements. In other words, whereas the first category "vertically" aggregates observations and predictions independently before the overall comparison, the second category compares first "horizontally" time series data pairs, which can then be further processed statistically. Illustrative listings of these two types of indicators are provided in Figure 1. Note that in this framework, we have grouped indicators, which address aggregate traits of the predictions (such as total number of actions, median state durations, etc.) along with indicators, which address the proximity of predicted probability distributions to those of the measured ones (such as Jensen-Shannon divergence).

It can be argued that while a superior performance in terms of aggregate indicators is specifically desired in simulation studies geared at performance levels over longer periods of time (such as conventional use of building performance simulation models for estimation of annual energy demands), the indicators resulting from interval-by-interval contrast of predictions and measurements are of more interest in studies, in which short-term performance predictions play a central role (e.g., predictive building systems control).

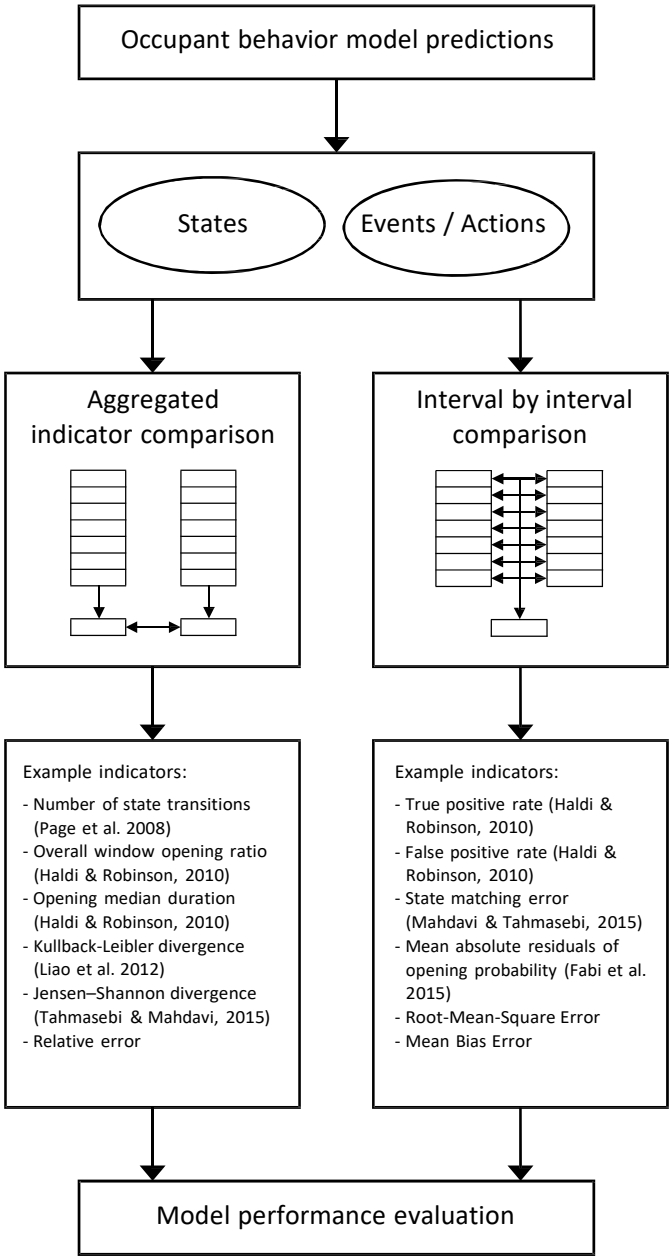

Figure 1. Categories, aggregation structures, and example indicators for occupant behaviour model evaluation

For the purpose of the current case study, we used the following indicators to evaluate the predictive performance of window operation models. Note that whereas the first three selected indicators in the following list belong to the interval-by-interval comparison category, the last four are typical for the aggregated indicator category:

– Fraction of correct open state predictions [%]: This is the number of correctly predicted open state intervals divided by the total number of open state intervals.

- Fraction of correct closed state predictions [%]: This is the number of correctly predicted closed state intervals divided by the total number of closed state intervals.
- Fraction of correct state predictions [%]: This is the number of correctly predicted interval states divided by total number of intervals.
- Overall fraction of open state [%]: This is the total window opening time divided by the observation time.
- Mean number of actions per day [d^{-1}] averaged over the observation time.
- Open state durations' median and interquartile range [hour].
- Closed state durations' median and interquartile range [hour].

