Prediction of outdoor air temperature and humidity using Xgboost

Xiaoming Ma¹², Cong Fang¹, Junping Ji¹³,*
1 School of Environment and Energy, Peking University Shenzhen Graduate School, Shenzhen, China
2 College of Environmental Sciences and Engineering, Peking University, Beijing, China
3 Energy Analysis and Environmental Impacts Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
E-mail: jackyji@pku.edu.cn

Abstract. Outdoor temperature and humidity prediction plays an important role in HVAC intelligent control with huge energy saving potential and identification of the urban heat island effect. The aim of this paper is to develop and evaluate a Xgboost model for the prediction of outdoor air temperature and humidity using acquired data from Shenzhen, China. Datasets fetched from sensor real-time collection and meteorological station interface are utilized as observations, through the construction of Xgboost to predict outdoor temperature and humidity in predictive horizon of 1-3 hours. The effectiveness of the prediction is verified by comparing the prediction with the measured outdoor air temperature and humidity. In addition, statistical tools such as R² and root mean square error (rmse) are applied to evaluate the performance of the model. The results show the excellent performance of Xgboost in accurately predicting outdoor temperature and humidity by comparison between the measured and predicted outdoor air temperature (R² > 0.73, rmse < 1.77) and air humidity (R² > 0.81, rmse < 6.33).

1. Introduction
With HVAC energy consumption accounting for more than 40% of a commercial building’s whole energy use and rapid development of artificial intelligent optimization technologies, data-driven algorithms are implemented by many building energy efficiency researchers to the HVAC control for optimizing its energy consumption. The temperature and humidity around the building affect the transfer of energy in the middle of outdoor and building entity, as well as the energy consumption of the HVAC system, therefore it’s essential to predict the trend of temperature and humidity as basis for intelligent control algorithm. G. Gao et al.[1] have implemented the prediction of outdoor building environments including temperature and humidity as the parameters of the deep neural network algorithm-based strategy to reduce the energy use of HVAC.

In addition, prediction of air temperature and humidity outside is applied to energy-saving technologies based on perceived environmental information for minimizing the energy consumption of HVAC system at a predictable level. T. Wei et al.[2] take into consideration of multi-horizon prediction of outdoor environment data for developing a data-driven approach that utilizes the deep reinforcement learning(DRL) algorithm, to intelligently learn the optimized strategy for controlling the building HVAC system. P. Fazenda et al.[3] has applied a scattered and constant supervisory control method on the fundamental reinforcement learning, which significantly learns the schedule mode of temperature setpoints with consideration of inputs including outdoor weather forecast. Their research shows that it is necessary to incorporate dynamic prediction of the outdoor weather such as air temperature and
humidity to self-adaptive control of building HVAC systems.

As demonstrated in figure 1, the prediction of air outdoor temperature can be utilized as an input to the data-driven HVAC control algorithm. The predictive optimization algorithm provides the best solution for lower energy consumption operation of HVAC, as shown by Provata et al.[4], the optimization algorithm reduces the energy consumption at the benchmark level.

With human activities improving, relevant urban researchers have studied the urban heat island effect. Kolokotsa et al. and Maragkogiannis et al.[5, 6] have discovered the urban heat island effect respectively by using the collected measurement data as well as the results of simulation. Kolokotsa et al. has utilized hardware devices to estimate the change of outdoor air temperature in Chania. The analysis of the obtained results shows that the urban heat island effect has appeared in Chania city. Maragkogiannis et al. uses a simulation model to confirm the rise in outdoor temperature in the city’s middle area.

As a result of determining the urban heat island effect’s impact on energy consuming, Bueno et al.[7] have developed a weather generator for urban areas which evaluates hourly urban canopy air temperature and humidity data locally from a weather station nearby based on neighbourhood regional energy consumption balances. Furthermore, Smargiassi et al.[8] has addressed the impact of urban heat islands on indoor air temperature with a method developed for the prediction of indoor temperature with outdoor temperature related to it. The results of their studies shows that it is necessary to predict the variation of outdoor air temperature that is beyond the comfort range and affects the health of urban residents to identify the urban heat island effect.

