HVAC System Performance Modeling Using Component-Based Machine Learning

Seyed Azad Nabavi¹, Ueli Saluz¹, Marco Wolf¹, and Philipp Geyer¹
¹ Sustainable Building Systems Group, Leibniz University Hannover, Germany

Abstract
Heating Ventilation and Air Conditioning (HVAC) systems are responsible for a significant portion of building energy consumption, accounting for up to 38% and 12% of global energy consumption. Predicting energy consumption for HVAC systems is highly important in the early design phases due to their significant impact on energy use and user comfort. However, it is a challenging task due to the complex and dynamic nature of these systems requiring the effort of building simulation. The current state-of-the-art methods for modeling HVAC systems use one data-driven model created by machine learning for the whole HVAC system. In this way, the behavior of the individual HVAC components is neglected and the developed model will have limited explainability and generalizability. The novelty of this study is breaking down HVAC systems into three component categories, which are zone components, secondary HVAC components, and primary HVAC components to not only predict the energy performance of HVAC systems but also the dependencies among them. Then, we apply a component-based machine learning approach to create data-driven models for the HVAC components’ performance. A random forest regression algorithm as a machine learning component serves to predict the performance of HVAC system components in buildings. Machine learning models developed for each component is performing predictions in a hierarchy of receiving/delivering information from other components. In this hierarchical component model, the zone components model informs the secondary HVAC components model of heat distribution components, and the secondary HVAC components model informs the primary model of the heat supply. The random forest approach was highly accurate in predicting component performance, with $R^2$ values of 0.99 and 0.97 for peak heating demand and annual heating demand in the zones, and 0.99 and 0.87 for the maximum design flow rate and UA value of Secondary-HVAC systems. The primary model also had high performance, with an $R^2$ value of 0.98. Compared to conventional data-driven models generated by machine learning, this component-based approach allows for better error tracking and offers explainability. Furthermore, the ML models have high generalizability due to the limited number of parameters and their reusability in further systems.

Highlights
- Developing component-based machine learning in HVAC system modeling.
- Dividing HVAC systems into three components including zone components, primary HVAC components, and secondary HVAC components.
- Developing ML models for each component to address not only component performance but also dependencies between components.
- Connect ML models in a hierarchical order to transfer information/prediction among the HVAC system components.

Introduction
The high energy consumption and CO2 emissions of buildings are major concerns in the effort to mitigate climate change. Buildings account for 33% of final energy consumption and 40% of direct and indirect CO2 emissions globally (Nejat et al., 2015). Within buildings, heating, ventilation, and air conditioning (HVAC) systems are responsible for a significant portion of energy consumption, accounting for 40-60% of the total energy usage (Solano et al., 2021). Therefore, forecasting the energy consumption of HVAC systems has become a crucial topic due to its impact on energy conservation strategies and the thermal comfort of building occupants (Afram & Janabi-Sharifi, 2015). For energy conservation connected to heating and cooling demand, there are vital interactions with the building shape and design of the envelope defined in early design phases. Therefore, accurately forecasting HVAC energy consumption in these design phases is essential for energy-efficient building operations.

The application of load forecasting for HVAC systems takes place at two stages: design and operation. During the design stage, load forecasting is used to choose energy supply equipment and pipeline sizes to ensure the HVAC system runs efficiently and comfortably. Furthermore, load forecasting during the operation stage is essential to efficiently run the system and to support energy-saving technologies such as cold and heat storage systems. Optimization processes at both stages require accurate load analysis to determine the optimal operating conditions and setting points (Qian et al., 2020).

For load forecasting in the design stage, the main influencing factors are building physical parameters, outdoor meteorological parameters, indoor
environmental parameters, and usage type. In this stage, the common load forecasting method is simulation modeling. However, on the one hand, it is difficult to obtain the influencing factors in the design stage. On the other hand, different degrees of assumptions and simplification in the process of building the model is necessary. However, design assumptions and simplifications result in significant differences between the forecasting results and the actual situation (Qian et al., 2020).

