A simulation-based overheating indicator for 20 million dwellings nationwide

Yannis Merlet, Adrien Toesca, Camille Bacon, Pascal Schetelat, Anais Machard

A simulation-based overheating indicator for 20 million dwellings nationwide

Yannis Merlet[1], Adrien Toesca[1*], Camille Bacon[1], Pascal Schetelat[1], Anaïs Machard[1].

1 Department of Energy and Environment, Centre Scientifique et Technique du Bâtiment, France
[*]Corresponding author's email: adrien.toesca@cstb.com

Abstract. Climate change leads to an overall increase in temperature, and this can lead to risks for the health, and specifically for housing building occupants. Public policies about the mitigation of the overheating risk in housing building may have to be implemented and should be able to rely on a precise and exhaustive description of the initial state of the building capacity to mitigate overheating.
This paper presents a methodology used to calculate an overheating indicator for each housing at a country scale (around 20 million housing buildings).
Because the model developed to calculate this overheating indicator must be fast enough to apply it on 20 million buildings, without giving up much on the accuracy of detailed simulation models, a machine learning approach has been chosen.
The overheating indicator was first calculated on a large sample of housing building using dynamic thermal simulation. The results of those simulation made it possible to identify the most influential parameters of a housing on overheating: its geometry, its inertia, its windows, and the insulation of its walls and roof. Then, we trained a random forest regressor to have an accurate statistical model with almost instant computing performance.
As a result, we could evaluate the overheating indicator for all the 20 million housings in France providing much information about population health risks.
Moreover, the evaluation of the influence of the parameters on each housing buildings on the overheating performance can be performed, thus underlying which actions must be promoted to mitigate risks.

Keywords: Overheating in buildings, Climate change mitigation, Urban heat island.

1 INTRODUCTION

A warming climate and a rise of the frequence and severity of heatwave in France will lead to an increased risk of death for the population [1]. As the population is spending most of its time indoor, it is crucial for public policies to be able to evaluate the vulnerability of the housing stock.

Overheating is a key metric to assess the vulnerability of housing to heat stress. While instantaneous indicators, such as those reviewed by [2] or the adaptative thermal comfort

approach [3], assess discomfort for a specific moment, they don't fully capture long-term thermal performance. To address this limitation, integrated metrics are essential.

Among these integrated metrics, the Discomfort Degree Hours (DDH) indicator is a relevant approach to evaluate overheating over long periods. As highlighted by [4], DDH aggregates the duration and intensity of discomfort, making it suitable to assess summer thermal performance. It has been widely used in scientific studies, including recent work of [5] studying overheating in buildings in hot and humid climates.

This approach has also been adopted in the French thermal regulation RE2020 [6], highlighting its applicability in evaluating overheating at the building scale. This indicator provides a relevant framework for characterizing the current state of the French housing stock in terms of summer thermal performance.

Multiple approaches are using simulation to evaluate the energy performance of dwellings at a large geographic scale. [7] mentioned that various statistical methods are used to speed up simulations and make them compatible with nationwide energy simulations. The specific topic of overheating or summer thermal comfort on a large geographic area was discussed in various papers, each time with limitations. Most of the time, overheating is calculated on typologies (often coming from TABULA database) and then generalized on the whole country. In the Netherlands 9216 fictive designs of dwellings are used to represent the whole building stock [8]. Overheating was assessed as well in the UK Building stock with the same approach using 8 typologies with the addition of multiple occupant scenario (around 600 simulations) [9].

To the knowledge of the authors, no simulation based overheating assessment at a large geographic scale is using a large set (100 000+ buildings) of real buildings as inputs.

This paper presents the methodology used to assess the overheating in 20 million dwellings for the whole continental France. The overheating is assessed thanks to a statistical model that is trained on 420 000 simulations of real dwelling. The statistical model has 14 selected input parameters and is validated against dynamic thermal simulations with a good precision (r^2=0.89).

The novelty of this approach consists in the use of large-scale database of EPC to generate 420000 buildings to simulate, and then the use of such a large training set to train a nation-wide precise model. The ability to take into account as an input the effect of the urban heat island in our indicator available for this geographic area is also novelty brought by our work.

This work aims to tackle the question of how to elaborate an accurate overheating indicator nationwide for each dwelling that can be used in both public policies and further scientific work?

The paper is organized as follow: the following section describe the methodology to elaborate the overheating prediction model. Then the results of the overheating prediction are presented to demonstrate the accuracy of the prediction. Finally, the limitations are discussed.