For the purpose of predicting outdoor air temperature accurately and constantly, Gobakis et al.[9] implemented a neural network in Athens in summer. The prediction outcome has been feeding into the geographic information system application to demonstrate the heat island effect in further few hours. Papantoniou et al.[10] present the implementation and verification of neural network based identification models for the prediction of outdoor air temperature in four European cities.

As the parametric model based algorithm requires large amount of dataset to train the model, such as neural network, the model efficiently fitting small amount of dataset has not been discussed. Apart from that, there has been a lot of research on implementing and evaluating algorithms for outdoor temperature prediction but humidity prediction has not been covered.

To fill this gap, this research aims to develop and evaluate a prediction model based on Xgboost, which can use data from sensors deployed outside the building and API serving interface to local meteorological station to predict the outdoor temperature and humidity.

The whole research is presented in three more parts. Part 2 illustrates the explanation of the methods applied in outdoor air temperature and humidity prediction. Part 3 introduces the outcomes of the method implemented in Shenzhen, China. Finally, Part 4 presents the conclusions and future research prospects.

2. Predictive algorithm of outdoor air temperature and humidity

The prediction algorithm is developed based on the Xgboost because it can adapt to a large number of cumulative experimental data to improve the performance of the data accumulation model and also have excellent performance in small datasets. The full name of Xgboost is extreme Gradient Boosting, T. Chen[11], who formally proposed the algorithm for the first time in 2014. The basic principle is to combine several low-precision base models into a high-precision model. The loss function of Xgboost has the characteristics of expansibility, high accuracy and good fitting effect. The model of the lifting tree is iterated by the base model to reduce the loss error. A typical structure of a Xgboost is shown in figure 2.

Compared with the neural network used by Gobakis et al., Mihalakou et al., Kolokotrodi et al., Papantoniou et al.[9, 10, 12, 13, 14] to predict outdoor temperature, Xgboost is more generalized to different amounts of datasets and its performance in training set and test set is more consistent. In addition, although there are many software that can be used for the prediction of air temperature, the computing time required for software programs to run is longer than the time required for Xgboost model reasoning which is unable to meet the needs of real-time prediction scenarios. Furthermore, compared with weather forecasts which are applicable to a wider range of areas, Xgboost-based predictions can be used in specific areas of cities with locally required data.
Figure 1. Implementation of an optimization algorithm to minimize energy using within prediction of outdoor temperature and humidity.

2.1 Applied methodology
The method followed developed a specific method for this study, as shown in figure 3, and then evaluated the outdoor air temperature prediction in Shenzhen, China. The following data is selected as features to Xgboost:

- Time (months, days, hours)
- Outdoor temperature
- Outdoor relative humidity
- Solar Radiation
- Regional weather forecasts

The extreme gradient lifting (Xgboost) algorithm developed by Chen and Guestrin [11] is a new gradient integration machine implementation. The algorithm is derived from the idea of "Boosting". It combines the prediction results of a group of "weak" learners and integrates a "strong" learner through cumulative training steps. Xgboost has features of preventing overfitting as well as optimizing computing resources which is achieved by adding a regular term to the objective function with the optimal speed of calculation maintained by parallel computing in training process.

The cumulative learning process in Xgboost is explained as below. Firstly, the start learner is fitted to the whole space of the input data. Then the second model is fitted with the residual of the first learner in order to improve weak learning ability of the previous learner. The accumulation process is repeated several times until the stop condition is satisfied. The final predictive result of the algorithm is calculated by the addition of the predictions of all learners. The predictive function in step t is as follows[15]:

\[ f^{(t)}_i = \sum_{k=1}^{t} f_k(x_i) = f^{(t-1)}_i + f_i(x_i) \]  

(1)

Where \( f_i(x_i) \) is the learner in step t, \( f^{(t)}_i \) and \( f^{(t-1)}_i \) are the predicted values in steps t and t-1, and \( x_i \) is the input observation. In order to prevent the over-fitting of the training data from affecting the calculation speed of the model, the Xgboost model uses the following equation to evaluate the model:
Figure 2. Typical structure of a tree-based Xgboost with n estimators.