Data-driven models based on machine learning techniques have become increasingly important in the energy modeling of HVAC systems due to their ability to handle non-linear and complex problems based on the dataset of the systems. These models have proven to be robust and efficient, providing accurate prediction results in a shorter time frame. As a result, they have gained significant attention from designers and stakeholders, who benefit from their quick and reliable prediction results in making informed decisions (Yu et al., 2022). Several research studies have proposed different methods for HVAC load forecasting and energy consumption prediction.

Qian et al. (2020) proposed a combined method of simulation and transfer learning to improve the forecasting accuracy of heating and cooling loads. Liu et al., (2019) introduced a combination of Autoencoder and Deep Deterministic Policy Gradient (DDPG) algorithm for short-term HVAC system energy consumption prediction. The Bayesian Network technique was used by Tian et al. (2019) to select the most energy-efficient primary HVAC systems, and they conducted a survey to explore the important factors considered by designers in selecting primary HVAC systems. For newly designed HVAC systems, Woods & Bonnema (2019) proposed a regression model plug-in in EnergyPlus to estimate the influences of newly designed HVAC systems on building energy performance.

Deng & Chen (2019) applied Artificial Neural Network (ANN) to model the relationship between occupant behavior and HVAC energy consumption. The occupant behavior adds a significant component of complexity to HVAC systems modeling. To simplify the HVAC system’s complexity regarding the environmental factors and occupant behavior, Sha et al. (2019) proposed a simplified data-driven method to forecast the energy consumption of HVAC systems using time-series-related features of the outside-inside temperature difference and schedules (Sha et al., 2019).

These studies demonstrate the potential of machine learning and data-driven methods in improving the forecasting accuracy and control of HVAC energy consumption in buildings. However, the limited generalization and black-box nature of ML methods bring about limited Generalizability(reusability), interpretability, and explainability. Generalizability refers to the performance of ML models in forecasting new and unseen case studies. ML models’ generalizability is significantly important as these prediction models are applied in non-existing buildings in the early design phase (Chen et al., 2022). Moreover, machine learning models as a type of black box model provide limited guiding information for the design processes. In this regard, explainability and interpretability are about providing insights into how ML models work in order to check and improve prediction results (Geyer et al., 2021).

To overcome the downsides of monolithic black-box ML models, the study uses a novel component-based machine learning method for HVAC systems. This method breaks the HVAC system down into separate data-driven models instead of representing the whole system with one data-driven model. For prediction, these data-driven components are put together and form a deep ML structure aligned with the system. The components have significantly less complexity due to the limited number of variables. Decreasing the complexity of ML models improves the transferability of developed ML models to new cases. Moreover, the proposed method provides information that improves the interpretability and explainability of ML models and allows designers and engineers to reason much better using the models. Moreover, unlike monolithic models, CBML models provide error traceability as these models include an ML model for each component that provides the opportunity to get performance metrics for each component model.

**Methods**

Decision support for the HVAC system design is a multi-variate problem that requires models with the capability of dealing with multi-dimensionality. Moreover, providing support for designers and decision-makers requires fast models to evaluate the performance of HVAC systems for many design options in a short time with the limited amount of information given in the early design phase. Hence, a nimble model with high accuracy in forecasting multi-dimensional problems is required to provide assistance for designers and decision-makers regarding HVAC systems.

For tackling problems with numerous variables Random Forest Regression (RFR) algorithms are highly competent. A large group of decision trees makes up the random forest model, and this method performs efficiently and effectively when it comes to multi-variate regression tasks. The random forest algorithm typically achieves good results without requiring feature selection, and it also has the ability to intuitively determine which features are more significant in producing the final outcome (Li, 2020).