2 METHODOLOGY

2.1 General methodology

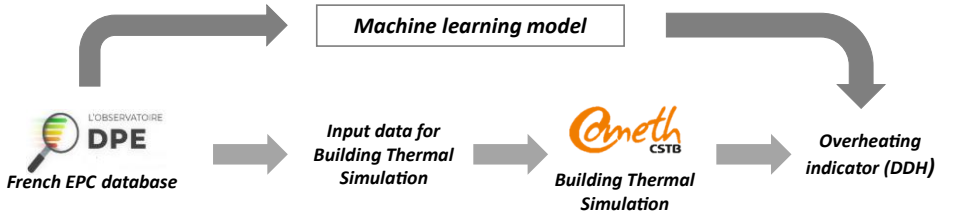

Fig 1: Schematic representation of the general methodology

Fig 1 is a schematic representation of the general methodology of this work. The methodology begins with data derived from Energy Performance Certificates (EPC). These certificates are used as input parameters for Dynamic Thermal Simulations (DTS). The COMETh DTS simulation engine, which is the computational core of the RE2020 French regulation, is used to simulate building performance and predict the overheating indicator. So, the inclusion of a valid EPC is a prerequisite for any dwelling to be included in the training dataset.

The overheating indicator used in this study is the Discomfort Degree Hour (DDH), a metric that quantifies thermal discomfort by aggregating the intensity and duration of indoor temperatures exceeding a threshold. The DDH is calculated using the temperature threshold defined in the RE2020. This threshold is based on the EN 16978 [10], and it ranges from 26°C to 28°C during daytime and is fixed to 26°C during the night. Eq. (1) explains how DDH is calculated, with T_i, the operative temperature inside the dwelling, T_{comf} the temperature threshold and N the number of hours over the RE2020 adaptative comfort period (i.e. summer period).

$$DDH\ [°C.h] = \sum_{i=1}^{N} (T_i - T_{comf}) \quad if\ T_i > T_{comf} \tag{1}$$

To calculate an overheating indicator for the entire French residential building stock, DTS like those performed with COMETh are computationally expensive and impractical at this scale. To address this limitation, a machine learning model is developed as a by-pass for DTS, allowing for the rapid estimation of overheating. The model is trained on a dataset derived from COMETh simulation and predict the overheating indicator using the EPC data as input.

2.2 Creation of training set

To ensure that the predictive models accurately estimate overheating, a diverse training dataset was built. This section describes the step taken to select and prepare the dataset,

focusing on climatic diversity, altitude considerations and geographical representativeness of residential buildings across metropolitan France.

The training dataset was built using residential buildings for which valid EPCs were available. EPCs provide the key input parameters required for DTS, including building morphology, energy systems, surface areas and the thermal performance of building envelopes. Those characteristics are mapped and enriched to fill the DTS inputs of Cometh engine. The simulation parameters used are conventional and follow the assumptions defined by the RE2020 regulation. It includes among other occupancy patterns, interactions with windows, use of solar shading devices…

To consider diversity in climate, the simulation incorporates the meteorological data corresponding to the eight climate zones defined in the RE2020 regulation. Each climate zone corresponds to one weather station from French meteorological office (Meteo France). These files, built using the ISO 15927-4 standard, provide hourly data for a standard year, including temperature, solar radiation, humidity, and wind conditions. In addition, to assess building overheating, the regulation includes meteorological data from the 2003 deadly heatwave within the typical year, to build an extreme year. In this study, only the extreme years were used. Temperature and specific humidity reduction were computed using the RE2020 rules depending on the altitude of each building.

In urban contexts, local climate effects such as the Urban heat Island (UHI) phenomenon can exacerbate overheating. To account for this, the Urban Weather Generator (UWG) model [11] is used to adapt the RE2020 extreme weather files to urban environments. UWG inputs are computed using the BDNB [12].

In collective housing, the thermal performance of a dwelling depends on its position within the building. To account for this, each selected collective housing unit was duplicated and adjusted to represent two distinct dwelling: top floor, which is more exposed to overheating due to direct roof exposure and intermediate floor. The only difference between the two simulations is the consideration of thermal exchange through the roof for top floor and adiabatic condition for the intermediate one. This approach ensures that the dataset captures the impact of dwelling position on thermal performance.

The training dataset contains approximately 420 000 residential dwellings, distributed across all climate zones of metropolitan France, considering urban, rural, and mountainous areas alike.

2.3 Selection of input parameters for the prediction model

The selection of the parameters used to train the overheating prediction model had two main purposes. First, the prediction model should have the necessary parameters to provide an accurate prediction of the overheating when compared to the simulation. Secondly, the parameters should describe the building physics involved in overheating in housing.