From the above indictors, the fraction of correct open state predictions (as "true positive rate"), fraction of open state, mean number of actions per day, median open state duration, and median closed state duration have been suggested in previous studies to evaluate the predictive performance of window operation models. We added three indictors to the previous work, namely fraction of correct closed state predictions to express models' state prediction performance, and the interquartile range of open state and closed state durations to capture the spread of window states' durations.

To ensure the robustness, transparency, and integrity of model evaluation procedures, the selection of reliable, expressive, and consistent model performance metrics is indispensable. Related future efforts in this direction are thus of utmost importance.

3.7 *Results*

The obtained values of evaluation indicators for different window operation models are given in Table 1 (without considering the models' feedback) and Table 2 (under consideration of the models' feedback via calibrated building performance model). These values are obtained from model executions in the whole validation period (a full calendar year). To better illustrate the performance of models in terms of different evaluation indicators, Figure 2 to Figure

4 show the models' prediction errors under consideration of their feedback. Note that in these Figures, models' relative error percentages are displayed in a logarithmic scale: For instance, a value of 1 read from the y-axis denotes a relative error of 10% in the evaluation indicator with reference to the benchmark. This mode of representation facilitates a better visibility of the differences in models' behaviour. In addition to the graphical representation of data, Table 3 provides a numeric overview of the relative deviations of predictions from corresponding observations.

Table 1. The values of evaluation statistics obtained from model executions without feedback (Tahmasebi & Mahdavi, 2016)

Models	Fraction of correct open state [%]	Fraction of correct closed state [%]	Fraction of correct states [%]	Fraction of open state [%]	Actions per day [d^{-1}]	Opening duration [hour]		Closing Duration [hour]	
						Median	IQR	Median	IQR
Observed	100.0	100.0	100.0	4.1	0.28	1.8	5.3	23.5	55.3
A_o	71.8	39.2	40.5	61.3	0.01	1180.0	2803.2	452.8	1442.3
A_o*	26.0	98.7	95.7	2.3	0.10	4.9	4.1	23.9	96.6
B_o	47.5	84.4	82.9	16.9	5.37	0.5	0.5	0.5	0.8
C_o	61.3	70.1	69.7	31.2	0.09	44.3	102.6	97.3	212.5
C_o*	22.2	97.9	94.8	2.9	0.15	4.2	4.7	76.3	157.5
A_c	80.9	46.4	47.8	54.7	0.01	1380.1	1318.2	635.0	974.1
A_c*	30.8	98.8	95.9	2.4	0.10	4.8	5.5	22.0	106.5
B_c	42.0	95.1	92.9	6.4	0.29	3.7	5.8	42.4	81.1
C_c	55.0	80.6	79.6	20.9	0.17	5.2	26.1	56.7	118.7
C_c*	33.7	97.5	94.9	3.8	0.22	3.2	5.6	54.2	110.1
D	32.0	98.7	96.0	2.6	0.35	0.8	2.3	1.8	18.0
E	51.5	97.8	95.9	4.2	0.14	7.8	5.0	17.8	48.1
F	45.3	93.7	91.7	7.9	0.95	0.8	2.8	1.0	15.0
G	100.0	0.0	4.1	100.0	0.0	8760.0	0.0	-	-
H	0.0	100.0	95.9	0.0	0.0	-	-	8760.0	0.0

Table 2. The values of evaluation statistics obtained from model executions with feedback (Tahmasebi & Mahdavi, 2016)

