Journal of Physics: Conference Series **3140** (2025) 022002 doi:10.1088/1742-6596/3140/2/022002

Performance-based drift detection for active machine learning model adaption: A comparative analysis across 35 HVAC devices

K Derzsi, F Stinner, L Patrun, J Klingebiel, and D Müller

RWTH Aachen University, E.ON Energy Research Center, Institute for Energy Efficient Buildings and Indoor Climate, Mathieustraße 10, 52072 Aachen, Germany

E-mail: kai.derzsi@eonerc.rwth-aachen.de

Abstract. Building operations contribute significantly to global CO_2 emissions, making optimal energy system control crucial for climate change mitigation. While machine learning models can predict building system behavior to enable advanced control strategies, model performance is prone to deterioration due to concept drift in real-world data streams. This study investigates active model adaptation with concept drift detection across 35 HVAC devices in 14 German non-residential buildings on two years of monitoring data. We compared static models against various combinations of machine learning algorithms and performance-based drift detection methods. Results revealed model adaptation effectiveness correlates with baseline performance characteristics. Improvements were observed for devices exhibiting wider model performance distributions, while negative effects were discovered for devices with narrower ranges. Notably, drift detection timing proved more critical than retraining frequency. Active model adaptation achieved improvements of 6.44 to 35.58 % across device types, with significant variations based on machine learning algorithm and concept drift detection method combinations.

1. Introduction

Non-residential building operations make up 35 to 40 % of CO_2 emissions across Europe and the USA [1]. Furthermore, an estimated 90 % of building areas are subject to incorrect control implementation, leaving approximately 34 % of potential primary energy savings in heating, ventilation, and air conditioning (HVAC) operations unexploited [2].

Model-based prediction of building system component behavior facilitate deeper system understanding, while enabling advanced control strategies including model predictive control (MPC), Demand Response, and optimization frameworks that maximize energy efficiency, cost savings, occupant comfort, and emissions reduction [3]. MPC has demonstrated potential to reduce CO_2 emissions by 15 to 50 % in non-residential buildings, but practical implementation faces significant challenges due to highly individual characteristics of building systems [4].

The proliferation of building automation systems (BAS), Internet of Things (IoT) devices, and advanced metering infrastructure (AMI) has generated unprecedented access to operational building data [3], [5]. Leveraging these resources, data-driven approaches offer high accuracy with reduced engineering costs by extracting implicit physical relationships from operational data. [6], [7].

Research has primarily focused on model development, with the majority of predictive models remaining limited to static information [8]. However, real-world building data streams are subject to concept drift, where underlying relationships evolve dynamically over time due to various factors including aging effects, seasonal variations, changes in occupant behavior, installation of new appliances, and equipment degradation [9]. These changes lead to deteriorating performance of traditional static machine learning models in practical applications, necessitating model adaption [10].
Concept drift detection, though primarily studied for classification tasks outside of building automation, offers promising solutions for maintaining model performance through active model adaptation.
Lima et al. [11] conducted a comprehensive evaluation of seven drift detection methods paired with ten regression models across synthetic and real datasets, analyzing performance using mean squared error and statistical tests. Their findings revealed that Page-Hinkley Test (PHT) and Kolmogorov-Smirnov Windowing (KSWIN) excelled on synthetic datasets, while Early Drift Detection Method (EDDM) and variants of the Hoeffding Drift Detection Method (HDDM) performed best on real-world bike-sharing data.
In the building energy domain, Toquica et al. [12] evaluated both passive and active retraining approaches for photovoltaic power generation prediction from simulated data, and residential power demand and indoor temperature forecasting from measurement data. Active model adaptation using concept drift detection methods Adaptive Windowing (ADWIN) and Klinkenberg Method achieved the lowest average errors. However, performance differences between various retraining strategies remained modest ranging from 2 to 5 %.
Mariano-Hernandez et al. [13] investigated electricity demand forecasting in university buildings across multiple years of operation. Passive retraining every 24 hours achieved the best overall results, while active model adaptation using KSWIN produced comparable performance reducing retraining frequency by more than 50 %. ADWIN demonstrated further reductions in retraining events at competitive model performance.
Existing studies demonstrate the potential of active model adaptation for predicting building system component behavior. However, these studies are typically restricted to a small number of individual devices or buildings, yielding results that favor different machine learning algorithms and concept drift detection methods that may not generalize well.