**Data Generation and Processing**

In the literature, the reliability of EnergyPlus in HVAC system simulation has, for instance, been validated using measurement data from the FLEXLAB project to estimate the prediction accuracy of EnergyPlus in simulating HVAC systems (Haves et al., 2019). The EnergyPlus HVAC system simulation results have an error range of up to 9.4% root mean squared error averaged in overall cases. Moreover, the EnergyPlus has an integrated solution of load (zone component), system
(secondary HVAC component), and plant (primary HVAC component) which enables EnergyPlus to model complex HVAC systems (Zhou et al., 2014). Therefore, we assume EnergyPlus is able to capture the relevant characteristics of the HVAC systems in buildings. Furthermore, the prepared simulation data, generated by EnergyPlus, in our study serve as input data to the component-based machine learning models. These data are used as preliminary data sources for developing the component-based ML model while the developed CBML models are generalizable to measured data from real buildings. In this study, we have simulated 1500 building configurations to achieve generalizability by covering typical office building characteristics, especially, using a variation of pre-validated HVAC systems, ASHP-Baseboard, Boiler-Baseboard, and DHW-Baseboard, in EnergyPlus. In this paper, we considered each building as a set of zones. Each zone is heated by a secondary HVAC system. Hence, for training the zone component and secondary HVAC component models, we have in total of 6002 samples. In the simulated data, the design space matrix for these simulations considers the influences of orientation, number of floors, floor height, façade type, primary HVAC systems type, and heating setpoints in the performance of the HVAC systems. In Figure 1, the variation of input variables for the simulations is represented.

![Building Orientation](image1)

![Façade Type](image2)

![Primary HVAC-System Types](image3)

![Number of Floors](image4)

In this study, the primary HVAC system type, building orientation (0 to 180 degrees), and heating setpoint (21-24 °C) are inputs to the primary component ML model. In the zone component, the cardinal direction of the exterior walls and windows is important as it determines the solar heat gain of the zones. Hence, we have considered the cardinal orientation of the zone surface times the area of that face. The cardinal direction includes north, east, west, and south. In this way, both the direction of the face and the area of the face is considered in the inputs. Moreover, each building has an energy standard that defines the U-Value for walls, slabs, and windows.

The Energy Standards include five categories;

- **1970s**: Constructed buildings prior to 1970.
- **GEG**: Constructed buildings according to Building Energy Code in Germany.
- **NZEB**: Constructed buildings according to Net-Zero energy building Standards.
- **Passive house**: Constructed buildings according to Passive House Standards.

The average U-Value for building components includes base plates, flat roofs, sloped roofs, above-ground bearing exterior walls, below-ground bearing exterior walls non-bearing exterior walls, and windows are defined based on the energy standards. In Figure 1-f, the averaged value of U-Values for each category is demonstrated. Moreover, the floor height and the number of floors are considered in zone-component modeling.

---

1. Gebäudeenergiegesetz, GEG
Finally, the collected information from the simulations for each variable has different dimensions. In this regard, we scaled the input and output variables to bring them to a 0 to 10 range. The schematic overview of the buildings, zones, and secondary and primary systems is presented in Figure 2. As illustrated in this figure, each building has a primary system. In these buildings, each zone is connected to a secondary system. Heating demand for all secondary systems in a building is supplied by a primary system.

![Figure 2 - A schematic overview of the buildings, zones, secondary HVAC systems, and primary HVAC systems.](image)

**Hierarchy of the proposed method**

The CBML model in this study includes three components of data-driven modeling for zone-component heating demand, secondary HVAC components in each zone, and primary HVAC components for each building, respectively. These ML models are connected in a hierarchical structure. This hierarchical structure is demonstrated in Figure 3. This structure begins with the zone information layer representing the zones’ characteristics. In the next component (zone component), a machine learning model receives the zone information and forecasts the Annual Heating Demand (AHD) and Peak Heating Demand (PHD) for each zone. At the secondary system component, the predicted annual heat demand, the predicted peak heating demand, and the zone information are input into the ML model. This component is composed of two ML models for forecasting the Maximum Design Flow Rate (MDFR) and UA of the secondary systems, respectively. The first ML model forecasts the maximum design flow rate which later is inserted as an input to the second ML model to forecast UA for each secondary HVAC system. Lastly, at the primary HVAC system component, the sum of the UA value of all secondary HVAC systems, the heating setpoint for each building, and primary HVAC system types in each building are imported as inputs to the ML model to forecast the design capacity and design outlet temperature of the primary HVAC system.

The developed ML models for the components are evaluated using cross-validation with five folds which means the data set is divided into 5 sections and the models are run five times. Each time one subset of the data section is selected as a test set. Then, the average performance metrics are estimated for the components ML model.

![Figure 3 – Hierarchical structure of the proposed method.](image)
Figure 4 - Correlation diagram of the input variables and target variables in the zone-component ML model.