[13] showed that overheating in free floating buildings is mainly due to three mechanisms related to building properties: solar gains, inertia, and natural ventilation. Some parameters related to the climate where selected. **Table 1** sums up the input parameters used in our prediction model and the building physics phenomena covered by each of these

parameters. Conductive heat gains are less dominant factor of overheating in buildings compared to solar heat gains.

Table 1: List of used parameters for the overheating prediction model developed in this paper

	Solar gains	Inertia	Natural ventilation	Climate	Geometry
Wall U value	x				
Roof U value	x				
Windows U value	x				
Inertia class		x			
Windows area	x				
Shading devices type	x				
Height of the building	x		x		x
Dual-aspect housing			x		x
Living surface		x			x
Volume		x			x
Altitude				x	
Weather				x	
UHI				x	
Construction year	x	x			

The selection of those parameters is sufficient to an accurate prediction, as it is shown in the section 3.1. The addition of further parameters was either not improving the performance of the prediction or adding further uncertainties to the inputs of the prediction model. Uncertainties of those input parameters is not assessed here, as discussed in section 4.1.

2.4 Prediction model

Estimators (or metamodeling methods) depends on the type of model that the estimator should replicate, on the number and type of input parameters, on the number of outputs of the model and on the number of samples used to train the model. In this work, the dynamic thermal simulation that is estimated with the prediction model is a non-linear model, with 14 inputs (continuous and discrete). The prediction model has been trained on a dataset of 420k samples, and the prediction model has an only output which is the predicted DDH.

The choice of a model for our prediction model was based on trial-and-error approach within the models that seemed suitable for such a use. First tried model was a linear

regression model with a Lasso error coefficient. The performance of the prediction with a linear regression was not sufficient to describe the non-linear component of overheating.

The selected estimator is a Random Forest model for our overheating prediction model. Generic random forest models are already widely used in building physics energy modelling thanks to their precision, the efficiency of the training of such models, and their interpretability [14].

Hyperparameters for the generation of the random forest trees were selected empirically and are detailed in **Table 2**. The Python library Scikit-learn [15] was used in our work to train the model with the Random Forest Classifier.

Table 2: Hyperparameters of the random forest prediction model

Number of estimators	500
Max depth	50
Min. samples at a node	2

3 RESULTS

This section presents the accuracy of the predictive models compared to DTS.

3.1 Model accuracy

Once trained, the predictive performance of the machine learning models was evaluated using cross-validation techniques. The dataset was split using 80% for training and 20% for the evaluation. The metric employed to quantify the accuracy and robustness of the models is the coefficient of determination (R^2): this metric assesses how well the predicted values align with the DTS results.

The results of the three models for the three types of dwelling: individual houses, collective housing on running and top floor, are summarized in **Fig 2**. The first line of the figure illustrates the distribution of DDH [°C.h] obtained from DTS (in blue) and those predicted by the data-driven models (in orange) for each of the three housing types. The close overlap of these distributions shows the models ability to replicate the range of overheating indictors observed within the simulations.

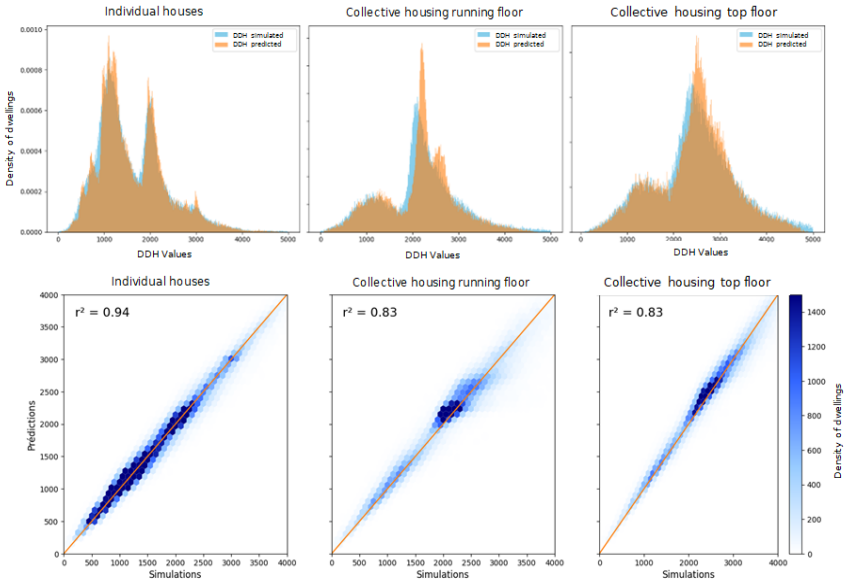

Fig 2: Comparison of predicted and simulated Discomfort Degree Hours (DDH) [°C.h] for the three trained models (individual houses, collective housing running floor and top floor. Comparison of distribution on the top line and scatter plots of predicted versus simulated DDH, with diagonal line (x=y) indicating perfect agreement and r^2 values demonstrating model accuracy.