Models	Fraction of correct open state [%]	Fraction of correct closed state [%]	Fraction of correct states [%]	Fraction of open state [%]	Actions per day [d^{-1}]	Opening duration [hour]		Closing Duration [hour]	
						Median	IQR	Median	IQR
Observed	100.0	100.0	100.0	4.1	0.28	1.8	5.3	23.5	55.3
A_o	44.0	85.2	83.5	16.0	0.05	18.6	59.0	152.2	308.8
A_o*	47.2	96.9	94.9	4.9	0.21	5.7	5.3	22.4	66.0
B_o	41.8	88.4	86.5	12.9	5.2	0.5	0.5	0.5	0.8
C_o	54.2	78.2	77.2	23.1	0.07	37.1	91.2	133.7	313.2
C_o*	30.9	97.5	94.7	3.7	0.18	4.5	4.9	56.4	120.9
A_c	41.3	86.0	84.2	15.1	0.04	19.8	93.1	172.5	408.2
A_c*	44.4	97.5	95.3	4.2	0.18	5.4	5.4	23.6	76.2
B_c	44.6	96.4	94.3	5.3	0.31	2.8	5.9	38.3	76.3
C_c	47.9	83.9	82.5	17.4	0.16	3.7	22.8	63.0	128.5
C_c*	35.4	97.2	94.7	4.1	0.24	3.2	5.8	45.8	97.6
D	36.0	97.6	95.1	3.8	1.25	0.3	0.3	0.5	2.5
E	54.3	95.8	94.1	6.3	0.23	6.8	6.0	18.8	47.9
F	44.1	94.8	92.8	6.8	1.78	0.3	0.5	0.5	1.3
G	100.0	0.0	4.1	100.0	0.0	8760.0	0.0	-	-
H	0.0	100.0	95.9	0.0	0.0	-	-	8760.0	0.0

Table 3. Relative devation of the predictions from the observed behaviour in terms of 5 evluation indicators obtained from model executions with feedback

Models	Model type	Coefficients	Adjustment for the absence of nighttime ventilation	Relative deviation from observed behaviour [%]				
				Open state predictions	State preditions	Fraction of open state	Number of actions	Median opening duration
A_o	Stochastic	Original	No	56.0	16.5	289.7	81.9	962.9
B_o				58.2	13.5	213.6	1775.9	71.4
C_o				45.8	22.8	464.1	74.7	2017.4
A_o*	Stochastic	Original	Yes	52.8	5.1	20.0	25.9	225.1
C_o*				69.1	5.3	9.6	34.4	155.6
A_c	Stochastic	Calibrated	No	58.7	15.8	268.6	84.4	1033.1
B_c				55.4	5.7	28.3	13.0	57.7
C_c				52.1	17.5	323.3	40.8	112.6
A_c*	Stochastic	Calibrated	Yes	55.6	4.7	3.1	35.2	209.9
C_c*				64.6	5.3	0.1	15.2	84.1
D	Non-stochastic	-	-	64.0	4.9	7.3	352.1	85.7
E				45.7	5.9	52.7	18.2	285.7
F				55.9	7.2	65.0	541.9	85.7

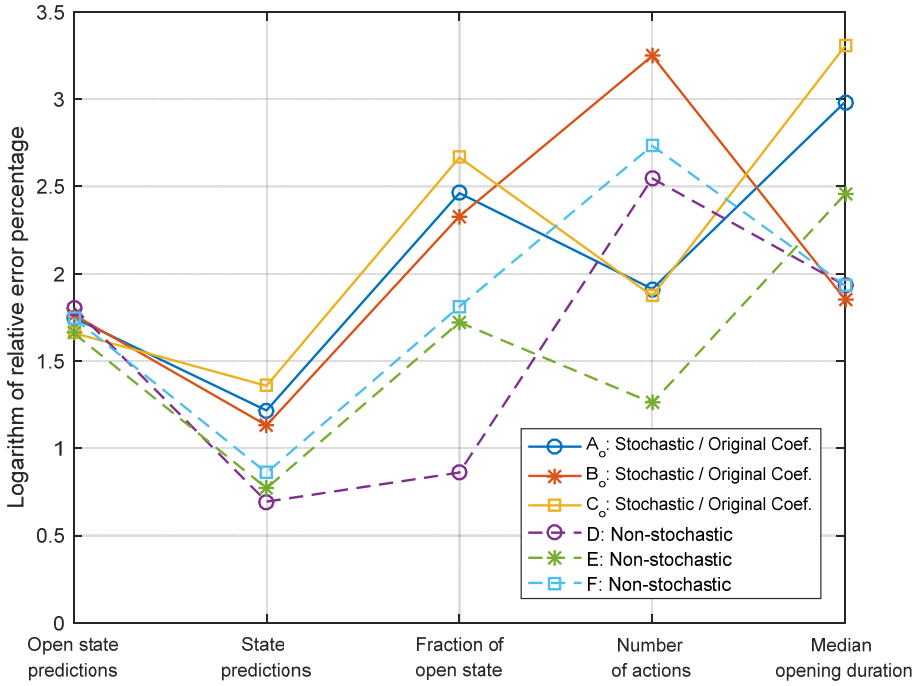

Figure 2. Errors of stochastic window operation models with original coefficients and no adjustment (A_o, B_o, and C_o) as well as non-stochastic models D, E, and F in terms of 5 evaluation statistics (Tahmasebi & Mahdavi, 2016)

