Using a Machine Learning Algorithm to predict Summer Overheating in Buildings

J Herz[1*], M Hartner[1], S Carrigan[1] and O Kornadt[1]

[1] Bauphysik / Energetische Gebäudeoptimierung, RPTU Kaiserslautern-Landau, Kaiserslautern, Germany

[*]E-mail: jakob.herz@rptu.de

Abstract. In the scope of warming cities, this work aims to develop a software application that predicts overheating of buildings in summer with an accuracy comparable to thermal building simulations, but with a significantly reduced operational complexity and computational cost. The application will be suited for engineering offices as a hands-on assessment tool for energy consulting and when planning new or refurbished buildings.

The approach is to use a Random Forest Regression (RFR) machine learning algorithm to generate predictions for the air temperature and operational temperature in a room according to given features (room geometry, window specifications, shading, wall insulation, etc.). The training of the RFR is conducted on the basis of numerous thermal building simulations. These simulations were performed using TRNSYS, which has been thoroughly validated and provides a great range of possibilities for a trained user. In order to provide a sufficient multitude and variety of training data, automation software for TRNSYS was developed over the course of this work.

In this contribution, we report on the first results of predicting summer overheating using machine learning with the developed application. More specifically, the RFR results are compared to simulation data.

1. Introduction

1.1 Motivation

The German standard DIN 4108-2 [1] regulates summer thermal protection for buildings, aligning with the Buildings Energy Act (GEG). Existing methods for determining summer thermal protection requirements include thermal building simulation and the normative solar input method. While thermal building simulation is reliable, it is time-consuming and costly. The solar input method, though simpler, sacrifices accuracy and often overlooks specific local conditions. In practice, approximately 90% of projects use the solar input method, despite its limitations. Both methods fail to account for future climate change, which is crucial to avoid costly and energy intensive retrofitting of the building technology in the future.

The project aims to develop a fast, accurate, and cost-effective software tool for evaluating summer thermal protection in construction and refurbishment projects. It envisions a consulting scenario where clients and consultants use a simple online form to select from current and future (predicted by climate models,) weather data analyze building site conditions and discuss the results of various refinement options without the need of time-consuming traditional simulations.

1.2 Prediction tool

The tool should serve as a hands-on assessment tool for energy consulting and therefore must not only generate results fast and accurate but visualise effects of changing relevant aspects of a building site (building materials, window area, floor area, ceiling height, etc.) too. Hence the goal of the software is to predict the *air temperature* (*TAIR*) and the *operative room temperature (TOP)* hourly throughout the year for a room, using the latest weather data available from the Deutscher Wetterdienst (DWD) [2] to attribute the surrounding microclimate. Afterwards, the *overtemperature degree hours* are calculated in compliance with DIN 4108-2 [1].

The Random Forest Regression (RFR) machine learning algorithm was chosen as a prediction tool and numerous automated thermal simulations for various parameter and climate combinations were conducted in TRNSYS [3], a widely recognised tool in building performance analysis, to synthesize the needed amount of training data.

1.3 Choice of ML algorithm

The RFR of the *scikit-learn* Python package [4,5] was selected as the machine learning algorithm for this study due to its robustness, interpretability, and efficiency in handling complex, non-linear relationships within the data. RFR constructs multiple decision trees during training and outputs the mean prediction, enhancing accuracy and reducing overfitting [6],which is particularly important when dealing with high-dimensional data such as the thermal characteristics of buildings. It is well-suited for capturing interactions between features like room geometry and environmental conditions, and it provides insights into feature importance, aiding in informed decision-making for energy consulting and building design.

However, it is important to note that RFR has certain limitations. It does not extrapolate well beyond the range of the training data, due to its nature of predicting an average of the values seen in the training phase. Secondly, it is unfeasible to use it as a “simulation engine“ and predict one sample at the time as the prediction time does not scale linearly with the size of the test dataset. By leveraging the strengths of RFR and addressing its limitations through careful data preparation, this study aims to develop a practical and efficient tool for predicting time series for *TAIR* and *TOP* in buildings.