Additionally, the second line of **Fig 2** provides scatter plots comparing predicted DDH values to the simulated ones for each model. The diagonal reference line (x=y) indicates perfect agreement, and the R^2 values are displayed to confirm the strong correlation between predicted and simulated overheating values. These graphics further highlight the reliability of the predictive models in estimating overheating across diverse building types.

These results stress the utility of machine learning models in bypassing the computationally intensive DTS, enabling large-scale overheating assessments of the French residential building stock.

The prediction model is then used to assess overheating for all of the 20 million dwellings in France, with DDH value ranging from 200 to 5000°C.h depending on the typology and the climate, as discussed in the next section.

3.2 Importance of each input parameters in the prediction

Feature importance in Random Forest models highlights the relative contribution of each input variable to the model's predictions, normalized to sum to 1 [16] . This allows us to identify the most influential parameters and their impact on the target variable.

Fig 3 presents the feature importances for the three Random Forest models developed in this study. Colors group parameters influencing similar physical effects (**Table 1**).

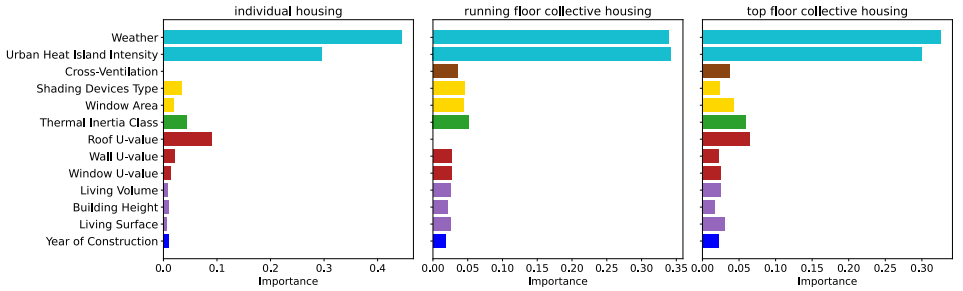

Fig 3: Feature importance for Random Forest models across the three housing types: individual houses, intermediate-floor collective dwellings, and top-floor collective dwellings

Fig 3 highlights the dominant influence of climatic parameters on the model predictions particularly climatic severity (linked to climate zone and altitude) and urban heat island (UHI) effects. UHI is less relevant for individual housing, which are typically located in less urbanized areas.

Architectural parameters have a secondary influence, with variations by housing type:

- Individual housing: Roof insulation is the most significant parameter, followed by inertia class and solar shading, while the cross-ventilation characteristic has negligible influence.
- Top floor collective housing: Roof insulation remains dominant, followed by inertia class and window-to-wall area percentage. The cross-ventilation housing character is more influential, ranking fourth among architectural parameters.
- Running floor collective housing: As there is no roof thermal exchange, roof insulation is irrelevant. The most influential parameters are inertia class, solar shading, window-to-wall ratio percentage, and the cross-ventilation housing characteristic.

Finally, other parameters such as wall and widows' insulation, geometry, and year of construction exert greater influence in collective housing compared to individual housing.

4 DISCUSSION AND CONCLUSION

The results are showing a great performance of prediction of the DDH with our statistical model compared to the simulations. Those prediction have been done on the 20 million residential buildings in France, and consider local climate, and 14 input parameters detailed in section 2.3. This nationwide indicator can enhance the relevance of public policies and is precise enough to generate statistics nationwide. However, some limitations apply to those results.

First, in the process of generating the training dataset, described section 2.2, uncertainties are mainly due to the conversion of EPCs to dynamic thermal simulation models : as the

level of details in the models is different, a converter is mapping and infer some of the inputs, thus adding uncertainties to the input parameters. The uncertainty was not quantified in this work, thus limiting the possible interpretations of comparisons between close values of predicted DDH.

To limit misinterpretations of predicted DH, uncertainties should be calculated and provided at each prediction, which is a current limitation and a perspective of our work.