2. Use cases

For this study, comprehensive real-world monitoring datasets were sourced from 14 non-residential buildings in the city of Aachen, North Rhine-Westphalia, Germany. The data spans two full years from January 10$^{\text{th}}$, 2021, to January 8$^{\text{th}}$, 2023 with measurements recorded at 15-minute intervals. The building portfolio consists of seven schools, two fire stations, three event buildings, one kindergarten, and one administrative building with diverse usage characteristics, from which operational data was collected across 35 distinct HVAC devices, comprising 17 air handling units (AHU), 12 boilers (Bo), and 6 heat exchangers (HEx). The datasets contain time variables, weather data, and technical parameters specific to each HVAC type. AHU data incorporates heat consumption, valve positions, supply, exhaust and return air temperatures, ventilation duct pressures, and indoor temperatures. Bo datasets contain supply, return, and boiler temperatures. HEx records include primary and secondary temperatures, thermal energy consumption, and control valve positions. The data availability differed across the 35 technical installations, resulting in varying parameter sets for individual devices.
The time series data exhibit consistent patterns across all HVAC devices, including pronounced seasonal variations strongly correlating with weather conditions and responsiveness to control adjustments. During the observed period of operation, three distinct concept drift patterns emerged: sudden drift characterized by abrupt operational changes, recurring drift manifested

as seasonal variations, and incremental drift evidenced by progressive shifts in parameter relationships as device conditions evolved with seasonal transitions.

3. Methodology

The monitoring datasets from the individual HVAC devices were utilized to develop data-driven models based on different machine learning algorithms for short-term forecasting device-specific technical parameters over a 24-hour period ahead. The models were trained on the first complete year of available data from January 10th, 2021 to January 2nd, 2022, encompassing a full heating season. Prequential evaluation was conducted on instances from the second year of available data from January 3rd, 2022 to January 2nd, 2023, simulating real-world deployment under data stream conditions. The initial models served as static reference benchmarks for comparison with active model adaptation approaches.

3.1. Data-driven models

Prior to model development, Savitzky-Golay filters were applied to raw data to identify and exclude periods of inactivity. Weather data and technical parameters were scaled through min-max normalization, enabling consistent comparison across HVAC devices of varying scales. Time variables, including day of week, week of year, and minute of day, were extracted from timestamps, with cyclical encoding applied to the former two variables. The most relevant features for each device-specific regression task were selected through forward feature selection based on Spearman correlation analysis. Target features were selected based on specific HVAC devices type: supply air temperature for AHU, return flow temperature for Bo, and valve position and temperatures depending on data availability for HEx.

The machine learning models were implemented using the Python package scikit-learn [14]. Random Forest Regressor (RF) was employed in standard implementation due to demonstrated effectiveness and robustness in handling complex time series regression tasks with minimal hyperparameter tuning requirements. Multilayer Perceptron (MLP) with default configuration was selected as a second approach, leveraging its single hidden layer architecture with Rectified Linear Unit (ReLU) activation function and Adam optimizer. Additionally, an optimized MLP (OPT) was developed to evaluate the impact of parameter tuning on model performance, specifically in the context of performance-based concept drift detection and active model adaptation. Bayesian optimization was utilized for device-specific hyperparameter tuning across various neural network architectures.

3.2. Active model adaption and concept drift detection

Model adaptation was implemented as an active approach using batch learning, where explicit concept drift detection triggers model retraining from scratch. Performance-based drift detection monitored model degradation through statistically significant increases in daily absolute error metrics. Upon drift detection, model retraining utilized all data instances between current and previous drift events, with a 28-day minimum dataset threshold to ensure sufficient training data and accommodate temporal concept evolution. A 7-day drift break following drift detection prevented data leakage and excessive loss of prediction periods by temporarily suspending subsequent retraining triggers. Following drift detection and model retraining, the new model is deployed for predictions, while the previous model is discarded.