Table 1. Hyper parameters for training the Xgboost.

| Hyper parameter name | Value | Description               |
|----------------------|-------|---------------------------|
| eta                  | 0.05  | learning rate             |
| max_depth            | 5     | max depth of a tree       |
| subsample            | 1     | sample ratio of training data |
| colsample_bytree     | 1     | sample ratio of features  |
| alpha                | 0.2   | L1 regularization term    |
| n_estimators         | 10000 | number of estimators for boosting |
| scale_pos_weight     | 1     | weight value of labels    |

\[
\text{Obj}^{(t)} = \sum_{k=1}^{n} l(y_k, y_{tk}) + \sum_{k=1}^{t} \Omega(f_k)
\]

Where \(l\) represents the loss function, \(n\) represents the amount of input data, and \(\Omega\) represents the regular term, which is defined as follows:

\[
\Omega(f) = \gamma T + \frac{1}{2} \lambda ||\omega||^2
\]

Where \(\omega\) represents the leaf score, \(\lambda\) is the hyper parameter of regular term, and \(\gamma\) represents the minimum loss that the leaf node needs to split. More specific information and calculation process of Xgboost can be required in Chen and Guestrin [11].

In this research tree model is the booster of Xgboost, therefore it’s not required to normalize the input data during data preprocessing. In this paper, the measurement data and regional weather forecasts from April 28 to October 19 is used to train the Xgboost. Over the next period with predictive horizon of 1-3
hours, the trained Xgboost was used to predict outdoor air temperature and humidity. The Xgboost is retrained for updating model parameters to fit latest data once a week. Therefore, it takes into account the latest changes of outdoor environments. In the reasoning stage, the renewed Xgboost is implemented to predict the outdoor air temperature and humidity in the next 1-3 hours.

Before the training stage, some hyper parameters have been defined with certain values and used in whole research. As these hyper parameters play an important role in the prediction outcomes of the Xgboost, they are demonstrated in table 1.

Dataset of initial stage for training is 125-day constant outdoor temperature and humidity data in Shenzhen and the training set and the test set are respectively 90% and 10% of the dataset. The evaluation of the Xgboost is realized by calculating error analysis indicators such as:

- R-square (R²).
- Root mean square error (rmse).

These two error analysis indicators assist to assess the overall performance of predictions. The stats library of python was used for statistical analysis. The calculation process of R-squared and rmse are as follows:

\[
R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}
\]  

\[
rmse = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}
\]

3. Implementation of predictive algorithm

The Xgboosts are trained and evaluated in Shenzhen, China. Historical data accumulated by sensors measuring and fetching from API interface to meteorological station are stored in database located in Peking University Shenzhen Graduate School. The time frame for outdoor air temperature and humidity prediction is 1-3 hours. The inputs for training and testing were obtained from this database.

3.1. Outdoor temperature prediction

The performance of the Xgboost predicting outdoor temperature 1–3 hours ahead is illustrated in figure 4 and in figure 5 a regression result demonstrates the deviation of prediction and measured value of outdoor air temperature. In figure 4, the Xgboost has excellent performance in predicting outdoor air temperature which is consistent with the pattern of measured temperature. Furthermore the regression result (figure 5) for the prediction horizon of 3 hours shows that the Xgboost predicts accurately the temperature of most values even with small amount of data for training. And the predictive performance can be improved in minimum and maximum values with more data accumulated. The statistical analysis confirms the effectiveness of outdoor air temperature prediction of the Xgboost. In table 2, the R2 indicator is above 0.7 for all the predictive time frame which implies that the predicted values fit the measured ones with the consideration of small amount of dataset. The rmse indicator increases with the expansion of the predictive time frame which implies that the predicted values fit the measured ones with the consideration of small amount of dataset. The rmse indicator increases with the expansion of the predictive time frame, but increase of the error is not significant from 2 hours to 3 hours.