We used a correlation diagram of the input variables and the output variable to graphically demonstrated the correlation matrix representing the correlation between variables. As noticeable in Figure 4, the UA of the windows and exterior walls have the greatest impact on both annual heat demand and peak heat demand, respectively. For the peak heat demand of the zones, the most important input variables are the UA of the windows (0.54), direction times the area of the zone faces (0.44), the zone’s volume (0.42), and the UA of the exterior walls and floor area of the zones equally important (0.4).

Moreover, for the annual heat demand of the zones, the most significant input variables are the UA of the exterior windows and walls (both 0.42), direction times the area of the zone faces (0.35), the zone’s volume (0.24), and the floor area of the zones (0.22). Moreover, the UA of the interior walls has the lowest correlation with both annual heat demand and peak heat demand.

Secondary HVAC system components

At the secondary system components, the zone information, annual heating demand, and peak heat demand are the input variable of the ML model. Firstly, the random forest regressor predicts the design maximum flow rate value of each secondary system. Then, the predicted design maximum flow rate value is added to the input variables and imported to the next ML model for predicting the UA value of the secondary HVAC system. The correlation between the inputs and outputs of both ML models in these components is presented in Figure 5.

As demonstrated in Figure 5, the design’s maximum flow rate, peak heat demand, and annual heat demand are highly correlated with the design’s maximum flow rate. It shows the linear relationship between the design maximum flow rate and peak heat demand and annual heat demand.

Regarding the UA value correlation matrix, the most important variable affecting the UA value of secondary HVAC systems are peak heat demand, annual heating demand, design maximum flow rate value, UA value for windows, direction times area, and finally the UA value for the exterior walls.

Figure 5 - Correlation diagram of the input variables and target variables in the secondary system ML model.

Primary HVAC system components

In this component, we have developed an ML model to predict the design capacity of the primary HVAC system considering setpoint temperature, design UA value of all secondary systems in a building, building orientation, and primary HVAC system type as input variables. The primary HVAC system types are central air source heat pump, gas boiler, and district hot water. A heatmap demonstration of the correlation between the input and output variables is presented in Figure 6.

Figure 6 - Correlation diagram of the input variables and target variables in the primary system ML model.

Result and Discussion

Our proposed method is composed of three ML models for forecasting three components of HVAC systems including zones, secondary HVAC systems, and primary HVAC systems. RFR is used to forecast annual heat demand and peak heat demand of each zone in a building. Also, this ML method is used to predict the maximum design flow rate and UA value of the secondary HVAC systems, and the capacity and outlet temperature of the primary HVAC systems.

We used several performance measures including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Average Percentile Error (MAPE), and R-Squared, to assess how well the proposed methods performed. These performance measures are calculated for each component and presented in Table 1.

Table 1: Performance Metrics.

<table>
<thead>
<tr>
<th>Component</th>
<th>Metric</th>
<th>Unit</th>
<th>MAE</th>
<th>RMSE</th>
<th>MAPE</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zone</td>
<td>PHD</td>
<td>W</td>
<td>666</td>
<td>1368</td>
<td>5.8%</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>AHD</td>
<td>GJ</td>
<td>6.36</td>
<td>11.8</td>
<td>21%</td>
<td>0.97</td>
</tr>
</tbody>
</table>
Zone Component results

In this study, firstly, we used RFR models to forecast the annual heat demand and peak heat demand of zones. The test error plot of the developed models for forecasting annual and peak heat demands are demonstrated in Figure 7 and Figure 8. Moreover, the feature importance analysis of the RFR model is represented in Figure 9. The UA value of exterior walls, windows, and interior walls have, respectively, the highest impact on both peak and annual heat prediction.

![Figure 7 - Test error plot of peak heat demand prediction](image1)

![Figure 8 - Test error for annual heat demand prediction](image2)

In these Figure 7 and Figure 8, the R-squared value and the error distribution plots are presented for the test set. As noticeable in these figures, the RFR approach has a high performance in forecasting both annual and peak heat demand.