Four performance-based drift detection methods were considered, drawing from the largest group of concept drift detection approaches [15], [10]. ADWIN, Drift Detection Method (DDM), and EDDM were selected as established state-of-the-art methods in current research [16], [17], while KSWIN was added due to superior performance in comparative studies [18], [19] and specific efficacy in building energy contexts [13]. All methods were implemented using standard implementations from the scikit-multiflow Python package [20].

3.2.1. Adaptive Windowing The ADWIN algorithm monitors model error stability through an adaptive window W of length n that grows with each new data instance until drift is detected. W splits into two sufficiently large sub-windows W_0 and W_1 of lengths n_0 and n_1, comparing differences in expected values of a performance metric $\mu_{\hat{W}_0}$ and $\mu_{\hat{W}_1}$ to a threshold ϵ_{cut}:

$$|\mu_{\hat{W}_0} - \mu_{\hat{W}_1}| \geq \epsilon_{\text{cut}} \tag{1}$$

The threshold ϵ_{cut} is established by computing the harmonic mean of the lengths of the sub-windows m, the observed variance of performance metric values within the adaptive window σ^2_W, and a confidence level δ':

$$\epsilon_{\text{cut}} = \sqrt{\frac{2}{m} \cdot \sigma^2_W \cdot \ln(\frac{2}{\delta'})} + \frac{2}{3 \cdot m} \cdot \ln(\frac{2}{\delta'}) \tag{2}$$

Upon drift detection, W is replaced and reduced to the most recent sub-window W_1, representing the new concept. [21]
ADWIN was parameterized with a confidence level δ of 0.002 and applied to daily MAE values.

3.2.2. Drift Detection Method The DDM algorithm processes error rates in a growing landmark window, assuming stability during stationary concepts. For each timestep, the error rate e_i and the standard deviation of the error rate σ_i are calculated, tracking the observed minima e_{min} and σ_{min}. The drift detection is issued based on the confidence interval of 99 % corresponding to the respective trigger condition [22]:

$$e_i + \sigma_i \geq e_{\text{min}} + 3 \cdot \sigma_{\text{min}} \tag{3}$$

For regression tasks, continuous error metrics are transformed into a binary format introducing a threshold value E_{thr}. The threshold is established by applying the Innerquartile Range (IQR) to the absolute errors, using the Mean Absolute Error (MAE) with the first quartile Q_1 and the third quartile Q_3 from the current model evaluated on training data representing the current concept [11]:

$$E_{\text{thr}} = MAE + 1.5 \cdot (Q_3 - Q_1) \tag{4}$$

3.2.3. Early Drift Detection Method EDDM was derived from DDM to enhance concept drift detection in the presence of gradual drift by monitoring error distances in terms of timesteps between errors rather than error rates. Within landmark windows, the mean error distance e'_i and the standard deviation of the error distance σ'_i are calculated at each timestep i, tracking maximum values e'_{max} and σ'_{max}. Drift detection is based on a confidence interval of 95 % [23]:

$$e'_i + 2 \cdot \sigma'_i < 0.90 \cdot (e'_{\text{max}} + 2 \cdot \sigma'_{\text{max}}) \tag{5}$$

EDDM requires at least 30 errors to occur before error distance statistics are considered for drift detection, ensuring stable calculations. For regression tasks, continuous errors are converted to binary format using the established IQR method.

3.2.4. Kolmogorov-Smirnov Windowing The original KSWIN algorithm for data distribution-based concept drift detection utilizing the Kolmogorov-Smirnov Test [18], was adapted as a performance-based drift detection method applied to daily MAE distributions. This method operates on the principle that concept drift manifests as changes in probability distributions. Data flows through a sliding window W of length n, with two distinct sub-windows: R containing

the r most recent instances and S comprising r uniformly sampled non-recent instances. The Kolmogorov-Smirnov Test evaluates whether both sub-windows originate from identical distributions. The test calculates Kolmogorov-Smirnov Distance d as the maximum difference between empirical distribution functions $F_{R,r}(x)$ and $F_{S,r}(x)$, comparing it against a threshold derived from significance level α and window size r. When d exceeds this threshold, the null hypothesis is rejected, signaling concept drift:

$$d = \max_x |F_{R,r}(x) - F_{S,r}(x)| > \sqrt{\frac{-ln(\alpha)}{r}} \tag{6}$$

The implementation used α of 0.005, n of 100 and r of 30, representing approximately four months and one month of daily MAE values, respectively.