Looking deeper into figure 4, it’s illustrated that the most predicted range of outdoor air temperature has less error than maximum predicted values especially in some days with extreme high maximum temperature which implies that the Xgboost achieves to predict the steady change of temperature and accurately predict extreme values of outdoor air temperature in most situation but has potential for improving the performance is to capture sudden variation pattern of outdoor temperature.

3.2. Outdoor humidity prediction

The performance of the Xgboost predicting outdoor humidity 1–3 hours ahead is demonstrated in figure 6 for the prediction and in figure 7 a regression outcome shows the deviation of prediction and measured
Figure 3. Followed methodology for preliminary and continuous training of the Xgboost.

Figure 4. Continuous outdoor temperature prediction based on the Xgboost. (From 09-June)

value of outdoor air humidity. In figure 6, compared with the prediction of outdoor temperature, the prediction of outdoor air humidity which has better performance in minimum and maximum values and is also consistent with the variation mode of measured values. Furthermore the regression outcome (figure 7) for the prediction time frame of 3 hours shows that the Xgboost predicts accurately the humidity of most values and the predictive performance can be improved in minimum values with more data in that range accumulated. The statistical analysis of error indicators proves the humidity prediction ability of the Xgboost. In table 3, the value of R2 indicator is above 0.8 for all time frames which implies
Figure 5. Regression result comparison between measured and predicted temperature (3-hour predictive horizon).

Table 2. Statistical analysis of temperature prediction performance of Xgboost

| Predictive horizon | 1 hour | 2 hours | 3 hours |
|--------------------|--------|---------|---------|
| $R^2$              | 0.88   | 0.75    | 0.73    |
| rmse               | 1.18   | 1.70    | 1.77    |

Figure 6. Continuous outdoor humidity prediction based on the Xgboost. (From 09-June)

Figure 7. Regression result comparison between measured and predicted humidity (3-hour predictive horizon).
Table 3. Statistical analysis of temperature prediction performance of Xgboost

| Predictive horizon | 1 hour | 2 hours | 3 hours |
|--------------------|--------|---------|---------|
| R²                 | 0.92   | 0.86    | 0.81    |
| rmse               | 3.89   | 5.26    | 6.33    |

that the predicted values fit the measured ones. The rmse increases with the expansion of the predictive time frame, and the error increase from 1 to 2 hours is as significant as from 2 to 3 hours. As is shown in figure 6, the most predicted outdoor air temperature values have less error than minimum predicted values especially in the day with extreme low minimum humidity which implies that the Xgboost achieves to predict the steady change of humidity and accurately predict minimum outdoor air humidity in most stable situation but the potential for improving the performance is to capture sudden variation pattern of outdoor humidity. Over long period of days, the variation trend of humidity is captured accurately.

4. Conclusions and future research prospects
In this manuscript, it is introduced that the prediction of outdoor air temperature and humidity can be realized accurately by using Xgboost. The performance of Xgboost is evaluated by using the measurement data collected by the sensor and weather forecasts fetched from API. Error analysis indicators such as R² and RMSE are utilized to estimate the performance of the Xgboost. The results demonstrate a satisfied ability of the Xgboost to predict outdoor air temperature and humidity with a predictive horizon of 1-3 hours and as prediction horizon expands from 1 to 3 hours, the error of prediction outdoor air temperature and humidity increase. In further research, the prediction time range will be extended to 4-24 hours to verify the prediction performance of the model in wider prediction horizons. In addition, as more meteorological data are collected, such as wind speed, rainfall ratio and atmospheric pressure, Xgboost will be trained for predicting wind speed, rainfall ratio, atmospheric pressure and other meteorological indicators.