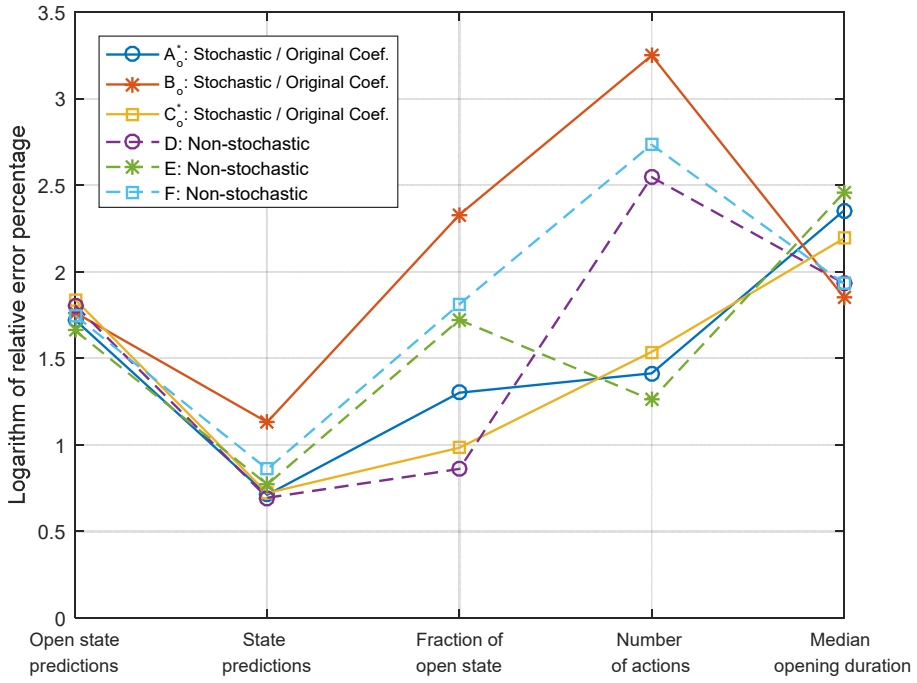

Figure 3. Errors of stochastic window operation models with original coefficients and adjusted to buildings without night time ventilation (A_o*, B_o, and C_o*) as well as non-stochastic models D, E, and F in terms of 5 evaluation statistics (Tahmasebi & Mahdavi, 2016)