2. Methods

In this study, synthetic data generated through thermal simulations using TRNSYS [3] were employed. This approach theoretically allows for an almost limitless supply of data, which is crucial for training effective machine learning (ML) models. The primary challenge lies in generating data with high information density and variance within the feature space, ensuring that the model can generalise well to unseen scenarios. For clarity in this paper *features* are defined as the input for the ML model while the inputs for the simulations are called *parameters.*

Table 1. Training data is generated by simulations of rooms. The rooms differ within the limits of the following features:

Feature	Description	Lower limit	Upper limit	Unit
Room characteristics				
Length	Cubic measures	3,0	10,0	m
Width		2,5	6,0	m
Height		2,4	4,0	m
Floor	1: ground floor, 2: middle floor, 3: top floor	1	3	-
U_ext_wall	Thermal transmittances of room-bounding structures	0,1	3,0	$Wm^{-2}K^{-1}$
U_roof[a]		0,1	3,0	$Wm^{-2}K^{-1}$
U_ground_floor[a]		0,1	3,0	$Wm^{-2}K^{-1}$
U_adjacent[b]		0,5	3,0	$Wm^{-2}K^{-1}$
Ventilation	Total air exchange rate.	0,25	5,0	h^{-1}
Internal_Gain	Heat contributions from internal sources.	0	50	$kJh^{-1}m^{-2}$
Capacity	Thermal capacity per unit ground area.	0	150	$Whm^{-2}K^{-1}$
Window characteristics (for each window)				
Win_wall_ratio	Ratio of window compared to the facade area.	0,001	0,97	-
Frame_win_ratio	Ratio of frame area compared to the window area.	0,01	0,4	-
U_window	Thermal transmittances of window.	0,7	5,6	$Wm^{-2}K^{-1}$
G_window	Total solar energy transmittance.	0,14	0,84	-
Fc	External shading factor on solar gains.	0	1	-
Environmental factors				
RADOS[c]	Total short-wave radiation on the vertical window	0	1000	Wm^{-2}
T	Ambient air temperature.	-20	37,2	°C
N	Cloudiness of the sky.	0	8	1/8

[a] Used, even if room is in a middle story
[b] Not a feature of the current RFR but considered in the simulations.
[c] Calculated by RADOS Module

2.1 Defining the feature space

To align with the requirements of the German standard DIN4108-2 [1], which outlines essential parameters for building energy efficiency during the cooling period, we defined a standardized room configuration. This configuration is essential for training an RFR, a supervised machine learning algorithm that requires comprehensive training data covering the entire feature space of future predictions. A standard room (see fig. 1a) in this study is assumed to be a cuboid with one or two facades, each containing a window. The room can be located on the ground floor, middle floor, or top floor of a building. Per definition [1], the air temperature within the room is assumed not to fall below 20°C and no active cooling is to be applied. This simplification allows for a consistent comparison of thermal behaviours while maintaining realistic boundary conditions. The selected features are categorised into three groups: room characteristics, window specifications, and environmental conditions. All features, as outlined in Table 1, vary within predefined limits, which were chosen based on physical necessities, DIN requirements, or use cases suggested by the industrial partner in this research project.

To determine the solar radiation energy input for vertical surfaces, a RADiation On Surface (RADOS) module was developed. This module converts the beam and the diffuse radiation on a horizontal surface, provided by the German Test Reference Year (TRY) weather data files [2,7], into radiation on a tilted surface. It follows the descriptions of Chapter 1 and 2 of „Solar Engineering of Thermal Processes, Photovoltaics and Wind" [8], where the solar radiation incidence angle is calculated in terms of the slope and azimuth of a surface [9] and the diffuse radiation is estimated by the Reindl model [10].

2.2 Automated simulations

To efficiently generate the necessary volume of data, a TRNSYS-Python API was developed. This API automates the creation of input files for TRNSYS building simulations and executes the simulations, bypassing the time-consuming manual process of using TRNSYS's graphical user interface (GUI). A reference building model, designed as a cuboid with 24 standard rooms, was initially constructed and saved as a *.b18* file, serving as a template modifiable via Python.