In application aspect, Xgboost will be integrated into microcontrollers which can further reduce the cost of implementing energy management. Apart from that, outdoor weather prediction based on Xgboost will be part of input data for reinforcement control algorithm which can be used to save energy consuming of building facilities.

References
[1] G. Gao, J. Li, Y. Wen, “Energy-efficient thermal comfort control in smart buildings via deep reinforcement learning,” arXiv:1901.04693v1, 2019.
[2] T. Wei, Y. Wang, and Q. Zhu, “Deep reinforcement learning for building HVAC control,” in Design Automation Conference (DAC), 2017 54th ACM/EDAC/IEEE. IEEE, 2017, pp. 1–6.
[3] P. Fazenda and et al. Using reinforcement learning to optimize occupant comfort and energy usage in hvac systems. Journal of Ambient Intelligence and Smart Environments, pages 675{690, 2014.
[4] D. Kolokotsa, D. Rovas, E. Kosmatopoulos, K. Kalaitzakis, A roadmap towardsintelligent net zero- and positive-energy buildings, Sol. Energy 85 (December(12)) (2011) 3067–3084.
[5] D. Kolokotsa, A. Psomas, E. Karapidakis, Urban heat island in southern Europe:the case study of Hania, Crete, Sol. Energy 83 (October (10)) (2009)1871–1883.
[6] K. Maragkogiannis, D. Kolokotsa, E. Maravelakis, A. Konstantaras, Combiningterrestrial laser scanning and computational fluid dynamics for the study ofthe urban thermal environment, Sustain. Cities Soc. 13 (2014) 207–216.
[7] Nakano, A., Bueno, B., Norford, L. and Reinhart, C.F. (2015) Urban weather generator – A novel workflow for integrating urban heat island effect within urban design process. In Proceedings of
Building Simulation 2015, 14th Conference of International Building Performance Simulation Association Hyderabad, India, December 7-9, 2015, pp. 1901-1908.

[8] A. Smargiassi, M. Fournier, C. Griot, Y. Baudouin, T. Kosatsky, Prediction of the indoor temperatures of an urban area with an in-time regression mapping approach, J. Expo. Sci. Environ. Epidemiol. 18 (3) (2008) 282–288.

[9] K. Gobakis, D. Kolokotsa, A. Synnefa, M. Saliari, K. Giannopoulou, M. Santamouris, Development of a model for urban heat island prediction using neural network techniques, Sustain. Cities Soc. 1 (July (2)) (2011) 104–115.

[10] S. Papantoniou, D. Kolokotsa, Prediction of outdoor air temperature using Neural Networks; application in 4 European cities, Third Internation Conference on Countermeasures to Urban Heat Island (2014)

[11] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM; 2016. p. 785–94.

[12] G. Mihalakakou, M. Santamouris, D.N. Asimakopoulos, The total solar radiation time series simulation in Athens, using neural networks, Theor. Appl. Climatol. 66 (3–4) (2000) 185–197.

[13] G. Mihalakakou, M. Santamouris, D. Asimakopoulos, Modeling ambient air temperature time series using neural networks, J. Geophys. Res. 103 (1998)509–517.

[14] M. Kolokotroni, M. Davies, B. Croxford, S. Bhuiyan, A. Mavrogianni, A validated methodology for the prediction of heating and cooling energy demand for buildings within the Urban Heat Island: case-study of London, Sol. Energy 84(12) (2010) 2246–2255.

[15] Fan J., Wang X., Wu L., Zhou H., Zhang F., Yu X., Lu X., Xiang Y. Comparison of support vector machine and extreme gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China Energy Convers. Manage., 164 (2018), pp. 102-111, 10.1016/j.enconman.2018.02.087