4. Results

Prequential validation was performed on 35 HVAC device-specific datasets using three machine learning algorithms as static references and in combination with four concept drift detection methods for active model adaptation totaling 525 simulations. Performance was evaluated through the Coefficient of Variation of the Root Mean Square Error (CVRMSE), while computational efficiency was measured by the number of models trained during validation.

Figure 1 illustrates CVRMSE against model count for all simulations. Performance varied substantially by device type. AHU devices showed CVRMSE values ranging from 7.31 to 79.69 %, while Bo devices demonstrated better consistency with values between 4.17 and 27.12 %. HEx devices exhibited the widest variation, spanning from 3.63 to 113.67 %.

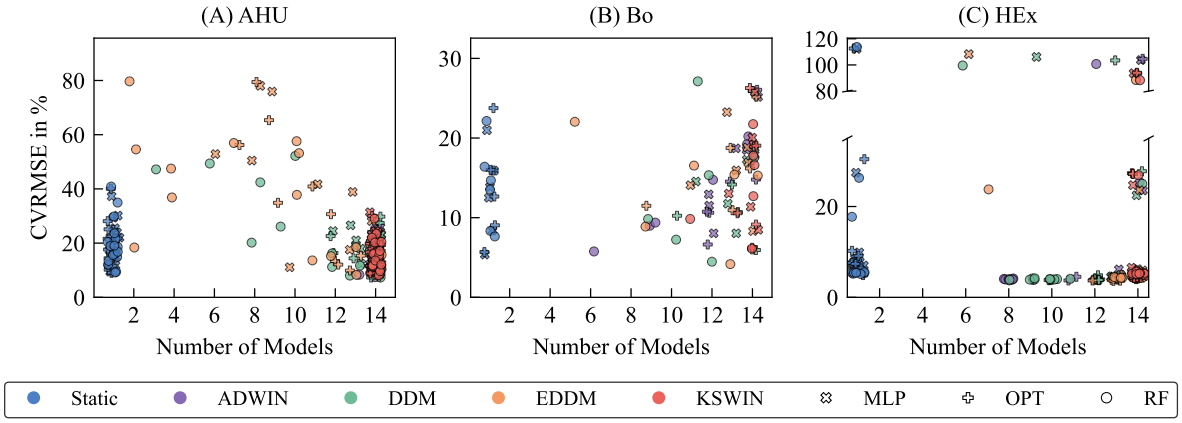

Figure 1. Average model performance against number of models by machine learning algorithm and concept drift detection method combinations.

The number of trained models typically ranged from 8 to 14 across all device categories, with adaptation methods showing device-specific effectiveness. For AHU, DDM and EDDM occasionally required fewer retrainings but sometimes sacrificed accuracy, while ADWIN and KSWIN consistently achieved moderate improvements with 14 deployed models. Bo devices showed similar patterns to AHU but with less variation in model count. HEx devices displayed more pronounced differences among methods, with DDM and ADWIN achieving strong performance efficiently at 8 to 11 models, EDDM requiring 12-14 models with slightly degraded results, and KSWIN consistently needing 14 models with less favorable outcomes.

Among algorithms, OPT achieved the best average performance (16.65 %), followed closely by

RF (17.31 %) and MLP (17.39 %.) RF demonstrated greater efficiency by requiring fewer models on average (8.62) compared to OPT (9.78) and MLP (9.91), particularly for Bo and HEx.

Figure 2 shows average performance improvements for each machine learning algorithm and concept drift detection method combination compared to the respective static reference models. Improvement patterns varied significantly by device type. AHU devices showed predominantly positive improvements ranging from 7.00 to 26.49 %, with ADWIN consistently producing the strongest enhancements, followed by KSWIN. EDDM showed inconsistent performance, while the RF-DDM combination demonstrated weaker improvements. Bo devices showed a contrasting pattern, with most adaptation approaches degrading performance. MLP algorithms were particularly susceptible to deterioration when retrained on focused current contexts.