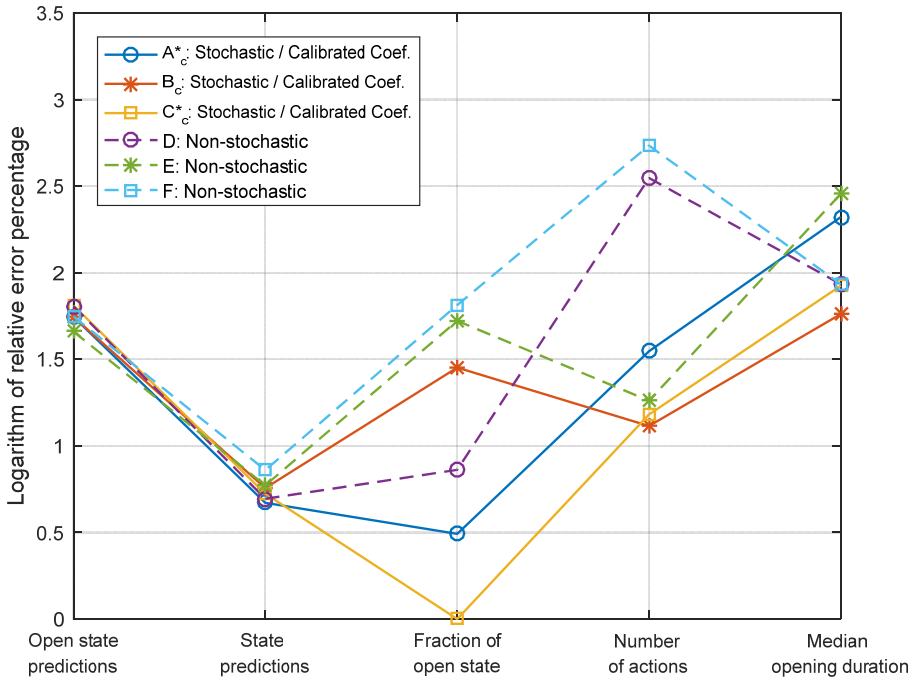

Figure 4. Errors of stochastic window operation models with calibrated coefficients and adjusted to buildings without night time ventilation (Ac*, Bc, and Cc*) as well as non-stochastic models D, E, and F in terms of 5 evaluation statistics (Tahmasebi & Mahdavi, 2016)

3.8 *Discussion*

A fundamental question with regard to the application of behavioural models concerns their capability in reproducing empirical observations. We may thus first ask if the models could, in the present case, provide acceptable approximations of the observations. As mentioned before, most behavioural models use some indoor environmental data as independent variables. However, empirical evaluation of such models typically ignores action consequences for the indoor environment. To address this very problem, in the presented case study, the model evaluation was conducted using two alternatives, namely with and without inclusion of models' feedback. Given the respective results shown in Table 1 and Table 2, the performance of the models relative to each other appears to be independent of feedback provision. However, without considering the models' feedback (in this case, regarding indoor temperature), the evaluation process may generate misleading results. For example, without feedback, model A largely overestimates the fraction of open state and opening duration, as the measured indoor

temperatures do not fall below the dead-band defined in this model to close the windows. This tendency can be seen less dramatically in the fraction of open state predicted by model C. Ignoring feedback also hides the tendency of the non-stochastic models D and F to predict an unrealistically large number of actions. As such, according to these models, windows are operated as soon as the temperature falls below or rises above a certain threshold. With included feedback, this would result in a large number of opening and closing actions. However, without considering the models' feedback, opening of the window does not reduce the indoor air temperature and is therefore not followed by a closing action.

Given these circumstances, it can be inferred that validation efforts pertaining to window operation models (or any behavioural model with indoor environmental input), which neglect the models' feedback would be inconclusive. Therefore, the use of calibrated simulation models is more likely to provide a dependable analysis of the window operation models' performance.

With this in mind, let us return to our model performance comparison in the case at hand. Assuming a threshold of ±20% for the relative error of model predictions as a reasonable benchmark, we must conclude that without adjustments (night-time ventilation, calibrated coefficients), none of the studied models performs satisfactorily (see Table 2 and as well as Figure 2). Only regarding the indicator "fraction of correct state predictions" do the non-stochastic models meet this criterion. However, the night-time ventilation adjustment markedly improves the performance of the stochastic models A_o* and C_o* (see Figure 3). Furthermore, calibrating the coefficients of stochastic models via observational data results in a significant improvement of their predictive performance. Specifically, for indicators "fraction of correct state predictions", "predicted fraction of open state", and "the number of daily actions", these models' relative errors remain roughly under 30% (see Figure 4).

Note that, as stressed before, the presented case study was based on a limited set of empirical data obtained from one office area. While we consider the underlined shortcomings

valid and worthy of serious attention in future studies, we do not suggest the findings can be extrapolated to the modelling efforts in different contexts. Ongoing and future – more extensive – cross-sectional investigations in this area are expected to utilise a larger empirical foundation and thus lead to more representative and inclusive model evaluations. Specifically, while calibration of occupant behaviour models is not feasible in majority of building performance simulation efforts, similar external validation studies can also contribute toward a repository of coefficients for the use of existing occupant behaviour models in different contexts.