Throughout the study, a Python package was developed to mimic the organizational structure of TRNSYS's multizone building environment. This package includes classes representing various TRNSYS components, such as *LAYERS, SCHEDULES, CONSTRUCTIONS, ZONES* etc. or *EXTENSIONS* like the *WINPOOL* and *BuildingGeometry*. Users can modify building parameters by following the logic of the *TRNBuild* GUI, and the package allows for the extraction of room-specific features from the building model, even if they are not directly set as parameters. These features include:

- *Length, Width, Height*: Calculated as distances between vertices in the *BuildingGeometry* class.
- *U_construction*: Calculated in the *CONSTRUCTION* class using the specified thickness and conductivity values from the *LAYER* class.
- *Capacity*: Determined in the *CONSTRUCTION* class according to norm DIN 4108-6 [11], using thickness and conductivity, capacity, and density values from the LAYER class.
- *U_window, G_window*: Extracted from the *WINPOOL* class.
- *RADOS*: Derived from TRY weather data files, with slope and azimuth extracted from the building model.

To achieve high information density, TRNSYS's capability to schedule parameters like ventilation, internal gains, and frame-to-window ratio was leveraged. These parameters were varied periodically over the simulation time, generating multiple feature values within a single simulation run. Each simulation spans an entire year, producing 210.240 data points for 24 rooms.

To introduce further variance, specific features such as floor level, ventilation rates, internal gains, window-to-wall ratio, frame-to-window ratio, *U_window, G_window*, and shading factor (*Fc*) were varied across rooms, while cubic measurements and thermal insulation properties remained consistent.

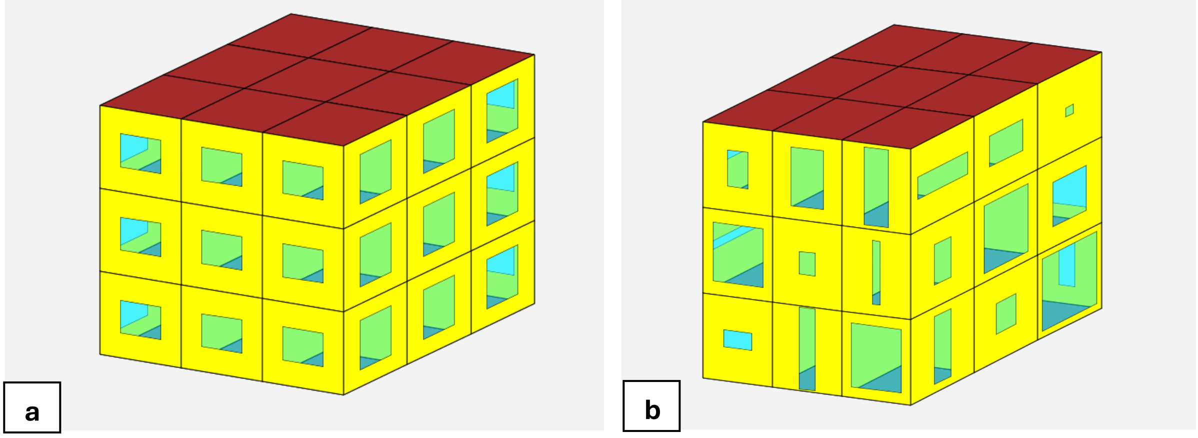

Figure 1: TRNSYS Multizone building models, which were used to generate synthetic training data. **a)** Reference building to initiate the automation. **b)** Building model after randomization.

2.3 Randomised simulations

The automated simulation routine is fundamental to covering the extensive feature space required for this study. RFR do not require grid-like data, as individual decision trees within the forest do not consider all features simultaneously. However, training data should roughly cover the entire feature space, with features being independent of one another. To achieve this, we implemented randomisation strategies in the selection of input parameters for the simulations:

- *Scaling Building Dimensions:* Random vectors were generated within the predefined limits for length, width, and height. Scaling factors were calculated for each axis and applied to every vertex in the *BuildingGeometry* class, ensuring that the room dimensions varied realistically within the constraints. A randomly scaled building is depicted in figure 1b.