Secondary HVAC system component

After forecasting annual and peak heat demand at the zone component, we use an RFR approach to forecast the maximum design flow rate. In simulation tools, the maximum design flow rate is calculated based on the peak and annual heat demand of the zones. Similarly, ML model predictions are highly correlated to these variables. As noticeable in Figure 10, the prediction accuracy of RFR in the test set is significantly high. This shows the high capability of ML models in replacing simulation tools.
In the next step, using the predicted maximum design flow rate, annual heat demand and peak heat demand, and zone features as inputs to RFR we predicted the UA value for each secondary component in the zones. Predicting UA value for the secondary HVAC components is a multivariate and complex problem. However, the RFR model forecasted this value with a high accuracy considering the complexity and multi-dimensionality features of the UA values. In Figure 12, the feature importance analysis of the RFR models in forecasting maximum design flow rate and UA-secondary HVAC is demonstrated. As presented in Figure 12 (a), the prediction of the maximum design flow rate is highly correlated to the peak heat demand value of the zones. Moreover, the most important feature in forecasting UA values of the secondary HVAC systems, as depicted in Figure 12 (b), are maximum design flow rate, UA value of windows, peak heat demand, and annual heat demand features, respectively.

Moreover, the correlation between the UA value of windows and the UA value of secondary HVAC systems shows the great role of the windows in heat loss and respectively the need for bigger secondary HVAC systems with higher U-values.

**Primary HVAC system component**

In modeling the primary HVAC systems, we forecast the design capacity of the primary HVAC systems. For forecasting this variable, we used the sum of UA values for all secondary HVAC systems, primary HVAC system type, heating temperature setpoint, and the orientation of the buildings. In Figure 13, the test error distribution plot of the design capacity prediction results is demonstrated. The RFR model shows high performance in forecasting the capacity of the primary HVAC systems. Moreover, the feature importance analysis of the primary HVAC system is demonstrated in Figure 14.

---

**Figure 10** - Test error plot of maximum design flow rate prediction

**Figure 11** - Test error for UA prediction of secondary HVAC systems

**Figure 12** - Feature importance analysis of the input variable in maximum design flow rate (a) prediction and UA of secondary HVAC (b) prediction.
As noticeable in Figure 14, the maximum design flow rate and UA of secondary HVAC systems have the highest importance in forecasting Primary HVAC capacity, respectively. The building orientation and the heating setpoint temperature are the third and fourth important features in forecasting the Primary HVAC system’s capacity.

In the zone components, there are a wide variety of zone configurations that provide a high generalizability of the trained ML model. Similarly, in the secondary HVAC components, the UA value and the maximum design flow rate is representing most of the water-based secondary HVAC systems. Finally, the primary HVAC systems are represented by the capacity that represents most of the primary HVAC systems. Hereby, the proposed ML models for the components have high generalizability as the components have a limited variety of configurations, unlike monolithic HVAC system models.

In future works, we will compare the CBML approach with a monolithic approach in terms of forecasting performance, explainability, and generalizability.

Conclusion
In this study, the proposed approach for modeling HVAC systems involves breaking HVAC systems down into three components, each of which is modeled using machine learning. This approach allows for better error tracking and explanation of the component models. The proposed ML models for each component have high generalizability due to the limited range of configurations of the zone, secondary HVAC, and primary HVAC components. In this study, each building is composed of a set of zones, secondary HVAC systems, and a primary HVAC system. Accordingly, we developed an ML model for each component. These components are connected in a hierarchical pattern as the zones are modeled then the results are transferred to the next component which is the secondary HVAC system. Finally, all secondary systems ML results are transferred to the primary HVAC system in a building to forecast the primary HVAC systems’ capacity.

We used a random forest regression algorithm as a component-based machine learning model to forecast building HVAC systems’ component performance. The random forest regression approach has forecasted the HVAC system components with high accuracy according to the estimated performance metrics for each component. The R-squared value for peak heating demand and annual heating demand in the zone components is 0.99, and 0.97, respectively. Moreover, in the secondary HVAC components, the R-squared values are 0.99 and 0.87 in forecasting the UA value and maximum design flow rate of the secondary HVAC systems. Finally, the ML component has high performance in forecasting primary HVAC systems’ capacity prediction with an R-squared value of 0.98.

Acknowledgments
We gratefully acknowledge the German Research Foundation’s (DFG) support for funding the project under grant GE 1652/3-2 in the Researcher Unit FOR 2363 and under grant GE 1652/4-1 as a Heisenberg professorship.

References