Aside from these specific case study results regarding the performance of the selected models, we would like to highlight a number of observations that are relevant to the model evaluation discussion in general:

- As noted earlier, a general problem in both development and evaluation of behavioural models pertains to the paucity of empirical data. For instance, models A and B were solely based on office buildings in UK (15 in case of model A and 1 in case of model B), whereas model C was based on one office building in Switzerland. Moreover, the monitoring period for data collection was rather limited in case of models B (four months).
- Earlier in the paper, we suggested that a sound model evaluation process requires the availability of clear and detailed model documentations. This condition is often ignored and was not also fully met in our case study. For instance, in case of model A, the treatment of night time ventilation was not clearly described. Likewise, in case of model C, it was not clear that the parameter included for closing window upon last departure does not suffice to make the model with original coefficients applicable for buildings without night time ventilation.
- As suggested previously, model developers should ideally conduct an internal validation via separate developmental and evaluative data sets. In the present case study, this

was not done in case of models A and B. In case of model C, the publication introducing the model suggests that a "cross-validation" was performed. Note that only the publication related to model C included some model validation metrics. However, the types, coverage, scope, and suitability of performance metrics for behavioural models remains an open challenge.

- We suggested that a sound model documentation should entail comments on the applicability of the proposed models (e.g., with regard to building type, location, climate, deployment scenario). The documentations of the models selected for our case study did not provide such comments.

All in all, the above illustrative external evaluation study underlines a number of challenges in the evaluation process of behavioural models. These include the paucity of underlying empirical information that is of sufficiently high quality and of representative nature, shortcomings in model documentation, model input requirements that cannot be met in realistic model deployment situations, problems associated with model coefficients and their calibration, lack of a set of comprehensive, adequate, and universally accepted model performance metrics, and – last but not least – the problem of feedback, i.e., the inclusion of the predicted actions' impact on environmentally relevant model input variables.

4 CONCLUSIONS

Building performance assessment tools and methods can be significantly improved in their coverage and applicability if they are enriched with high-resolution representations of inhabitants. Many recent model development efforts have explored the potential of detailed mathematical formalisms for such representations. However, rigorous external evaluation processes are needed to ensure the usability and reliability of occupancy-related behavioural models. Given the lack of related general procedures and guidelines, we formulate a number of rele-

vant conditions and requirements. Furthermore, we presented a demonstrative model evaluation study involving a number of recently proposed window operation models. Thereby, our concern was not only to highlight the observed large deviations from reality underlined in this specific case. Rather, as a paradigmatic model case, the external window operation evaluation study provided us with the opportunity to point to the need for clear documentation of associated uncertainties with existing behavioural models in different deployment scenarios as well as development of more generally applicable occupancy-related models. Definition and pursuit of rigorous model validation procedures in the behavioural modelling field may be seen as work in progress. As a consequence, both model developers and potential users would be well-advised to be careful with regard to introduction and application of behavioural models pertaining to inhabitants' actions in buildings. Specifically, statements concerning models' validity and overall applicability in the building delivery process would be of little credibility without comprehensive empirical backing and careful model testing procedures.

5 ACKNOWLEDGEMENTS

The research presented in this paper benefited from the authors' participation in the ongoing efforts of the IEA-EBC Annex 66 (Definition and Simulation of Occupant Behaviour in Buildings) and the associated discussions.

REFERENCES

D'Oca S., Fabi V., Corgnati S.P., Andersen R.K., 2014. Effect of thermostat and window opening occupant behavior models on energy use in homes, BUILD SIMUL (2014) 7: 683–694, DOI 10.1007/s12273-014-0191-6.

EN 15251, 2007. Standard EN 15251–2007: Indoor environmental input parameters for design and assessment of energy performance of buildings addressing indoor air quality, thermal environment, lighting and acoustics.

EnergyPlus, 2016. https://energyplus.net/. [Accessed 23 May 2016].

ESP-r, 2016. http://www.esru.strath.ac.uk/Programs/ESP-r.htm [Accessed 23 May 2016].

Fabi V., Andersen R.K. & Corgnati S., 2015. Verification of stochastic behavioural models of occupants' interactions with windows in residential buildings, Building and Environment, 94(1), pp 371–383, doi:10.1016/j.buildenv.2015.08.016.

Feng X., Yan D., Hong T., 2015. Simulation of occupancy in buildings, Energy and Buildings 87 (2015), 348-359, doi:10.1016/j.enbuild.2014.11.067.

Gaetani I., Hoes P. & Hensen J.L.M. 2016. Occupant behavior in building energy simulation: Towards a fit-for-purpose modeling strategy, Energy and Buildings doi:10.1016/ j.enbuild.2016.03.038.

Haldi F. & Robinson D., 2009. Interactions with window openings by office occupants. Building and Environment, 44(2009), pp 2378-2395, doi:10.1016/ j.buildenv.2009.03.025.