Secondly, the prediction model is trained with an extensive dataset of dynamic thermal simulation that were validated as it is the current French thermal regulation. Another step of validation of our prediction model is the validation against experimental data and real measurements. Such a dataset to carry out the validation is hard to find as such a large-scale indicator needs a variety of measurements from multiple typologies to validate it for the whole building stock.

The validation process with real measurements is a scientific work of its own, but other limitations can be addressed in further version of the overheating indicator. As the indicator will be published in the French National Building Database (BDNB), more feedback will help in the refinement of the indicator.

5 REFERENCES

[1] A. Gasparrini *et al.*, « Projections of temperature-related excess mortality under climate change scenarios », *Lancet Planet. Health*, vol. 1, nº 9, p. e360-e367, déc. 2017, doi: 10.1016/S2542-5196(17)30156-0.

[2] D. Enescu, « A review of thermal comfort models and indicators for indoor environments », *Renew. Sustain. Energy Rev.*, vol. 79, p. 1353-1379, nov. 2017, doi: 10.1016/j.rser.2017.05.175.

[3] R. de Dear, J. Xiong, J. Kim, et B. Cao, « A review of adaptive thermal comfort research since 1998 », *Energy Build.*, vol. 214, p. 109893, mai 2020, doi: 10.1016/j.enbuild.2020.109893.

[4] S. Carlucci et L. Pagliano, « A review of indices for the long-term evaluation of the general thermal comfort conditions in buildings », *Energy Build.*, vol. 53, p. 194-205, oct. 2012, doi: 10.1016/j.enbuild.2012.06.015.

[5] T. J. Matongo, G. R. H. Ngock, E. Yamb, L. Mba, B. S. Diboma, et J. G. Tamba, « Assessing the severity of thermal discomfort in a building in the course of hot and humid climate », *F1000Research*, vol. 13, p. 962, nov. 2024, doi: 10.12688/f1000research.154075.2.

[6] *Annexe III - Arrêté du 4 août 2021 relatif aux exigences de performance énergétique et environnementale des constructions de bâtiments en France métropolitaine et portant approbation de la méthode de calcul prévue à l'article R. 172-6 du code de la construction et de l'habitation - Légifrance*. Consulté le: 3 février 2025. [En ligne]. Disponible sur: https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000043936431

[7] A. Foucquier, S. Robert, F. Suard, L. Stéphan, et A. Jay, « State of the art in building modelling and energy performances prediction: A review », *Renew. Sustain. Energy Rev.*, vol. 23, p. 272-288, juill. 2013, doi: 10.1016/j.rser.2013.03.004.

[8] M. Hamdy, S. Carlucci, P.-J. Hoes, et J. L. M. Hensen, « The impact of climate change on the overheating risk in dwellings—A Dutch case study », *Build. Environ.*, vol. 122, p. 307-323, sept. 2017, doi: 10.1016/j.buildenv.2017.06.031.

[9] P. Symonds *et al.*, « Development of an England-wide indoor overheating and air pollution model using artificial neural networks », *J. Build. Perform. Simul.*, vol. 9, nº 6, p. 606-619, nov. 2016, doi: 10.1080/19401493.2016.1166265.

[10] *CEN/TR 16798-2:2019*.

[11] B. Bueno, L. Norford, J. Hidalgo, et G. Pigeon, « The urban weather generator », *J. Build. Perform. Simul.*, vol. 6, nº 4, p. 269-281, juill. 2013, doi: 10.1080/19401493.2012.718797.

[12] C. Bacon *et al.*, « La base de données nationales des bâtiments (BDNB) ». 2021. doi: 61dc7157488f8cdb4283e3c3.

[13] A. Machard, « Towards mitigation and adaptation to climate change : Contribution to Building Design », phdthesis, Université de La Rochelle, 2021. Consulté le: 15 janvier 2025. [En ligne]. Disponible sur: https://theses.hal.science/tel-03675251

[14] Y. Chen, M. Guo, Z. Chen, Z. Chen, et Y. Ji, « Physical energy and data-driven models in building energy prediction: A review », *Energy Rep.*, vol. 8, p. 2656-2671, nov. 2022, doi: 10.1016/j.egyr.2022.01.162.

[15] F. Pedregosa *et al.*, « Scikit-learn: Machine Learning in Python », *J. Mach. Learn. Res.*, vol. 12, nº 85, p. 2825-2830, 2011.

[16] L. Breiman, « Manual on setting up, using, and understanding random forests v3. 1 », *Stat. Dep. Univ. Calif. Berkeley CA USA*, vol. 1, nº 58, p. 3-42, 2002.