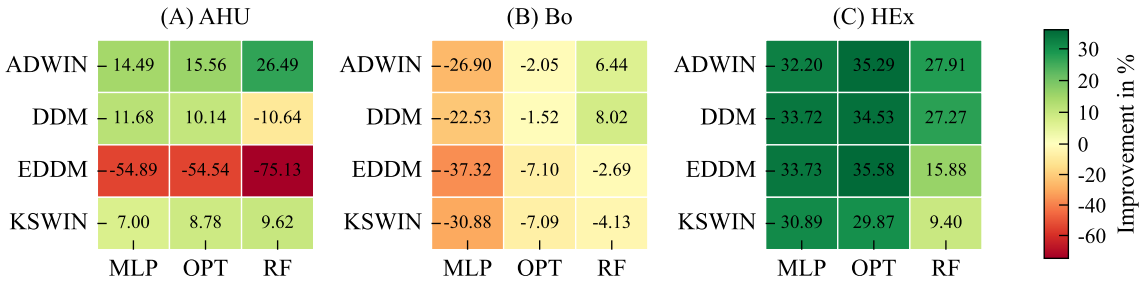

Figure 2. Average model performance improvement compared to the static benchmark model by machine learning algorithm and concept drift detection method combinations.

Only RF combined with DDM or ADWIN yielded improvements of 8.02 % and 6.44 %, respectively. The stable baseline performance in Bo devices suggests model adaptation may be counterproductive when initial models capture essential patterns. HEx devices exhibited the highest and most consistent improvements at 9.40 to 35.58 %. Neural network approaches benefited more substantially from model adaptation, with OPT consistently outperforming MLP. For RF, DDM and ADWIN achieved considerably higher gains than EDDM and KSWIN.

5. Discussion and conclusion

In summary, improvements scaled proportionally with error ranges, delivering greater benefits in AHU and HEx devices exhibiting wider performance distributions. Conversely, for Bo devices, which demonstrated narrower performance ranges and stronger baseline accuracy, model adaptation frequently proved counterproductive. Except for AHU devices, no direct relationship between model count and overall performance was observed, suggesting that the timing of drift detection is more important than retraining frequency. Among concept drift detection methods, DDM triggered fewer adaptations, ADWIN displayed balanced behavior while maintaining strong performance, and KSWIN showed high sensitivity, consistently triggering the most retrainings with mixed results. EDDM exhibited the strongest performance degradation for AHU and Bo devices, which can likely be attributed to suboptimal identification of drift events. Future research will further explore concept drift detection events to provide valuable insights into underlying system changes, potentially enabling more targeted model adaption.

Acknowledgments

We gratefully acknowledge the financial support provided by the BMWK (Federal Ministry for Economic Affairs and Climate Action), promotional reference 03SBE0006A.