Hulley S. B., Cummings S. R., Browner W. S., Grady D. G., and Newman T. B., Designing clinical research: Lippincott Williams & Wilkins, 2013.

IDA ICE, 2016. http://www.equa.se/en/ida-ice [Accessed 23 May 2015].

Langevin J., Wen, J., Gurian P.L., 2015. Simulating the human-building interaction: Development and validation of an agent-based model of office occupant behaviors, Building and Environment 88 (2015) 27-45, doi:10.1016/j.buildenv.2014.11.037.

Liang X., Hong T., Shen G., 2016. Improving the accuracy of energy baseline models for commercial buildings with occupancy data, Applied Energy 179 (2016), 247-260.

Liao C., Lin C., Barooah P., 2012. Agent-based and graphical modelling of building occupancy, Journal of Building Performance Simulation 5(1), 2010, 5–25.

Mahdavi, A. 2011. The Human Dimension of Building Performance Simulation, Proceedings of the 12th Conference of the International Building Performance Simulation Association, K16 - K33, ISBN: 978-0-646-56510-1.

Mahdavi, A. 2015. Common fallacies in representation of occupants in building performance simulation, Proceedings of Building Simulation Applications 2015 - 2nd IBPSA-Italy Conference, pp 1-7, Bozen-Bolzano University Press, ISBN: 978-88-6046-074-5.

Mahdavi, A., Tahmasebi, F, 2015. The inter-individual variance of the defining markers of occupancy patterns in office buildings: a case study, Proceedings of BS2015, 2243-2247.

Mahdavi, A. & Tahmasebi, F. 2016. The deployment-dependence of occupancy-related models in building performance simulation, Energy and Buildings 117 (2016) 313–320, doi:10.1016/j.enbuild.2015.09.065.

O'Brien, W, Gunay H.B, 2016. Occupant behaviour diversity modelling and its applications, Proceedings of eSim 2016, 76-92.

Oleckno W. A. and Anderson B., Essential epidemiology: principles and applications: Waveland, 2002.

Parys, W., Saelens D., Hens H., 2011. Coupling of Dynamic Building Simulation with Stochastic Modelling of Occupant Behaviour in Offices – A Review-Based Integrated Methodology. Journal of Building Performance Simulation 4 (4): 339–358. doi: 10.1080/19401493.2010.524711.

Rijal H.B, Tuohy P., Humphreys M.A., Nicol J.F., Samuel A. & Clarke J., 2007. Using results from field surveys to predict the effect of open windows on thermal comfort and energy use in buildings. Energy and Buildings, 39 (2007), pp 823-836, doi:10.1016/j.enbuild.2007.02.003.

Shipworth D., Huebner G., Schweiker M., Kingma BRM, Diversity in Thermal Sensation: drivers of variance and methodological artefacts, Proceedings of 9thWindsor Conference, Cumberland Lodge, Windsor, UK, 7-10 April 2016.

Schweiker M., Haldi F., Shukuya M. & Robinson D., 2012. Verification of stochastic models of window opening behaviour for residential buildings, Journal of Building Performance Simulation, 5(1), pp 55-74, doi:10.1080/19401493.2011.567422.

Schweiker M., Wagner A. 2016. Exploring potentials and limitations of the adaptive thermal heat balance framework, Proceedings of 9thWindsor Conference, Cumberland Lodge, Windsor, UK, 7-10 April 2016.

Tahmasebi F. & Mahdavi A. 2012. Optimization-based simulation model calibration using sensitivity analysis, 7th Conference of IBPSA-CZ, Brno, Czech Republic.

Tahmasebi F. & Mahdavi A. 2015. The sensitivity of building performance simulation results to the choice of occupants' presence models: a case study. Journal of Building Performance Simulation, (2015), doi:10.1080/19401493.2015.1117528.

Tahmasebi F. & Mahdavi A. 2016. An inquiry into the reliability of window operation models in building performance simulation, Building and Environment 105 (2016), 343–357 DOI:10.1016/j.buildenv.2016.06.013.

Yun G.U. & Steemers K., 2008. Time-dependent occupant behaviour models of window control in summer. Building and Environment, 43(2008), pp 1471-1482, doi:10.1016/ j.buildenv.2007.08.001.