- *Scaling Windows:* Vertices for exterior walls and their respective windows were identified. Random vectors were generated within the plane of the wall, constrained to the wall's centre and corners. Scaling factors were then calculated and applied to every window vertex, adjusting the window dimensions accordingly (see fig. 1b).

- *Creating Schedules*: A pool of weekly schedules was created, where every three hours, a random percentage of a features value range was chosen to simulate variations in parameters such as ventilation internal gains or frame to window ratio

- *Creating constructions:* Random U-value (*u_target*) within the defined limits were chosen (see Table 1). As long as the termination criterion $u = u\_target \pm 0{,}05\ \mathrm{Wm^{-2}K^{-1}}$ is not achieved a construction was created. Its thickness must exceed *0,12* m and its layers can be picked out of structural, insulation and/or coating layers (see below), although insulation layers must be coated.

- *Picking layers*: *TRNBuild* provides a library of construction materials. The library contains coating material, found within the wood, stucco or building board materials, insulation material, defined by a conductivity λ ≤ 0.06 $Wm^{-1}K^{-1}$ and structure material like stonework or concrete.

- *Picking weather*: A new TRY file is randomly selected before every simulation.

2.4 Data preprocessing and training

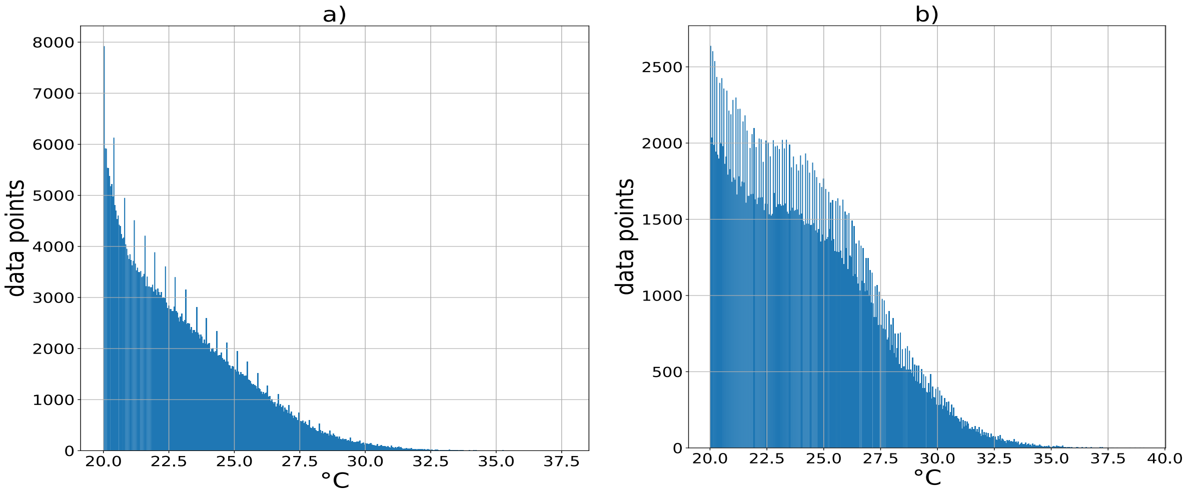

Figure 2. Histograms of the *TAIR* target values in the middle room dataset before (**a**) and after (**b**) the data preprocessing. Both show a significant underrepresentation of target values above *30°* C. For both plots data points where *TAIR* = 20 °C were omitted.

The simulated *TAIR* and *TOP* data follow an annual cycle, leading to an imbalanced distribution of target values (see fig. 2a) and a significantly reduced prediction accuracy for minority values. Specifically, the constraint on room air temperature (see section 2.1) results in nearly half of the generated data points having a value of 20° C, which lead us to generally exclude them. To further address this issue, the following preprocessing steps were applied and resulted into figure 2b:

- Only rooms with over temperature degree hours of 400 or more and plausible window-to-floor ratios were considered.