References

[1] Hamilton I, Kennard H and Rapf O 2022 2022 Global Status Report for Buildings and Construction: Towards a Zero-emission, Efficient and Resilient Buildings and Construction Sector Report United Nations Environment Programme Nairobi
[2] Waide P, Ure J, Karagianni N, Smith G and Bordass B The scope for energy and CO2 savings in the EU through the use of building automation technology: Final Report URL https://leonardo-energy.pl/wp-content/uploads/2017/07/The-scope-for-energy-savings-from-energy-management.pdf
[3] Zhang L, Wen J, Li Y, Chen J, Ye Y, Fu Y and Livingood W 2021 *Applied Energy* **285** URL https://ideas.repec.org/a/eee/appene/v285y2021ics0306261921000209.html
[4] Drgoňa J, Arroyo J, Cupeiro Figueroa I, Blum D, Arendt K, Kim D, Ollé E P, Oravec J, Wetter M, Vrabie D L and Helsen L 2020 *Annual Reviews in Control* **50** 190–232 ISSN 1367-5788 URL https://www.sciencedirect.com/science/article/pii/S1367578820300584
[5] Yildiz B, Bilbao J, Dore J and Sproul A 2017 *Applied Energy* **208** 402–427 ISSN 03062619 URL https://linkinghub.elsevier.com/retrieve/pii/S0306261917314265
[6] Deb C and Schlueter A 2021 *Renewable and Sustainable Energy Reviews* **144** 110990 ISSN 13640321 URL https://linkinghub.elsevier.com/retrieve/pii/S1364032121002823
[7] Chen Y, Guo M, Chen Z, Chen Z and Ji Y 2022 *Energy Reports* **8** 2656–2671 ISSN 23524847 URL https://linkinghub.elsevier.com/retrieve/pii/S2352484722001615
[8] Marinakis V 2020 *Energies* **13** 1555 ISSN 1996-1073 URL https://www.mdpi.com/1996-1073/13/7/1555
[9] Jagait R K, Fekri M N, Grolinger K and Mir S 2021 *IEEE Access* **9** 98992–99008 ISSN 2169-3536 URL https://ieeexplore.ieee.org/document/9476011/
[10] Bayram F, Ahmed B S and Kassler A 2022 *Knowledge-Based Systems* **245** 108632 ISSN 09507051 URL https://linkinghub.elsevier.com/retrieve/pii/S0950705122002854
[11] Lima M, Filho T S and De A Fagundes R A 2021 A Comparative Study on Concept Drift Detectors for Regression *Intelligent Systems* vol 13073 ed Britto A and Valdivia Delgado K (Springer International Publishing) pp 390–405 ISBN 978-3-030-91701-2 978-3-030-91702-9 URL https://link.springer.com/10.1007/978-3-030-91702-9_26
[12] Toquica D, Agbossou K, Malhamé R, Henao N, Kelouwani S and Cardenas A 2020 *Energies* **13** 2250 ISSN 1996-1073 URL https://www.mdpi.com/1996-1073/13/9/2250
[13] Mariano-Hernández D, Hernández-Callejo L, Solís M, Zorita-Lamadrid A, Duque-Pérez O, Gonzalez-Morales L, García F S, Jaramillo-Duque A, Ospino-Castro A, Alonso-Gómez V and Bello H J 2022 *Sustainability* **14** 5857 ISSN 2071-1050 URL https://www.mdpi.com/2071-1050/14/10/5857
[14] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M and Duchesnay E 2011 *Journal of Machine Learning Research* **12** 2825–2830
[15] Lu J, Liu A, Dong F, Gu F, Gama J and Zhang G 2018 *IEEE Transactions on Knowledge and Data Engineering* 1–1 ISSN 1041-4347, 1558-2191, 2326-3865 (*Preprint* 2004.05785) URL http://arxiv.org/abs/2004.05785
[16] Wares S, Isaacs J and Elyan E 2019 *SN Applied Sciences* **1** 1412 ISSN 2523-3963, 2523-3971 URL http://link.springer.com/10.1007/s42452-019-1433-0
[17] Lima M, Neto M, Filho T S and De A Fagundes R A 2022 *IEEE Access* **10** 45410–45429 ISSN 2169-3536 URL https://ieeexplore.ieee.org/document/9762269/
[18] Raab C, Heusinger M and Schleif F M 2020 *Neurocomputing* **416** 340–351 ISSN 09252312 (*Preprint* 2007.05432) URL http://arxiv.org/abs/2007.05432
[19] Hinder F, Vaquet V, Brinkrolf J and Hammer B 2023 On the Hardness and Necessity of Supervised Concept Drift Detection: *Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods* (SCITEPRESS - Science and Technology Publications) pp 164–175 ISBN 978-989-758-626-2 URL https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0011797500003411
[20] Montiel J, Read J, Bifet A and Abdessalem T 2018 *Journal of Machine Learning Research* **19** 1–5 URL http://jmlr.org/papers/v19/18-251.html
[21] Bifet A and Gavaldà R 2007 Learning from Time-Changing Data with Adaptive Windowing *Proceedings of the 2007 SIAM International Conference on Data Mining* (Society for Industrial and Applied Mathematics) pp 443–448 ISBN 978-0-89871-630-6 978-1-61197-277-1 URL https://epubs.siam.org/doi/10.1137/1.9781611972771.42
[22] Gama J, Medas P, Castillo G and Rodrigues P 2004 Learning with Drift Detection *Advances in Artificial Intelligence – SBIA 2004* vol 3171 ed Bazzan A L C and Labidi S (Springer Berlin Heidelberg) pp 286–295 ISBN 978-3-540-23237-7 978-3-540-28645-5 URL http://link.springer.com/10.1007/978-3-540-28645-5_29
[23] Baena-Garcıa M, Gavalda R and Morales-Bueno R 2006