- Data was limited to the period from 1. May to 31. August to focus on the most relevant summer months.

- The dataset was split into two subsets: one for corner rooms (with two windows) and one for middle rooms (with one window), as the number of features differs between them. Consequently, a separate RFR estimator was trained for each subset.

Before training the RFR to predict the operational temperature, the room temperature must first be predicted. This is necessary to comply with DIN 4108-2 [1], which allows for dynamic adjustment of ventilation based on *TAIR* and ambient temperature *T*. This poses a challenge, as predictions can only be made on complete datasets, not on individual data points (see section 1.3). To address this, the feature *TAIR_before* was introduced as a short-term memory for the RFR model. During training, *TAIR_before* is set to the room temperature from the previous time step. For predictions, *TAIR_before* is updated with the results from the last prediction, and the process is repeated until a stopping criterion is met. *TOP* uses then the obtained *TAIR* values as additional feature to incorporate the now gained short-term memory but to avoid the time-intensive convergence method. For the training of the *TOP* estimator the choice of data was only constraint by considering rooms that reach 400 or more overtemperature degree hours.

3. Results and discussion

At the current state 2400 rooms were simulated, resulting in more than $300{\cdot}10^3$ data points for training a single RFR estimator. Figure 3a and 3b compares the predictions of our trained model with simulation results for a middle room. The room, with an area of 12 m^2, is on the upper floor, with a 4.8 m^2 south-facing window using sun protection glass and external shading. The roof and window are insulated, while the walls are not.

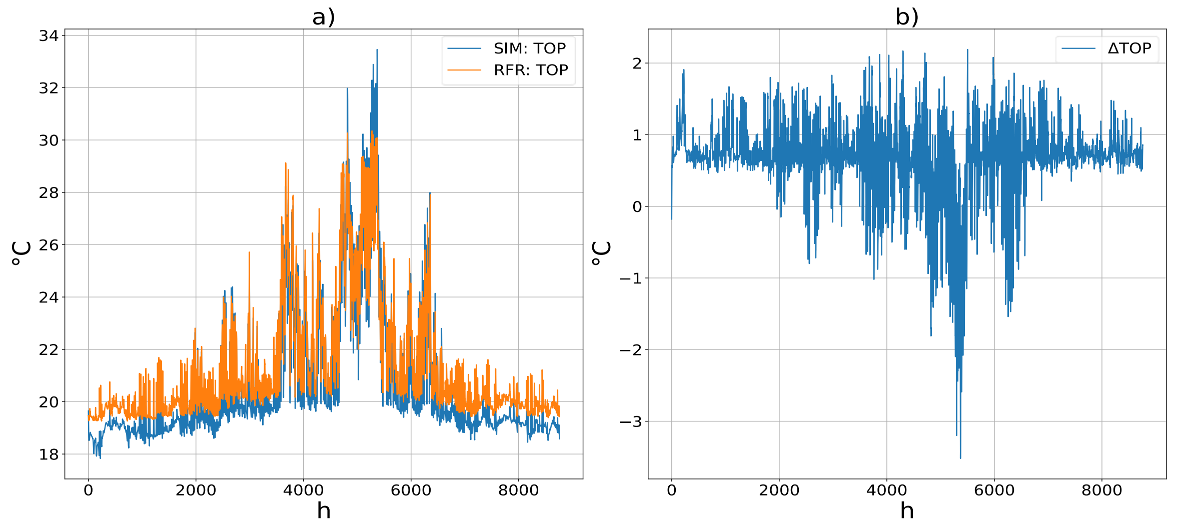

Figure 3. a) Comparison of the operational temperature (*TOP*) when predicted by the RFR and by a simulation (SIM) of a middle room. **b)** Differences between the RFR prediction and the simulated TOP.

The RFR predictions for *TOP* closely follow the simulated results, particularly in the 22° C to 30° C range. However, the model struggles to predict values above 30° C, likely due to imbalanced training data. Additionally, the predictions exhibit an overestimation of temperatures below 20.5° C. It is attributed to the current method of preprocessing of *TAIR* data (see 2.4). For the sake of improved accuracy of high temperature predictions data points of the heating period are underrepresented in the training data. This shows the necessity to further investigate and improve the method of data preprocessing. The model also shows sensitivity to the short-term memory feature, *TAIR_before*, indicating that the temporal coherence of predictions is not fully captured.

4. Conclusion and outlook

In this study, we developed a method to predict air temperature (*TAIR*) and operative temperature (*TOP*) in a room throughout the year using a Random Forest Regressor (RFR). To synthesise the necessary training data, we developed software to automatically simulate randomised buildings. To capture temporal sensitivity, we introduced the feature *TAIR_before* during data preprocessing, providing the estimator with a short-term memory. Our model demonstrated qualitative validation in the temperature range of 22° C to 30° C. However, to enhance the model's applicability as a tool for predicting overheating in buildings, further improvements in accuracy, particularly for higher temperature values, are necessary.

Future work should focus on implementing over- and under-sampling algorithms to address imbalanced training data, generating new training data with a bias towards warmer rooms and without the constraint of a minimum air temperature of 20° C, refining the short-term memory mechanism, and reducing the number of features to create a more generalized model capable of predicting temperatures in non-standardized rooms. By addressing these aspects, the RFR model can be further refined to serve as a practical and efficient tool for energy consulting and building design, ensuring buildings are better equipped to handle future climatic conditions.

5. Acknowledgements

This publication resulted from research in cooperation with bow ingenieure GmbH supported by the Central Innovation Programme for SMEs (ZIM) of the German Federal Ministry of Economics and Climate Action.

The authors acknowledge the assistance of an AI language model in refining and editing this paper. The AI provided revisions to improve clarity and overall language quality. The authors are responsible for the final content and interpretations presented in this work.

References

[1] "DIN 4108-2:2013-02, Wärmeschutz und Energie-Einsparung in Gebäuden_- Teil_2: Mindestanforderungen an den Wärmeschutz,".

[2] "Wetter und Klima - Deutscher Wetterdienst - TRY - Testreferenzjahre (TRY)," 2/11/2025, https://www.dwd.de/DE/leistungen/testreferenzjahre/testreferenzjahre.html;jsessionid=7D7898857DB2864279A8D52B58C760D6.live11042?nn=507312.

[3] S. e. a. Klein, *TRNSYS 18: A Transient System Simulation Program*, Solar Energy Laboratory, University of Wisconsin, Madison, USA, 2017.

[4] Scikit-learn: Machine Learning in Python, "RandomForestRegressor," https://github.com/scikit-learn/scikit-learn/blob/3f89022fa/sklearn/ensemble/_forest.py#L1453.

[5] F. Pedregosa, G. Varoquaux, A. Gramfort et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[6] T. K. Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1, 278-282 vol.1, 1995.

[7] "Projektbericht Ortsgenaue Testreferenzjahre von Deutschland für mittlere, extreme und zukünftige Witterungsverhältnisse," Deutescher Wetterdienst, 2017.

[8] N. Blair, W. Beckman, and J. Duffie, *Solar Engineering of Thermal Processes, Photovoltaics and Wind, Fifth Edition*, wiley, 2021.

[9] J. E. Braun and J. C. Mitchell, "Solar geometry for fixed and tracking surfaces," *Solar Energy*, vol. 31, no. 5, pp. 439–444, 1983.

[10] D. T. Reindl, W. A. Beckman, and J. A. Duffie, "Evaluation of hourly tilted surface radiation models," *Solar Energy*, vol. 45, no. 1, pp. 9–17, 1990.

[11] "DIN V 4108-6:2003-06, Wärmeschutz und Energie-Einsparung in Gebäuden_- Teil_6: Berechnung des Jahresheizwärme- und des Jahresheizenergiebedarfs,".